

# ON PERFECT CLUSTERING OF HIGH DIMENSION, LOW SAMPLE SIZE DATA

SOHAM SARKAR<sup>1</sup> AND ANIL K. GHOSH<sup>2</sup>

*Theoretical Statistics and Mathematics Unit  
Indian Statistical Institute  
203, B. T. Road, Kolkata 700108*

## Abstract

Popular clustering algorithms based on usual distance functions (e.g., Euclidean distance) often suffer in high dimension, low sample size (HDLSS) situations, where concentration of pairwise distances has adverse effects on their performance. In this article, we use a dissimilarity measure based on the data cloud, called MADD, which takes care of this problem. MADD uses the distance concentration phenomenon to its advantage, and as a result, clustering algorithms based on MADD usually perform better for high dimensional data. Using theoretical and numerical results, we amply demonstrate it in this article.

We also address the problem of estimating the number of clusters. This is a very challenging problem in cluster analysis, and several algorithms have been proposed for it. We show that many of these existing algorithms have superior performance in high dimensions when MADD is used instead of the Euclidean distance. We also construct a new estimator based on penalized Dunn index and prove its consistency in the HDLSS asymptotic regime, where the sample size remains fixed and the dimension grows to infinity. Several simulated and real data sets are analyzed to demonstrate the importance of MADD for cluster analysis of high dimensional data.

**Keywords:** Dunn index, hierarchical clustering, high dimensional consistency, k-means clustering, pairwise distances, Rand index.

## 1 Introduction

Let  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$  be a sample of  $n$  unlabeled observations coming from different populations. The aim of cluster analysis is to divide this sample into several groups of ‘similar’ observations. In practice, one uses an appropriate measure of similarity (or, dissimilarity) between a pair of observations, and a clustering algorithm is developed based on that. When all measurement variables are continuous, a popular choice for the dissimilarity index is the Euclidean distance or the squared Euclidean distance. Popular clustering algorithms like  $k$ -means,  $k$ -medoids and hierarchical clustering (see [Hastie et al. 2009](#); [Duda et al. 2012](#)) generally use dissimilarity indices based on the Euclidean distance. Spectral clustering algorithms (see [von Luxburg 2007](#)) often use similarity index based on

---

<sup>1</sup>Email:sohamsarkar1991@gmail.com

<sup>2</sup>Email:akghosh@isical.ac.in

the radial basis function, which is a decreasing function of the Euclidean distance. These algorithms work well when the sample size is sufficiently large. But, like other nonparametric methods, they often perform poorly in high dimension, low sample size (HDLSS) situations.

To demonstrate this, we consider an example (call it Example A), with two  $d$ -dimensional normal distributions  $\mathcal{N}_d(\mathbf{0}_d, \sigma_1^2 \mathbf{\Sigma}_d)$  and  $\mathcal{N}_d(\boldsymbol{\mu}_d, \sigma_2^2 \mathbf{\Sigma}_d)$ , where  $\mathbf{0}_d = (0, \dots, 0)^\top$ ,  $\boldsymbol{\mu}_d = (1, -1, \dots, (-1)^{d+1})^\top$ , and  $\mathbf{\Sigma}_d = ((\sigma_{ij}))_{d \times d}$  is a block diagonal matrix with  $\sigma_{ii} = 1$  for  $i = 1, \dots, d$ ,  $\sigma_{(2i-1)2i} = \sigma_{2i(2i-1)} = 0.98$  for  $i = 1, \dots, \lfloor d/2 \rfloor$  ( $\lfloor t \rfloor$  is the largest integer  $\leq t$ ) and  $\sigma_{ij} = 0$  otherwise. Taking  $\sigma_1^2 = 0.5$  and  $\sigma_2^2 = 2$ , we generated 50 observations from each distribution. Figure 1(a) shows the central regions of these two distributions with coverage probabilities 0.25, 0.5, 0.75 and 0.9 when  $d = 2$ . We used the average linkage method (AvgL) as well as the  $k$ -means algorithm (kM) based on Euclidean distance to estimate two clusters in the sample consisting of 100 observations. For  $i = 1, \dots, n$ , let  $\mathcal{C}(\mathbf{x}_i)$  be the actual cluster label of  $\mathbf{x}_i$  and  $\delta(\mathbf{x}_i)$  be the cluster label assigned to  $\mathbf{x}_i$  by a clustering algorithm  $\delta$ . We measure the performance of  $\delta$  using the Rand index (see Rand 1971)

$$\mathcal{R}(\delta) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbb{I} \left[ \mathbb{I}\{\delta(\mathbf{x}_i) = \delta(\mathbf{x}_j)\} + \mathbb{I}\{\mathcal{C}(\mathbf{x}_i) = \mathcal{C}(\mathbf{x}_j)\} = 1 \right], \quad (1)$$

where  $\mathbb{I}\{\cdot\}$  is the indicator function. Note that  $\delta$  leads to perfect clustering if  $\delta(\mathbf{x}_i) = \pi\{\mathcal{C}(\mathbf{x}_i)\}$  for all  $i = 1, \dots, n$  and a suitable permutation  $\pi$  of  $\{1, \dots, \max\{\mathcal{C}(\mathbf{x}_1), \dots, \mathcal{C}(\mathbf{x}_n)\}\}$ . In that case, we have  $\mathcal{R}(\delta) = 0$ . Higher values of Rand index indicate more deviation from perfect clustering. We repeated our experiment 100 times, and the average Rand index of an algorithm over these 100 trials was computed for  $d = 2^r$ , with  $r = 1, \dots, 11$ . In this example, separation between the two populations is quite evident (see Figure 1(a)), and it increases with the dimension. So, for a good clustering algorithm, the Rand index is expected to shrink to 0 as  $d$  increases. But, that was not the case for AvgL and kM. Both of them had miserable performance for all values of  $d$  (see Figure 2(a)). The spectral clustering algorithm proposed by Shi and Malik (2000) (Spect) also had higher Rand index when a similarity measure based on radial basis function was used (see Figure 2(a)).

We carried out another experiment with observations generated from three distributions with disjoint supports, viz.,  $\mathcal{U}_d(0, 0.5)$ ,  $\mathcal{U}_d(1, 1.5)$  and  $\mathcal{U}_d(2, 2.5)$ . Here  $\mathcal{U}_d(a, b)$  denotes the  $d$ -dimensional uniform distribution over the region  $\{\mathbf{x} \in \mathbb{R}^d : a\sqrt{d} \leq \|\mathbf{x}\| \leq b\sqrt{d}\}$ . Figure 1(b) shows the supports of these three distributions for  $d = 2$ . We generated 50 observations from each distribution, and different clustering algorithms were used to divide these 150 observations into three different clusters.

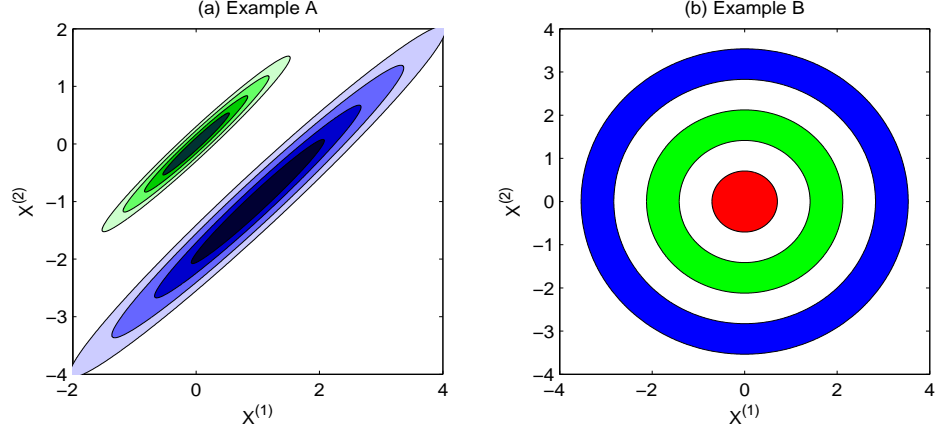


Figure 1: (a) Central regions of the two normal populations in Example A and (b) Supports of the three non-overlapping populations in Example B (for  $d = 2$ ).

In this example (call it Example B) also, all these three methods, especially AvgL and kM, had poor performance (see Figure 2(b)).

Clustering algorithm based on maximal data pilling (MDP) distance (Ahn et al. 2012), which is especially designed for high dimensional data, performed well in Example A for  $d \geq 2^7$ , but it performed poorly for all smaller values of  $d$ . In Example B, its performance was even worse. It had much higher Rand index for all values of  $d$  considered here.

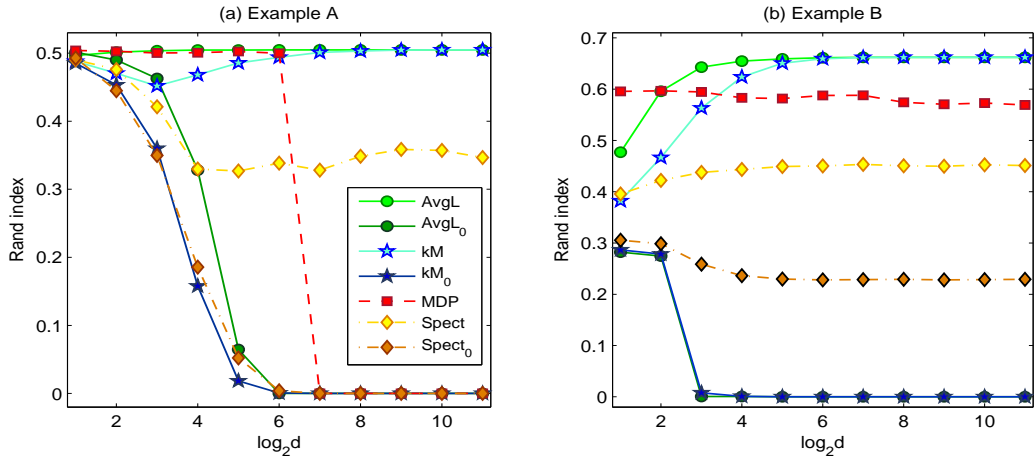


Figure 2: Rand indices for different algorithms in Examples A and B.

Failure of these popular algorithms shows the necessity to develop new methods for clustering high dimensional data. In this article, we use a new dissimilarity index, called MADD (defined in Section 2), for this purpose. Notice that both AvgL and kM yielded excellent results when MADD

was used as the dissimilarity measure (see the curves corresponding to AvgL<sub>0</sub> and kM<sub>0</sub> in Figure 2). In both examples, they led to perfect clustering in high dimensions. It is well known that both AvgL and kM based on Euclidean distance are not much useful for finding non-convex clusters in the data, but Example B clearly shows that their MADD versions can overcome this limitation. Spectral clustering algorithm of Shi and Malik (2000) also performed well when a decreasing function of MADD (defined in Section 2) was used as the similarity measure. While the use of MADD led to perfect clustering for large  $d$  in Example A, it significantly reduced the Rand index in Example B as well (see the curves corresponding to Spect<sub>0</sub> in Figure 2).

The reasons behind the failure of Euclidean distance based clustering and the excellence of MADD based clustering are investigated in Section 2. In this connection, we prove the high dimensional consistency of some clustering algorithms based on MADD. Simulation studies are also carried out to demonstrate the superiority of these MADD based algorithms. We consider the problem of estimating the number of clusters from the data in Section 3. This is an important problem in cluster analysis, and several methods are available for it. We observe that many of these methods perform better when MADD is used for their constructions. We investigate the high dimensional behavior of these MADD based estimation methods under appropriate regularity conditions. We also construct a new estimator based on penalized Dunn index and prove its high dimensional consistency. Empirical performances of different estimation methods are evaluated using simulation studies. Two benchmark data sets are analyzed in Section 4 for further comparison among different estimation methods and clustering algorithms. Finally, Section 5 gives a brief summary of the work and ends with some related discussions. All proofs and mathematical details are given in the Appendix.

## 2 Clustering algorithms based on MADD

Suppose that the whole sample  $\mathcal{X} = \cup_{i=1}^{k_0} \mathcal{X}_i$  consists of  $n$  unlabeled observations, where  $\mathcal{X}_i$  denotes the collection of  $n_i$  observations ( $\sum_{i=1}^{k_0} n_i = n$ ) from the  $i$ -th ( $i = 1, \dots, k_0$ ) population. In Example A, we had  $k_0 = 2$  and  $n_1 = n_2 = 50$ . In this example, for two observations  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})^\top$  and  $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(d)})^\top$  from the second population,  $d^{-1} \|\mathbf{X} - \mathbf{Y}\|^2 = d^{-1} \sum_{q=1}^d (X^{(q)} - Y^{(q)})^2$ , being an average of an  $m$ -dependent sequence (with  $m = 1$ ) of identically distributed random variables, converges to  $E(X^{(1)} - Y^{(1)})^2 = 2\sigma_2^2 = 4$  in probability. But, if  $\mathbf{X}$  comes from the first

population and  $\mathbf{Y}$  comes from the second population,  $d^{-1}\|\mathbf{X} - \mathbf{Y}\|^2$  converges in probability to  $\sigma_1^2 + \sigma_2^2 + 1 = 3.5$ . Due to this concentration of pairwise distances, for large  $d$ , all observations in  $\mathcal{X}_2$  had their neighbors in  $\mathcal{X}_1$ . So, clustering algorithms based on the Euclidean distance failed to put them in the same cluster. Hall et al. (2005) proved the concentration of Euclidean distance assuming weak dependence among the component variables and provided an idea about the high dimensional geometry of the data cloud consisting of observations from two distributions. They also pointed out the adverse effects of distance concentration on some popular classifiers. In Example A, we observe its adverse effects on clustering algorithms. In Example B also, we have similar convergence of pairwise distances, which is shown in the following lemma.

**Lemma 1.** *If  $\mathbf{X} \sim \mathcal{U}_d(a_1, b_1)$ ,  $\mathbf{Y} \sim \mathcal{U}_d(a_2, b_2)$ , and they are independent, then  $d^{-1}\|\mathbf{X} - \mathbf{Y}\|^2$  converges in probability to  $b_1^2 + b_2^2$  as the dimension  $d$  tends to infinity.*

From Lemma 1, it is clear that for large values of  $d$ , all observations in  $\mathcal{X}_2$  and  $\mathcal{X}_3$  had their nearest neighbors in  $\mathcal{X}_1$  only. So, AvgL and kM algorithms had miserable performances in this example as well. However, these two algorithms produced excellent results in Examples A and B when, instead of Euclidean distance, we used a new dissimilarity index given by  $\rho(\mathbf{x}, \mathbf{y}) = (n - 2)^{-1} \sum_{\mathbf{z} \in \mathcal{X} \setminus \{\mathbf{x}, \mathbf{y}\}} \left| \|\mathbf{x} - \mathbf{z}\| - \|\mathbf{y} - \mathbf{z}\| \right|$ . Following our above discussion, one can show that in both of these examples,  $d^{-1/2}\rho(\mathbf{X}, \mathbf{Y}) \xrightarrow{P} 0$  as  $d \rightarrow \infty$  if and only if  $\mathbf{X}$  and  $\mathbf{Y}$  come from the same population. Otherwise, it converges to a positive constant. So, clustering algorithms based on  $\rho$  had better performance in high dimensions.

This type of dissimilarity index based on **Mean Absolute Difference of Distances** (called MADD) can be constructed using other distance functions as well. In this article, we consider distance functions of the form  $\varphi_{h,\psi}(\mathbf{x}, \mathbf{y}) = h\left\{d^{-1} \sum_{q=1}^d \psi(|x^{(q)} - y^{(q)}|)\right\}$ , where  $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  and  $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  are continuous, monotonically increasing functions with  $h(0) = \psi(0) = 0$  such that  $\varphi_{h,\psi}$  is a distance in  $\mathbb{R}^d$ . Clearly, this class of distance functions include all  $\ell_p$  distances (upto a scalar constant) with  $p \geq 1$ . We define the general version of MADD as

$$\rho_{h,\psi}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-2} \sum_{\mathbf{z} \in \mathcal{X} \setminus \{\mathbf{x}, \mathbf{y}\}} \left| \varphi_{h,\psi}(\mathbf{x}, \mathbf{z}) - \varphi_{h,\psi}(\mathbf{y}, \mathbf{z}) \right|. \quad (2)$$

Using  $h(t) = \sqrt{t}$  and  $\psi(t) = t^2$ , we get  $\rho_{h,\psi}(\mathbf{x}, \mathbf{y}) = d^{-1/2}\rho(\mathbf{x}, \mathbf{y})$ , and we call it  $\rho_0$ . MADD has some nice properties as a dissimilarity index. Lemma 2 shows that it is a semi-metric.

**Lemma 2.** *If  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  contains  $n \geq 3$  observations,  $\rho_{h,\psi}$  is a semi-metric on  $\mathcal{X}$ .*

MADD is not a metric since it is possible to get  $\mathbf{x} \neq \mathbf{y}$  such that  $\rho_{h,\psi}(\mathbf{x}, \mathbf{y}) = 0$ . But, when  $\mathcal{X}$  consists of continuous random vectors, for  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$  and  $\mathbf{x} \neq \mathbf{y}$ ,  $\rho_{h,\psi}(\mathbf{x}, \mathbf{y}) > 0$  holds with probability one. So, for all practical purposes, it behaves like a metric.

Since  $\varphi_{h,\psi}$  satisfies the triangle inequality, one can show that  $\rho_{h,\psi}(\mathbf{x}, \mathbf{y}) \leq \varphi_{h,\psi}(\mathbf{x}, \mathbf{y})$  for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ . Thus, closeness in terms of  $\varphi_{h,\psi}$  (e.g., Euclidean distance) also indicates closeness in terms of MADD, but not the converse. In particular, for high dimensional data, unlike the Euclidean distance, MADD usually provides small dissimilarities among observations from the same population, and that helps us to develop better clustering algorithms.

## 2.1 High dimensional behavior of MADD

To study the high dimensional behavior of  $\rho_{h,\psi}$  and associated clustering algorithms in details, we assume that  $\mathcal{X}$  consists of  $n$  independent observations on the measurement vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})^\top$  coming from a mixture of  $k_0$  populations, where  $n_i = |\mathcal{X}_i| \geq 2$  for all  $i = 1, \dots, k_0$ . We also make the following assumption.

(A1) For independent observations  $\mathbf{X}$  and  $\mathbf{Y}$  from  $i$ -th and  $j$ -th populations ( $1 \leq i, j \leq k_0$ ),  $d^{-1} \sum_{q=1}^d \{\psi(|X^{(q)} - Y^{(q)}|) - E\psi(|X^{(q)} - Y^{(q)}|)\} \xrightarrow{P} 0$  as  $d \rightarrow \infty$ .

This assumption regarding weak convergence of the sequence  $\{\psi(|X^{(q)} - Y^{(q)}|) : q \geq 1\}$  is pretty common in the HDLSS literature. A sufficient condition for (A1) is  $Var\{\sum_{q=1}^d \psi(|X^{(q)} - Y^{(q)}|)\} = \mathbf{o}(d^2)$ . If the component variables are independent and identically distributed (*i.i.d.*) with  $E\psi(|X^{(1)} - Y^{(1)}|) < \infty$ , then (A1) holds. For sequences of dependent and non-identically distributed random variables, we need some additional conditions. Several sufficient conditions have been used by many researchers. For instance, [Hall et al. \(2005\)](#) assumed a  $\rho$ -mixing condition on the measurement variables and uniform boundedness of their fourth order moments to study the high dimensional behavior of some classifiers based on the Euclidean distance (i.e., when  $\psi(t) = t^2$ ). [Jung and Marron \(2009\)](#) used some slightly weaker conditions to establish the high dimensional consistency of their estimated principle component directions. Similar conditions were used by [Ahn et al. \(2012\)](#) and [Biswas et al. \(2014\)](#) for high dimensional asymptotics. [Biswas et al. \(2015\)](#) derived some sufficient conditions for the weak convergence of  $\{\psi(|X^{(q)} - Y^{(q)}|) : q \geq 1\}$ . Some sufficient conditions based on mixingales were derived by [Andrews \(1988\)](#) and [de Jong \(1995\)](#).

Suppose that  $\mathbf{X}$  and  $\mathbf{Z}$  are independent observations from the  $i$ -th and the  $\ell$ -th populations ( $1 \leq i, \ell \leq k_0$ ). Then, using (A1) and the continuity of  $h$ , one gets  $|\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}^*(i, \ell)| \xrightarrow{P} 0$ , where  $\varphi_{h,\psi}^*(i, \ell) = h\{d^{-1} \sum_{q=1}^d E\psi(|X^{(q)} - Z^{(q)}|)\}$ . So, if  $\mathbf{X}$  and  $\mathbf{Y}$  are from  $i$ -th and  $j$ -th populations, we have  $|\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) - \rho_{h,\psi}^*(i, j)| \xrightarrow{P} 0$ , where  $\rho_{h,\psi}^*(i, j) = (n-2)^{-1}[(n_i-1)|\varphi_{h,\psi}^*(i, j) - \varphi_{h,\psi}^*(i, i)| + (n_j-1)|\varphi_{h,\psi}^*(i, j) - \varphi_{h,\psi}^*(j, j)| + \sum_{\ell \neq i, j} n_\ell |\varphi_{h,\psi}^*(i, \ell) - \varphi_{h,\psi}^*(j, \ell)|]$ . This is formally stated below.

**Lemma 3.** *Suppose that we have  $n$  independent observations from  $k_0$  populations satisfying (A1). If  $\mathbf{X}$  and  $\mathbf{Y}$  come from the  $i$ -th and the  $j$ -th ( $1 \leq i, j \leq k_0$ ) populations, respectively, and  $h$  is continuous, then  $|\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) - \rho_{h,\psi}^*(i, j)| \xrightarrow{P} 0$  as  $d \rightarrow \infty$ .*

Clearly,  $\rho_{h,\psi}^*(i, j) = 0$  if  $i = j$  and  $\rho_{h,\psi}^*(i, j) \geq 0$  for  $i \neq j$ . However, for good performance of clustering algorithms based on MADD, one would like to choose  $h$  and  $\psi$  such that  $\rho_{h,\psi}^*(i, j) > 0$  for  $i \neq j$ . Lemma 4 guides us to some suitable choices in this regard.

**Lemma 4.** *If  $h$  and  $\psi$  are strictly increasing, and  $\psi'(t)/t$  is a non-constant monotone function on  $(0, \infty)$ , then for any  $i \neq j$ ,  $\rho_{h,\psi}^*(i, j) = 0$  if and only if the  $i$ -th and the  $j$ -th populations have the same one-dimensional marginals.*

There are several choices of  $\psi$  satisfying the properties mentioned in Lemma 4 (see, e.g., [Baringhaus and Franz 2010](#); [Biswas et al. 2015](#)). Some of them (e.g.,  $\psi(t) = t$ ,  $\psi(t) = t/(1+t)$ ,  $\psi(t) = 1 - e^{-t}$ ) lead to distance functions in  $\mathbb{R}$ . For such choices of  $\psi$ , it is enough to take  $h(t) = t$  to make  $\varphi_{h,\psi}$  a distance in  $\mathbb{R}^d$ . In these cases, we have  $\rho_{h,\psi}^*(i, j) > 0$  unless the two populations have identical marginal distributions. For the Euclidean distance (i.e.,  $\psi(t) = t^2$ ),  $\psi$  does not satisfy the property mentioned in Lemma 4. But Lemma 5 shows that even in that case,  $\rho_{h,\psi}^*(i, j)$  turns out to be positive for a large class of examples.

**Lemma 5.** *Let  $\boldsymbol{\mu}_i$  and  $\boldsymbol{\Sigma}_i$  be the mean vector and the dispersion matrix of the  $i$ -th population ( $i = 1, \dots, k_0$ ), respectively. For  $h(t) = \sqrt{t}$  and  $\psi(t) = t^2$ ,  $\rho_{h,\psi}^*(i, j)$  takes the value 0 if and only if  $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$  and  $\text{trace}(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j) = 0$ .*

Lemmas 4 and 5 show that for suitable choices of  $h$  and  $\psi$ , we usually have  $\rho_{h,\psi}^*(i, j) > 0$  for all values of  $d$ . So, it is reasonable to make the following assumption.

(A2) *For every  $1 \leq i \neq j \leq k_0$ ,  $\liminf_{d \rightarrow \infty} \rho_{h,\psi}^*(i, j) > 0$ .*

Note that (A2) holds in Examples A and B discussed in Section 1. It says that the separation between two populations is not asymptotically negligible. We will use this assumption for investigating the high dimensional behavior of MADD based algorithms.

## 2.2 High dimensional behavior of MADD based clustering

We know that AvgL begins with  $n$  groups, each consisting of a single observation. At each step, it chooses two closest groups and merges them into a single one. To measure closeness,  $\Delta(C_i, C_j) = (|C_i||C_j|)^{-1} \sum_{\mathbf{z} \in C_i, \mathbf{w} \in C_j} \|\mathbf{z} - \mathbf{w}\|$  is used as the distance between two groups  $C_i$  and  $C_j$ . AvgL stops merging when the pairwise distance between any two groups is bigger than a certain threshold or a specified number of groups is attained. The final groups thus formed are considered as the estimated clusters. Note that in Example A, for  $\mathbf{X}_1, \mathbf{X}_2 \stackrel{i.i.d.}{\sim} \mathcal{N}_d(\mathbf{0}_d, \sigma_1^2 \Sigma_d)$  and  $\mathbf{Y}_1, \mathbf{Y}_2 \stackrel{i.i.d.}{\sim} \mathcal{N}_d(\boldsymbol{\mu}_d, \sigma_2^2 \Sigma_d)$ , we have  $\Pr(\|\mathbf{X}_1 - \mathbf{X}_2\| < \|\mathbf{X}_1 - \mathbf{Y}_1\| < \|\mathbf{Y}_1 - \mathbf{Y}_2\|) \rightarrow 1$  as  $d \rightarrow \infty$  (see the discussion at the beginning of Section 2). So, for large values of  $d$ , after the first 49 steps, all observations from the first population were merged into a single group, and in each subsequent step, one observation from the second population was added to it. As a result, when AvgL ended with two estimated clusters, one of them had a single observation from the second population, and the other had the rest of the observations. This led to a Rand index of 0.505. A similar phenomenon occurred in Example B as well, where two of the three clusters estimated by AvgL had one observation each from the third population (i.e.,  $\mathcal{U}_d(2, 2.5)$ ), while the third cluster contained the rest. As a result, the Rand index turned out to be 0.662. The same phenomenon was observed even when single or complete linkage was used instead of AvgL.

We observed a diametrically opposite behavior for AvgL<sub>0</sub>, the MADD version of AvgL based on  $\rho_0$ , where  $\rho_0$  is used in place of  $\|\cdot\|$  to define  $\Delta(C_i, C_j)$ . In Example A, as  $d \rightarrow \infty$ , both  $\rho_0(\mathbf{X}_1, \mathbf{X}_2)$  and  $\rho_0(\mathbf{Y}_1, \mathbf{Y}_2)$  converge to 0, while  $\rho_0(\mathbf{X}_1, \mathbf{Y}_1)$  converges to a positive constant. So, any linkage method based on  $\rho_0$  leads to perfect clustering (i.e., zero Rand index) as  $d$  increases. Similar phenomenon occurs in Example B as well. This property of AvgL( $h, \psi$ ), the MADD version of AvgL based on  $\rho_{h, \psi}$ , is asserted by the following theorem.

**Theorem 1.** *Suppose that we have independent observations from  $k_0$  populations satisfying (A1). If  $h$  and  $\psi$  satisfy (A2), and AvgL( $h, \psi$ ) is used to estimate  $k_0$  clusters in the data, its Rand index converges to 0 in probability as  $d$  tends to infinity.*

For a given  $k$ , kM algorithm aims at finding  $k$  groups  $C_1, \dots, C_k$  with centers  $\mathbf{m}_1, \dots, \mathbf{m}_k$  such that  $\Phi(C_1, \dots, C_k) = \sum_{j=1}^k \sum_{i: \mathbf{x}_i \in C_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2$  is minimized. In practice, it starts with an initial choice of  $k$  groups, and then at each step an observation  $\mathbf{x}$  is assigned to the group having the center

closest to it. Group centers are updated accordingly. This iterative process is terminated when no groups are modified further. Using the convergence results for Euclidean distance, one can show that in Examples A and B, for large  $d$ ,  $\Phi$  is minimized when we have the same type of estimated clusters as obtained by AvgL. Since  $\Phi(C_1, \dots, C_k) = \sum_{r=1}^k (2|C_r|)^{-1} \sum_{\mathbf{z}, \mathbf{w} \in C_r} \|\mathbf{z} - \mathbf{w}\|^2$ , for the MADD version of kM (denoted by  $\text{kM}(h, \psi)$ ), we minimize  $\Phi^*(C_1, \dots, C_k) = \sum_{r=1}^k (2|C_r|)^{-1} \sum_{\mathbf{z}, \mathbf{w} \in C_r} \rho_{h, \psi}^2(\mathbf{z}, \mathbf{w})$ . Again we start with  $k$  initial groups and use an iterative algorithm. At each step, distance of an observation  $\mathbf{x}$  from a group  $C_j$  is computed as  $\rho_{h, \psi}^{(0)}(\mathbf{x}, C_j) = |C_j|^{-1} \sum_{\mathbf{z} \in C_j} \rho_{h, \psi}^2(\mathbf{x}, \mathbf{z})$ , and it is assigned to the group  $C_{\tilde{k}}$ , where  $\tilde{k} = \operatorname{argmin}_j \rho_{h, \psi}^{(0)}(\mathbf{x}, C_j)$ . This is done for all observations, and the process is repeated until convergence. From Lemma 3, it is clear that for  $k_0$  estimated clusters  $C_1, \dots, C_{k_0}$ ,  $\Phi^*(C_1, \dots, C_{k_0})$  attains its minimum value if and only if we have perfect clustering. So,  $\text{kM}(h, \psi)$  had excellent performance in Examples A and B, specially for large  $d$ . This perfect clustering property of  $\text{kM}(h, \psi)$  is asserted by the following theorem.

**Theorem 2.** *Suppose that we have independent observations from  $k_0$  populations satisfying (A1). If  $h$  and  $\psi$  satisfy (A2), and  $\text{kM}(h, \psi)$  is used to estimate  $k_0$  clusters in the data, its Rand index converges to 0 in probability as  $d$  tends to infinity.*

Theorems 1 and 2 show the perfect clustering property of  $\text{AvgL}(h, \psi)$  and  $\text{kM}(h, \psi)$  when  $\liminf_{d \rightarrow \infty} \rho_{h, \psi}^*(i, j) > 0$  for all  $i \neq j$ . This holds when  $\liminf_{d \rightarrow \infty} d^{-1} \sum_{q=1}^d \{2E\psi(|X^{(q)} - Y^{(q)}|) - E\psi(|X_1^{(q)} - X_2^{(q)}|) - E\psi(|Y_1^{(q)} - Y_2^{(q)}|)\} > 0$ , where  $\mathbf{X}_1, \mathbf{X}_2$  are from  $i$ -th population and  $\mathbf{Y}_1, \mathbf{Y}_2$  are from  $j$ -th population (see the proof of Lemma 4). So, in some sense, (A2) assumes that the total signal increases at least at the order of  $d$ . As pointed out by one of the reviewers, this is quite restrictive in practice. We can relax this condition if we make a slightly stronger assumption on  $h$ . Let us assume that for any pair of independent observations  $\mathbf{X}$  and  $\mathbf{Z}$ ,  $\text{Var}\{\sum_{q=1}^d \psi(|X^{(q)} - Z^{(q)}|)\} = \mathbf{O}(\vartheta^2(d))$ . Then the perfect clustering property of  $\text{AvgL}(h, \psi)$  and  $\text{kM}(h, \psi)$  can be proved under the following assumption.

(A2<sup>o</sup>) *For every  $i \neq j$ ,  $\rho_{h, \psi}^*(i, j) d/\vartheta(d) \rightarrow \infty$  as  $d \rightarrow \infty$ .*

Note that if the component variables are *i.i.d.* with  $E\psi^2(|X^{(1)} - Z^{(1)}|) < \infty$ , then  $\vartheta(d) \asymp d^{1/2}$  (i.e.,  $\vartheta(d)$  and  $d^{1/2}$  are of the same asymptotic order). Under appropriate moment condition, we have the same asymptotic order of  $\vartheta(d)$  for  $m$ -dependent sequence of random variables as well. Also, under weak mixing conditions on the component variables, we have  $\vartheta(d) = \mathbf{o}(d)$  (see, e.g., [Lin and Lu](#)

1996, Chap. 2). In all such situations,  $d/\vartheta(d) \rightarrow \infty$ , and hence (A2) implies (A2°). Theorem 3 shows the perfect clustering property of  $\text{AvgL}(h, \psi)$  and  $\text{kM}(h, \psi)$  under this weaker assumption (A2°) when  $h$  is Lipschitz continuous.

**Theorem 3.** *Suppose that we have independent observations from  $k_0$  populations satisfying (A2°). Also assume that  $h$  is Lipschitz continuous and  $\psi'(t)/t$  is a non-constant monotone function. Then, Rand indices of  $\text{AvgL}(h, \psi)$  and  $\text{kM}(h, \psi)$  converge to 0 as  $d$  tends to infinity.*

If  $\mathbf{X}$  and  $\mathbf{Y}$  are two independent observations from the  $i$ -th and the  $j$ -th populations, under Lipschitz continuity of  $h$ , we have  $\rho_{h, \psi}(\mathbf{X}, \mathbf{Y}) = \rho_{h, \psi}^*(i, j) + \mathbf{O}_p(\vartheta(d)/d)$  (see the proof of Theorem 3). While  $\vartheta(d)/d$  can be interpreted as the order of stochastic variation (noise),  $\rho_{h, \psi}^*(i, j)$  can be viewed as the signal. Theorem 3 shows the perfect clustering property of  $\text{AvgL}(h, \psi)$  and  $\text{kM}(h, \psi)$  when this signal-to-noise ratio diverges. Similar results may hold even when  $h$  is not Lipschitz continuous. For instance, in the case of  $\rho_0$ , where  $h(t) = \sqrt{t}$  does not satisfy the Lipschitz condition, we have the following result.

**Theorem 4.** *Suppose that we have independent observations from  $k_0$  populations, where the  $i$ -th ( $i = 1, \dots, k_0$ ) population has mean  $\boldsymbol{\mu}_i$  and dispersion matrix  $\boldsymbol{\Sigma}_i$  that satisfies  $\liminf_{d \rightarrow \infty} \text{tr}(\boldsymbol{\Sigma}_i)/\vartheta(d) > 0$ . For every  $i \neq j$ , if  $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2/\vartheta(d) \rightarrow \infty$  and/or  $|\text{tr}(\boldsymbol{\Sigma}_i) - \text{tr}(\boldsymbol{\Sigma}_j)|/\vartheta(d) \rightarrow \infty$ , then Rand indices of  $\text{AvgL}_0$  and  $\text{kM}_0$  converge to 0 as  $d \rightarrow \infty$ .*

Therefore, if  $\vartheta(d) \asymp d^{1/2}$  (i.e., in cases of weak dependence among component variables), the perfect clustering property of  $\text{AvgL}_0$  and  $\text{kM}_0$  holds when  $d^{-1/2}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 \rightarrow \infty$  and/or  $d^{-1/2}|\text{tr}(\boldsymbol{\Sigma}_i) - \text{tr}(\boldsymbol{\Sigma}_j)| \rightarrow \infty$  as  $d \rightarrow \infty$ .

The spectral clustering algorithm of Shi and Malik (2000) also failed to perform well in Examples A and B considered in Section 1. Note that spectral clustering methods deal with an edge-weighted graph with nodes  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  and a symmetric weight matrix  $\mathcal{W} = ((w_{ij}))_{d \times d}$ , where  $w_{ij}$  represents similarity between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . The matrix  $\mathcal{W}$  is usually computed from a similarity matrix  $\mathcal{S}$ , and different methods are available for it (see von Luxburg 2007). Often  $\mathcal{S} = ((s_{ij}))$  itself is used as  $\mathcal{W}$ , and one popular choice is the radial basis function  $s_{ij} = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\}$ , where  $\sigma$  is a tuning parameter that controls the degree of similarity. These algorithms implicitly assume that  $s_{ij}$  will be large (respectively, small) if  $\mathbf{x}_i$  and  $\mathbf{x}_j$  belong to the same population (respectively, different populations). Since that was not the case in Examples A and B, Spect had poor performance.

However, we do not have this problem if  $s_{ij}$  is defined using  $\rho_{h,\psi}^2(\mathbf{x}_i, \mathbf{x}_j)$  instead of  $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ . So,  $\text{Spect}(h, \psi)$ , spectral clustering based on  $\rho_{h,\psi}$ , is expected to perform well, especially for large values of  $d$ . We observed the same for  $\text{Spect}_0$  (spectral clustering based on  $\rho_0$ ) in Examples A and B.

MDP clustering algorithm (Ahn et al. 2012) largely depends on the data piling property, which occurs only when the dimension exceeds the sample size. So, as expected, it performed poorly in both examples for smaller values of  $d$ . Surprisingly, it failed in Example B even when  $d$  was large. A simple investigation explains this artifact. MDP clustering algorithm estimates the clusters by using binary splits at each step. For observations from two populations with mean vectors  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$  and dispersion matrices  $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$  satisfying  $d^{-1}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \rightarrow \nu_{12}$ ,  $d^{-1}\text{trace}(\boldsymbol{\Sigma}_1) \rightarrow \sigma_1^2$  and  $d^{-1}\text{trace}(\boldsymbol{\Sigma}_2) \rightarrow \sigma_2^2$  as  $d \rightarrow \infty$ , this algorithm perfectly separates the observations in the HDLSS set up when

$$\nu_{12} + \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} > \min \left\{ \frac{G + n_1}{Gn_1} \sigma_1^2, \frac{G + n_2}{Gn_2} \sigma_2^2 \right\}, \quad (3)$$

where  $G$  is a pre-specified minimum number of observations in a cluster. Following Ahn et al. (2012), we used  $G = 5$  throughout this article. Recall that in Example B, all three populations had the same location, and condition (3) was violated for each pair of populations.

### 2.3 Comparison of clustering algorithms using simulated datasets

We analyzed some simulated data sets for further evaluation of different clustering algorithms. In each example, we generated the data set by taking 50 observations from each population, and different algorithms were used on these data sets assuming the number of clusters to be known. For these examples, we considered  $d = 100, 200$  and  $500$ , and each experiment was repeated 100 times. Average Rand indices of different algorithms were computed over these 100 trials, and they are reported in Tables 1 and 2. MDP clustering needs the number of eigen-vectors  $T$  to be specified. We used  $T = 1, 2, 3$  as in Ahn et al. (2012), and reported the best results. For MADD, we used  $h(t) = \sqrt{t}, \psi(t) = t^2$ ;  $h(t) = t, \psi(t) = t$ ; and  $h(t) = t, \psi(t) = 1 - e^{-t}$ . The MADD indices for these three cases will be denoted by  $\rho_0, \rho_1$  and  $\rho_2$ , respectively. These three choices led to similar results in Examples 1–6 (descriptions are given below). So, in Table 1, results are reported for  $\rho_0$  only.

**Example-1:** Observations were generated from three Gaussian distributions with the same scatter matrix  $\boldsymbol{\Sigma}_d^\circ = ((0.5^{|i-j|}))_{d \times d}$  but different mean vectors  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$  and  $\boldsymbol{\mu}_3$ , respectively. We took  $\boldsymbol{\mu}_1 = \mathbf{0}_d$ , while  $\boldsymbol{\mu}_2$  (respectively,  $\boldsymbol{\mu}_3$ ) had the first  $d/2$  elements equal to 0.75 (respectively,  $-0.75$ ) and the rest equal to 0.

**Example-2:** We used observations from four normal distributions,  $\mathcal{N}_d(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_d^\circ)$ ,  $\mathcal{N}_d(\boldsymbol{\beta}, 4\boldsymbol{\Sigma}_d^\circ)$ ,  $\mathcal{N}_d(-\boldsymbol{\alpha}, \boldsymbol{\Sigma}_d^\circ)$  and  $\mathcal{N}_d(-\boldsymbol{\beta}, 4\boldsymbol{\Sigma}_d^\circ)$ , which differed in their locations and scales. The mean vector  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)^\top$  had  $\alpha_i = 1$  and  $\alpha_i = 0.5$  for even and odd values of  $i$ , respectively. We took  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top$  with  $\beta_i = (-1)^i \alpha_i$  for  $i = 1, \dots, d$ , and  $\boldsymbol{\Sigma}_d^\circ$  as in Example-1.

**Example-3:** We considered three uniform distributions with disjoint supports  $S_1$ ,  $S_2$  and  $S_3$ , where  $S_i = \{\mathbf{x} \in \mathbb{R}^d : i - 1 \leq \mathbf{x}'\boldsymbol{\Sigma}_d^{\circ-1}\mathbf{x} \leq i - 1/2\}$  for  $i = 1, 2, 3$ , and  $\boldsymbol{\Sigma}_d^\circ$  is as in Example-1. Figure 3(a) shows the supports of the three distributions for  $d = 2$ .

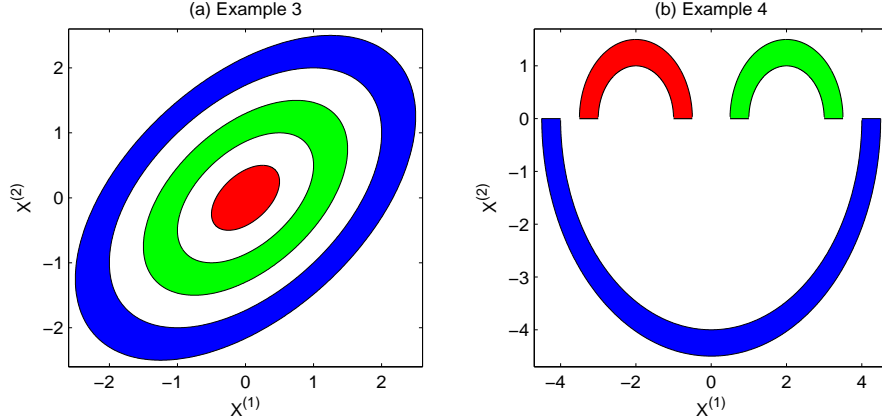


Figure 3: Different populations in Examples 3 and 4 when  $d = 2$ .

**Example-4:** Define three sets  $S_1^\circ = \{(x, y) : y \geq 0, 1 \leq \sqrt{(x-2)^2 + y^2} \leq 1.5\}$ ,  $S_2^\circ = \{(x, y) : y \geq 0, 1 \leq \sqrt{(x+2)^2 + y^2} \leq 1.5\}$  and  $S_3^\circ = \{(x, y) : y \leq 0, 4 \leq \sqrt{x^2 + y^2} \leq 4.5\}$  (see Figure 3(b)). We generated  $d/2$  independent observations from the uniform distribution on  $S_i^\circ$  to get  $d$  components of an observation from the  $i$ -th population ( $i = 1, 2, 3$ ).

**Example-5:** Observations were generated from two auto-regressive processes  $X^{(t)} = 0.75 + 0.25X^{(t-1)} + \varepsilon_t$  and  $X^{(t)} = 0.25 + 0.75X^{(t-1)} + \varepsilon_t$  for  $t = 1, \dots, d$ . In both cases, we had  $\varepsilon_t \sim \mathcal{N}(0, 1)$  for every  $t$ . The distribution of  $X^{(0)}$  was taken to be  $\mathcal{N}(1, 16/15)$  and  $\mathcal{N}(1, 16/7)$  in these two cases to make the processes stationary.

**Example-6:** Let  $\mathcal{S}_d$  be the  $d$ -dimensional unit sphere with center at the origin, and  $\mathcal{C}_d$  be the largest hypercube inscribed in it. We considered two uniform distributions, one on  $\mathcal{S}_d$  and the other on  $\mathcal{C}_d$ . Note that if  $\mathbf{X}$  comes from the first population, then  $\Pr(\mathbf{X} \in \mathcal{C}_d) \rightarrow 0$  as  $d \rightarrow \infty$ . So, the two populations become completely separated in high dimensions.

Table 1: Average Rand indices of different clustering algorithms in Examples 1–6

	$d$	AvgL	AvgL <sub>0</sub>	kM	kM <sub>0</sub>	MDP	Spect	Spect <sub>0</sub>
Ex-1	100	0.2906	0.0865	<b>0.0185</b>	0.0367	0.5801	0.2512	0.1851
	200	0.2168	0.0104	0.0201	0.0095	<b>0.0000</b>	0.2419	0.1953
	500	0.0429	<b>0.0000</b>	0.0074	<b>0.0000</b>	<b>0.0000</b>	0.2330	0.1919
Ex-2	100	0.7335	0.0502	0.4206	<b>0.0067</b>	0.6204	0.2609	0.0415
	200	0.7361	0.0115	0.6206	<b>0.0001</b>	0.0714	0.2462	0.0440
	500	0.7378	<b>0.0000</b>	0.6982	<b>0.0000</b>	0.0018	0.2187	0.0434
Ex-3	100	0.6616	<b>0.0000</b>	0.6608	<b>0.0000</b>	0.5885	0.4492	0.2348
	200	0.6619	<b>0.0000</b>	0.6617	<b>0.0000</b>	0.5826	0.4513	0.2298
	500	0.6619	<b>0.0000</b>	0.6619	<b>0.0000</b>	0.5761	0.4515	0.2286
Ex-4	100	0.2346	<b>0.0000</b>	0.2762	<b>0.0000</b>	0.5841	0.0745	0.0417
	200	0.2330	<b>0.0000</b>	0.2769	<b>0.0000</b>	<b>0.0000</b>	0.0714	0.0361
	500	0.2327	<b>0.0000</b>	0.2748	<b>0.0000</b>	<b>0.0000</b>	0.0752	0.0478
Ex-5	100	0.5047	0.3516	0.5027	<b>0.2271</b>	0.4895	0.4801	0.2584
	200	0.5048	<b>0.0762</b>	0.5040	0.0784	0.4863	0.4813	0.1231
	500	0.5048	<b>0.0028</b>	0.5048	0.0060	0.4795	0.4726	0.0119
Ex-6	100	0.5047	<b>0.0000</b>	0.5042	<b>0.0000</b>	0.4857	0.5000	<b>0.0000</b>
	200	0.5048	<b>0.0000</b>	0.5048	<b>0.0000</b>	0.4836	0.4992	<b>0.0000</b>
	500	0.5048	<b>0.0000</b>	0.5048	<b>0.0000</b>	0.4803	0.4992	<b>0.0000</b>

Bold figures indicate the best result in each example.

Table 1 clearly shows that both AvgL<sub>0</sub> and kM<sub>0</sub> performed much better than AvgL and kM. In all six examples, they led to perfect clustering, especially for higher values of  $d$ . MDP clustering algorithm performed poorly for  $d = 100$ . For  $d = 200$  and 500, it performed well in Examples 1, 2 and 4, but in the other three examples, where the population distributions did not differ in their means, it had miserable performance. Spect<sub>0</sub> also performed better than Spect. In Example-6, when all other clustering algorithms had Rand indices close to 0.5, those based on  $\rho_0$  led to perfect clustering for all values of  $d$  considered here. Among them, overall performance of AvgL<sub>0</sub> and kM<sub>0</sub> was much better than Spect<sub>0</sub>.

Next, we considered two examples, where clustering based on  $\rho_0$ ,  $\rho_1$  and  $\rho_2$  led to widely varying results (see Table 2). Descriptions of these two data sets are given below.

**Example-7:** Observations were generated from four normal distributions having the same mean  $\mathbf{0}_d$  and diagonal dispersion matrices. For the first (respectively, second) population, the first  $d/2$  diagonal elements were 1 (respectively, 9) and the rest were 9 (respectively, 1). The scatter matrix of the third (respectively, fourth) population had 1 and 9 (respectively, 9 and 1) at even and odd places along the diagonal, respectively.

**Example-8:** We considered two populations where all the measurement variables were *i.i.d.* For the first population, they were distributed as  $\mathcal{N}(0,3)$ , while they had standard  $t_3$  ( $t$  with 3 d.f) distribution for the second population. So, the two populations had the same mean vector and dispersion matrix, but they differed in their shapes.

Table 2: Average Rand indices of different clustering algorithms in Examples 7 and 8

	$d$	AvgL	AvgL <sub>0</sub>	AvgL <sub>1</sub>	AvgL <sub>2</sub>	kM	kM <sub>0</sub>	kM <sub>1</sub>	kM <sub>2</sub>	MDP	Spect	Spect <sub>0</sub>	Spect <sub>1</sub>	Spect <sub>2</sub>
Ex-7	100	0.7366	0.4831	0.4914	0.0044	0.4432	0.4102	0.2721	<b>0.0001</b>	0.5868	0.7034	0.3765	0.1732	0.0907
	200	0.7364	0.4873	0.3168	0.0001	0.4593	0.4082	0.0935	<b>0.0000</b>	0.6252	0.6990	0.3767	0.1310	0.0646
	500	0.7370	0.4776	0.0471	<b>0.0000</b>	0.4522	0.4048	0.0192	<b>0.0000</b>	0.6173	0.6975	0.3756	0.0903	0.0540
Ex-8	100	0.5048	0.5020	0.3883	0.1309	0.5048	0.4955	0.3132	<b>0.0845</b>	0.5021	0.5049	0.4894	0.3127	0.0956
	200	0.5048	0.5021	0.2837	0.0251	0.5048	0.4930	0.2087	<b>0.0157</b>	0.5027	0.5048	0.4801	0.2138	0.0188
	500	0.5049	0.5003	0.1109	0.0002	0.5049	0.4888	0.0889	<b>0.0000</b>	0.5029	0.5047	0.4818	0.0878	<b>0.0000</b>

Bold figures indicate the best result in each example.

Table 2 shows that AvgL, kM, MDP and Spect, all had miserable performance in these two examples. Even MADD clustering algorithms failed when  $\rho_0$  was used, but those based on  $\rho_1$  (denoted by AvgL<sub>1</sub>, kM<sub>1</sub> and Spect<sub>1</sub>) and  $\rho_2$  (denoted by AvgL<sub>2</sub>, kM<sub>2</sub> and Spect<sub>2</sub>) had improved performance. In these examples, we have  $\rho_{h,\psi}^*(i,j) = 0$  for all  $i \neq j$  when  $\rho_0$  is used, but they are positive for  $\rho_1$  and  $\rho_2$ . That was the reason for their improved performance. Among these two choices,  $\rho_2$ , which is based on a bounded  $\psi$  function, yielded better results. We also observed similar phenomenon when the  $t$  distribution in Example-8 was replaced by the standard Cauchy distribution. In that case, clustering algorithms based on  $\rho_2$  had Rand indices close to 0 for all choices of  $d$ , but those for all other methods were close to 0.5. This shows the robustness of MADD clustering algorithms based on bounded  $\psi$  functions against heavy tailed distributions.

### 3 Estimation of the number of clusters

So far we have assumed  $k_0$  to be known for our analysis. But in practice, one needs to estimate  $k_0$ . Several estimation methods have been proposed for it (see [Calinski and Harabasz 1974](#); [Hartigan 1975](#); [Krzanowski and Lai 1985](#); [Kaufman and Rousseeuw 1990](#); [Tibshirani et al. 2001](#); [Sugar and James 2003](#); [Wang 2010](#)). Brief descriptions of some of these methods, that we use in this article, are given below. These estimation methods can be used with any base clustering algorithm. Throughout this

article, we use average linkage or  $k$ -means algorithm (either based on Euclidean distance or based on MADD) for base clustering.

KL statistic (Krzanowski and Lai 1985): For a given  $k$ , if  $C_1, \dots, C_k$  are the  $k$  clusters estimated by the base clustering algorithm, then the KL statistic is defined as  $\text{KL}(k) = |\text{DIFF}(k)/\text{DIFF}(k+1)|$ , where  $\text{DIFF}(k) = (k-1)^{2/d}W_{k-1} - k^{2/d}W_k$ , and  $W_k = \sum_{j=1}^k (2|C_j|)^{-1} \sum_{\mathbf{z}, \mathbf{w} \in C_j} \|\mathbf{z} - \mathbf{w}\|^2$  is the within group sum of squares.  $\text{KL}(k)$  is computed for a range of values  $\{2, \dots, K\}$  of  $k$  and  $\hat{k}_{KL} = \text{argmax}_{2 \leq k \leq K} \text{KL}(k)$  is used to estimate  $k_0$ .

GAP statistic (Tibshirani et al. 2001): For any fixed  $k$ , the GAP statistic is defined as  $\text{GAP}(k) = B^{-1} \sum_{b=1}^B \log(W_k^{(b)}) - \log(W_k)$ , where  $W_k$  is as defined above, and  $W_k^{(b)}$  is the within group sum of squares computed using the  $b$ -th bootstrap sample ( $b = 1, \dots, B$ ) generated from a reference distribution. The number of clusters  $k_0$  is estimated by  $\hat{k}_G = \min\{k : \text{GAP}(k) \geq \text{GAP}(k+1) - s_{k+1}\}$ , where  $s_k = \sqrt{(1+B^{-1})sd_k}$ , and  $sd_k$  is the standard deviation of  $\log(W_k^{(b)})$ . Unlike the KL statistic,  $\text{GAP}(k)$  can be defined for  $k = 1$  as well.

JUMP statistic (Sugar and James 2003): For  $k \geq 1$ , the JUMP statistic is defined as  $\text{JUMP}(k) = \hat{d}_k^{-t} - \hat{d}_{k-1}^{-t}$ , where  $\hat{d}_0^{-t} = 0$ ,  $\hat{d}_k = d^{-1} \sum_{i=1}^n \min_{r=1, \dots, k} (\mathbf{x}_i - \mathbf{m}_r)^\top \mathbf{\Gamma}^{-1} (\mathbf{x}_i - \mathbf{m}_r)$  for  $k \geq 1$ , and  $\mathbf{m}_r$  is the center of the  $r$ -th cluster. The number of clusters is estimated by  $\hat{k}_J = \text{argmax}_{1 \leq k \leq K} \text{JUMP}(k)$ . The authors suggested to use  $\mathbf{\Gamma} = \mathbf{I}_d$  (the  $d \times d$  identity matrix) and  $t = d/2$ . Note that for  $k$ -means clustering with  $\mathbf{\Gamma} = \mathbf{I}_d$ , we get  $\hat{d}_k = d^{-1} \sum_{i=1}^n \min_{r=1, \dots, k} \|\mathbf{x}_i - \mathbf{m}_r\|^2 = d^{-1} \sum_{j=1}^k \sum_{\mathbf{z} \in C_j} \|\mathbf{z} - \mathbf{m}_j\|^2 = d^{-1} \sum_{j=1}^k (2|C_j|)^{-1} \sum_{\mathbf{z}, \mathbf{w} \in C_j} \|\mathbf{z} - \mathbf{w}\|^2 = W_k/d$ .

Cross-validated Rand index (Wang 2010): The whole sample  $\mathcal{X}$  is randomly divided into three parts  $\mathcal{X}^{(1)}, \mathcal{X}^{(2)}$  and  $\mathcal{X}^{(3)}$  of sizes  $m, m$  and  $n - 2m$ , respectively. For any given  $k$ , the first two parts are used to develop two clustering algorithms  $\delta_1 = \delta_{\mathcal{X}^{(1)}, k}$  and  $\delta_2 = \delta_{\mathcal{X}^{(2)}, k}$ , which are then used on  $\mathcal{X}^{(3)}$  to estimate clustering instability given by  $\text{INS}(k) = \binom{n-2m}{2}^{-1} \sum_{\mathbf{x} \neq \mathbf{y} \in \mathcal{X}^{(3)}} \mathbb{I} \left[ \mathbb{I}\{\delta_1(\mathbf{x}) = \delta_1(\mathbf{y})\} + \mathbb{I}\{\delta_2(\mathbf{x}) = \delta_2(\mathbf{y})\} = 1 \right]$ . This process is repeated  $B$  times, and the results are aggregated. The author proposed two methods for aggregation. In one method (call it  $\text{CV}_a$ ), the average instability over  $B$  repetitions is computed, and  $k_0$  is estimated by minimizing this average instability with respect to  $k$ . In the other method (call it  $\text{CV}_v$ ), for each repetition, the number of clusters is estimated by minimizing  $\text{INS}(k)$  over  $k$ , and finally the modal value of the minimizers is used as the estimator of  $k_0$ .

We use another method based on the DUNN index (Dunn 1973). For fixed  $k$ , let  $C_1, \dots, C_k$  be the clusters estimated by a base clustering algorithm. Define  $\Delta_0(C_i) = \{|C_i|(|C_i| - 1)\}^{-1} \sum_{\mathbf{z}, \mathbf{w} \in C_i} \|\mathbf{z} - \mathbf{w}\|$  and  $\Delta(C_i, C_j) = (|C_i||C_j|)^{-1} \sum_{\mathbf{z} \in C_i, \mathbf{w} \in C_j} \|\mathbf{z} - \mathbf{w}\|$  (instead of average, other suitable measures can also be used). The DUNN index is given by  $D(k) = B_k^\circ / W_k^\circ$ , where  $W_k^\circ = \max_{1 \leq i \leq k} \Delta_0(C_i)$  and  $B_k^\circ = \min_{1 \leq i < j \leq k} \Delta(C_i, C_j)$ , respectively. Dunn (1973) used this index for cluster validation, but here we use it to estimate  $k_0$  by  $\hat{k}_D = \operatorname{argmax}_{2 \leq k \leq K} D(k)$ .

When AvgL or kM is used for base clustering, we use usual versions of these statistics. But, when  $\text{AvgL}(h, \psi)$  or  $\text{kM}(h, \psi)$  is used for base clustering, we use  $\rho_{h, \psi}$  in place of  $\|\cdot\|$  to define  $W_k$  for KL, GAP and JUMP statistics, and to define  $\Delta_0, \Delta$  for the DUNN index.

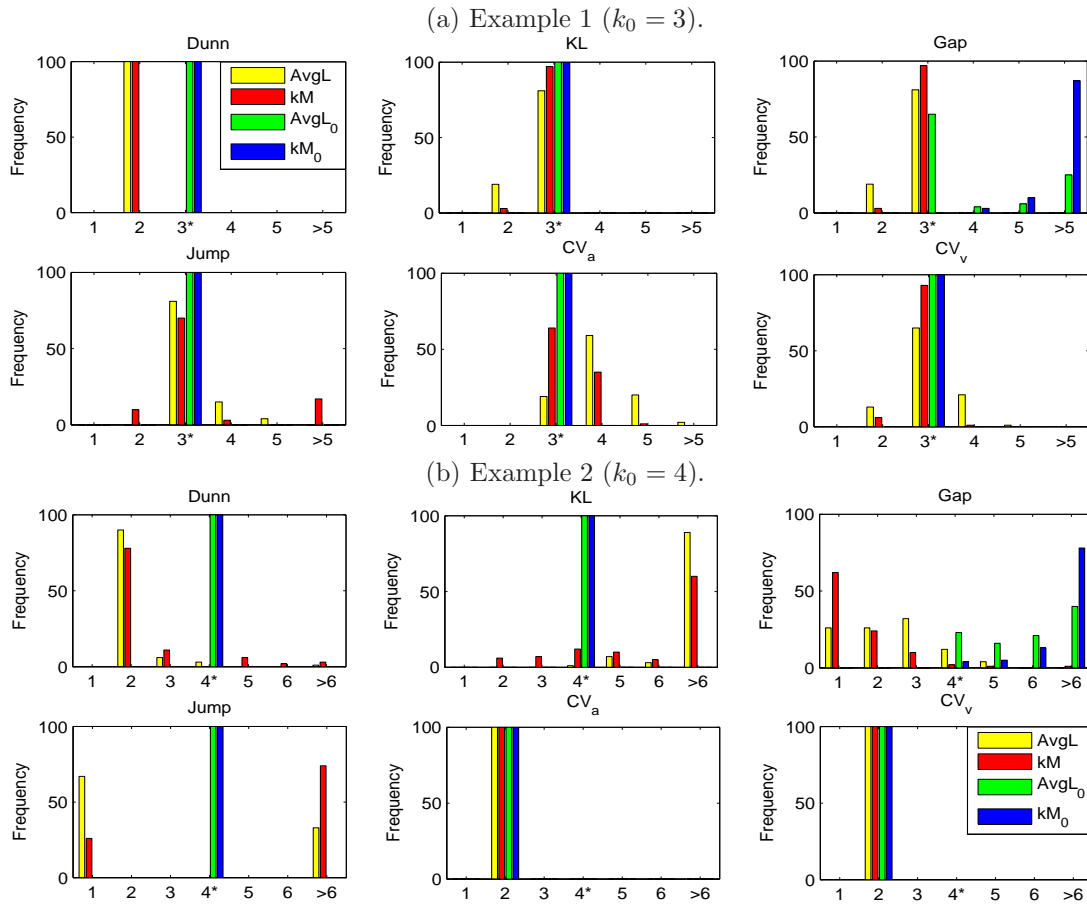


Figure 4: Barplots for  $k_0$  estimated by different methods in Examples 1 and 2.

To investigate the performance of these estimation methods, we considered the examples used in Section 2.3 with  $d = 500$ . For  $CV_a$  and  $CV_v$ , we used  $B = 100$ , and  $m$  was taken to be the largest

multiple of 5 not exceeding  $n/3$ . For the GAP statistic, we used  $B = 100$ , and bootstrap samples were generated from the uniform distribution on the range of the measurement variables. When  $d$  was large (in the order of hundreds or more), the use of  $t = d/2$  led to poor performance by the JUMP statistic. So, we tried several values of  $t$ , and based on our empirical experience, we selected  $t = 1$  when MADD was used. However, we were unable to find any such  $t$  when the Euclidean norm was used. In such cases, we performed our experiments with several choices of  $t$  and here we report the best results. Throughout this article, we consider values of  $k$  in the range  $\{1, \dots, 12\}$ . However, only GAP and JUMP statistics are defined for  $k = 1$ .

Figure 4 shows barplots for the number of clusters estimated by different methods in Examples 1 and 2. From this figure, it is clear that barring the GAP statistic, all other methods worked better when  $\rho_0$  was used. The results based on  $\rho_0, \rho_1$  and  $\rho_2$  were almost similar. We observed this phenomenon in Examples 3–6 as well. So, for reporting the performance of these methods in Examples 1–6, we considered the results based on  $\rho_0$  only (see Table 3). For the GAP statistic, there was no clear winner. For this method, we used both the Euclidean distance and  $\rho_0$  in all examples, and in each case, the best result Table 3 clearly shows that except for Examples 1 and 6, GAP statistic performed poorly throughout.  $CV_a$  and  $CV_v$  also underestimated  $k_0$  in Examples 2 and 3. But KL statistic, JUMP statistic and DUNN index correctly estimated  $k_0$  on all occasions.

Results for Examples 7 and 8 are given in Table 4. Since the performance of the GAP statistic was inferior to other methods, those results are not reported in Table 4. In these two examples, MADD versions of different methods did not have satisfactory performance when  $\rho_0$  was used, but those based on  $\rho_1$  and  $\rho_2$ , particularly the latter ones, had much improved performance. This is consistent with what we observed in Section 2.3.

The success of KL statistic, JUMP statistic and DUNN index in all examples motivated us to carry out a theoretical investigation regarding their high dimensional behavior. For this investigation, we make some assumptions on asymptotic orders of  $\rho_{h,\psi}$ . For two independent observations  $\mathbf{X}$  and  $\mathbf{Y}$  from  $i$ -th and  $j$ -th populations, let  $\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) \stackrel{P}{\asymp} \phi_{ij}(d)$ , i.e., as  $d \rightarrow \infty$ ,  $\Pr(\rho_{h,\psi}(\mathbf{X}, \mathbf{Y})/\phi_{ij}(d) \text{ remains bounded away from } 0 \text{ and } \infty) \rightarrow 1$ . Here we assume that

$$(A3) \quad \phi_{ii}(d) \asymp \phi_-(d) \text{ for every } i = 1, \dots, k_0 \text{ and } \phi_{ij}(d) \asymp \phi_+(d) \text{ for every } 1 \leq i \neq j \leq k_0, \text{ where } \phi_-(d) = \mathbf{o}(\phi_+(d)).$$

Table 3: Frequency distribution for the estimated number of clusters in Examples 1–6

	$k$	AvgL											kM												
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Ex-1	DUNN*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	KL*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	GAP <sup>?</sup>	0	19	<b>81</b>	0	0	0	0	0	0	0	0	0	0	3	<b>97</b>	0	0	0	0	0	0	0	0	0
	JUMP*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	CV <sub>a</sub> *	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	CV <sub>v</sub> *	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
Ex-2	DUNN*	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0
	KL*	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0
	GAP <sup>?</sup>	0	0	0	<b>23</b>	16	21	15	12	5	1	5	2	0	0	0	<b>4</b>	5	13	20	14	18	10	10	6
	JUMP*	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0
	CV <sub>a</sub> *	0	100	0	<b>0</b>	0	0	0	0	0	0	0	0	0	100	0	<b>0</b>	0	0	0	0	0	0	0	0
	CV <sub>v</sub> *	0	100	0	<b>0</b>	0	0	0	0	0	0	0	0	0	100	0	<b>0</b>	0	0	0	0	0	0	0	0
Ex-3	DUNN*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	KL*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	GAP <sup>?</sup>	0	0	<b>12</b>	13	41	18	13	3	0	0	0	0	0	0	<b>0</b>	2	20	35	23	18	1	1	0	0
	JUMP*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	CV <sub>a</sub> *	0	100	<b>0</b>	0	0	0	0	0	0	0	0	0	0	100	<b>0</b>	0	0	0	0	0	0	0	0	0
	CV <sub>v</sub> *	0	100	<b>0</b>	0	0	0	0	0	0	0	0	0	0	100	<b>0</b>	0	0	0	0	0	0	0	0	0
Ex-4	DUNN*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	KL*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	GAP <sup>?</sup>	0	20	<b>19</b>	10	15	9	9	4	9	1	2	2	0	3	<b>27</b>	23	24	9	4	6	2	0	1	1
	JUMP*	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0
	CV <sub>a</sub> *	0	9	<b>91</b>	0	0	0	0	0	0	0	0	0	0	9	<b>91</b>	0	0	0	0	0	0	0	0	0
	CV <sub>v</sub> *	0	9	<b>91</b>	0	0	0	0	0	0	0	0	0	0	9	<b>91</b>	0	0	0	0	0	0	0	0	0
Ex-5	DUNN*	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	KL*	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	GAP <sup>?</sup>	0	<b>19</b>	12	29	28	10	2	0	0	0	0	0	0	<b>0</b>	1	17	47	28	7	0	0	0	0	0
	JUMP*	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	CV <sub>a</sub> *	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	CV <sub>v</sub> *	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
Ex-6	DUNN*	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	KL*	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	GAP <sup>?</sup>	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	JUMP*	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	CV <sub>a</sub> *	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0
	CV <sub>v</sub> *	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0

Figures in bold indicate frequencies corresponding to  $k_0$ . \* Results obtained using methods based on  $\rho_0$ .  
<sup>?</sup> Both the Euclidean distance and  $\rho_0$  were used, and the best result is reported

If  $\mathbf{X}$  and  $\mathbf{Y}$  come from the same population, under (A1) we have  $\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) = \mathbf{o}_P(1)$ . So,  $\phi_-(d)$  should decrease to 0 as  $d$  increases. It also follows from (A1) and (A2) that if  $\mathbf{X}$  and  $\mathbf{Y}$  are from the  $i$ -th and the  $j$ -th populations ( $i \neq j$ ),  $|\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) - \rho_{h,\psi}^*(i, j)| \xrightarrow{P} 0$ , where  $\liminf_{d \rightarrow \infty} \rho_{h,\psi}^*(i, j) > 0$ . So,  $\phi_+(d)$  remains bounded away from 0 as  $d$  increases. Thus, (A3) holds trivially under (A1) and (A2). Note that under (A2<sup>o</sup>) also, we have  $\phi_-(d)/\phi_+(d) = \mathbf{O}(\vartheta(d)/d \rho_{h,\psi}^*(i, j)) = \mathbf{o}(1)$ .

Table 4: Frequency distribution for the estimated number of clusters in Examples 7 and 8

	$k$	$\rho_0$									$\rho_1$									$\rho_2$												
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	
Ex-7	AvgL	DUNN*	0	100	0	<b>0</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		PD*	58	42	0	<b>0</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		KL*	0	10	21	<b>21</b>	26	11	4	3	2	2	0	5	1	<b>58</b>	19	4	4	3	3	3	0	0	0	<b>100</b>	0	0	0	0	0	0
		JUMP*	100	0	0	<b>0</b>	0	0	0	0	0	0	0	0	0	<b>13</b>	11	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
		CV <sub>a</sub> *	0	1	1	<b>0</b>	2	0	2	2	3	89	0	1	0	<b>40</b>	55	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CV <sub>v</sub> *	0	38	5	<b>10</b>	10	8	2	1	0	26	0	13	0	<b>50</b>	36	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	kM	DUNN*	0	99	1	<b>0</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		PD*	0	100	0	<b>0</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		KL*	0	60	14	<b>4</b>	7	6	2	3	2	2	0	0	5	<b>63</b>	6	2	1	6	9	8	0	0	0	<b>100</b>	0	0	0	0	0	0
		JUMP*	100	0	0	<b>0</b>	0	0	0	0	0	0	0	0	0	<b>15</b>	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CV <sub>a</sub> *	0	32	8	<b>1</b>	2	3	2	5	1	46	0	2	0	<b>69</b>	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CV <sub>v</sub> *	0	74	21	<b>2</b>	3	0	0	0	0	0	0	18	1	<b>79</b>	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ex-8	AvgL	DUNN*	0	<b>85</b>	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		PD*	37	<b>63</b>	0	0	0	0	0	0	0	0	1	<b>99</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		KL*	0	<b>18</b>	17	14	22	14	8	5	1	1	0	<b>85</b>	12	1	0	1	1	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0
		JUMP*	0	<b>0</b>	16	26	32	13	6	4	2	1	0	<b>94</b>	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CV <sub>a</sub> *	0	<b>32</b>	0	0	0	0	0	1	3	4	60	0	<b>93</b>	2	0	0	1	0	0	3	1	0	0	<b>100</b>	0	0	0	0	0	0
		CV <sub>v</sub> *	0	<b>94</b>	2	0	0	0	0	0	0	0	4	0	<b>99</b>	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	kM	DUNN*	0	<b>78</b>	16	6	0	0	0	0	0	0	0	<b>99</b>	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		PD*	57	<b>42</b>	1	0	0	0	0	0	0	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		KL*	0	<b>22</b>	25	20	10	13	6	2	1	1	0	<b>96</b>	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		JUMP*	1	<b>2</b>	44	33	14	5	1	0	0	0	0	<b>96</b>	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CV <sub>a</sub> *	0	<b>15</b>	0	0	0	0	1	2	2	2	78	0	<b>95</b>	1	0	0	0	0	0	0	1	3	0	<b>100</b>	0	0	0	0	0	0
		CV <sub>v</sub> *	0	<b>83</b>	3	0	0	3	0	2	5	4	0	0	<b>100</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figures in bold indicate frequencies corresponding to  $k_0$ . \* Results are obtained using MADD versions.

Under this assumption, MADD versions of KL statistic, JUMP statistic and DUNN index have some nice properties in high dimensions if an appropriate base clustering algorithm is used. To make it clear what we mean by an appropriate base clustering algorithm, we now introduce the concept of perfect and order preserving (POP) clustering.

**Definition 1.** For any fixed  $k$ , let  $C_1^{(k)}, \dots, C_k^{(k)}$  be  $k$  clusters estimated using a clustering algorithm on  $\mathcal{X} = \cup_{i=1}^{k_0} \mathcal{X}_i$ , which consists of observations from  $k_0$  classes. We call the algorithm perfect and order preserving (POP) at  $k_0$  if the following conditions hold.

- (a) The clustering algorithm is perfect, i.e., for  $k = k_0$ ,  $C_i^{(k_0)} = \mathcal{X}_{\pi(i)}$  for every  $i = 1, \dots, k_0$  and some permutation  $\pi$  of  $\{1, \dots, k_0\}$ .
- (b) For any  $k < k_0$  and for every  $i = 1, \dots, k_0$ , there exists  $j \leq k$  such that  $C_i^{(k_0)} \subseteq C_j^{(k)}$ .
- (c) For any  $k > k_0$  and for every  $i = 1, \dots, k$ , there exists  $j \leq k_0$  such that  $C_i^{(k)} \subseteq C_j^{(k_0)}$ .

Figure 5 demonstrates a POP clustering (at 4) by taking only one value of  $k$  smaller than 4 and one value of  $k$  bigger than 4. But, one should notice that property (b) (respectively, property (c))

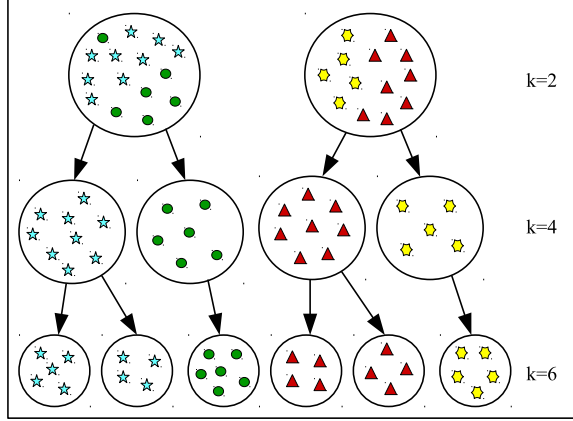


Figure 5: A clustering algorithm which is POP at 4.

has to hold for all  $k < 4$  (respectively,  $k > 4$ ). It is easy to check that any hierarchical algorithm is order preserving (i.e., satisfies (b) and (c)). So, if it leads to perfect clustering (i.e., satisfies (a)), it becomes POP at  $k_0$ . In Theorems 1–4, we have seen that  $\text{AvgL}(h, \psi)$  and  $\text{KM}(h, \psi)$  become perfect with probability tending to one as  $d$  tends to infinity. Using Lemmas 3–5, one can show that, they become POP at  $k_0$  with probability tending to one as the dimension diverges. Assuming that such a POP algorithm is used for base clustering, the following theorem shows the high dimensional behavior of estimators based on the MADD versions of DUNN index, KL statistic and JUMP statistic.

**Theorem 5.** *Suppose that there are observations from  $k_0 \geq 2$  populations which satisfy (A3), and also assume that the base clustering algorithm is POP at  $k_0$ .*

*Then (i)  $\hat{k}_D^* \xrightarrow{P} k_0$ , (ii)  $\hat{k}_{KL}^* \xrightarrow{P} k_0$  and (iii)  $\Pr(\hat{k}_J^* \geq k_0) \rightarrow 1$  as  $d \rightarrow \infty$ .*

*(Here  $\hat{k}_D^*$ ,  $\hat{k}_{KL}^*$  and  $\hat{k}_J^*$  are the number of clusters estimated by MADD versions of DUNN index, KL statistic and JUMP statistic (with  $t = 1$ ), respectively.)*

Theorem 5 shows the high dimensional consistency of  $\hat{k}_{KL}^*$  and  $\hat{k}_D^*$ . But since  $D(1)$  and  $KL(1)$  are not defined, they cannot detect the presence of a single cluster. JUMP statistic can be used in such situations, but Theorem 5 only shows that  $\Pr(\hat{k}_J^* \geq k_0) \rightarrow 1$  as  $d \rightarrow \infty$ . So, it can overestimate  $k_0$  in some cases. To overcome these limitations, we define a penalized version of DUNN index (PD). For any fixed  $k$ , it is given by  $\text{PD}(k) = B_k^\circ / W_k^\circ - k\zeta(d)$ , where  $B_k^\circ$  (for  $k \geq 2$ ) and  $W_k^\circ$  have the same meaning as in the DUNN index,  $B_1^\circ \stackrel{\text{def}}{=} B_2^\circ$  and  $\zeta$  is the penalty function. We estimate  $k_0$  by maximizing  $\text{PD}(k)$  with respect to  $k$  and denote it by  $\hat{k}_{PD}$  ( $\hat{k}_{PD}^*$  when MADD versions are used). The following theorem shows the high dimensional consistency of  $\hat{k}_{PD}^*$  for suitable choices of  $\zeta$ .

**Theorem 6.** *Suppose that there are observations from  $k_0 \geq 1$  population(s), which satisfy (A3), and the penalty function  $\zeta(d) \rightarrow \infty$  in such a way that  $\phi_-(d)\zeta(d)/\phi_+(d) \rightarrow 0$  as  $d \rightarrow \infty$ . If the base clustering algorithm is POP at  $k_0$ , then  $\hat{k}_{PD}^* \xrightarrow{P} k_0$  as  $d \rightarrow \infty$ .*

We have already seen that under (A1) and (A2), while  $\phi_+(d)$  remains bounded away from 0,  $\phi_-(d)$  converges to 0 as  $d$  increases. So, if we assume  $\phi_-(d) = \mathbf{O}(d^{-\alpha_0})$  for some  $\alpha_0 > 0$ , one can use any  $\zeta$  such that  $1/\zeta(d)$  decreases to zero at a slower rate than  $\mathbf{O}(d^{-\alpha_0})$ . For instance, one can use  $\zeta(d) = \lambda \log(d)$  for a suitable choice of the parameter  $\lambda$ . Some sufficient conditions for  $\phi_-(d) = \mathbf{O}(d^{-\alpha_0})$  are given in the Appendix (see Lemma 6 and the remark after the proof of Lemma 6). Throughout this article, we used  $\zeta(d) = \lambda \log(d)$ , where  $\lambda = 0.015$  was chosen based on our empirical experience. This choice of  $\zeta$  worked well in all simulated and real data sets we analyzed in this article.

In Examples 1–6,  $\hat{k}_{PD}^*$  had same results as obtained using  $\hat{k}_D^*$ . Use of  $\rho_0$ ,  $\rho_1$  and  $\rho_2$  yielded similar results in these examples. In Examples 7 and 8, however, the MADD version of PD did not have satisfactory results when  $\rho_0$  was used. In many cases, it failed to identify the underlying clusters, and  $\hat{k}_{PD}^*$  turned out to be 1. Using  $\rho_1$  and  $\rho_2$ , we got better results in these two examples (see the results corresponding to PD\* in Table 4). Among these two choices, the latter one yielded better results. For further evaluation of the performance of  $\hat{k}_{PD}^*$ , we generated 100 observations from a uniform distribution on the 500-dimensional unit hypercube, and repeated the experiment 100 times. In all these 100 cases, it successfully identified the presence of a single cluster in the data set for all three choices of  $\rho_{h,\psi}$ .

Note that Theorems 5 and 6 show the consistency of  $\hat{k}_{KL}^*$ ,  $\hat{k}_D^*$  and  $\hat{k}_{PD}^*$  when all within cluster separations are of the same asymptotic order and so are the between cluster separations, i.e.,  $\phi_{ii}(d) \asymp \phi_-(d)$  for all  $i$  and  $\phi_{ij}(d) \asymp \phi_+(d)$  for all  $i \neq j$ . If that is not the case but  $\max_i \phi_{ii}(d) = \mathbf{o}(\min_{i \neq j} \phi_{ij}(d))$ , these methods may detect  $k'_0 (< k_0)$  super-clusters in the data, each consisting of one or more clusters (can be proved using similar arguments as used in the proofs of Theorems 5 and 6). In that case, instead of stopping after one step, we need to repeat the algorithm on each of the estimated super-clusters. One can use the penalized DUNN index (with appropriate penalty function) for this purpose and stop splitting a super-cluster when  $\hat{k}_{PD}^*$  turns out to be 1. One can check that this repetitive use of PD consistently estimates  $k_0$ . However, we did not use this repetitive method in this article.

## 4 Analysis of benchmark data sets

We analyzed two benchmark data sets, ‘Lymphoma’ data and ‘Control Chart’ data, for further evaluation of our proposed methods. Lymphoma data set was first analyzed by [Alizadeh et al. \(2000\)](#) for identification of distinct types of lymphoma, and it is available in the R package `spls`. Control Chart data set can be obtained from the [UCI Machine Learning Repository \(https://archive.ics.uci.edu/ml/datasets.html\)](https://archive.ics.uci.edu/ml/datasets.html).

### 4.1 Lymphoma data

This data set contains expression levels of 4026 genes for 42 diffuse large B-cell lymphoma (DLBCL), 9 follicular lymphoma (FL) and 11 chronic lymphocytic leukemia (CLL) samples. A plot of these 62 observations is given in Figure 6.

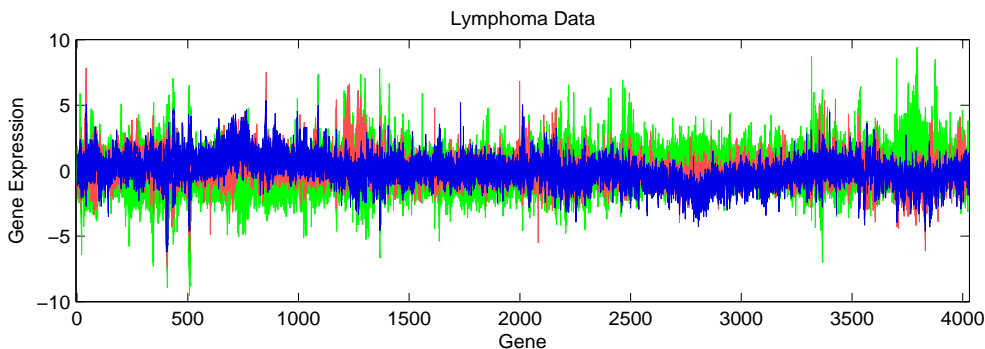


Figure 6: Gene expression of 4026 genes for DLBCL(■), FL(■) and CLL(■).

We used different methods to estimate  $k_0$  and the results are given in Table 5. This table shows that all methods, except the GAP statistic, identified two clusters. When we used different clustering algorithms to estimate these two clusters, all of them put almost all DLBCL samples in one cluster and the rest in another cluster (see Figure 7(a)). This indicates that it is very hard to distinguish between FL and CLL samples, which can be seen in Figure 6 as well. This claim is also justified by the behavior of FL, which can sometimes present itself as CLL (see [https://en.wikipedia.org/wiki/B-cell\\_chronic\\_lymphocytic\\_leukemia](https://en.wikipedia.org/wiki/B-cell_chronic_lymphocytic_leukemia)).

Since it was known that the observations were actually from three populations, we used different clustering algorithms to find three clusters in this data set as well. In Figure 7(b), one can see that both AvgL and kM failed to identify the three populations. But, AvgL<sub>0</sub> and kM<sub>0</sub> successfully differentiated between observations from FL and CLL classes. The method based on MDP led to

Table 5: Number of clusters estimated by different methods in ‘Lymphoma’ data

	DUNN	PD	KL	GAP	JUMP	$CV_a$	$CV_v$
AvgL	2	2	2	12	2	2	2
AvgL <sub>0</sub>	2	2	2	7	2	2	2
kM	2	2	2	12	2	2	2
kM <sub>0</sub>	2	2	2	7	2	2	2

perfect clustering, but spectral clustering algorithms did not perform well. Since all three choices of  $\rho_{h,\psi}$  (i.e.,  $\rho_0$ ,  $\rho_1$  and  $\rho_2$ ) led to similar results in this data set, here we have reported the results for MADD versions based on  $\rho_0$  only.

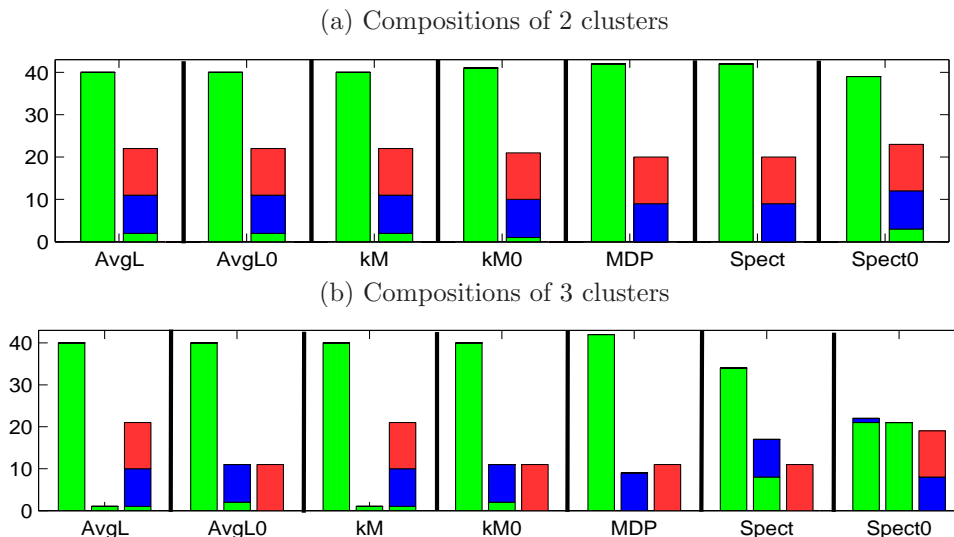


Figure 7: Compositions of (a) two and (b) three estimated clusters for Lymphoma data. Each bar corresponds to a single cluster consisting of DLBCL (■), FL (■) and CLL (■) samples

## 4.2 Control chart data

This data set contains 60 dimensional observations from 6 classes, viz., normal(N), cyclic(C), increasing trend(IT), decreasing trend(DT), upward shift(US) and downward shift(DS). We have 100 observations from each class. Figure 8 depicts a representation of the 6 classes.

DUNN index and PD could find only two clusters in this data set, but most of their MADD versions identified three or more clusters, as did most other methods (see Table 6). GAP statistic again overestimated  $k_0$ . JUMP statistic also overestimated  $k_0$  in some cases, but when  $\rho_2$  was used,  $\hat{k}_j^*$  turned out to be 1. It also turned out to be 6 in some cases, but those estimated clusters did not correspond to the six classes, as one can see from Figure 10.

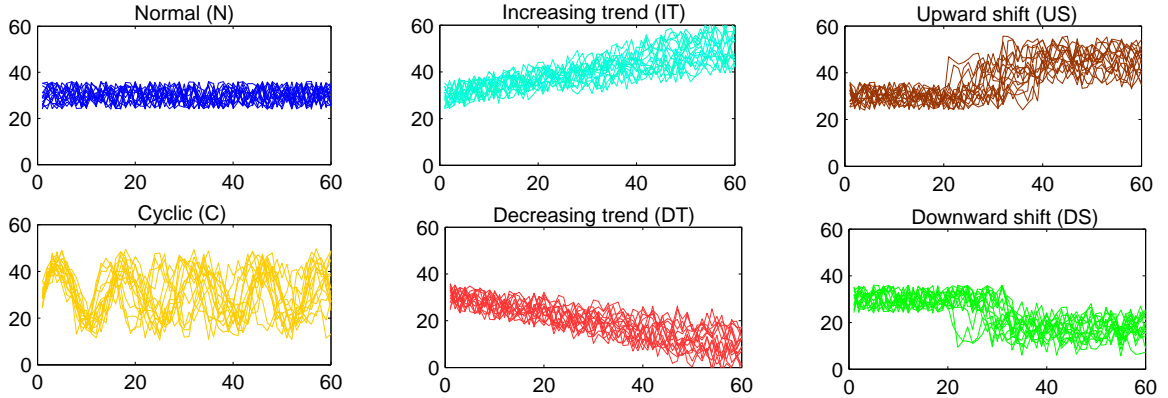


Figure 8: Six classes in ‘Control Chart’ data

Table 6: Number of clusters estimated by different methods in ‘Control Chart’ data

	DUNN	PD	KL	GAP	JUMP	$CV_a$	$CV_v$
AvgL	2	2	3	10	8	3	3
AvgL <sub>0</sub>	3	3	10	8	6	3	3
AvgL <sub>1</sub>	3	3	10	10	10	3	3
AvgL <sub>2</sub>	3	2	11	7	1	4	4
kM	2	2	3	9	3	3	3
kM <sub>0</sub>	3	3	3	10	10	3	2
kM <sub>1</sub>	3	3	3	10	6	3	3
kM <sub>2</sub>	4	2	6	7	1	4	4

Since most of the methods identified two or three clusters in this data set, at first we used different clustering algorithms for finding those two or three clusters. These results are shown in Figure 9. MDP clustering had poor performance in this example. Since the dimension was smaller than the sample size, it was quite expected in view of the results reported in Figure 1 and Tables 1–2. So, results for MDP clustering are not reported here. Results for  $\rho_0$  and  $\rho_1$  were almost similar, but those for  $\rho_2$  were somewhat different. So, we reported the results based on  $\rho_0$  and  $\rho_2$  only.

Figure 9(a) shows that when different clustering algorithms were used to divide the data set into two groups, most of them put the observations from classes IT and US in one cluster and the rest in the other cluster. Methods based on  $\rho_2$  led to different cluster formations. Spect put the observations from the class IT in a cluster and the rest in another cluster.

When we divided the data set into three clusters, AvgL, AvgL<sub>0</sub>, kM and kM<sub>0</sub> formed one cluster mainly consisting of N and C samples; one cluster mainly consisting of DT and DS samples, while the third cluster was formed mainly by IT and US samples as before (see Figure 9(b)). Again,  $\rho_2$

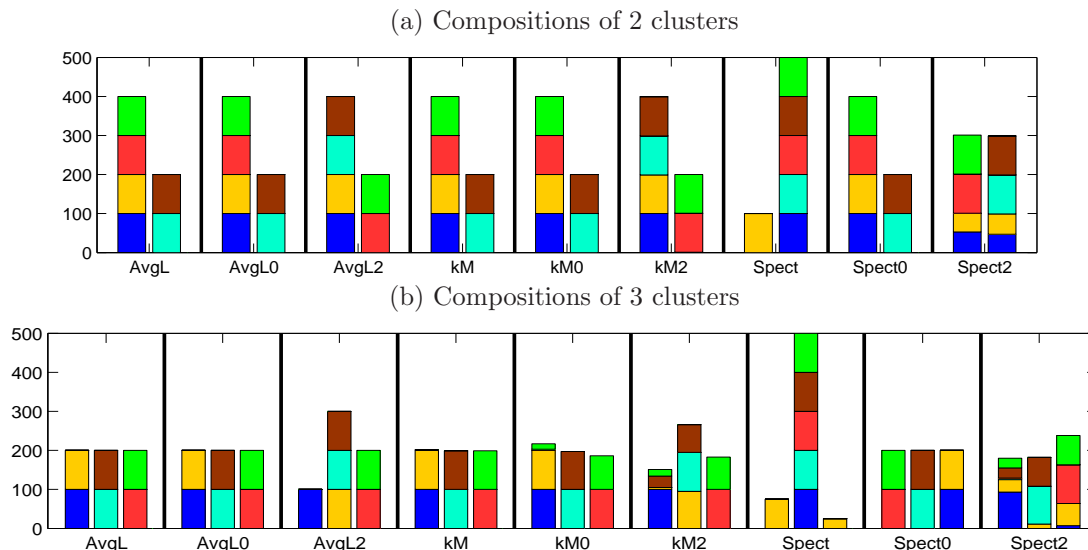


Figure 9: Compositions of (a) two and (b) three estimated clusters. Each bar corresponds to a single cluster, comprising of observations from different classes (normal(■), cyclic(■), increasing trend(■), decreasing trend(■), upward shift(■) and downward shift(■)).

led to slightly different formation of clusters. Performance of AvgL and AvgL<sub>0</sub> was slightly better than kM and kM<sub>0</sub>. Spect performed poorly, but Spect<sub>0</sub> performed much better. It led to the same clusters as obtained by AvgL and AvgL<sub>0</sub>.

Figure 10(a) shows the clusters estimated by different methods when the observations were divided into four clusters. In this case, AvgL and kM divided the cluster containing N and C samples to form two new clusters, one containing half of the C samples, and the other containing the rest. However, AvgL<sub>2</sub> and kM<sub>2</sub> successfully separated N and C samples. Performances of Spect and Spect<sub>0</sub> were similar, but Spect<sub>2</sub> yielded different results.

We also divided the data set into six clusters. In that situation, clustering algorithms based on MADD (both  $\rho_0$  and  $\rho_2$ ) performed better than their Euclidean counterparts (see Figure 10(b)). For instance, while AvgL and kM put many of the N and C samples in the same cluster, AvgL<sub>0</sub>, kM<sub>0</sub>, AvgL<sub>2</sub> and kM<sub>2</sub> successfully separated normal (N), cyclic (C), upward (IT, US) and downward (DT, DS) patterns. However, none of them could completely distinguish between IT and US samples or DT and DS samples. This is quite expected from the plot of the observations in Figure 8. Spectral clustering algorithms failed to have satisfactory performance in this case.

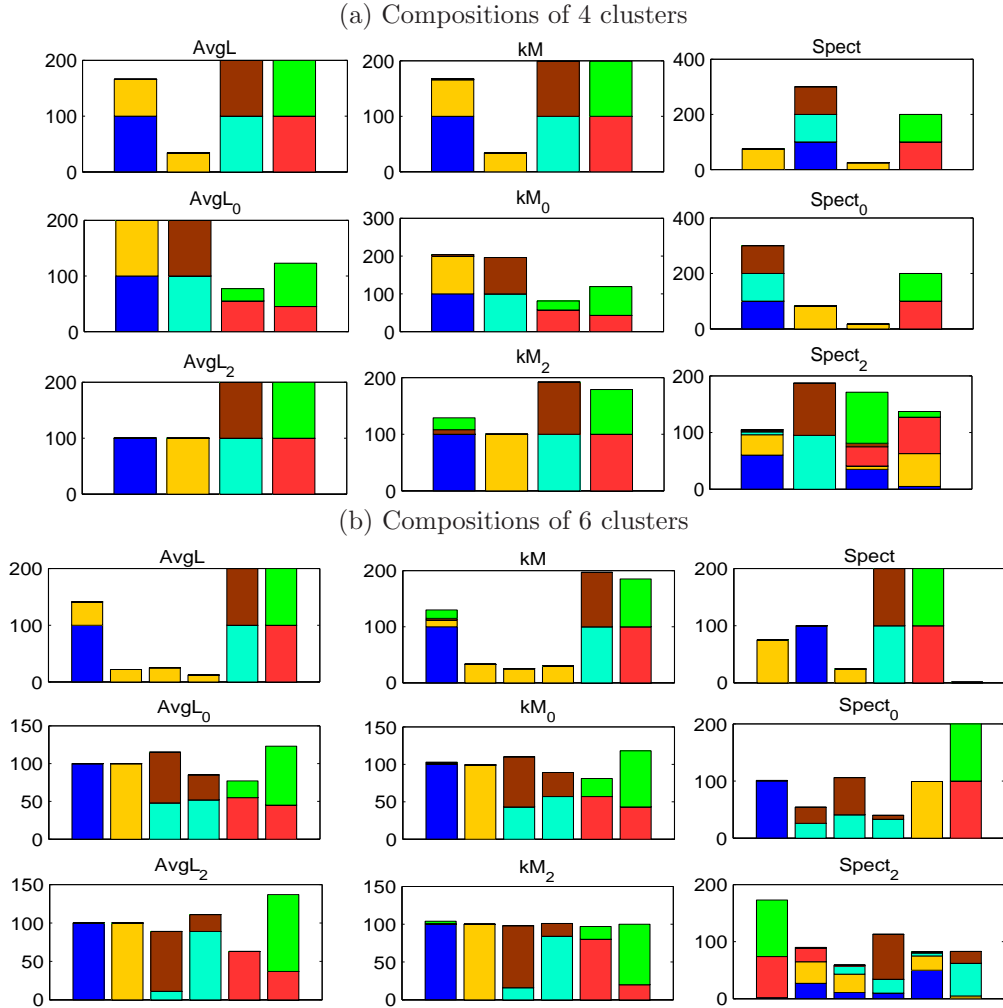


Figure 10: Compositions of (a) four and (b) six estimated clusters. Each bar corresponds to a single cluster, comprising of observations from different classes (normal(■), cyclic(■), increasing trend(■), decreasing trend(■), upward shift(■) and downward shift(■)).

## 5 Concluding remarks

In high dimensions, concentration of Euclidean distance often leads to poor performance by clustering algorithms based on it. In this article, we have used a data driven dissimilarity measure, called MADD, which takes care of this problem. Clustering algorithms based on MADD can lead to perfect clustering for HDLSS data even when those based on Euclidean distance perform miserably. We have amply demonstrated it in this article using theoretical as well as numerical results. While MDP clustering performs poorly for HDLSS data with populations not differing in their locations,

MADD based clustering algorithms can have excellent performance even when the populations have the same location and scale. Using suitable transformation  $\psi$  on each covariate, MADD is able to distinguish between populations with different marginal distributions. However, instead of applying  $\psi$  on each co-ordinate, one can divide  $\mathbf{X}$  into disjoint blocks  $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d_0)})^\top$  and define  $\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) = h\{\sum_{q=1}^{d_0} \psi(\|\mathbf{X}^{(q)} - \mathbf{Z}^{(q)}\|)\}$ . MADD can be defined accordingly. If the sizes of these blocks are uniformly bounded,  $\rho_{h,\psi}^*(i, j)$  turns out to be positive unless the  $i$ -th and the  $j$ -th populations have the same block distributions. Naturally one would like to have nearly independent blocks, but a suitable algorithm needs to be developed for this purpose.

For most of the data sets analyzed in this article, the spectral clustering algorithm of [Shi and Malik \(2000\)](#) also worked better when a MADD based similarity measure was used. We observed the same for the spectral clustering algorithm of [Ng et al. \(2002\)](#) as well, but to save space, we decided not to report them in this article. Throughout this article, we have used AvgL for hierarchical clustering. However, other linkage methods like single linkage, complete linkage, Ward’s linkage or centroid linkage (see, e.g., [Duda et al. 2012](#); [Johnson and Wichern 2014](#)) can also be used. One can prove the perfect clustering property of MADD versions of these linkage algorithms following the same line of arguments as used in the proofs of Theorems 1, 3 and 4.

We have also considered the problem of estimating the number of clusters and seen that the methods based on JUMP statistic and KL statistic usually perform better in high dimensions when their MADD versions are used. We have also successfully used MADD versions of DUNN index and penalized DUNN index for this purpose. Under appropriate regularity conditions, the methods based on KL statistic, DUNN index and penalized DUNN index turn out to be consistent in HDLSS asymptotic regime when AvgL( $h, \psi$ ) or kM( $h, \psi$ ) is used for base clustering. But, the choice of penalty function  $\zeta$  in the penalized DUNN index still remains an issue to be resolved. Throughout this article, we have used  $\zeta(d) = \lambda \log(d)$ , which was chosen based on our empirical experience. But, a suitable data driven choice of  $\zeta$  may further improve the empirical performance of different clustering algorithms.

## Acknowledgment

We are grateful to Dr. J. Ahn for providing us with the codes for MDP clustering.

## Appendix: Proofs and mathematical details

**Proof of Lemma 1:** Since  $\mathbf{X} \sim \mathcal{U}_d(a_1, b_1)$ , the distribution function of  $R = d^{-1/2}\|\mathbf{X}\|$  is given by  $F_R(r) = (r^d - a_1^d)/(b_1^d - a_1^d)$  for  $a_1 \leq r \leq b_1$ , 0 for  $r < a_1$  and 1 for  $r > b_1$ . So,  $R$  has the density  $f_R(r) = dr^{d-1}/(b_1^d - a_1^d)$  for  $a_1 \leq r \leq b_1$  and 0 otherwise. Therefore,

$$\begin{aligned} E(d^{-1}\|\mathbf{X}\|^2) &= E(R^2) = \frac{d}{d+2} \frac{b_1^{d+2} - a_1^{d+2}}{b_1^d - a_1^d} \rightarrow b_1^2, \text{ and} \\ E(d^{-2}\|\mathbf{X}\|^4) &= E(R^4) = \frac{d}{d+4} \frac{b_1^{d+4} - a_1^{d+4}}{b_1^d - a_1^d} \rightarrow b_1^4 \end{aligned} \quad (4)$$

as  $d \rightarrow \infty$ . This implies  $\text{Var}(d^{-1}\|\mathbf{X}\|^2) \rightarrow 0$  and hence  $d^{-1}\|\mathbf{X}\|^2 \xrightarrow{P} b_1^2$  as  $d \rightarrow \infty$ . Similarly, we have  $d^{-1}\|\mathbf{Y}\|^2 \xrightarrow{P} b_2^2$  as  $d \rightarrow \infty$ .

Now, it is enough to show that  $d^{-1}\langle \mathbf{X}, \mathbf{Y} \rangle \xrightarrow{P} 0$  as  $d \rightarrow \infty$ . Here  $\mathbf{X}$  and  $\mathbf{Y}$  are independent, and they are spherically symmetric about  $\mathbf{0}$  (see Fang et al. 1990). So, we have  $E(d^{-1}\langle \mathbf{X}, \mathbf{Y} \rangle) = d^{-1} \sum_{q=1}^d E(X^{(q)})E(Y^{(q)}) = 0$ , and  $E(X^{(q)}X^{(q')}) = E(Y^{(q)}Y^{(q')}) = 0$  for all  $q \neq q'$ . Therefore,  $\text{Var}(d^{-1}\langle \mathbf{X}, \mathbf{Y} \rangle) = E(d^{-2}\langle \mathbf{X}, \mathbf{Y} \rangle^2) = d^{-1}E(X^{(1)2})E(Y^{(1)2}) = d^{-1}E(d^{-1}\|\mathbf{X}\|^2)E(d^{-1}\|\mathbf{Y}\|^2)$ . We have proved that  $E(d^{-1}\|\mathbf{X}\|^2) \rightarrow b_1^2$  and  $E(d^{-1}\|\mathbf{Y}\|^2) \rightarrow b_2^2$  as  $d \rightarrow \infty$ . So,  $\text{Var}(d^{-1}\langle \mathbf{X}, \mathbf{Y} \rangle) \rightarrow 0$  and hence  $d^{-1}\langle \mathbf{X}, \mathbf{Y} \rangle \xrightarrow{P} 0$  as  $d \rightarrow \infty$ .  $\square$

**Proof of Lemma 2:** Non-negativity of  $\rho_{h,\psi}$  is obvious and symmetry comes from the fact that  $\varphi_{h,\psi}$  is symmetric. When  $n = 3$ , we get

$$\begin{aligned} |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_3) - \varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_3)| &= |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_3) - \varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_2) + \varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_2) - \varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_3)| \\ &\leq |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_2) - \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_2)| + |\varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_1) - \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_1)| \end{aligned}$$

When  $n \geq 4$ , for any  $k = 4, \dots, n$ , we get

$$\begin{aligned} |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_k)| &= |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_k) + \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_k)| \\ &\leq |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_k)| + |\varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_k)| \end{aligned}$$

Combining these two facts, we have

$$\begin{aligned} &\sum_{k \neq 1,2} |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_k)| \\ &\leq \sum_{k \neq 1,3} |\varphi_{h,\psi}(\mathbf{x}_1, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_k)| + \sum_{k \neq 2,3} |\varphi_{h,\psi}(\mathbf{x}_2, \mathbf{x}_k) - \varphi_{h,\psi}(\mathbf{x}_3, \mathbf{x}_k)| \end{aligned} \quad \square$$

**Proof of Lemma 3:** The proof follows from our discussion preceding the statement of the lemma. Hence it is omitted.

**Proof of Lemma 4:** If the  $i$ -th and the  $j$ -th populations have the same marginal distributions, then  $\varphi_{h,\psi}^*(i, \ell) = \varphi_{h,\psi}^*(j, \ell)$  for all  $\ell = 1, \dots, k_0$ . As a result, we have  $\rho_{h,\psi}^*(i, j) = 0$ .

For the only if part, first observe that  $\rho_{h,\psi}^*(i, j) \geq (n_i - 1)|\varphi_{h,\psi}^*(i, j) - \varphi_{h,\psi}^*(i, i)| + (n_j - 1)|\varphi_{h,\psi}^*(i, j) - \varphi_{h,\psi}^*(j, j)|$ . Now, if the right side is zero, we have  $\varphi_{h,\psi}^*(i, j) = \varphi_{h,\psi}^*(i, i)$  and  $\varphi_{h,\psi}^*(i, j) = \varphi_{h,\psi}^*(j, j)$ . Since  $h$  is a one-to-one function, this implies  $d^{-1} \sum_{q=1}^d \{2E\psi(|X_1^{(q)} - Y_1^{(q)}|) - E\psi(|X_1^{(q)} - X_2^{(q)}|) - E\psi(|Y_1^{(q)} - Y_2^{(q)}|)\} = 0$ , where  $\mathbf{X}_1, \mathbf{X}_2$  and  $\mathbf{Y}_1, \mathbf{Y}_2$  are independent observations from the  $i$ -th and the  $j$ -th populations, respectively. Now, since  $\psi'(t)/t$  is strictly monotone, each summand in the left side is positive, and it is zero if and only if the respective marginal distributions are equal (see [Baringhaus and Franz \(2010\)](#); [Biswas et al. \(2015\)](#)). Thus,  $\rho_{h,\psi}^*(i, j) = 0$  implies that the  $i$ -th and the  $j$ -th populations have the same marginals.  $\square$

**Proof of Lemma 5:** If  $\rho_{h,\psi}^*(i, j) = 0$ , then for  $\ell = 1, \dots, k_0$ , we have  $\varphi_{h,\psi}^*(i, \ell) = \varphi_{h,\psi}^*(j, \ell)$ . Now, for  $h(t) = \sqrt{t}$  and  $\psi(t) = t^2$ , we have  $\varphi_{h,\psi}^*(i, \ell) = d^{-1/2} \sqrt{tr(\boldsymbol{\Sigma}_i) + tr(\boldsymbol{\Sigma}_\ell) + \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell\|^2}$  and  $\varphi_{h,\psi}^*(j, \ell) = d^{-1/2} \sqrt{tr(\boldsymbol{\Sigma}_j) + tr(\boldsymbol{\Sigma}_\ell) + \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_\ell\|^2}$ . So,  $\rho_{h,\psi}^*(i, j) = 0$  if and only if  $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell\|^2 + tr(\boldsymbol{\Sigma}_i) = \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_\ell\|^2 + tr(\boldsymbol{\Sigma}_j)$  for every  $\ell = 1, \dots, k_0$ . Therefore, taking  $\ell = i$  and  $\ell = j$ , we get  $tr(\boldsymbol{\Sigma}_i) = tr(\boldsymbol{\Sigma}_j)$  and  $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| = 0$ . On the other hand, if  $tr(\boldsymbol{\Sigma}_i) = tr(\boldsymbol{\Sigma}_j)$  and  $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$ , it is easy to check that  $\rho_{h,\psi}^*(i, j) = 0$ .  $\square$

**Proof of Theorem 1:** From Lemma 3 and (A2), we have  $\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) \xrightarrow{P} 0$  when  $\mathbf{X}, \mathbf{Y}$  come from the same population, but when they are from different populations, we have  $\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) > 0$  for all but finitely many  $d$ . So, for every  $k$  and  $i \neq j$ , we get

$$\Pr\left(\max_{\mathbf{X}, \mathbf{Y} \in \mathcal{X}_k} \rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) < \min_{\mathbf{X} \in \mathcal{X}_i, \mathbf{Y} \in \mathcal{X}_j} \rho_{h,\psi}(\mathbf{X}, \mathbf{Y})\right) \rightarrow 1 \text{ as } d \rightarrow \infty. \quad (5)$$

Therefore, at the first step of  $\text{AvgL}(h, \psi)$ , two members of the same population merge together with probability converging to 1 as  $d \rightarrow \infty$ . Now at any step  $r$  ( $2 \leq r < n - k_0$ ), given that observations from the same population were merged together at each of the  $(r - 1)$  previous steps, any cluster  $C$  becomes a subset of  $\mathcal{X}_k$  for some  $k$ , and we have

$$\Pr\left(\max_k \max_{C, C' \subset \mathcal{X}_k} \Delta(C, C') < \min_{i \neq j} \min_{C \subset \mathcal{X}_i, C' \subset \mathcal{X}_j} \Delta(C, C')\right) \rightarrow 1 \text{ as } d \rightarrow \infty, \quad (6)$$

where  $\Delta(C, C') = (|C||C'|)^{-1} \sum_{\mathbf{X} \in C, \mathbf{Y} \in C'} \rho_{h,\psi}(\mathbf{X}, \mathbf{Y})$ . Therefore, two clusters containing observations from the same population will merge with probability tending to 1 as  $d \rightarrow \infty$ . Since  $k_0$  is known, these two facts together prove the result.  $\square$

**Proof of Theorem 2:** Note that for any  $k$ ,  $|C_k|^{-1} \sum_{\mathbf{Z} \in C_k, \mathbf{Z} \neq \mathbf{X}} \rho_{h,\psi}^2(\mathbf{X}, \mathbf{Z}) \xrightarrow{P} 0$  as  $d \rightarrow \infty$  if and only if  $\mathbf{X}$  and all observations in  $C_k$  are from the same population (follows from the proof of Theorem 1). So, if each  $C_k$  ( $k = 1, \dots, k_0$ ) contains observations from the same population,  $\Phi^*(C_1, \dots, C_{k_0}) \xrightarrow{P} 0$  as  $d \rightarrow \infty$ . Otherwise, we have  $\liminf_{d \rightarrow \infty} \Phi^*(C_1, \dots, C_{k_0}) > 0$  (follows from (A2)). So, when  $k_0$  is known, for the minimization of  $\Phi^*(C_1, \dots, C_{k_0})$ , each  $C_k$  must contain all observations from a single population with probability converging to one as the dimension increases. This proves the convergence of the Rand index to zero.  $\square$

**Proof of Theorem 3:** Let  $\mathbf{X}$  and  $\mathbf{Z}$  be independent observations from  $i$ -th and  $\ell$ -th populations ( $i, \ell = 1, \dots, k_0$ ), and define  $V_d = d^{-1} \sum_{q=1}^d \psi(|X^{(q)} - Z^{(q)}|)$ . Since  $(V_d - E(V_d))/\sqrt{\text{Var}(V_d)} = \mathbf{O}_P(1)$ , we have  $V_d - E(V_d) = \mathbf{O}_P(\vartheta(d)/d)$ . Since  $h$  is Lipschitz continuous, this implies  $|\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}^*(i, \ell)| = |h(V_d) - h(E(V_d))| \leq C_0 |V_d - E(V_d)| = \mathbf{O}_P(\vartheta(d)/d)$ . So, for an independent observation  $\mathbf{Y}$  from the  $j$ -th population, we get  $|\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}(\mathbf{Y}, \mathbf{Z})| = |\varphi_{h,\psi}^*(i, \ell) - \varphi_{h,\psi}^*(j, \ell)| + \mathbf{O}_P(\vartheta(d)/d)$  as  $d \rightarrow \infty$ . Since the number of observations is finite, we get  $\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) = \rho_{h,\psi}^*(i, j) + \mathbf{O}_P(\vartheta(d)/d)$ . Now, for all  $i = 1, \dots, k_0$ ,  $\rho_{h,\psi}^*(i, i) = 0$ , while for all  $i \neq j$ ,  $\rho_{h,\psi}^*(i, j)$  has asymptotic order higher than that of  $\vartheta(d)/d$ . Therefore, for  $\mathbf{X}, \mathbf{Y}$  from the same population and  $\mathbf{X}', \mathbf{Y}'$  from different populations we get  $\Pr(\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) < \rho_{h,\psi}(\mathbf{X}', \mathbf{Y}')) \rightarrow 1$  as  $d \rightarrow \infty$ . Now, the proof follows using the same line of arguments as used in the proofs Theorems 1 and 2.  $\square$

**Proof of Theorem 4:** Let  $\mathbf{X}$  and  $\mathbf{Z}$  be two independent observations from  $i$ -th and  $\ell$ -th populations ( $i, \ell = 1, \dots, k_0$ ). Note that for  $\rho_0$ , we use  $h(t) = \sqrt{t}$  and  $\psi(t) = t^2$ . Therefore, taking  $V_d = d^{-1} \sum_{q=1}^d (X^{(q)} - Z^{(q)})^2$ , we get  $\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}^*(i, \ell) = \sqrt{V_d} - \sqrt{E(V_d)} = (V_d - E(V_d))/(\sqrt{V_d} + \sqrt{E(V_d)})$ , where  $E(V_d) = d^{-1} \{ \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell\|^2 + \text{tr}(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_\ell) \} \geq d^{-1} \text{tr}(\boldsymbol{\Sigma}_i)$ . So,  $\sqrt{dE(V_d)/\vartheta(d)}$  remains bounded away from 0, and hence  $\sqrt{\vartheta(d)}/(\sqrt{dV_d} + \sqrt{dE(V_d)})$  remains bounded as  $d \rightarrow \infty$ . Now,  $(V_d - E(V_d))/\sqrt{\text{Var}(V_d)} = \mathbf{O}_p(1)$  implies  $(V_d - E(V_d)) = \mathbf{O}_p(\vartheta(d)/d)$ . Again,  $1/(\sqrt{V_d} + \sqrt{E(V_d)}) = \mathbf{O}_p(\sqrt{d/\vartheta(d)})$ . So,  $\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) = \varphi_{h,\psi}^*(i, \ell) + \mathbf{O}_P(\sqrt{\vartheta(d)/d})$ , and hence we have  $\rho_0(\mathbf{X}, \mathbf{Y}) = \rho_0^*(i, j) + \mathbf{O}_P(\sqrt{\vartheta(d)/d})$ . So, following the proof of Theorem 3, one can show that AvgL<sub>0</sub> and kM<sub>0</sub> will have the perfect clustering property if for every  $i \neq j$ ,  $\sqrt{d}\rho_0^*(i, j)/\sqrt{\vartheta(d)} \rightarrow \infty$  or  $d\rho_0^{*2}(i, j)/\vartheta(d) \rightarrow \infty$

as  $d \rightarrow \infty$ . Now, from the proof of Lemma 5, it follows that if  $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2/\vartheta(d) \rightarrow \infty$  and/or  $|tr(\boldsymbol{\Sigma}_i) - tr(\boldsymbol{\Sigma}_j)|/\vartheta(d) \rightarrow \infty$ , then  $d\rho_0^{*2}(i,j)/\vartheta(d) \rightarrow \infty$  as  $d \rightarrow \infty$ .  $\square$

All estimation methods that we discuss henceforth are based on  $\rho_{h,\psi}$ . So, we have  $W_k = \sum_{j=1}^k (2|C_j|)^{-1} \sum_{\mathbf{z}, \mathbf{w} \in C_j} \rho_{h,\psi}^2(\mathbf{z}, \mathbf{w})$ ,  $\Delta_0(C_i) = \{|C_i|(|C_i| - 1)\}^{-1} \sum_{\mathbf{z}, \mathbf{w} \in C_i} \rho_{h,\psi}(\mathbf{z}, \mathbf{w})$ , and  $\Delta(C_i, C_j) = (|C_i||C_j|)^{-1} \sum_{\mathbf{z} \in C_i, \mathbf{w} \in C_j} \rho_{h,\psi}(\mathbf{z}, \mathbf{w})$ .

**Proof of Theorem 5:** (i) Since the base clustering algorithm is POP at  $k_0$ , for any  $k < k_0$ , there exists at least one estimated cluster which contains observations from two different populations, and no two clusters contain observations from the same population. So, under (A3), we have  $\Delta_0(C_i) \stackrel{P}{\asymp} \phi_+(d)$  for some  $i$  and  $\Delta(C_i, C_j) \stackrel{P}{\asymp} \phi_+(d)$  for every  $i \neq j$ . Thus,  $B_k^\circ = \min_{1 \leq i < j \leq k} \Delta(C_i, C_j) \stackrel{P}{\asymp} \phi_+(d)$  and  $W_k^\circ = \max_{1 \leq i \leq k} \Delta_0(C_i) \stackrel{P}{\asymp} \phi_+(d)$ , and hence we get  $D(k) = B_k^\circ/W_k^\circ \stackrel{P}{\asymp} 1$ .

For  $k > k_0$ , no cluster contains observations from two different populations, while there exists at least two clusters which contain observations from the same population. So,  $\Delta_0(C_i) \stackrel{P}{\asymp} \phi_-(d)$  for every  $i$  and  $\Delta(C_i, C_j) \stackrel{P}{\asymp} \phi_-(d)$  for some  $i \neq j$ . Thus,  $B_k^\circ \stackrel{P}{\asymp} \phi_-(d)$ ,  $W_k^\circ \stackrel{P}{\asymp} \phi_-(d)$ , and hence  $D(k) \stackrel{P}{\asymp} 1$ .

For  $k = k_0$ , each cluster contains observations from same population and two different clusters contain observations from two different populations. This implies that  $\Delta_0(C_i) \stackrel{P}{\asymp} \phi_-(d)$  for every  $i$  and  $\Delta(C_i, C_j) \stackrel{P}{\asymp} \phi_+(d)$  for every  $i \neq j$ . So, we have  $B_k^\circ \stackrel{P}{\asymp} \phi_+(d)$  and  $W_k^\circ \stackrel{P}{\asymp} \phi_-(d)$  and hence  $D(k) \stackrel{P}{\asymp} (\phi_+(d)/\phi_-(d))$ .

Combining these three cases, and noting that  $\phi_-(d) = \mathbf{o}(\phi_+(d))$ , we get  $\Pr(D(k_0) > D(k) \forall k \neq k_0) \rightarrow 1$  as  $d \rightarrow \infty$ . This implies  $\hat{k}_D^* \xrightarrow{P} k_0$  as  $d \rightarrow \infty$ .

(ii) For all  $k \geq 2$ , the KL statistic can be written as  $KL(k) = |(\Lambda_{k-1} - \Lambda_k)/(\Lambda_k - \Lambda_{k+1})|$ , where  $\Lambda_k = k^{2/d}W_k$ . Since the base clustering algorithm is POP at  $k_0$ , from our discussion in part (i), it follows that  $W_k \stackrel{P}{\asymp} \phi_+(d)$  for  $k < k_0$  and  $W_k \stackrel{P}{\asymp} \phi_-(d)$  for  $k \geq k_0$ . Note that, for any fixed  $k$ ,  $k^{2/d} \rightarrow 1$  as  $d \rightarrow \infty$ . So, for all  $k \geq 1$ ,  $\Lambda_k \stackrel{P}{\asymp} W_k$ . Now, it is easy to check that  $KL(k_0) \stackrel{P}{\asymp} (\phi_+(d)/\phi_-(d))$  and  $KL(k) \stackrel{P}{\asymp} 1$  for all other values of  $k$ . Therefore,  $\Pr(KL(k_0) > KL(k) \forall k \neq k_0) \rightarrow 1$  as  $d \rightarrow \infty$ , and hence we have  $\hat{k}_{KL}^* \xrightarrow{P} k_0$  as  $d \rightarrow \infty$ .

(iii) Note that  $\hat{d}_k = W_k \stackrel{P}{\asymp} \phi_+(d)$  and  $\phi_-(d)$  for  $1 \leq k < k_0$  and  $k \geq k_0$ , respectively (follows from our discussion in parts (i) and (ii)). So, we have  $JUMP(k) = \hat{d}_k^{-1} - \hat{d}_{k-1}^{-1} \stackrel{P}{\asymp} 1/\phi_+(d)$  or  $1/\phi_-(d)$  according as  $k < k_0$  or  $k \geq k_0$ . This implies that  $\Pr(\hat{k}_J^* < k_0) \rightarrow 0$  as  $d \rightarrow \infty$ .  $\square$

**Proof of Theorem 6:** First consider the case  $k_0 \geq 2$ . While proving part (i) of Theorem 5, we have shown that in this case,  $D(k_0) \stackrel{P}{\asymp} (\phi_+(d)/\phi_-(d))$  and  $D(k) \stackrel{P}{\asymp} 1$  for all other choice of  $k \geq 2$ . Also, we have  $W_1^\circ \stackrel{P}{\asymp} \phi_+(d)$  and  $B_1^\circ \stackrel{def}{=} B_2^\circ \stackrel{P}{\asymp} \phi_+(d)$ , which implies  $D(1) \stackrel{def}{=} B_1^\circ/W_1^\circ \stackrel{P}{\asymp} 1$ . So, for any  $k \neq k_0$ ,  $PD(k_0) - PD(k) = D(k_0) - D(k) - (k_0 - k)\zeta(d) \xrightarrow{P} \infty$  as  $d \rightarrow \infty$  (since  $\zeta(d) = \mathbf{o}(\phi_+(d)/\phi_-(d))$ ). Thus,  $\hat{k}_{PD}^* \xrightarrow{P} k_0$  as  $d \rightarrow \infty$ .

When  $k_0 = 1$ , we have  $W_k^\circ \stackrel{P}{\asymp} \phi_-(d)$ ,  $B_k^\circ \stackrel{P}{\asymp} \phi_-(d)$  for every  $k \geq 1$ . So,  $PD(1) - PD(k) = D(1) - D(k) + (k - 1)\zeta(d) \xrightarrow{P} \infty$  as  $d \rightarrow \infty$  (since  $D(k) \stackrel{P}{\asymp} 1$  for  $k \geq 1$  and  $\zeta(d) \rightarrow \infty$  as  $d \rightarrow \infty$ ) and hence  $\hat{k}_{PD}^* \xrightarrow{P} k_0$ .  $\square$

**Lemma 6.** Let  $\mathbf{X}$  and  $\mathbf{Y}$  be two independent observations from the  $i$ -th population. For an independent observation  $\mathbf{Z}$  from the  $j$ -th population ( $j = 1, \dots, k_0$ ), assume that  $\text{Var}\{\sum_{q=1}^d \psi(|X^{(q)} - Z^{(q)}|)\} = \mathbf{O}(d^{2-\epsilon_0})$  for some  $\epsilon_0 > 0$ . If  $h$  is Hölder continuous with exponent  $\gamma$ , then  $\rho_{h,\psi}(\mathbf{X}, \mathbf{Y}) = \mathbf{O}_P(d^{-\alpha_0})$  with  $\alpha_0 = \gamma\epsilon_0/2$ .

**Proof:** Define  $V_d = d^{-1} \sum_{q=1}^d \psi(|X^{(q)} - Z^{(q)}|)$  and  $V'_d = d^{-1} \sum_{q=1}^d \psi(|Y^{(q)} - Z^{(q)}|)$ , for some  $\mathbf{Z} \neq \mathbf{X}, \mathbf{Y}$ . Note that  $V_d - V'_d = (V_d - EV_d) - (V'_d - EV'_d)$ . Now, write

$$V_d - EV_d = \frac{V_d - EV_d}{\sqrt{\text{Var}(V_d)}} \sqrt{\text{Var}(V_d)}.$$

The first term on the right side is  $\mathbf{O}_P(1)$ , and under the given condition, the second term is  $\mathbf{O}(d^{-\epsilon_0/2})$ . So, we have  $V_d - EV_d = \mathbf{O}_P(d^{-\epsilon_0/2})$ . Similarly, one gets  $V'_d - EV'_d = \mathbf{O}_P(d^{-\epsilon_0/2})$ . Thus,  $V_d - V'_d = \mathbf{O}_P(d^{-\epsilon_0/2})$ . Now, since  $h$  is Hölder continuous with exponent  $\gamma$ , we get  $|\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}(\mathbf{Y}, \mathbf{Z})| = |h(V_d) - h(V'_d)| \leq C_0 |V_d - V'_d|^\gamma = \mathbf{O}_P(d^{-\gamma\epsilon_0/2}) = \mathbf{O}_P(d^{-\alpha_0})$ . Since  $n$  is finite, this in turn proves that  $\sum_{\mathbf{Z} \neq \mathbf{X}, \mathbf{Y}} |\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}(\mathbf{Y}, \mathbf{Z})| = \mathbf{O}_P(d^{-\alpha_0})$ .  $\square$

**Remark 1.** For  $\rho_0$  (i.e.,  $h(t) = \sqrt{t}$  and  $\psi(t) = t^2$ ),  $\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}(\mathbf{Y}, \mathbf{Z}) = h(V_d) - h(V'_d) = (V_d - V'_d)(2\sqrt{\xi_d})^{-1}$ , where  $\xi_d$  lies between  $V_d$  and  $V'_d$ . Also,  $V_d = d^{-1} \|\mathbf{X} - \mathbf{Z}\|^2$  remains bounded away from 0 in probability (and so does  $V'_d$ ). Therefore,  $\xi_d^{-1/2} = \mathbf{O}_P(1)$ , and hence  $\varphi_{h,\psi}(\mathbf{X}, \mathbf{Z}) - \varphi_{h,\psi}(\mathbf{Y}, \mathbf{Z}) = \mathbf{O}_P(d^{-\epsilon_0/2})$ . So, the Hölder continuity of  $h$  is only sufficient, but not necessary.

## References

Ahn, J., Lee, M. H. and Yoon, Y. J. (2012) Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, **22**, 443–464.

- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, James, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
- Andrews, D. W. K. (1988) Laws of large numbers for dependent nonidentically distributed random variables. *Econometric Theory*, **4**, 458–467.
- Baringhaus, L. and Franz, C. (2010) Rigid motion invariant two-sample tests. *Statistica Sinica*, **20**, 1333–1361.
- Biswas, M., Mukhopadhyay, M. and Ghosh, A. K. (2014) A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, **101**, 913–926.
- (2015) On some exact distribution-free one-sample tests for high dimension low sample size data. *Statistica Sinica*, **25**, 1421–1435.
- Calinski, R. B. and Harabasz, J. (1974) A dendrite method of cluster analysis. *Communications in Statistics*, **3**, 1–27.
- Duda, R. O., Hart, P. E. and Stork, D. G. (2012) *Pattern Classification*. Wiley, New York.
- Dunn, J. C. (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, **3**, 32–57.
- Fang, K. T., Kotz, S. and Ng, K. W. (1990) *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London.
- Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B*, **67**, 427–444.
- Hartigan, J. A. (1975) *Cluster Algorithms*. Wiley, New York.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.
- Johnson, R. A. and Wichern, D. W. (2014) *Applied Multivariate Statistical Analysis*. Prentice-Hall, New Jersey.
- de Jong, R. M. (1995) Laws of large numbers for dependent heterogeneous processes. *Econometric Theory*, **11**, 347–358.
- Jung, S. and Marron, J. (2009) PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, **37**, 4104–4130.
- Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Krzanowski, W. J. and Lai, Y. T. (1985) A criterion for determining the number of clusters in a data set. *Biometrics*, **44**, 23–34.
- Lin, Z. and Lu, C. (1996) *Limit theory for mixing dependent random variables*, vol. 378 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht; Science Press Beijing, New York.
- von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and Computing*, **17**, 395–416.

- Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002) On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, **2**, 849–856.
- Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.
- Shi, J. and Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 888–905.
- Sugar, C. A. and James, G. M. (2003) Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, **98**, 750–763.
- Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B*, **63**, 411–423.
- Wang, J. (2010) Consistent selection of the number of clusters via crossvalidation. *Biometrika*, **97**, 893–904.