Journal of Multivariate Analysis 124 (2014) 465–479

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Asymptotics of hierarchical clustering for growing dimension

Petro Borysov [a,b,*], Jan Hannig [b], J.S. Marron [b]

[a] *SAS Institute, Cary, NC, 27513, USA*

[b] *Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC, 27599, USA*

ARTICLE INFO

Article history:
Received 13 November 2012
Available online 2 December 2013

AMS subject classification:
62H30

Keywords:
Hierarchical clustering
Linkage function
Clustering behavior

ABSTRACT

Modern day science presents many challenges to data analysts. Advances in data collection provide very large (number of observations and number of dimensions) data sets. In many areas of data analysis an informative task is to find natural separations of data into homogeneous groups, i.e. clusters. In this paper we study the asymptotic behavior of hierarchical clustering in situations where both sample size and dimension grow to infinity. We derive explicit signal vs noise boundaries between different types of clustering behaviors. We also show that the clustering behavior within the boundaries is the same across a wide spectrum of asymptotic settings.

1. Introduction

A major challenge to modern data analysis is the increasing commonality of very large (in both number of observations and number of dimensions) data sets. There are databases that contain thousands or even millions of observations. For example, in the field of drug discovery, such databases are created by modern automated methods called high-throughput screening that allow performance of biological assays of a large number of compounds. At the same time each chemical compound is represented by thousands of variables (descriptors). Another important example of data with high dimension comes from gene expression measurements. It is possible to capture gene expression levels for thousands of genes per subject. But the measurements made for each subject can be expensive which constrains the sample size to low hundreds. In this paper we use the term *sample size* to represent the number of observations (subjects) in the data, and the term *dimension* to represent the number of variables (measurements) for each observation.

In many areas of data analysis there exist natural separation of data in homogeneous groups, i.e. clusters. For example, one of the basic assumptions of Quantitative Structure–Activity Relationship modeling, which is one of the key computational drug discovery approaches, is that similar molecules have similar activities. Due to the nature of the synthesis and testing, many chemical compounds in the data set tend to belong to several local clusters in the chemistry space (for details see [21]).

In this paper we will study the behavior of the hierarchical clustering in situations where both sample size and dimension grow to infinity. We will show important insights about the well known differences between linkage functions and will gain new understanding as to how distances between clusters compare asymptotically. In particular we will study the behavior of the clusters in a wide array of contexts. We will develop a theory which explicitly finds signal vs noise boundaries between different types of clustering behaviors. Also we will show that the clustering behavior within the boundaries is the same across the wide spectrum of asymptotic settings.

The rest of the paper is organized as follows. Section 2 introduces hierarchical clustering and defines distances between clusters. Section 3 demonstrates the major points using a simple two cluster example for clarity. General asymptotic settings are introduced in Section 4. The main results with examples are presented in Section 5. Simulation study is in Section 6. Mathematical proofs are in the Appendix.

[*] Corresponding author at: SAS Institute, Cary, NC, 27513, USA.
E-mail addresses: pborysov@yahoo.com, peter.borysov@sas.com (P. Borysov).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.11.010

2. Clustering

The clustering process separates data into subsets (clusters) so that data points from the same cluster are similar to each other. There are two main types of clustering: hierarchical and partitional [20]. Partitional clustering methods simply divide a data set into non-overlapping subsets so that each data point will be in one of the subsets. One of the most popular partitional clustering procedures is the K-means clustering method [11]. It partitions data into a pre-specified number K of clusters such that each observation is assigned to the cluster with the nearest center. Hierarchical clustering [7] algorithms group the data by creating a cluster tree (dendrogram) based on the specific distances between points and similarity or dissimilarity measures. For a detailed discussion about distance functions and similarity measures see Section 2.1. In the rest of the paper the focus will be on hierarchical clustering.

A standard dendrogram consists of vertical and horizontal line segments that create a tree by connecting clusters into successively larger clusters. The height of each horizontal line segment is determined by the distance between clusters that are connected. Each level of the dendrogram provides a grouping of the data points into disjoint clusters. The dendrogram also provides useful graphical representation of the grouping hierarchy which highly increases interpretability. See [7] for a good review of hierarchical clustering and some examples of dendrograms.

2.1. Linkage functions

The result of hierarchical clustering is usually impacted by the choice of distance measure between the points and clusters. This choice will influence the clustering since some data points could be closer to each other if one distance measure is used, but farther away for another. There are many standard choices for a distance between points, such as L^1, L^2, and L^∞. In this paper we only consider the Euclidean (L^2) distance.

Once a distance between points is selected, distance between clusters is determined by the linkage function. In many cases it is a function of the pairwise distances between data points. After the distances between all clusters are computed, then the two closest clusters are merged together to form a bigger cluster. There are many linkage functions available, but in this paper we investigate three common linkage functions: single, average and Ward's.

Suppose X and Y are two clusters of size N_X and N_Y, $d(x, y)$ is the distance between data points $x \in X$ and $y \in Y$. Then the linkage functions can be defined as follows:

1. Single linkage [13]. Define distance $D_{single}(X, Y)$ between clusters X and Y as the distance between the closest points in clusters X and Y, i.e.

$$D_{single}(X, Y) = \min_{x\in X, y\in Y} d(x, y). \tag{1}$$

By using only the two closest points single linkage sometimes fails to recover compact clusters, but in return it has the ability to isolate outliers as singleton clusters on the dendrogram since these data points will be far from their nearest neighbor. For properties and performance in Monte Carlo Studies see [4,14]. Single linkage clustering is consistent, in some sense, for high density clusters in one dimension [6].

2. Average linkage [19]. In this linkage the distance between two clusters is the average distance between pairs of observations from each cluster. More formally

$$D_{average}(X, Y) = \frac{1}{N_X N_Y} \sum_{x\in X} \sum_{y\in Y} d(x, y). \tag{2}$$

The clusters are merged based on the distances between all members of the two clusters and, in contrast to single linkage, the two closest data points cannot cause the clusters to merge if other data points are not similar enough.

3. Ward's linkage [23]. The distance between two clusters for this linkage is the increase in the total within-cluster sum of squares as the result of joining two clusters. It is given by

$$D_{ward}(X, Y) = (2\,[SS(X \cup Y) - (SS(X) + SS(Y))])^{1/2} \tag{3}$$

where SS of a set of values is the sum of squared deviations from the centroid of the cluster. For a cluster X with N_X points the $SS(X)$ is described by the following expression:

$$SS(X) = \sum_{i=1}^{N_X} \left\| X_i(d) - \frac{1}{N_X} \sum_{j=1}^{N_X} X_j(d) \right\|_2^2$$

where $\| \cdot \|_2$ is Euclidean distance. An equivalent expression to (3) is given by

$$D_{ward} = \sqrt{\frac{2N_X N_Y}{N_X + N_Y}} \|\bar{X} - \bar{Y}\|_2, \tag{4}$$

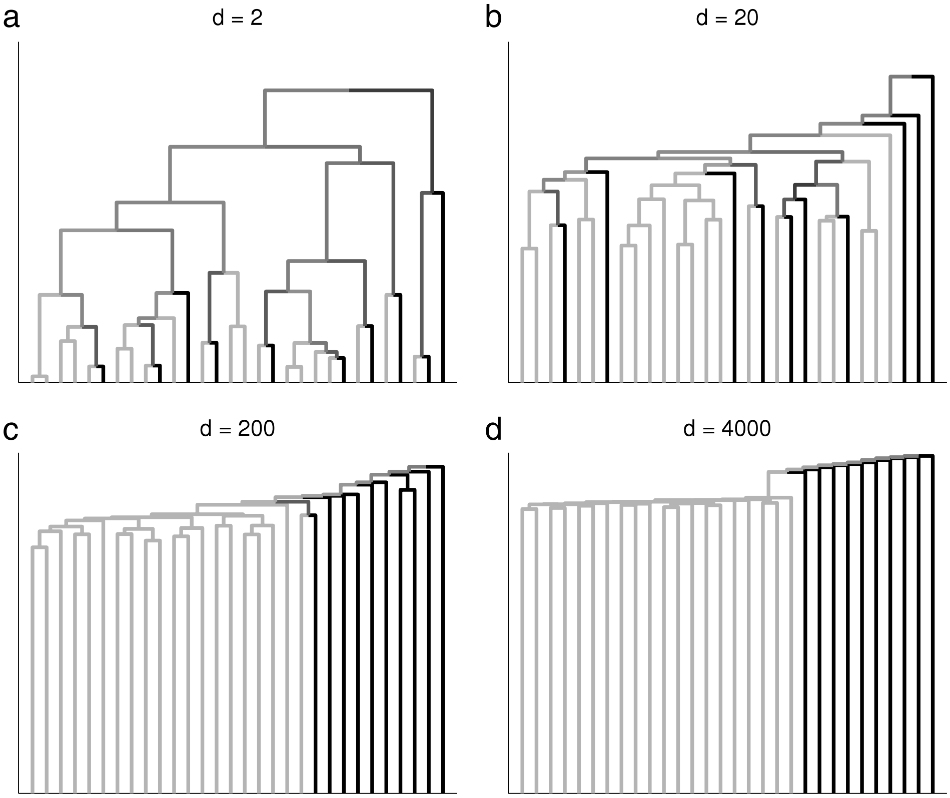


Fig. 1. Example based on the mixture of $n = 20$ observations (gray) from $N(0, 1)$ and $m = 10$ observations (black) from $N(0, 1.45)$ distributions with increasing dimension: (a) $d = 2$; (b) $d = 20$; (c) $d = 200$; (d) $d = 4000$ and average linkage function. This illustrates asymptotic clustering behavior when number of dimensions $d \to \infty$.

where $\bar{X}$, $\bar{Y}$ are the centroids of the clusters X and Y. As noted in [14], the tree produced by Ward's linkage tends to be influenced by outliers and tends to produce clusters with approximately the same number of observations. Given two pairs of clusters whose centers are equally apart, Ward's linkage will merge the smaller ones.

3. Motivating example

The following example illuminates the asymptotic behavior of hierarchical clustering in case of growing dimension and fixed sample size. In this example the data set is obtained by joining two samples from different distributions together. The example has been designed so that the ability of hierarchical clustering procedure to identify clusters will increase when the dimension of the data grows.

Both samples are generated from d-dimensional multivariate normal distribution with zero mean vector, but different covariance matrices. The first sample has 20 data points and identity covariance matrix $\Sigma_{cl1} = I_d$. The second sample has 10 data points and diagonal covariance matrix $\Sigma_{cl2} = 1.45 I_d$.

Panels (a)–(d) in Fig. 1 display dendrograms based on 30 observations with $d = 2, 20, 200$ and 4000 dimensions respectively, using the average linkage function. Data points that belong to the first sample are colored gray, points that belong to the second sample are colored black. Panel (a) shows the dendrogram based on two-dimensional data. In this case the clustering algorithm did not create homogeneous clusters. In panel (b) the dimension of the data is increased to $d = 20$. Now it is possible to see a group of 9 observations from the first sample (gray) that are joined together before the resulting cluster is combined with the observation from the second sample (black). The case of data with $d = 200$ dimensions is shown in panel (c). Now the two samples are almost perfectly separated, except one observation from the first sample is combined with the observations from the second sample. In panel (d) the dendrogram is constructed with data that has $d = 4000$ dimensions. Note that there is a perfect separation between the two samples. All observations from the first sample are combined before any observation from the second sample is added to them. In addition, the distances between data points and between clusters converge. This happens because in the case of high dimension both samples have a special geometric representation that was described in [5]. They showed that the distances between black data points converge to $(2.9d)^{1/2}$, distances between just gray points converge to $(2d)^{1/2}$ and distances between gray and black data points converge to $(3.9d)^{1/2}$ as the dimension $d \to \infty$. Hence in the last case of the toy example, the distances between data points in the second cluster are larger than all other distances. Finally, one can see that all points from the second sample are added to the cluster of points from the first sample one by one and none of the black subclusters are created in the process.

Overall comparison of four cases of data with different dimensions shown in Fig. 1 helps to build the intuition and understanding of clustering behavior as a function of signal vs noise and dimension.

4. Asymptotics

In statistics, asymptotic theory is a framework under which the properties of estimators and statistical procedures are evaluated. It has been very useful over the years for revealing the underlying structure in complicated settings.

There are several general asymptotic settings:

(A1) *Classical Asymptotics.* Results in the classical asymptotic setting are derived under the assumption that sample size $n \to \infty$, while dimension of the data remains fixed. Two well known examples are the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). See [22] for an overview of general results in the classical asymptotic setting.

(A2) *Moderate Dimension Asymptotics.* In this setting asymptotic results address the case where $n \to \infty$ with d also growing, but at a slower rate, $d/n \to 0$. This setting was investigated by [9,29,15,16].

(A3) *Random Matrix Theory Asymptotics.* In this setting the number of observations $n \to \infty$ and the number of dimensions $d \to \infty$ grow at the same rate, i.e. $d/n \to \gamma \in (0, \infty)$. Random matrix theory gained attention in the 1950s due to work of Eugene Wigner. In [24,25] he derived *Wigner's semicircle law* which states the limiting distribution of the eigenvalues of the square random matrix. In the case of a large sample covariance matrix the *Marčenko–Pastur law* [12] gives the limiting distribution of the eigenvalues.

(A4) *High Dimension Moderate Sample Size Asymptotics (HDMSS).* In this setting both $n \to \infty$ and $d \to \infty$, but d grows at a faster rate than n, i.e. $d/n \to \infty$. For example, a special case of HDMSS asymptotic setting was studied by [3], where $n \to \infty$ and $d \sim e^n$. Other recent papers in this include [26–28].

(A5) *High Dimension Low Sample Size Asymptotics (HDLSS).* In this setting asymptotic results are obtained by letting the number of dimensions $d \to \infty$ while keeping the sample size n fixed. An early paper that studied this setting was [2]. In recent years this area has been studied by [5,1,10,17]. For example, [5] showed that under the appropriate assumptions after proper rescaling the points in the sample are asymptotically located at the vertices of a deterministic simplex where each edge has fixed length. Thus, the randomness in the HDLSS data only comes from the random rotation of the hyperplane that is generated by the data.

In Section 5 we explore clustering behavior across the asymptotic settings (A2), (A3), (A4) and (A5).

5. Asymptotic boundaries in clustering behavior

In this section we present the main results of this paper such as properties of asymptotic behavior of hierarchical clustering algorithms. We consider data generated from a mixture of two Gaussian distributions. Theorem 1 presents asymptotic bounds for the single and average linkages, and Theorem 2 for Ward's linkage.

Suppose for $d = 1, 2, \ldots,$ we have $N = n + m$ observations from the mixture of two d-dimensional Gaussian distributions $N_d(\vec{\mu}_1, \sigma_1^2 I_d)$ and $N_d(\vec{\mu}_2, \sigma_2^2 I_d)$, where I_d is the d-dimensional identity matrix, and

$$\vec{\mu}_i = \left(\mu_i^{(1)}, \mu_i^{(2)}, \ldots, \mu_i^{(d)}\right), \quad i = 1, 2,$$

are population mean vectors. The first mixture component has n observations and is denoted by

$$\mathbf{X}(d) = \{X_1(d), X_2(d), \ldots, X_n(d)\}.$$

The second mixture component

$$\mathbf{Y}(d) = \{Y_1(d), Y_2(d), \ldots, Y_m(d)\},$$

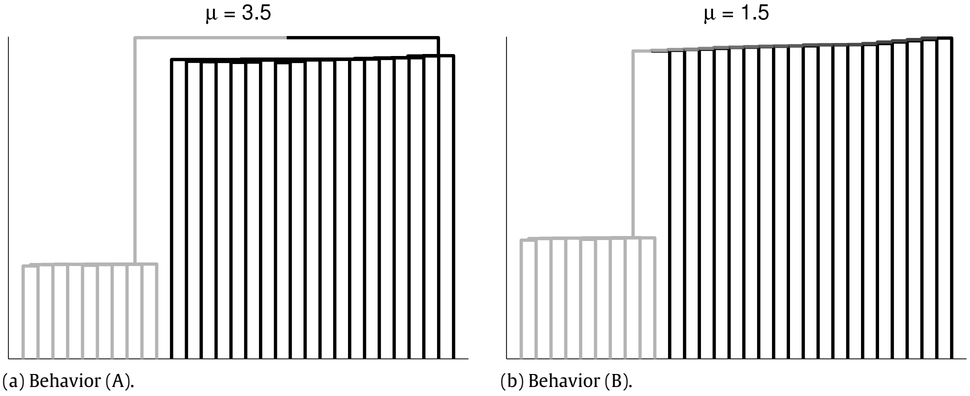
has m observations. Note, that the clustering algorithms do not use mixture labels and treat all data as one sample.

Additionally, assume that the mean vector of the difference $X_i(d) - Y_j(d)$ is not dominated by a few large values in the sense that for some $\epsilon > 0$

$$\sum_{k=1}^{d} \left(\mu_1^{(k)} - \mu_2^{(k)}\right)^4 = o(d^{2-\epsilon}), \quad \text{as } d \to \infty. \tag{5}$$

We consider two main asymptotic clustering behaviors, labeled (A) and (B), for single and average linkages and three, also including (AB), for Ward's linkage.

(A) In this clustering behavior the algorithm will join the points in a way so that points from $\mathbf{X}(d)$ and $\mathbf{Y}(d)$ are separated and only combined in the last step.



(a) Behavior (A).

(b) Behavior (B).

Fig. 2. Example based on the mixture of $n = 10$ observations from the 4000-dimensional $N(\vec{0}, I_{4000})$ ($\mathbf{X}(d)$ in gray on the left) and $m = 20$ observations from the $N(\mu\mathbb{1}_{4000}, 10I_{4000})$ ($\mathbf{Y}(d)$ in black on the right) distributions with single linkage function: (a) Behavior (A), $\mu = 3.5$; (b) Behavior (B), $\mu = 1.5$. This illustrates the difference between the clustering behaviors (A) and (B).

If mixture components have different variances, then the points from the mixture component with smaller variance will be combined before any points from another mixture component are joined. This behavior is a special case of behavior (A). An example of a hierarchical tree constructed from the data generated from mixture components with different variances is shown in panel (a) of Fig. 2. This behavior happens because asymptotically all distances within $\mathbf{X}(d)$ are smaller than all other distances. At the same time the distances between points within $\mathbf{Y}(d)$ are also smaller than distances between points in $\mathbf{X}(d)$ and $\mathbf{Y}(d)$.

(B) In this clustering behavior points from $\mathbf{X}(d)$ will be joined first and after that points from $\mathbf{Y}(d)$ will be added sequentially one by one.

This behavior is illustrated in panel (b) of Fig. 2. Now the distances between points within $\mathbf{X}(d)$ are still smaller than any distance in $\mathbf{Y}(d)$ and any distance between $\mathbf{X}(d)$ and $\mathbf{Y}(d)$. Also the distances between points within $\mathbf{Y}(d)$ are bigger than distances between points in $\mathbf{X}(d)$ and $\mathbf{Y}(d)$.

(AB) In this clustering behavior points from $\mathbf{X}(d)$ will be joined first and after that some points from $\mathbf{Y}(d)$ will be added sequentially one by one. Before all points are added at least one subcluster of $\mathbf{Y}(d)$ will be created.

This behavior is between (A) and (B) and is shown in Fig. 3 panel (b). In this case the distances between some subclusters of $\mathbf{Y}(d)$ become smaller than distances between the mixed cluster and points in $\mathbf{Y}(d)$.

The following theorems provide asymptotic signal vs noise boundaries for the hierarchical clustering behaviors (A) and (B).

Theorem 1 (*Single and Average Linkage Functions*)**.** *Suppose, without loss of generality, $\sigma_1^2 \leq \sigma_2^2$, $N = o(e^{d/2})$ and the assumption (5) is satisfied. Also suppose the hierarchical tree is constructed using either the single or the average linkage functions.*

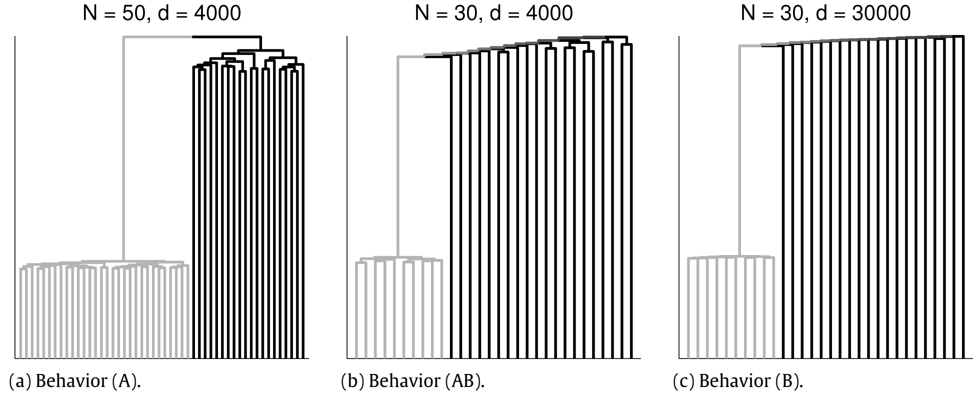
(a) *If $\liminf \|\vec{\mu}_1 - \vec{\mu}_2\|^2/d > |\sigma_2^2 - \sigma_1^2|$, then the probability of the clustering behavior* (A) *converges to* 1 *when $N \to \infty$ and $d \to \infty$.*
(b) *If $\limsup \|\vec{\mu}_1 - \vec{\mu}_2\|^2/d < |\sigma_2^2 - \sigma_1^2|$, then the probability of the clustering behavior* (B) *converges to* 1 *when $N \to \infty$ and $d \to \infty$.*

Additionally, if $\sigma_1^2 < \sigma_2^2$ then the cluster with smaller variance is combined during the first n steps.

It follows from Theorem 1 that the threshold between behaviors of clustering algorithms for single and average linkage functions depends on the distance between the centers of the sample components and their variability. The behavior (A) will happen when the mean vectors are far apart relative to the difference of the variances. The difference is illustrated in Fig. 2. Clustering behaviors of single and average linkage functions are identical asymptotically, therefore, only single linkage is presented. The asymptotic behavior of the hierarchical tree changes from (A) to (B) as the distance between the centroids of the mixture components decreases. The boundary that separates the behaviors (A) and (B) is sharp and the gap between them is of lower order, i.e. the probability of behavior (AB) converges to zero.

The asymptotic behavior of hierarchical clustering with single and average linkage functions requires only that $d/\log N$ grows to infinity, i.e. m and n can grow with various rates relative to the number of dimensions d. For example, cluster sizes can be fixed or they can be much larger than d. This means that the results of Theorem 1 are valid across the spectrum of asymptotic settings (A2)–(A5) discussed in Section 4. This phenomena is very different from the asymptotic results for PCA [18].

The proof of this theorem uses concentration inequalities with different versions of Chernoff bounds [8]. The bounds are applied to the distances between points, which have the χ_d^2 distribution if points are from the same sample components and non-central χ_d^2 otherwise. Detailed proof is in the Appendix.



(a) Behavior (A). (b) Behavior (AB). (c) Behavior (B).

Fig. 3. An example, similar to Fig. 2, based on the mixture of n observations from the d-dimensional $N_d(\vec{0}, I_d)$ and $m = 20$ observations from the $N_d(0.7\mathbb{1}_d, 10I_d)$ distributions and Ward's linkage function: (a) Behavior (A), $n = 30, d = 4000$; (b) Behavior (AB), $n = 10, d = 4000$; (c) Behavior (B), $n = 10, d = 30\,000$. This highlights how behaviors (A), (AB) and (B) compare.

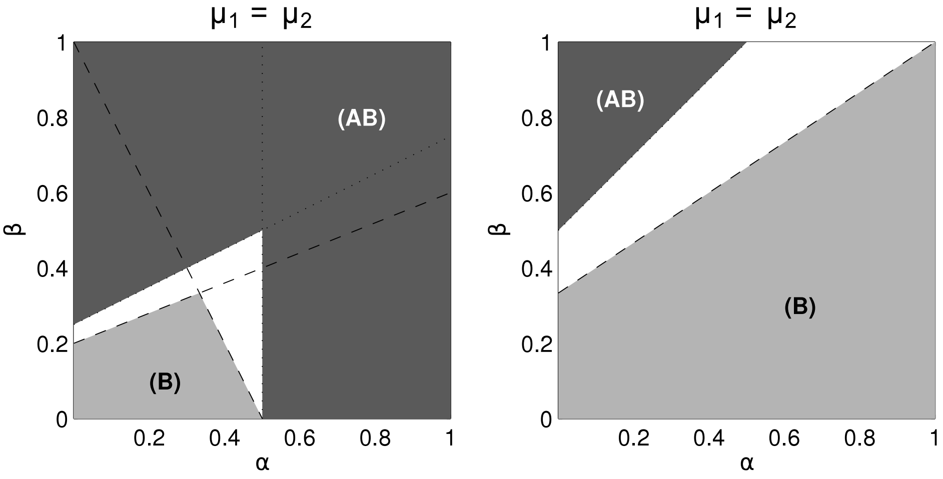


Fig. 4. Probability convergence regions for asymptotic clustering behaviors (B) and (AB), assuming $\limsup \left[n\|\vec{\mu}_1 - \vec{\mu}_2\|^2\right]/[d(\sigma_2^2 - \sigma_1^2)] < 1$, when $\vec{\mu}_1 = \vec{\mu}_2$ and $\vec{\mu}_1 \neq \vec{\mu}_2$ for $n = d^\alpha + o(d^\alpha)$, $m = d^\beta + o(d^\beta)$ and $d \to \infty$.

The situation is much different for Ward's linkage function as shown in the next theorem. The conditions for clustering behaviors are more complex compared to the single and average linkages because clustering behavior using Ward's distance depends much more critically on the sizes of the subclusters. The clustering behavior (AB) which lies between (A) and (B) now plays an important role.

Theorem 2 (*Ward's Linkage Function*)**.** *Suppose, without loss of generality, $\sigma_1^2 \le \sigma_2^2$, the assumption (5) is satisfied and the hierarchical tree is constructed using Ward's linkage function. Let n and m be the number of data points that are generated from the first and second sample component respectively. Additionally, suppose that $n = d^\alpha + o(d^\alpha)$ and $m = d^\beta + o(d^\beta)$ with $\alpha < 1, \beta < 1$.*

(a) *If $\liminf\left[n\|\vec{\mu}_1 - \vec{\mu}_2\|^2/d\right] > |\sigma_2^2 - \sigma_1^2|$, then the probability of the clustering behavior* (A) *converges to* 1 *when $n, m \to \infty$ and $d \to \infty$.*
(b) *If $\liminf\left[n\|\vec{\mu}_1 - \vec{\mu}_2\|^2/d\right] < |\sigma_2^2 - \sigma_1^2|$ and if either $\max(5\beta - 2\alpha, 2\alpha + \beta) < 1$ when $\vec{\mu}_1 = \vec{\mu}_2$, or $3\beta - 2\alpha < 1$ when $\vec{\mu}_1 \neq \vec{\mu}_2$, then the probability of the clustering behavior* (B) *converges to* 1 *when $n, m \to \infty$ and $d \to \infty$.*
(c) *If $\liminf\left[n\|\vec{\mu}_1 - \vec{\mu}_2\|^2/d\right] < |\sigma_2^2 - \sigma_1^2|$ and if either $\max(4\beta - 2\alpha, 2\alpha) > 1$ when $\vec{\mu}_1 = \vec{\mu}_2$, or $2\beta - 2\alpha > 1$ when $\vec{\mu}_1 \neq \vec{\mu}_2$, then the probability of the clustering behavior* (AB) *converges to* 1 *when $n, m \to \infty$ and $d \to \infty$.*

If $\sigma_1^2 < \sigma_2^2$ then the cluster with smaller variance is combined during the first n steps.

The roles of parameter α and β is to allow different relative growth of the first cluster size n, second cluster size m and number of dimensions d. The proof of Theorem 2 uses techniques similar to the proof of Theorem 1 and is relegated to Appendix.

Note, that in the case of Ward's linkage function, the threshold that separates the asymptotic behaviors (A) from (B) and (AB) also depends on the size of the cluster with smaller variance. This fact makes the behavior (A) more prevalent than in the single and average linkage cases, because the distance between centroids is magnified.

Behavior (B) is now less prevalent compared to the single and average linkages and also splits between (B) and (AB) in a complicated way. Conditions that separate (B) and (AB) depend on the relative rates of growth of m, n and d. These relationships are illustrated in Fig. 4. The first and second panels correspond to the cases when $\vec{\mu}_1 = \vec{\mu}_2$ and $\vec{\mu}_1 \neq \vec{\mu}_2$ respectively. Contrary to the single and average linkages, where the sample size and the number of dimensions grows such that $d/\log N \to \infty$ (asymptotic domains (A2)–(A5)), our proof for Ward's linkage requires that $d/n \to \infty$ and $d/m \to \infty$ (asymptotic domains (A4)–(A5)). Additionally, there is a gap between the behaviors (B) and (AB), the white area on both panels of Fig. 4. The gap between (B) and (AB) is created due to the techniques used in the proof and we conjecture that this gap is occupied by behavior (B).

The hierarchical trees with behaviors (A), (AB) and (B) are illustrated in Fig. 3 in panels (a), (b) and (c) respectively. The three panels show the change in clustering behaviors of Ward's linkage caused by the number of dimensions and the increase in the number of data points in the cluster with smaller variance. Since the distance between the centroids is magnified, it is difficult to observe behavior (B) when the size of the cluster with smaller variance is large. To show the difference between (A) and (AB) it was enough to reduce the number of observations sampled from $\mathbf{X}(d)$ from 30 to 10. Also behavior (AB) is much more prevalent compared to (B). In order to observe behavior (B) we had to increase the number of dimensions to 30 000 from 4000 for behavior (AB).

6. Simulation

This section illustrates the asymptotic clustering behaviors and the speed of convergence to probability one using simulation. Tables B.1 and B.2 demonstrate the differences between behaviors (A), (B) and (AB) described in Theorems 1 and 2 respectively. We also consider behavior (A*), which is a special case of behavior (A) where $\sigma_1^2 < \sigma_2^2$. Because it is the most extreme, and also the most interesting, we focus on the asymptotic setting (A5) (see Section 4) where the number of dimensions $d \to \infty$ while the numbers of data points in each cluster n and m remain constant. All estimated probabilities in Tables B.1 and B.2 are based on 1000 simulated realizations. For each realization we generate two clusters, one of n observations from the d-dimensional $N_d(\vec{0}, I_d)$ distribution, and the other of m observations from the $N_d(\mu \mathbb{1}_d, \sigma^2 I_d)$ distribution. After that we build a hierarchical tree based on the specified linkage function and then identify the corresponding clustering behavior. All dendrograms that cannot be assigned to one of the defined categories are classified as (Other).

Estimated probabilities of clustering behaviors for the single and average linkage functions are presented in Table B.1. As shown in Theorem 1 the asymptotic behaviors only depend on the distance between the two mean vectors and the differences between the variances. If the centers of the two clusters are far apart relative to the difference of the variances, then the probability of behavior (A) will converge to 1. Additionally, if variances are not equal, then all points from the cluster with the smaller variance will be combined before any two points from the other cluster are joined producing asymptotic behavior (A*). In the case where the distance between the mean vectors is relatively small, the probability of behavior (B), where the points from one cluster are combined first and after that points from another cluster are joined sequentially, will converge to 1. These statements are confirmed by simulation with the following parameters: $\mu \in \{1.5, 3, 3.5\}$ and $\sigma^2 \in \{1, 10\}$. When $\sigma^2 = 1$ the difference between the variances is zero and the probability of behavior (A) converges to 1 for any positive value of μ. When $\sigma^2 = 10$ the difference between the variances is positive, and asymptotic behavior will depend on the value of μ. In the case of $\mu = 3.5$ the distance between the centers of two clusters is larger than the difference of the variances. Also the variance of the first mixture component is smaller than the variance of the second. Hence, probability of behavior (A*) converges to 1. If $\mu = 1.5$ the distance between the mean vectors becomes relatively small. Therefore, hierarchical trees based on single and average linkages have asymptotic behavior (B). When $\mu = 3$ the parameters are located on the boundary between behaviors (A) and (B) defined by the equation $\|\vec{\mu}_1 - \vec{\mu}_2\|^2/d = |\sigma_2^2 - \sigma_1^2|$. For such combinations of parameters Theorem 1 does not specify asymptotic behavior, but simulation results suggest a mixture of the three behaviors (B), (AB) and (Other). We believe that with deeper asymptotics this mixture of behaviors could be precisely specified. Further investigation of this interesting phenomenon is left for future work.

Table B.2 presents estimated probabilities of clustering behaviors for dendrograms created using Ward's linkage function. There are two main differences compared to the simulation results for single and average linkages. First, the boundary that separates behaviors (A) from (B) and (AB) depends on the size of the cluster with smaller variance. Second, there is a combination of parameters μ, σ^2, n, m and d that leads to behavior (AB) with probability 1. The following are combinations of parameters considered in simulation and corresponding behaviors with estimated probabilities: $\mu \in \{0.7, 1.5, 3\}$, $\sigma^2 \in \{1, 10\}$ and $n \in \{10, 30\}$. Similarly to single and average linkage functions, when $\sigma^2 = 1$, the probability of behavior (A) converges to 1 for any value of μ and n. When $\sigma^2 = 10$ the combination of values μ and n will determine the asymptotic behavior. If $\mu = 1.5$ or $\mu = 3$ with $n = 10$ the probability of behavior (A*) converges to 1. Note that for single and average linkage functions when $\sigma^2 = 10$ the value of μ has to be greater than 3 to guarantee behavior (A*). It is not the case for Ward's linkage due to the presence of the parameter n in the boundary equation of Theorem 2. Similarly, the case when $\mu = 3$ is on the boundary for single and average linkages, but for Ward's linkage it is inside the convergence region of the behavior (A*). If $\mu = 0.7$ and $n = 10$ the asymptotic behavior is (B), but when $d < 10^5$, the probability of behavior (AB) is essentially 1. That is possible because the distance between two points from the second mixture component is relatively small. As the number of dimensions d grows the same distance becomes larger compared to other subclusters, hence we eventually converge to the behavior (B) with probability 1 (as indicated by Theorem 2). The case of $\mu = 0.7$, but $n = 30$

is similar to the previous combination of parameters, except the number of data points in the cluster with smaller variance increased from $n = 10$ to $n = 30$. This change alone was sufficient to switch from asymptotic behaviors (B) and (AB) to (A*).

7. Conclusions

In this paper we studied the behavior of hierarchical clustering in situations where both sample size and dimension grow to infinity. We developed a theory which explicitly identifies signal vs noise boundaries between different types of clustering behaviors.

In particular, we found that clustering behaviors of single and average linkage functions are asymptotically identical. The threshold that separates the two main behaviors depends on the distance between the population means of the two sample components and the difference between the variances. Additionally, the number of data points in both sample components is allowed to grow with various rates relative to the dimension d. For example, N can be fixed or grows much faster than d, i.e., $d/\log N \to \infty$. This setting includes asymptotic domains (A2)–(A5) described in Section 4.

The situation is more complicated for Ward's linkage function. Now the behavior (B) of the single and average linkage functions is divided into behaviors (AB) and (B). Also, the threshold that separates behaviors (A) from (AB) and (B) depends on the size of the cluster with the smaller variance, and the behavior (A) is more prevalent. Finally, our proof requires that the number of dimensions has to grow faster than the sample size when m, n and $d \to \infty$, which is the case for the asymptotic settings (A4)–(A5).

As the first exploration of asymptotics in this area, this paper has revealed a number of interesting open problems. We conjecture that our isotropic covariance assumption can be extended to reasonable conditions on the eigenvalues of the mixture components, so that parallel results can be obtained. For example, under suitable boundedness conditions on the eigenvalues, such as eigenvalues from the first and the second mixture component are bounded away from each other as well as uniformly bounded away from zero and infinity, we believe that much more complicated analogs of our theorems can be established.

The Gaussian mixture component assumption can also be relaxed. The bounds for the clustering behavior can be derived if the moment generating function of each mixture component is bounded and the second moment of each variable exists. Analogous proofs for Lemmas 2 and 3 would go through, however the exact boundary of the clustering behavior would have a much more complicated and case wise formulation. Conditions for asymptotic behaviors will have to be derived individually for each distribution family based on moment generating function properties and the relation between the parameters. We think that these provide interesting areas for future research.

Appendix A. Proofs

Lemma 1. *Let $W_1, \ldots, W_d$ be independent non-negative random variables with finite second moments. Define $S = \sum_{k=1}^{d}(W_k - EW_k)$ and $v = \sum_{k=1}^{d} EW_k^2$. Then for $t > 0$*

$$P(S \le -t) \le \exp\left(\frac{-t^2}{2v}\right). \tag{A.1}$$

Proof. The proof is based on the idea of Chernoff bound [8] and the fact that for all $u > 0$

$$e^{-u} \le 1 - u + u^{2/2}. \tag{A.2}$$

Using inequality (A.2) and the inequality $1 + x \le e^x$ for all x and fixed $k = 1, \ldots, d$

$$\begin{aligned} E\left[e^{-\lambda(W_k - EW_k)}\right] &= e^{\lambda EW_k} E\left[e^{-\lambda W_k}\right] \le e^{\lambda EW_k} E\left[1 - \lambda W_k + \frac{\lambda^2 W_k^2}{2}\right] \\ &= e^{\lambda EW_k} E\left[1 + \left(\frac{\lambda^2 W_k^2}{2} - \lambda W_k\right)\right] = e^{\lambda EW_k}\left[1 + \left(\frac{\lambda^2 EW_k^2}{2} - \lambda EW_k\right)\right] \\ &\le e^{\lambda EW_k} e^{(\lambda^2 EW_k^2)/2 - \lambda EW_k} = e^{(\lambda^2 EW_k^2)/2}. \end{aligned} \tag{A.3}$$

Next we combine inequality (A.3) with the idea of Chernoff bound with $\lambda > 0$ and independence of W_k:

$$\begin{aligned} P(S \le -t) &= P(e^{-\lambda S} \ge e^{\lambda t}) = E\left[e^{-\lambda \sum\limits_{k=1}^{d}(W_k - EW_k)}\right] e^{-\lambda t} \\ &= \prod_{k=1}^{d} E\left[e^{-\lambda(W_k - EW_k)}\right] e^{-\lambda t} \le \prod_{k=1}^{d} e^{(\lambda^2 EW_k^2)/2} e^{-\lambda t} \\ &\le \inf_{\lambda > 0} \exp\left[\frac{\lambda^2 v}{2} - \lambda t\right] = \exp\left(-\frac{t^2}{2v}\right). \quad \square \end{aligned}$$

A.1. Upper and lower concentration inequalities

Suppose $X_i(d) = (X_i^{(1)}, \ldots, X_i^{(d)})^T$ are independent data vectors from the sample $\mathbf{X}(d) = \{X_1(d), X_2(d), \ldots, X_n(d)\}$. Also suppose that sample $\mathbf{X}(d)$ is drawn from multivariate d-dimensional normal distribution $N_d(\vec{\mu}, \sigma^2 I_d)$, where I_d is the d-dimensional identity matrix, and $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^T$ is the population mean vector.

Lemma 2 (*Lower Tail*). *Define scalar $\mu^2 = \|\vec{\mu}\|^2/d$. Let $0 < a < \sigma^2 + \mu^2$. Then, for any $0 < i \leq n$,*

$$P\left(\|X_i(d)\|^2 < ad\right) \leq e^{-dC} \tag{A.4}$$

where $C = \left(a - \sigma^2 - \mu^2\right)^2 / (6\sigma^4 + 12\sigma^2\mu^2 + 2\sum_{k=1}^{d} \mu_k^4/d)$.

Proof. For any $k = 1, \ldots, d$ the ratio $\left(X_i^{(k)}/\sigma\right)^2$ has χ^2-distribution with 1 degree of freedom and noncentrality parameter $\gamma = \mu_k^2/\sigma^2$. Applying Lemma 1 with $W_k = X_i^{(k)}$, $S = \|X_i(d)\|^2 - \sigma^2 d - \mu^2 d$ and $v = 3\sigma^4 d + 6\sigma^2\mu^2 d + \sum_{k=1}^{d} \mu_k^4$ we get the result. □

Lemma 3 (*Upper Tail*). *Define scalar $\mu^2 = \|\vec{\mu}\|^2/d$ and let $a > \sigma^2 + \mu^2$. Then, for any $0 < i \leq n$,*

$$P\left(\|X_i(d)\|^2 > ad\right) \leq e^{-dC} \tag{A.5}$$

where $C = \left[a + \mu^2 - \sqrt{\sigma^4 + 4\mu^2 a} + \sigma^2 \log\left(\frac{\sigma^2 + \sqrt{\sigma^4 + 4\mu^2 a}}{2a}\right)\right] / (\sigma^2)$.

Proof. Define $Z = \|X_i(d)\|^2 = \sum_{k=1}^{d} X_i^{(k)}$. Since each $X_i^{(k)} \sim N(\mu_k, \sigma^2)$ for any $i = 1, \ldots, n$ and $k = 1, \ldots, d$, the ratio Z/σ^2 has noncentral χ^2-distribution with d degrees of freedom and noncentrality parameter $\gamma = \|\vec{\mu}\|^2/\sigma^2 = \mu^2 d/\sigma^2$. Therefore, using the idea of Chernoff bound with condition that $0 < t < 1/(2\sigma^2)$, we can get the result. □

Proof of Theorem 1(b). The proof consists of two parts. In the first part we derive conditions for the clustering behavior described in Section 5. In the second part of the proof we show that under the provided conditions the probability of clustering behavior converges to 1 for single and average linkage functions. Note that behavior B is not possible when $\sigma_1^2 = \sigma_2^2$, therefore without loss of generality we only consider the case of $\sigma_1^2 < \sigma_2^2$. This means that the data points from $\mathbf{X}(d)$ are combined during the first n steps.

Let $0 < a_1 < a_2 < a_3 < a_4$ and define the following events:

$$A_d = \left\{\max_{i,j} \|X_i(d) - X_j(d)\|^2 < \min_{i,j} \|Y_i(d) - Y_j(d)\|^2\right\},$$

$$B_d = \left\{\max_{i,j} \|X_i(d) - X_j(d)\|^2 < \min_{i,j} \|X_i(d) - Y_j(d)\|^2\right\},$$

$$C_d = \left\{\max_{i,j} \|X_i(d) - Y_j(d)\|^2 < \min_{i,j} \|Y_i(d) - Y_j(d)\|^2\right\}$$

and

$$E_d^1 = \left\{\max_{i,j} \|X_i(d) - X_j(d)\|^2 < a_1 d\right\},$$

$$E_d^2 = \left\{\min_{i,j} \|X_i(d) - Y_j(d)\|^2 > a_2 d\right\},$$

$$E_d^3 = \left\{\max_{i,j} \|X_i(d) - Y_j(d)\|^2 < a_3 d\right\},$$

$$E_d^4 = \left\{\min_{i,j} \|Y_i(d) - Y_j(d)\|^2 > a_4 d\right\}.$$

In this case

$$P[(B)] \geq P(A_d \cap B_d \cap C_d).$$

Hence, convergence in Theorem 1 can be shown by considering the complement of the event $A_d \cap B_d \cap C_d$.

$$\begin{aligned} P\left(A_d \cap B_d \cap C_d\right)^c &= 1 - P\left(A_d \cap B_d \cap C_d\right) \leq 1 - P\left(E_d^1 \cap E_d^2 \cap E_d^3 \cap E_d^4\right) \\ &= P\left(E_d^1 \cap E_d^2 \cap E_d^3 \cap E_d^4\right)^c = P\left(E_d^{1c} \cup E_d^{2c} \cup E_d^{3c} \cup E_d^{4c}\right) \\ &\leq P\left(E_d^{1c}\right) + P\left(E_d^{2c}\right) + P\left(E_d^{3c}\right) + P\left(E_d^{4c}\right) \to 0 \end{aligned} \tag{A.6}$$

as $n, d \to \infty$.

Each probability $P\left(E_d^{1c}\right)$, $P\left(E_d^{2c}\right)$, $P\left(E_d^{3c}\right)$, and $P\left(E_d^{4c}\right)$ can be bounded using results of Lemmas 2 and 3 and the following facts:

1. $X_i(d) - X_j(d) \sim N(\vec{0}, 2\sigma_1^2 I)$ for $i \neq j,\ 1 \leq i, j \leq n$,
2. $X_i(d) - Y_j(d) \sim N(\vec{\mu}_1 - \vec{\mu}_2, (\sigma_1^2 + \sigma_2^2)I)$ for all $1 \leq i \leq n$ and $1 \leq i \leq m$,
3. $Y_i(d) - Y_j(d) \sim N(\vec{0}, 2\sigma_2^2 I)$ for $i \neq j,\ 1 \leq i, j \leq m$,
4. For any random variables $Z_1, \ldots, Z_k$ and any b, $P(\max_n(Z_1, \ldots, Z_n) > b) \leq \sum_n P(Z_i > b)$ and $P(\min_n(Z_1, \ldots, Z_n) < b) \leq \sum_n P(Z_i < b)$.

Therefore,

$$\begin{aligned}
P\left(E_d^{1c}\right) &\leq [(n(n-1)/2)]\, e^{-d\left(a_1 - 2\sigma_1^2 + 2\sigma_1^2 \log\left[\frac{2\sigma_1^2}{a_1}\right]\right)/(4\sigma_1^2)} \\
&\leq e^{-d\left(a_1 - 2\sigma_1^2 + 2\sigma_1^2 \log\left[\frac{2\sigma_1^2}{a_1}\right]\right)/(4\sigma_1^2) + 2\log n}, \\
P\left(E_d^{2c}\right) &\leq e^{-d(a_2 - \sigma^2 - \mu^2)^2/(6\sigma^4 + 12\sigma^2\mu^2 + \tilde{\mu}) + \log nm}, \\
P\left(E_d^{3c}\right) &\leq e^{-d\left(a_3 + \mu^2 - \sqrt{\sigma^4 + 4\mu^2 a_3} + \sigma^2 \log\left(\frac{\sigma^2 + \sqrt{\sigma^4 + 4\mu^2 a_3}}{2a_3}\right)\right)/(2\sigma^2) + \log nm}, \\
P\left(E_d^{4c}\right) &\leq e^{-d(a_4 - 2\sigma_2^2)^2/(24\sigma_2^4) + 2\log m},
\end{aligned} \tag{A.7}$$

where $\mu^2 = \|\vec{\mu}_1 - \vec{\mu}_2\|^2/d$, $\tilde{\mu} = \sum_{k=1}^d \left(\mu_1^{(k)} - \mu_2^{(k)}\right)^4/d$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Lemmas 3 and 2 also provide conditions that have to be satisfied for the exponential bound to converge to zero: $a_1 > 2\sigma_1^2$, $a_2 < \sigma_1^2 + \sigma_2^2 + \mu^2$, $a_3 > \sigma_1^2 + \sigma_2^2 + \mu^2$, and $a_4 < 2\sigma_2^2$. Since $0 < a_1 < a_2 < a_3 < a_4$, these conditions can be simplified to $\sigma_1^2 < \sigma_2^2$ and $\mu^2 < \sigma_2^2 - \sigma_1^2$. Hence, the general condition is $\mu^2 < |\sigma_2^2 - \sigma_1^2|$.

The condition $\sum_{k=1}^d \mu_k^4 = o(d^{2-\epsilon})$ for $\epsilon > 0$ is necessary so that the denominator in (A.4) does not grow faster than the numerator.

Now suppose that all probabilities in (A.7) converge to zero. It follows that there exist $0 < a_1 < a_2$ such that $P\left(\max_{i,j} \|X_i(d) - X_j(d)\|^2 < a_1 d\right) \to 1$ and $P\left(\min_{i,j} \|X_i(d) - Y_j(d)\|^2 > a_2 d\right) \to 1$. Hence, with the probability converging to 1, the average distance between any point from Y and any sub-cluster of X is larger than $\min_{i,j} \|X_i(d) - Y_j(d)\|^2$ which in turn is larger than $\max_{i,j} \|X_i(d) - X_j(d)\|^2$. This means that during the first n steps of the hierarchical clustering procedure it is not possible to join any point from Y with any sub-cluster of X before all points from X are combined. Similarly, since by (A.7) $P\left(\min_{i,j} \|Y_i(d) - Y_j(d)\|^2 > a_4 d\right) \to 1$ and $a_1 < a_4$, it is not possible to join any two points from Y before all points from X are combined together. Therefore, with the probability converging to 1, all points from X are joined first.

Finally, the points from Y will be added one by one to the large cluster, i.e. no homogeneous sub-clusters of Y will be created. It is not possible to join two points from Y because with the probability converging to 1, the smallest distance between points in Y is larger than the average distance between points from X and Y. The proof for the single linkage function, where only the smallest distance between points is used, is analogous. □

Proof of Theorem 1(a). The flow of the proof is similar to the proof of Part 1 and Part 2 of Theorem 1 (b) provided above. The main difference is that in the first part of the proof we do not have a restriction on which mixture component is combined first. To show that the algorithm will separate data points from mixture components and only combine them during the last step define events:

$$\tilde{B}_d = \left\{\max_{i,j} \|X_i(d) - X_j(d)\|^2 < \min_{i,j} \|X_i(d) - Y_j(d)\|^2\right\},$$

$$\tilde{C}_d = \left\{\max_{i,j} \|Y_i(d) - Y_j(d)\|^2 < \min_{i,j} \|X_i(d) - Y_j(d)\|^2\right\}$$

and

$$E_d^1 = \left\{\max_{i,j} \|X_i(d) - X_j(d)\|^2 < ad\right\},$$

$$E_d^2 = \left\{\max_{i,j} \|Y_i(d) - Y_j(d)\|^2 < ad\right\},$$

$$E_d^3 = \left\{\min_{i,j} \|X_i(d) - Y_j(d)\|^2 > ad\right\}$$

for $0 < a < \infty$. Then $P[(A)] \geq P(\tilde{B}_d \cap \tilde{C}_d)$ and

$$P\left(\tilde{B}_d \cap \tilde{C}_d\right)^c \leq P\left(E_d^{1c}\right) + P\left(E_d^{2c}\right) + P\left(E_d^{3c}\right) \to 0$$

as $n, d \to \infty$. The bounds on $P\left(E_d^{1c}\right)$, $P\left(E_d^{2c}\right)$, and $P\left(E_d^{3c}\right)$ are given by

$$\begin{aligned} P\left(E_d^{1c}\right) &\leq e^{-d\left(a-2\sigma_1^2+2\sigma_1^2\log\left[\frac{2\sigma_1^2}{a}\right]\right)/(4\sigma_1^2)+2\log n}, \\ P\left(E_d^{2c}\right) &\leq e^{-d\left(a-2\sigma_2^2+2\sigma_2^2\log\left[\frac{2\sigma_2^2}{a}\right]\right)/(4\sigma_2^2)+2\log m} \\ P\left(E_d^{3c}\right) &\leq e^{-d(a-\sigma^2-\mu^2)^2/\left(6\sigma^4+12\sigma^2\mu^2+\tilde{\mu}\right)-\log nm} \end{aligned} \tag{A.8}$$

where $\mu^2 = \|\vec{\mu}_1 - \vec{\mu}_2\|^2/d$, $\tilde{\mu} = \sum_{k=1}^d \left(\mu_1^{(k)} - \mu_2^{(k)}\right)^4/d$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. In order for the exponential bounds to converge to zero the following conditions have to be satisfied: $a > 2\sigma_1^2$, $a > 2\sigma_2^2$, and $a < \sigma_1^2 + \sigma_2^2 + \mu^2$. Using the fact that $0 < a < \infty$ conditions are simplified to $\mu^2 > |\sigma_2^2 - \sigma_1^2|$.

The second part of the proof which shows that the probability of hierarchical clustering behavior (A) converges to 1 is almost identical to the second part of the proof of Theorem 1(a). Also the proof that the special case of behavior (A) converges to 1 (when the data points from the mixture component with the smaller variance are combined before any two points from the second component are joined) is similar to the proof of Theorem 1(a). □

Lemma 4. *Suppose $\sigma_1^2 < \sigma_2^2$, the assumption (5) is satisfied and the hierarchical tree is constructed using Ward's linkage function. Let n and m be the number of data points that are generated from the first and second sample component respectively. Additionally, suppose that $n = d^\alpha + o(d^\alpha)$ and $m = d^\beta + o(d^\beta)$ for $0 < \alpha < 1$ and $0 < \beta < 1$. Under the constraints,*

$$\max(5\beta - 2\alpha, 2\alpha + \beta) < 1 \quad \textit{when } \|\vec{\mu}_1 - \vec{\mu}_2\| = 0, \ \textit{or},$$

$$3\beta - 2\alpha < 1 \quad \textit{when } \|\vec{\mu}_1 - \vec{\mu}_2\| > 0$$

the following probabilities converge to zero, i.e.

$$P(E_1^c) = P(\max D^w(\mathbf{X}_1, \mathbf{X}_2) > a_1 d) \to 0, \tag{A.9}$$

$$P(E_2^c) = P(\min D^w(\mathbf{X}, Y_j) < a_2 d) \to 0, \tag{A.10}$$

$$P(E_3^c) = P(\max D^w(\mathbf{M}, Y_j) > a_3 d) \to 0, \tag{A.11}$$

$$P(E_4^c) = P(\min D^w(\mathbf{Y}_1, \mathbf{Y}_2) < a_4 d) \to 0, \tag{A.12}$$

where $D^w(\cdot, \cdot)$ is Ward's distance function, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}$ are subclusters of $\mathbf{X}(d)$, $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}$ are subclusters of $\mathbf{Y}(d)$ and cluster $\mathbf{M}$ is a union of cluster $\mathbf{X}(d)$ and some points from cluster $\mathbf{Y}(d)$.

Proof. The convergence in (A.9) is proved directly. Suppose that n_1 is the size of $\mathbf{X}_1$ and n_2 is the size of $\mathbf{X}_2$. Then $W = [2n_1n_2/(n_1+n_2)]^{1/2}(\bar{X}_1 - \bar{X}_2)$ are i.i.d. Multivariate Normal with zero mean vector and covariance $2\sigma_1^2 I_d$. Also each point of $\mathbf{X}(d)$ can be assigned to $\mathbf{X}_1$ or $\mathbf{X}_2$, or none of them, which leads to at most 3^n ways to create such clusters. Then,

$$\begin{aligned} P(\max D^w(\mathbf{X}_1, \mathbf{X}_2) > a_1 d) &= P\left(\max \frac{2n_1n_2}{n_1+n_2}\|\bar{X}_1 - \bar{X}_2\|^2 > a_1 d\right) \\ &\leq 3^n P\left(\|W\|^2 > a_1 d\right) \leq e^{-C_1 d + n\log 3}. \end{aligned} \tag{A.13}$$

Similarly, if m_1 is the size of $\mathbf{Y}_1$ and m_2 is the size of $\mathbf{Y}_2$, then the bound (A.12) is given by

$$\begin{aligned} P(\min D^w(\mathbf{Y}_1, \mathbf{Y}_2) < a_4 d) &= P\left(\min \frac{2m_1m_2}{m_1+m_2}\|\bar{Y}_1 - \bar{Y}_2\|^2 < a_4 d\right) \\ &\leq 3^m P\left(\|W\|^2 < a_4 d\right) \leq e^{-C_4 d + m\log 3}, \end{aligned}$$

where $a_1 > 2\sigma_1^2$ and $a_4 < 2\sigma_2^2$. The constants C_1 and C_4 are defined in (A.4) and (A.5) respectively.

Next we will show convergence of probability (A.10). In this case for any fixed n_1 the random variables $[2n_1/(n_1+1)]^{1/2}(\bar{X} - Y_j)$ will have the same normal distribution. Hence,

$$P(\min D^w(\mathbf{X}, Y_j) < a_2 d) \le \sum_{n_1=1}^{n} \binom{n}{n_1} mP\left(\frac{2n_1}{n_1+1}\|\bar{X} - Y_1\|^2 < a_2 d\right)$$
$$\le 2^n m \max_{n_1} P\left(\frac{2n_1}{n_1+1}\|\bar{X} - Y_1\|^2 < a_2 d\right).$$

Suppose that n_1 is fixed, then using result of Lemma 2 one can show that the bound converges to zero and it does not depend on n_1, i.e.

$$P\left(\frac{2n_1}{n_1+1}\|\bar{X} - Y_1\|^2 < a_2 d\right) \le e^{-C_2 d + \log m + n\log 2}.$$

Therefore,

$$P(\min D^w(\mathbf{X}, Y_j) < a_2 d) \le e^{-C_2 d + \log m + n\log 2}, \tag{A.14}$$

where

$$a_2 < \frac{2n}{n+1}\left(\sigma_1^2/n + \sigma_2^2 + \mu^2\right).$$

Now we will prove convergence (A.11) of the probability of the event that involves the distance between full cluster $\mathbf{X}(d)$ that was combined with m_1 points $\mathbf{Y}(d)$ and any remaining point from $\mathbf{Y}(d)$.

Suppose cluster $\mathbf{M}$ is a union of cluster $\mathbf{X}(d)$ and m_1 points from cluster $\mathbf{Y}(d)$ and has a size $n + m_1$. If number of points m_1 is fixed, then all $\binom{m}{m_1}(m - m_1)$ random variables $[2(n+m_1)/(n+m_1+1)]^{1/2}(\bar{M} - Y_j)$ are normally distributed with the same mean and variance.

$$P(\max D^w(\mathbf{M}, Y_j) > a_3 d) \le \sum_{m_1=0}^{m-1} \binom{m}{m_1}(m - m_1)P\left(\frac{2(n+m_1)}{n+m_1+1}\|\bar{M} - Y_1\|^2 > a_3 d\right)$$
$$\le 2^m m \max_{m_1} P\left(\frac{2(n+m_1)}{n+m_1+1}\|\bar{M} - Y_1\|^2 > a_3 d\right). \tag{A.15}$$

For any fixed m_1 the probability in (A.15) can be bounded using Lemma 3:

$$P\left(\frac{2(n+m_1)}{n+m_1+1}\|\bar{M} - Y_1\|^2 > a_3 d\right) \le e^{-dC(n,m_1)}, \tag{A.16}$$

where $C(n, m_1)$ is defined in (A.5) with

$$\mu^2 = \frac{2(n+m_1)n^2}{(n+m_1+1)(n+m_1)^2}\|\vec{\mu}_1 - \vec{\mu}_2\|^2/d,$$
$$\sigma^2 = \frac{2(n+m_1)}{n+m_1+1}\left[\frac{n\sigma_1^2 + m_1\sigma_2^2}{(n+m_1)^2} + \sigma_2^2\right]$$

and is positive for $\sigma^2 + \mu^2 < a_3 < 2\sigma_2^2$. Define $n = d^\alpha + o(d^\alpha)$ and $m_1 = d^\beta + o(d^\beta)$. Taylor series expansion of $C(n, m_1)$ suggests that $dC(n, m_1) \to \infty$ when $\max(4\beta - 2\alpha, 2\alpha) < 1$ for $\|\vec{\mu}_1 - \vec{\mu}_2\| = 0$ or $2\beta - 2\alpha < 1$ for $\|\vec{\mu}_1 - \vec{\mu}_2\| > 0$. If these conditions are satisfied for $m_1 = m$, then they will be satisfied for all $0 \le m_1 \le m$. Therefore, for any m_1, probability bound in (A.16), and, hence, in (A.15) will converge to zero if $\max(5\beta - 2\alpha, 2\alpha + \beta) < 1$ when $\|\vec{\mu}_1 - \vec{\mu}_2\| = 0$, or $\max(3\beta - 2\alpha, \beta) < 1$ when $\|\vec{\mu}_1 - \vec{\mu}_2\| > 0$. □

Proof of Theorem 2. First we show that under the certain set of conditions the probability of clustering behavior (B) for Ward's linkage function converges to one. The idea of the proof is similar to the proof of Theorem 1. Define events

$$A_d = \{\max D^w(\mathbf{X}_1, \mathbf{X}_2) < \min D^w(\mathbf{Y}_1, \mathbf{Y}_2)\},$$
$$B_d = \left\{\max D^w(\mathbf{X}_1, \mathbf{X}_2) < \min D^w(\mathbf{X}, Y_j)\right\}, \tag{A.17}$$
$$C_d = \left\{\max D^w(\mathbf{M}, Y_j) < \min D^w(\mathbf{Y}_1, \mathbf{Y}_2)\right\},$$

where $D^w(\cdot,\cdot)$ is Ward's distance between two subclusters, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}$ are some subclusters of $\mathbf{X}(d)$ and $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}$ are some subclusters of $\mathbf{Y}(d)$. Subcluster $\mathbf{M}$ is a mixed cluster that contains all n points from $\mathbf{X}(d)$ and some number of points from $\mathbf{Y}(d)$. Event A_d means that the largest Ward's distance between any two subclusters of $\mathbf{X}(d)$ is smaller than the minimum Ward's distance between any two subclusters of $\mathbf{Y}(d)$. Event B_d makes the maximum Ward's distance between subclusters of $\mathbf{X}(d)$ smaller than the minimum distance between any subcluster of $\mathbf{X}(d)$ and any data point from $\mathbf{Y}(d)$. Finally, event C_d

Table B.1
Simulated probabilities of asymptotic ($d \to \infty$) clustering behaviors based on the single and average linkage functions. Data is simulated from the mixture of $n = 10$ observations from the d-dimensional $N_d(\vec{0}, I_d)$ and $m = 20$ observations from the $N_d(\mu \mathbb{1}_d, \sigma^2 I_d)$ distributions. Clustering behaviors (A), (B) and (AB) are described in Section 5. Behavior (A*) is a special case ($\sigma_1^2 < \sigma_2^2$) of behavior (A). This confirms the results of Theorem 1.

μ	σ^2	d	Single linkage behaviors					Average linkage behaviors				
			(Other)	(A)	(A*)	(B)	(AB)	(Other)	(A)	(A*)	(B)	(AB)
1.5	1	5	0.994	0.006	0	0	0	0.892	0.108	0	0	0
1.5	1	10	0.931	0.069	0	0	0	0.466	0.534	0	0	0
1.5	1	100	0	1	0	0	0	0	1	0	0	0
1.5	1	1 000	0	1	0	0	0	0	1	0	0	0
1.5	1	10 000	0	1	0	0	0	0	1	0	0	0
1.5	10	5	0.889	0	0	0	0.111	0.997	0	0	0	0.003
1.5	10	10	0.398	0	0	0.047	0.555	0.801	0	0	0	0.199
1.5	10	100	0.001	0	0	0.992	0.007	0.006	0	0	0.257	0.737
1.5	10	1 000	0	0	0	1	0	0	0	0	1	0
1.5	10	10 000	0	0	0	1	0	0	0	0	1	0
3	1	5	0.119	0.881	0	0	0	0.032	0.968	0	0	0
3	1	10	0.003	0.997	0	0	0	0	1	0	0	0
3	1	100	0	1	0	0	0	0	1	0	0	0
3	1	1 000	0	1	0	0	0	0	1	0	0	0
3	1	10 000	0	1	0	0	0	0	1	0	0	0
3	10	5	0.911	0.001	0	0	0.088	0.997	0.001	0	0	0.002
3	10	10	0.805	0	0	0	0.195	0.977	0.001	0.001	0	0.021
3	10	100	0.855	0	0	0.002	0.143	0.968	0	0.001	0	0.031
3	10	1 000	0.844	0	0	0.005	0.151	0.972	0	0	0	0.028
3	10	10 000	0.840	0	0	0.011	0.149	0.971	0	0	0	0.029
3	10	100 000	0.848	0	0	0.003	0.149	0.970	0	0	0	0.030
3	10	1 000 000	0.829	0	0	0.011	0.160	0.966	0	0	0	0.034
3.5	1	5	0.014	0.986	0	0	0	0.012	0.988	0	0	0
3.5	1	10	0	1	0	0	0	0	1	0	0	0
3.5	1	100	0	1	0	0	0	0	1	0	0	0
3.5	1	1 000	0	1	0	0	0	0	1	0	0	0
3.5	10	5	0.945	0.002	0.001	0	0.052	0.988	0.009	0	0	0.003
3.5	10	10	0.901	0	0.006	0	0.093	0.983	0	0.005	0	0.012
3.5	10	100	0.859	0	0.135	0	0.006	0.506	0	0.494	0	0
3.5	10	1 000	0.003	0	0.997	0	0	0	0	1	0	0
3.5	10	10 000	0	0	1	0	0	0	0	1	0	0

is equivalent to the statement that the largest Ward's distance between any mixed cluster and any remaining point from $\mathbf{Y}(d)$ is smaller than the minimum Ward's distance between any two subclusters of $\mathbf{Y}(d)$.

Similarly to the proof of Theorem 1, for any $0 < a_1 < a_2 < a_3 < a_4 < \infty$, $P[(B)] \geq P(A_d \cap B_d \cap C_d)$ and

$$P\left(A_d \cap B_d \cap C_d\right)^c \leq P\left(E_d^{1c}\right) + P\left(E_d^{2c}\right) + P\left(E_d^{3c}\right) + P\left(E_d^{4c}\right) \to 0$$

where the conditions for the convergence of probabilities and events E_1, E_2, E_3 and E_4 are provided in Lemma 4. Using simple algebra, conditions on a_1, a_2, a_3 and a_4 derived in Lemma 4 are simplified to $\sigma_1^2 < \sigma_2^2$ and $\mu^2 < (\sigma_2^2 - \sigma_1^2)/n$. The case when the second cluster is combined first will lead to $\sigma_1^2 > \sigma_2^2$ and $\mu^2 < (\sigma_1^2 - \sigma_2^2)/n$. Hence, the general condition is $\mu^2 < |\sigma_2^2 - \sigma_1^2|/n$.

The proof of the fact that the probability of the clustering behavior (A) converges to 1 follows the steps of the second part of the proof of Theorem 1 and uses results of Lemma 4. Note, that in this case the relationships between n, m and d are simply $\alpha, \beta \in (0, 1)$ or $d/n \to \infty$ and $d/m \to \infty$ and the condition for clustering behavior (A) is $\mu^2 > |\sigma_2^2 - \sigma_1^2|/n$.

To prove that the probability of clustering behavior (AB) converges to 1, one has to show that the probability of the following events converges to one: $D_1 = \{$During the first n steps only points from $\mathbf{X}(d)$ are joined$\}$, $D_2 = \{$During the $(n + 1)$th step one point from $\mathbf{Y}(d)$ is combined with cluster $\mathbf{X}\}$ and $D_3 = \{$Ward's distance between some two points from $\mathbf{Y}(d)$ is smaller than Ward's distance between some point from $\mathbf{Y}(d)$ and mixed cluster $\mathbf{M}\}$. It was shown in the proof of part (a) of Theorem 2 and Lemma 4 that the $P(D_1 \cap D_2) \to 1$. Also, it was shown that $P(D_3^c) \to 0$ in a more general setting. Now, event D_3 does not require maximum or minimum distance. This fact makes the behavior (AB) easier to achieve than behavior (B). In particular, in addition to $\alpha < 1$ and $\beta < 1$, the required conditions are: $\max(4\beta - 2\alpha, 2\alpha + \beta) < 1$ when $\|\vec{\mu}_1 - \vec{\mu}_2\| = 0$, or $\max(2\beta - 2\alpha, \beta) < 1$ when $\|\vec{\mu}_1 - \vec{\mu}_2\| > 0$. □

Appendix B. Simulation results

See Tables B.1 and B.2.

Table B.2
Simulated probabilities of asymptotic ($d \to \infty$) clustering behaviors based on Ward's linkage function. Data is simulated from the mixture of n observations from the d-dimensional $N_d(\vec{0}, I_d)$ and $m = 20$ observations from the $N_d(\mu \mathbb{1}_d, \sigma^2 I_d)$ distributions. Clustering behaviors (A), (B) and (AB) are described in Section 5. Behavior (A*) is a special case ($\sigma_1^2 < \sigma_2^2$) of behavior (A). This confirms the results of Theorem 2.

μ	σ^2	n	d	(Other)	(A)	(A*)	(B)	(AB)
0.7	1	10	5	1	0	0	0	0
0.7	1	10	10	1	0	0	0	0
0.7	1	10	100	0.279	0.721	0	0	0
0.7	1	10	1 000	0	1	0	0	0
0.7	10	10	5	1	0	0	0	0
0.7	10	10	10	0.984	0	0	0	0.016
0.7	10	10	100	0.847	0	0	0	0.153
0.7	10	10	1 000	0.451	0	0	0	0.549
0.7	10	10	10 000	0	0	0	0	1
0.7	10	10	100 000	0	0	0	0.702	0.298
0.7	10	10	1 000 000	0	0	0	1	0
0.7	1	30	5	1	0	0	0	0
0.7	1	30	10	1	0	0	0	0
0.7	1	30	100	0.398	0.602	0	0	0
0.7	1	30	1 000	0	1	0	0	0
0.7	10	30	5	1	0	0	0	0
0.7	10	30	10	1	0	0	0	0
0.7	10	30	100	0.943	0	0.002	0	0.055
0.7	10	30	1 000	0.689	0	0.300	0	0.011
0.7	10	30	10 000	0.002	0	0.998	0	0
0.7	10	30	100 000	0	0	1	0	0
1.5	1	10	5	0.898	0.102	0	0	0
1.5	1	10	10	0.540	0.46	0	0	0
1.5	1	10	100	0	1	0	0	0
1.5	1	10	1 000	0	1	0	0	0
1.5	10	10	5	1	0	0	0	0
1.5	10	10	10	0.991	0.002	0	0	0.007
1.5	10	10	100	0.539	0	0.449	0	0.012
1.5	10	10	1 000	0	0	1	0	0
3	1	10	5	0.047	0.953	0	0	0
3	1	10	10	0	1	0	0	0
3	1	10	100	0	1	0	0	0
3	10	10	5	0.911	0.089	0	0	0
3	10	10	10	0.545	0.305	0.148	0	0.002
3	10	10	100	0	0	1	0	0

References

[1] J. Ahn, J.S. Marron, K.M. Muller, Y.-Y. Chi, The high-dimension, low-sample-size geometric representation holds under mild conditions, Biometrika 94 (2007) 760–766.

[2] G. Casella, J.T. Hwang, Limit expressions for the risk of james-stein estimators, Canad. J. Statist. 10 (1982) 305–309.

[3] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (2008) 849–911.

[4] L. Fisher, J.W. Van Ness, Admissible clustering procedures, Biometrika 58 (1971) 91–104.

[5] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, J. Roy. Statist. Soc. Ser. B 67 (2005) 427–444.

[6] J.A. Hartigan, Consistency of single linkage for high-density clusters, J. Amer. Statist. Assoc. 76 (1981) 388–394.

[7] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer, 2009.

[8] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.

[9] P.J. Huber, Robust regression: asymptotics, conjectures and Monte Carlo, Ann. Statist. 1 (1973) 799–821.

[10] S.K. Jung, J.S. Marron, PCA consistency in high dimension, low sample size context, Ann. Statist. 37 (6B) (2009) 4104–4130.

[11] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, pp. 281–297.

[12] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR-Sbornik 1 (1967) 457.

[13] L.L. McQuitty, Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies, Educational Psychol. Measurement 17 (1957) 207–229.

[14] G.W. Milligan, An examination of the effect of six types of error perturbation on fifteen clustering algorithms, Psychometrika 45 (1980) 325–342.

[15] S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency, Ann. Statist. 12 (1984) 1298–1309.

[16] S. Portnoy, Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation, Ann. Statist. 13 (1985) 1403–1417.

[17] X. Qiao, H.H. Zhang, Y. Liu, M.J. Todd, J.S. Marron, Weighted distance weighted discrimination and its asymptotic properties, J. Amer. Statist. Assoc. 105 (489) (2010) 401–414. http://dx.doi.org/10.1198/jasa.2010.tm08487.

[18] D. Shen, H. Shen, J.S. Marron, A General Framework for Consistency of Principal Component Analysis, 1–28, 2012. Retrieved from http://arxiv.org/abs/1211.2671.

[19] R.R. Sokal, C.D. Michener, A statistical method for evaluating systematic relationships, Univ. Kansas Sci. Bull. 38 (1958) 1409–1438.

[20] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, first ed., Addison-Wesley, 2006.

[21] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Molecular Inform. 29 (2010) 476–488.

[22] A.W. Van Der Vaart, Asymptotic Statistics, Cambridge Series on Statistical and Probabilistic Mathematics, Cambridge University Press, 2000.

[23] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc. 58 (1963) 236–244.

[24] E.P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions, Ann. of Math. (2) 62 (1955) 548–564.

[25] E.P. Wigner, On the distribution of the roots of certain symmetric matrices, Ann. of Math. (2) 67 (1958) 325–327.

[26] K. Yata, M. Aoshima, PCA consistency for non-Gaussian data in high dimension, low sample size context, Comm. Statist. Theory and Methods 38 (2009) 2634–2652.

[27] K. Yata, M. Aoshima, Intrinsic dimensionality estimation of high-dimension, low sample size data with D-asymptotics, Comm. Statist. Theory Methods 39 (2010) 1511–1521.

[28] K. Yata, M. Aoshima, Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, J. Multivariate Anal. 105 (2012) 193–215.

[29] V.J. Yohai, R.A. Maronna, Asymptotic behavior of M-estimators for the linear model, Annals of Statistics 7 (1979) 258–268.