Statistica Sinica **22** (2012), 443-464
doi:http://dx.doi.org/10.5705/ss.2010.148

CLUSTERING HIGH DIMENSION, LOW SAMPLE SIZE DATA USING THE MAXIMAL DATA PILING DISTANCE

Jeongyoun Ahn, Myung Hee Lee and Young Joo Yoon

University of Georgia, Colorado State University and Konkuk University

Abstract: We propose a new hierarchical clustering method for high dimension, low sample size (HDLSS) data. The method utilizes the fact that each individual data vector accounts for exactly one dimension in the subspace generated by HDLSS data. The linkage that is used for measuring the distance between clusters is the orthogonal distance between affine subspaces generated by each cluster. The ideal implementation would be to consider all possible binary splits of the data and choose the one that maximizes the distance in between. Since this is not computationally feasible in general, we use the singular value decomposition for its approximation. We provide theoretical justification of the method by studying high dimensional asymptotics. Also we obtain the probability distribution of the distance measure under the null hypothesis of no split, which we use to propose a criterion for determining the number of clusters. Simulation and data analysis with microarray data show competitive clustering performance of the proposed method.

Key words and phrases: Hierarchical clustering, high dimension, low sample size data, maximal data piling, singular value decomposition

1. Introduction

Clustering high dimension, low sample size (HDLSS) data is an important task in many application areas (Datta and Datta (2003); Loewenstein et al. (2008)), especially in the area of microarray gene expression data analysis. Pan (2006), Pan and Shen (2007), and Wang and Zhu (2008) take a model-based approach, focusing on dimension reduction via feature selection. The singular value decomposition is also popular in high dimensional clustering (Liu et al. (2003); Wall and Dyck, and Brettin (2001); Wall, Rechtsteiner, and Rocha (2003)), especially for the bi-clustering analysis that clusters both genes and samples.

A main interest in clustering is how to measure the distance between clusters. Some classical measures include single, complete, average, and centroid linkage to name a few. However, approaches based on the pairwise L_2 distance between data vectors, such as single and complete linkage methods, do not work well in HDLSS due to the fact that all observations are far apart from each other (Beyer et al. (1999); Hinneburg, Aggarwal, and Keim (2000)). In particular, it is known that minimum and maximum pairwise distances are approximately

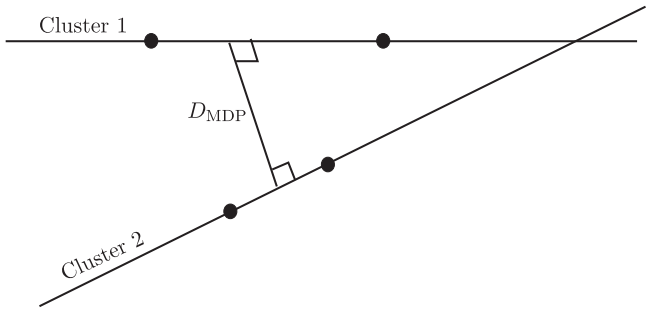

Figure 1. Illustration of the MDP distance between two clusters of size two. Note that the two subspaces are not intersecting in the 3-d space. D_{MDP} is the orthogonal distance between the affine subspaces (lines) generated by the data vectors in each cluster.

indistinguishable when the dimension of the data is much larger than the sample size (Steinbach, Ertoz, and Kumar (2003)).

In this paper we consider a distance measure that can only be defined for HDLSS data. Assume that there are N data vectors in $\mathbb{R}^d$ with $d > N$ and that the data are non-degenerate in the sense that they generate a subspace of dimension N. Since each data vector accounts for exactly one dimension in the subspace, each cluster generates a subspace with the dimension equal to the size of the cluster. This fact leads us to consider a distance that measures how far these subspaces are from each other. Thus we propose to use the orthogonal distance between affine subspaces generated by each cluster. Figure 1 illustrates this distance, denoted by D_{MDP}, when two clusters have two data points each. We call it the maximal data piling (MDP) distance since it is also the distance between projections by the MDP direction vector (Ahn and Marron (2010)) that is explained in detail in Section 2.1.

In Section 2.2, we propose a hierarchical clustering algorithm based on the MDP distance. The proposed algorithm begins with a single cluster containing all observations and makes successive splits. Ideally we wish to search the entire space of possible binary splits and choose the split yielding the largest MDP distance. A direct implementation of this idea is not practical for a decent size of N, as the size of the search space is $2^{N-1}-1$. Thus we suggest an approximating algorithm using the singular value decomposition (SVD) that effectively reduces the search space.

Asymptotic studies related to HDLSS data involve the dimension d tending to infinity. In Section 3.1 we investigate the large-d asymptotic properties of the MDP distance and the proposed clustering algorithm. Employing the HDLSS geometric representation by Hall, Marron, and Neeman (2005) and Ahn et al.

(2007), we show that the MDP distance is maximized at the true split and that the SVD and MDP are equivalent in the large-d limit.

It is a key component in clustering analysis to determine the number of clusters. AIC or BIC types of measures are common choices for model-based methods. In HDLSS clustering Liu et al. (2008) proposed a method with which one can calculate an empirical p-value for the significance of a binary split. We propose a testing procedure for the significance of a split in Section 3.2 that is based on the probability distribution of the MDP distance under the null hypothesis of no split.

A simulation study to investigate the performance of the proposed method is presented in Section 4. Section 5 applies the proposed method to five microarray data examples. Section 6 ends the paper with a discussion.

2. Maximal Data Piling Clustering

2.1. Maximal data piling direction vector

In this section we introduce maximal data piling in the binary discrimination setting where the labels are known. Suppose that we have a Class $+1$ sample $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and a Class -1 sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ in $\mathbb{R}^d$ and that $d \geq N-1 = m+n-1$. Also assume that the data vectors are linearly independent. Then there exist infinitely many direction vectors onto which the data vectors project to only two distinct values, one for each class. Ahn and Marron (2010) showed that there exists a unique direction vector that is optimal among these in the sense that it produces the largest distance between the projections. They named it the maximal data piling (MDP) direction vector since it maximizes not only the amount of data piling but also the distance between the piling sites. See Marron, Todd, and Ahn (2007) and Ahn and Marron (2010) for detailed discussions on the data piling phenomenon in HDLSS discrimination.

In what follows we introduce the mathematical formulation of the MDP. Let $\mathbf{X}_{(d\times m)} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ and $\mathbf{Y}_{(d\times n)} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ be data matrices for each class. We center each row and let $\mathbf{X}_c$ and $\mathbf{Y}_c$ denote their mean-subtracted centered version. Also let $\mathbf{C}_{(d\times N)} = [\mathbf{X}_c, \mathbf{Y}_c]$ denote the horizontal concatenation of the two matrices. Let $\mathbf{w} = \bar{\mathbf{x}} - \bar{\mathbf{y}}$ denote the group mean difference vector and let $\mathbf{A}^\dagger$ be the Moore–Penrose generalized inverse of a matrix $\mathbf{A}$. Note that $\mathbf{P} = \mathbf{C}\mathbf{C}^\dagger$ is the projection matrix to the column space of $\mathbf{C}$.

Consider the following optimization problem: find $\mathbf{v}$ that maximizes $(\mathbf{v}'\mathbf{w})^2$ subject to $\mathbf{C}'\mathbf{v} = \mathbf{0}$ and $\|\mathbf{v}\| = 1$. The solution is given by the projection of $\mathbf{w}$ onto the orthogonal complement of the column space of $\mathbf{C}$,

$$\mathbf{v}_{\text{MDP}} \propto (\mathbf{I}_d - \mathbf{P})\mathbf{w}, \tag{2.1}$$

where $\mathbf{I}_d$ is the d-dimensional identity matrix. Note that $\mathbf{v}_{\text{MDP}}$ has a specific position in the data space. Since both $\mathbf{w}$ and the column vectors of $\mathbf{C}$ are in the subspace generated by globally centered data, $\mathbf{v}_{\text{MDP}}$ lies in that $(N-1)$-dimensional subspace while being orthogonal to the $(N-2)$-dimensional subspace generated by the class-wise centered data vectors.

The distance between the projections of Class $+1$ and Class -1 data vectors, the MDP distance, is

$$D_{\text{MDP}} = |\mathbf{w}'\mathbf{v}_{\text{MDP}}| = \frac{\mathbf{w}'(\mathbf{I}_d - \mathbf{P})\mathbf{w}}{\|(\mathbf{I}_d - \mathbf{P})\mathbf{w}\|} = \|(\mathbf{I}_d - \mathbf{P})\mathbf{w}\|^{1/2}, \tag{2.2}$$

where $\|\cdot\|$ is the L_2 norm. An equivalent formula to (2.1) is

$$\mathbf{v}_{\text{MDP}} \propto \mathbf{Z}'^{\dagger}\boldsymbol{\ell}_0, \tag{2.3}$$

where $\mathbf{Z}$ is the centered data matrix obtained by subtracting the overall mean from the whole data matrix $[\mathbf{X}, \mathbf{Y}]$, and $\boldsymbol{\ell}_0 = [\mathbf{1}'_m, -\mathbf{1}'_n]'$ is the N-vector of class labels. The MDP distance derived from (2.3) is

$$D_{\text{MDP}} = \frac{2}{\|\mathbf{Z}'^{\dagger}\boldsymbol{\ell}_0\|}, \tag{2.4}$$

which is equivalent to (2.2). See Ahn and Marron (2010) for the derivation of the formulas.

2.2. Clustering with the maximal data piling distance

Given a data set with unknown class membership, the MDP clustering finds successive binary splits, each of which creates two clusters in such a way that the affine subspaces generated by them are as far away from each other as possible. At each split the optimization problem is to maximize the MDP distance D_{MDP} in (2.4) over all possible choices of the label vector.

Before introducing the clustering algorithm, we demonstrate that D_{MDP} is an appropriate measure for clustering, via simulation. We generated two clusters of size five from d-variate normal distributions with identity covariance where $d =$ 100, 500, 1,000, 2,500, 5,000. The underlying means of the two clusters were set apart by $.25\sqrt{d}$. We ignored the cluster membership, made all possible 5-5 splits, and calculated three distance measures for each split: the MDP distance, minimum pairwise distance between clusters (single linkage), and average pairwise distance (average linkage). Figure 2 depicts the proportion of times out of 1,000 repetitions that the distance from the true split is the largest among all splits, i.e., the true clusters are found by the distance measure in question. The MDP distance is the most effective among the three distance measures, especially when the dimension is large.

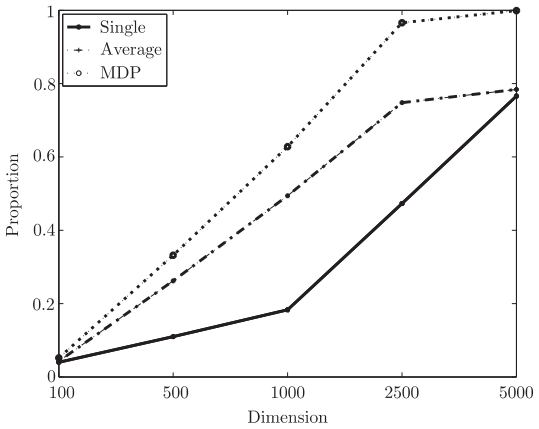

Figure 2. Comparison of the MDP distance with single and average linkages. Proportion of correct identification of the true clusters is shown. The MDP distance is the most efficient, especially for large dimensions.

This result implies that measures based on Euclidean distances between individual data points can be unstable for HDLSS data, as shown by Beyer et al. (1999). However, the MDP distance D_{MDP} depends on the affine subspaces generated by the data in a cluster, rather than specific locations of individual points. Later, in Theorem 1, we identify the condition under which a large-d asymptotic optimality of D_{MDP} is achieved.

With a slight abuse of notation, suppose that we have a cluster $\mathcal{C}$ with N observations split at a certain stage of hierarchical clustering. Then we look for the label vector $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_N)'$, where $\ell_i \in \{-1, +1\}$, that maximizes (2.4). Ideally we wish to search the space $\mathcal{L} = \{-1, +1\}^N \backslash \{-1\}^N \backslash \{+1\}^N$ exhaustively; however, this is computationally infeasible unless N is very small, say less than 10.

In order to circumvent heavy computation, we propose a heuristic for obtaining an approximate solution to the problem. We allow the labels to be continuous so that the modified task is to find $\hat{\boldsymbol{\ell}} \in \mathbb{R}^N$ that minimizes $\|\mathbf{Z}'^{\dagger}\hat{\boldsymbol{\ell}}\|$ where $\|\hat{\boldsymbol{\ell}}\| = 1$. Let the singular value decomposition of $\mathbf{Z}$ be $\mathbf{Z} = \mathbf{U}_{(d\times N)}\mathbf{S}_{(N\times N)}\mathbf{V}'_{(N\times N)}$, where $\mathbf{S} = \text{diag}\{s_1, \ldots, s_{N-1}, 0\}$. Then the optimization problem is

$$\min_{\hat{\boldsymbol{\ell}}\in\mathbb{R}^N, \|\hat{\boldsymbol{\ell}}\|=1} \hat{\boldsymbol{\ell}}'\mathbf{V}\mathbf{S}^{-2}\mathbf{V}'\hat{\boldsymbol{\ell}},$$

which is equivalent to

$$\max_{\hat{\boldsymbol{\ell}}\in\mathbb{R}^N, \|\hat{\boldsymbol{\ell}}\|=1} \hat{\boldsymbol{\ell}}'\mathbf{V}\mathbf{S}^{2}\mathbf{V}'\hat{\boldsymbol{\ell}}.$$

It is clear that the solution for $\hat{\boldsymbol{\ell}}$ is the first right singular vector of $\mathbf{Z}$, or the first eigenvector of $\mathbf{Z}'\mathbf{Z}$. Denoting the vector by $\mathbf{v}_1 = (v_1, \ldots, v_N)'$, we sort the

entries v_i, $i = 1, \ldots, N$, so $v_{j_1} \geq \cdots \geq v_{j_N}$. Then we search for the largest gap, say between v_{j_k} and $v_{j_{k+1}}$, at which we split the current cluster $\mathcal{C}$ into two with one containing the $j_1, \ldots, j_k$th observations and the other containing the $j_{k+1}, \ldots, j_N$th observations. Theorem 2 in Section 3.1 provides a theoretical justification for this approximation and identifies the condition under which $\mathbf{v}_1$ is asymptotically equivalent to the true label vector $\boldsymbol{\ell}_0$. The simulated example for Figure 4 also provides empirical evidence for the approximation.

Of practical concern, the final algorithm implements some fine adjustments to the basic prototype procedure explained in the previous paragraph. First we consider the possibility that subsequent eigenvectors are more informative than the first, especially when there are more than two clusters in the data. Therefore we propose to look at the first $T \geq 2$ eigenvectors, locate the largest gap in the sorted entries of each vector, and then choose the split corresponding to the gap that induces the largest MDP distance.

In order to see the effect of different choices of T, we ran the simulation in Section 4 with different $T = 1, 2$, and 3. In all simulation settings, the effect was minimal and hardly changed the mean error rates. It is always recommended to use a larger value of T since it yields a better chance to find the split with the largest MDP distance. In practice, a guessed number of clusters prior to the analysis can be a reasonable upper bound. One might look at sizes of eigenvalues to determine T, for example using a scree plot or a pre-determined proportion explained by the first few eigenvalues. We used $T = 2$ for the data examples in Section 5.

The second adjustment is to set a minimum size of a cluster. Due to outliers or the sparse nature of HDLSS data, the largest gap sometimes happens to be at the end of the sorted entries, which produces too small a cluster. This can be prevented by putting a constraint on the minimum size of a cluster, say at G. One may try several different values of G and choose the cluster solution with the best interpretation. Empirically we found that $G = 5$ works quite well, thus we use it for simulation and the data examples in this paper. Figure 3 summarizes the proposed algorithm.

The successive splitting process can be stopped according to some criterion, such as when (1) a pre-determined number of clusters is reached, (2) a produced cluster is too small, or (3) a current cluster is too small to divide. We can also evaluate the statistical significance of a split via hypothesis testing, as introduced in Section 3.2. These stopping rules are optional because the decision to discontinue the tree making process should depend on a specific problem. Also one may want to build a large clustering tree and prune it afterward.

Summary

Make successive binary splits in the following way. See Section 3.2 for a stopping rule based on hypothesis testing or the last paragraph of Section 2.2 for user-specified criteria. In order to find an optimal split of a current cluster of size N, obtain the first T eigenvectors of $\mathbf{Z}'\mathbf{Z}$ where $\mathbf{Z}$ is $d \times N$ mean-centered data matrix, denoted by $\mathbf{v}_1, \ldots, \mathbf{v}_T$. For $t = 1, \ldots, T$,

1. Discard the largest G and the smallest G elements of $\mathbf{v}_t$.
2. Find the largest gap between the sorted elements of $\mathbf{v}_t$.
3. Calculate the MDP distance for the split induced by the largest gap.

Choose the split that has the largest MDP distance.

Figure 3. Algorithm of the proposed method

3. Theoretical Properties

3.1. High dimensional asymptotics

In this section the optimality of the MDP method is established by utilizing the asymptotic geometric representation of HDLSS data by Hall, Marron, and Neeman (2005) and Ahn et al. (2007). These papers established the representation under different distributional settings, and later Jung and Marron (2009) did it in a unified framework. In this section we use the conditions in Hall, Marron, and Neeman (2005) since it is easier to discuss the geometry in their setting. Suppose that at a given stage of hierarchical clustering we have $N = m + n$ data vectors from two underlying clusters. Let $X^{(d)} = (X_1, \ldots, X_d)'$ and $Y^{(d)} = (Y_1, \ldots, Y_d)'$ denote the d-variate random vectors for the two clusters, respectively. Hereafter we suppress the use of $^{(d)}$ to simplify notation. Assume that their population structures satisfy the following conditions in Hall, Marron, and Neeman (2005) as d tends to infinity. Note that the condition (e) from Jung and Marron (2009) controls the degree of dependence among variables and modifies the original condition in Hall, Marron, and Neeman (2005) slightly so that it does not depend on the order of the variable entries.

(a) The fourth moments of the entries of the data vectors are uniformly bounded.

(b) $d^{-1} \sum_{j=1}^{d} \mathrm{Var}\,(X_j) \longrightarrow \sigma^2$.

(c) $d^{-1} \sum_{j=1}^{d} \mathrm{Var}\,(Y_j) \longrightarrow \tau^2$.

(d) $d^{-1} \sum_{j=1}^{d} \{E(Y_j) - \mathrm{E}\,(X_j)\}^2 \longrightarrow \mu^2$.

(e) There exists a permutation of the entries of the data vectors such that the sequence of the variables are ρ-mixing for functions that are dominated by quadratics.

Then, as d tends to infinity, the data vectors approximately form an N-polyhedron while each cluster forms a regular simplex with m and n vertices, denoted by $\mathcal{X}$ and $\mathcal{Y}$ respectively. The length of an edge connecting data vectors in $\mathcal{X}$ (or $\mathcal{Y}$) is approximately $\sqrt{2}\sigma$ (or $\sqrt{2}\tau$) after scaling by $\sqrt{d}$. The length of an edge connecting data vectors from different clusters is $\sqrt{\sigma^2+\tau^2+\mu^2}$ after scaling by $\sqrt{d}$.

Theorems 1 and 2 together show that, as d tends to infinity, the proposed hierarchical clustering method finds the optimal split if the underlying means of the two clusters are reasonably well separated compared to the within-cluster variation. Theorem 1 states that the MDP distance for the true split is the largest among all possible splits. Theorem 2 states that the first right singular vector $\mathbf{v}_1$ of $\mathbf{Z}$ is approximately equivalent to the true label vector $\boldsymbol{\ell}_0$. Even though the theorems imply that there are only two classes at each split, based on empirical evidence we conjecture that at each split the data have, approximately, one cluster at each node. These asymptotic results are consistent with our empirical findings in Figure 4. The proofs of the theorems are given in the Appendix.

Theorem 1. *Assume* (a)−(e) *are satisfied and that the minimum size of a cluster is set at* $G \le min\{m,n\}$. *Also suppose that*

$$\mu_0^2 := \mu^2 + \frac{\sigma^2}{m} + \frac{\tau^2}{n} > \max\left\{\frac{m+G}{mG}\sigma^2, \frac{n+G}{nG}\tau^2\right\}. \tag{3.1}$$

Then, in the large-d limit where the HDLSS geometric representation holds, the MDP distance (2.4) *is maximized when* $\boldsymbol{\ell}_0 = [\mathbf{1}_m', -\mathbf{1}_n']'$ *is the true label vector.*

If $\mathbf{u}_1$ is the first left singular vector of $\mathbf{Z}$, Theorem 2 states that $\mathbf{u}_1$ and $\mathbf{v}_{\text{MDP}}$ are approximately equivalent under milder conditions than those for Theorem 1.

Theorem 2. *Suppose assumptions* (a)−(e) *are satisfied and that*

$$\mu_0^2 > \left(\frac{1}{m}+\frac{1}{n}\right)\max\{\sigma^2,\tau^2\}. \tag{3.2}$$

Then, in the large-d limit where the HDLSS geometric representation holds, $\mathbf{u}_1$ *is equal to* $\mathbf{v}_{MDP}$ *which, in turn, implies that* $\mathbf{v}_1$*, the first right singular vector, is equivalent to the true label vector* $\boldsymbol{\ell}_0$.

In what follows we provide a justification of the SVD approximation to MDP via a simulation study. Two clusters of size five were generated from d-variate spherical Gaussian distributions. We set $\sigma^2 = 2$, $\tau^2 = 1$, and $\mu = 0.3, 0.4, 0.5, 0.6$ in the conditions (b)−(d) in Section 3.1. Figure 4 depicts the angle between $\mathbf{u}_1$ and $\mathbf{v}_{\text{MDP}}$ for $d = 10^2, \ldots, 10^5$. Note that in this setting the condition (3.2) in Theorem 2 suggests that $\mu^2 > (1/5 + 1/5) \times 2 - 2/5 - 1/5 = 0.2 = (0.4472)^2$.

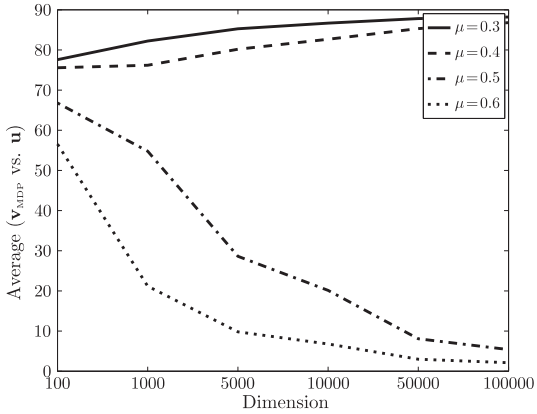

Figure 4. Angle between the first left singular vector $\mathbf{u}_1$ and the MDP direction vector $\mathbf{v}_{\text{MDP}}$ for Gaussian data when $\sigma^2 = 2, \tau^2 = 1, m = n = 5$.

It can be seen that as d increases $\mathbf{u}_1$ and $\mathbf{v}_{\text{MDP}}$ become close to each other when (3.2) is satisfied, but otherwise become nearly orthogonal.

An intuitive interpretation of the conditions (3.1) and (3.2) is provided as follows. Set $\sigma^2 = \tau^2 = 1$, $m = 2$, and $n = 1$. Note that the two conditions are equivalent under this particular setting. Suppose that we consider all possible 2-1 splits of the three data points. In the HDLSS limit, the three data vectors form a triangle. We focus on the lengths of the edges of the triangle, with lengths divided by $\sqrt{d}$. The edge connecting the data points in Class +1 has length $\sqrt{2}$ and the edge connecting data points from different classes has length $\sqrt{\mu^2 + 2}$. Theorems 1 and 2 state that the proposed method finds the optimal split if $\mu_0^2 = \mu^2 + 1/2 + 1 > 3/2$, i.e., $\mu^2 > 0$. Geometrically, this indicates that if the triangle forms an isosceles triangle with the shortest side being the line connecting the two data points from Class +1, the direction of the highest altitude aligns with $\mathbf{v}_{\text{MDP}}$ and also with $\mathbf{u}_1$.

It is worth noting that in general the condition (3.1) for Theorem 1 is stricter than the condition (3.2) for Theorem 2, while the two conditions are equivalent when $k = m = n$. This is because in Theorem 2 we have already determined the labels and thus essentially deal with a classification problem, not clustering. Unsupervised learning problems, such as clustering, need stricter conditions to ensure good performance than supervised learning such as classification.

In the proof of Theorem 2 in the Appendix, we establish that D_{MDP}^2/d approaches $\mu_0^2 = \mu^2 + \sigma^2/m + \tau^2/n$ as d tends to infinity. Thus D_{MDP} is an increasing function of the distance between classes and within-class variation, while it is a decreasing function of the sample size. In the next subsection, we investigate the

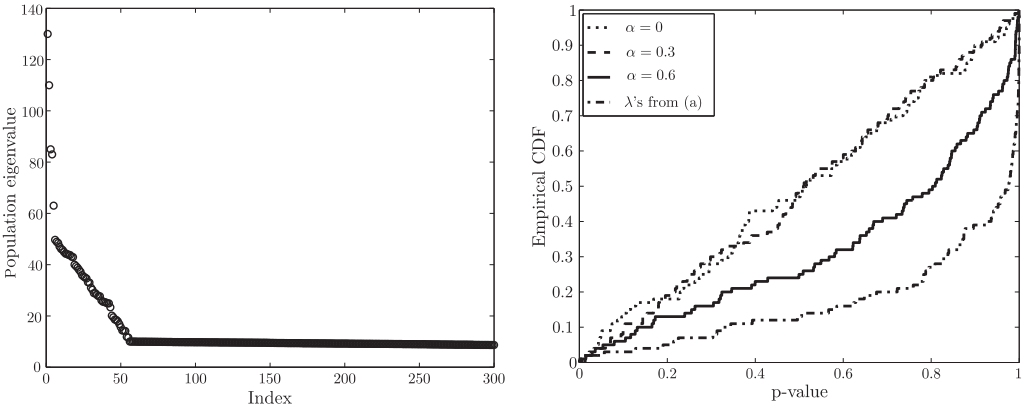

Figure 5. (a) First 300 population eigenvalues for the fourth simulation setting. (b) Empirical cumulative distribution function of the p-values under the four simulation settings.

behavior of D_{MDP} for an arbitrary dimension by studying its probability distribution.

3.2. Distribution of the MDP distance and a stopping criterion

At each step of hierarchical clustering, one important question is whether to continue the procedure or not. To this end it is useful to evaluate the significance of a split under the null hypothesis that there is only one underlying cluster. The following theorem shows that D^2_{MDP} follows a chi-square distribution under the null hypothesis if the data are Gaussian. Note that this theorem also implies that D^2_{MDP}/d converges to $\sigma^2/m + \tau^2/n$ in probability as $d \to \infty$, which is consistent with the discussion in the previous paragraph when $\mu^2 = 0$.

Theorem 3. *Suppose there are m (n) data vectors from Class $+1$ (-1) and that the underlying distributions are $\mathcal{N}_d(\mathbf{0}, \sigma^2\mathbf{I}_d)$ and $\mathcal{N}_d(\mathbf{0}, \tau^2\mathbf{I}_d)$, respectively. Then, D^2_{MDP} between the two classes is $(\sigma^2/m + \tau^2/n)\chi^2_{df}$, where $df = rank(\mathbf{I}_d - \mathbf{P}) = d - N + 2$.*

In what follows we develop a hypothesis test for the significance of a split based on Theorem 3. Suppose that we test whether two clusters of sizes m and n have the same underlying mean. The hypotheses are $H_0 : \mu^2 = 0$ vs. $H_1 : \mu^2 > 0$, where μ is from the condition (d). We suggest the following testing procedure.

Step 1. Calculate D^2_{MDP} based on the given two-cluster assignment.

Step 2. Estimate the within-class variances, σ^2 and τ^2, by $\hat{\sigma}^2 = \text{mean}\{\hat{\sigma}^2_j\}_{j=1}^d$ and $\hat{\tau}^2 = \text{mean}\{\hat{\tau}^2_j\}_{j=1}^d$, where $\hat{\sigma}^2_j$ and $\hat{\tau}^2_j$ are the sample variances of X_j and Y_j, respectively.

Step 3. Calculate the p-value for D^2_{MDP} by computing

$$P\left[\left(\frac{\hat{\sigma}^2}{m}+\frac{\hat{\tau}^2}{n}\right)\chi^2_{d-N+2} > D^2_{\text{MDP}}\right].$$

Applicability of this stopping criterion to more general settings is investigated in a simulation study. For $d = 500, m = 20, n = 15$, two class samples were generated from the same multivariate Gaussian distribution with mean zero. We used four covariance matrices $\mathbf{\Sigma}_j = \mathbf{R}\mathbf{\Lambda}_j\mathbf{R}', j = 1, \ldots, 4$, where $\mathbf{R}$ is an orthogonal matrix whose (l, k)th element is $\sqrt{2/(d+1)}\sin(lk\pi/(d+1))$. The first three had $\mathbf{\Lambda}_j = \text{diag}\{\lambda_1, \ldots, \lambda_d\} = \text{diag}\{d^{\alpha_j}, 1, \ldots, 1\}$ where $\alpha_1 = 0, \alpha_2 = .3$, and $\alpha_3 = .6$; the fourth was chosen to mimic the eigenvalue pattern often observed in data examples. Figure 5(a) displays the first 300 largest eigenvalues of $\mathbf{\Sigma}_4$; Figure 5(b) displays the empirical cumulative distribution function of p-values obtained under the four settings from 100 repetitions. As expected in the spherical case ($\alpha = 0$), the p-values followed a uniform distribution. A fair amount of deviation from sphericity ($\alpha = .3$) did not yield a severe change in the distribution of p-values. Larger deviation from sphericity ($\alpha = .6$) or more realistic situation such as the fourth model tended to make p-values higher.

This simulation study suggests that the testing procedure is conservative in the sense that it tends to prevent one from falsely declaring two clusters to be different. The simulation also implies that it is possible to underestimate the number of clusters using the current testing procedure. In order to avoid the possible underestimation problem, one may also employ a testing procedure such as SigClust of Liu et al. (2008). SigClust may have a higher power for detecting clustering structure for correlated data, even though it assumes Gaussian distributions with the same covariance matrix for each cluster. On the other hand, when two clusters are suspected to have unequal variances but with mild correlations between variables, the proposed test can give a good approximate p-value. We also note that the theoretical justification of the SigClust procedure was done asymptotically as $d \to \infty$, while the proposed test is exact under the assumptions in Theorem 3.

4. Simulation Study

Clustering performance of the proposed method in four different HDLSS settings was investigated. As competitors we chose the sparse K-means clustering (SK-means) of Witten and Tibshirani (2010) with variable selection property, Ward's hierarchical clustering method (Ward (1963)), a model-based clustering method (Mclust) of Fraley and Raftery (2002), and the pivoted QR decomposition method (p-QR) of Zha et al. (2001); the last of these we discuss in detail in Section 6.

Table 1. Results of simulation study. Average clustering errors are shown with standard errors in parentheses.

Setting	d	μ	SK-means		Ward		Mclust		p-QR		MDP	
I	1000	0.6	0.1841	(0.0155)	0.0938	(0.0074)	0.0935	(0.0069)	0.3179	(0.0101)	0.0278	(0.0040)
		0.8	0.0341	(0.0088)	0.0138	(0.0013)	0.0118	(0.0012)	0.3161	(0.0134)	0.0001	(0.0001)
		1	0	(0)	0.0013	(0.0003)	0.0003	(0.0002)	0.3125	(0.0144)	0	(0)
	2000	0.6	0.3403	(0.0108)	0.2618	(0.0145)	0.2618	(0.0145)	0.3377	(0.0080)	0.1109	(0.0106)
		0.8	0.0407	(0.0089)	0.0317	(0.0020)	0.0317	(0.0020)	0.2980	(0.0120)	0.0048	(0.0011)
		1	0.0041	(0.0030)	0.0076	(0.0007)	0.0075	(0.0007)	0.2891	(0.0145)	0	(0)
II	1000	0.6	0.0271	(0.0076)	0.1360	(0.0042)	0.1298	(0.0044)	0.3740	(0.0073)	0.0375	(0.0036)
		0.8	0.0004	(0.0002)	0.0287	(0.0016)	0.0138	(0.0015)	0.3513	(0.0074)	0.0009	(0.0003)
		1	0	(0)	0.0040	(0.0007)	0.0003	(0.0002)	0.3385	(0.0072)	0	(0)
	2000	0.6	0.0905	(0.0131)	0.2670	(0.0088)	0.2670	(0.0088)	0.3735	(0.0086)	0.1568	(0.0107)
		0.8	0.0003	(0.0002)	0.0737	(0.0035)	0.0725	(0.0035)	0.3369	(0.0089)	0.0042	(0.0007)
		1	0	(0)	0.0129	(0.0011)	0.0087	(0.0009)	0.3178	(0.0093)	0	(0)
III	1000	0.2	0.4232	(0.0102)	0.3913	(0.0076)	0.4088	(0.0070)	0.3358	(0.0092)	0.2129	(0.0116)
		0.25	0.4018	(0.0134)	0.3927	(0.0080)	0.4073	(0.0069)	0.2853	(0.0110)	0.0945	(0.0148)
		0.3	0.3571	(0.0167)	0.3830	(0.0081)	0.4042	(0.0071)	0.3033	(0.0131)	0.0277	(0.0102)
	2000	0.2	0.3829	(0.0149)	0.3901	(0.0077)	0.3968	(0.0077)	0.3011	(0.0114)	0.1146	(0.0151)
		0.25	0.2697	(0.0198)	0.3935	(0.0078)	0.4007	(0.0075)	0.3018	(0.0140)	0.0401	(0.0121)
		0.3	0.1728	(0.0189)	0.3681	(0.0093)	0.3869	(0.0088)	0.3013	(0.0152)	0.0042	(0.0042)
III	1000	0.2	0.5185	(0.0098)	0.5637	(0.0043)	0.5661	(0.0042)	0.3897	(0.0077)	0.3736	(0.0102)
		0.25	0.4415	(0.0123)	0.5274	(0.0064)	0.5339	(0.0065)	0.3543	(0.0077)	0.2013	(0.0121)
		0.3	0.3332	(0.0129)	0.4354	(0.0096)	0.4455	(0.0092)	0.3373	(0.0079)	0.0506	(0.0111)
	2000	0.2	0.3766	(0.0120)	0.5005	(0.0072)	0.5016	(0.0072)	0.3535	(0.0072)	0.2440	(0.0153)
		0.25	0.2402	(0.0109)	0.3650	(0.0089)	0.3680	(0.0088)	0.3271	(0.0077)	0.0282	(0.0077)
		0.3	0.1760	(0.0134)	0.2780	(0.0115)	0.2789	(0.0116)	0.3179	(0.0083)	0.0395	(0.0099)

In each setting we took $d = 1{,}000$ and 2,000, among which only 150 are relevant to the clustering task. The first setting is from Wang and Zhu (2008), and the others were designed so that the task becomes gradually more challenging. Each setting was repeated 100 times. In order to make a straightforward comparison, the number of clusters was fixed at the true value for each method.

1. **Setting I - Two clusters with identity covariance**
   The first 150 variables were independently from $\mathcal{N}(0,1)$ for the first cluster and $\mathcal{N}(\mu,1)$ for the second cluster, where $\mu = 0.6, 0.8, 1$. The remaining variables were independently $N(0,1)$ for both clusters. The two clusters were of size 85 and 15, respectively.
2. **Setting II - Three clusters with identity covariance**
   The first 150 variables were $\mathcal{N}(0,1)$ for the first cluster; for the second cluster, the first 75 variables were $\mathcal{N}(\mu,1)$ and the next 75 were $N(-\mu,1)$; for the third cluster, the 150 informative variables were $\mathcal{N}(\mu,1)$, $\mu = 0.6, 0.8, 1$. The noninformative variables were independently $N(0,1)$ for all three clusters. The sizes of the three clusters were 50, 30, and 20, respectively.

3. **Setting III - Two clusters with correlated variables**
   This is similar to Setting I, but informative variables are correlated. The first 150 variables were $\mathcal{N}_{150}(\mathbf{0}, \mathbf{\Sigma})$ for the first cluster and $\mathcal{N}_{150}(\boldsymbol{\gamma}, \mathbf{\Sigma})$ for the other, with diagonal entries 1 and off-diagonal entries 0.5. The mean vector $\boldsymbol{\gamma} = d^{1/2}\mu\mathbf{u}$ where $\mathbf{u}$ was the randomly selected eigenvector of $\mathbf{\Sigma}$ and $\mu = 0.2, 0.25, 0.3$. The two clusters were of size 85 and 15, respectively.
4. **Setting IV - Three clusters with correlated variables**
   This is similar to Setting II. The first 150 informative variables were $\mathcal{N}_{150}(\mathbf{0}, \mathbf{\Sigma})$ for the first cluster, $\mathcal{N}_{150}(\boldsymbol{\gamma}_1, \mathbf{\Sigma})$ for the second, and $\mathcal{N}_{150}(\boldsymbol{\gamma}_2, \mathbf{\Sigma})$ for the third. $\mathbf{\Sigma}$ was the same as in Setting 3, $\boldsymbol{\gamma}_1 = d^{1/2}\mu\mathbf{u}_1$, and $\boldsymbol{\gamma}_2 = d^{1/2}\mu\mathbf{u}_2$, where $\mathbf{u}_1$ and $\mathbf{u}_2$ were the randomly selected eigenvectors of $\mathbf{\Sigma}$ and $\mu = 0.2, 0.25, 0.3$. The sizes of the three clusters were 50, 30, and 20, respectively.

Table 1 displays the average error rates and their standard errors. The clustering error was determined by first finding the maximal correspondence between the true cluster labels and permutations of the estimated cluster labels. Then using that permutation, the error rate was calculated as the proportion of instances that were incorrectly assigned.

We can see that the MDP method performs significantly better than other methods in almost all settings. Note that in Settings I and II all methods are worse for $d = 2{,}000$ than for $d = 1{,}000$ when the clusters are not far apart.

The sparse K-means performed significantly worse in Setting I when μ was small, while in Setting II it was the best. Ward, Mclust, and p-QR performed poorly in all settings, and the former two were especially poor in Settings III and IV with correlated variables.

In Settings III and IV, the distance parameter μ controls the difficulty of the clustering, thus we expect that as μ increases, the error rates should decrease for a reasonable clustering method. The MDP method improved as μ went from .2 to .3, however, the improvement by other methods was minimal. The advantage of the MDP method in the presence of correlated variables has been also noted by Ahn and Marron (2010) for the classification problem.

5. Application to Cancer Microarray Data

Five microarray data sets are used to investigate the performance of the MDP clustering method. The first data set is the lung cancer data in Liu et al. (2008). Detailed description of this data set can be found in Bhattacharjee et al. (2001). We use the preprocessed version of the data for the analysis, see Liu et al. (2008) for details on the preprocessing. The preprocessed data set contains 2530 genes from 56 subjects from four clusters: 20 pulmonary carcinoid (Carcinoid), 13 colon cancer metastasis (Colon), 17 normal lung (Normal), and 6 small cell carcinoma samples (SmallCell).

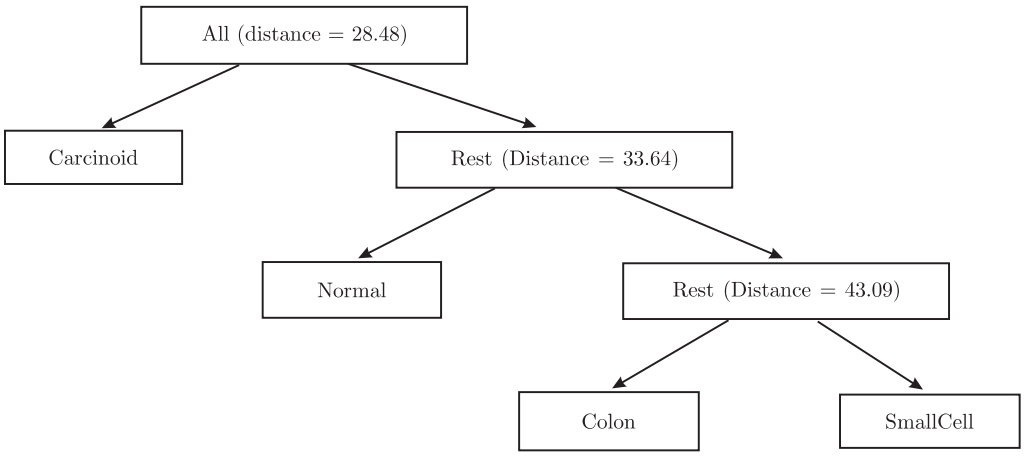

Figure 6. Tree diagram of binary splits of the lung cancer data by the MDP method. The MDP distance of each split is also shown.

Table 2. Clustering results of four microarray gene expression data sets.

	N	d	no. clusters	SK-means	Ward	Mclust	p-QR	MDP
Colon	62	2,000	2	12	30	31	25	15
Breast	49	7,129	2	22	22	N/A	23	18
Prostate	102	6,033	2	41	44	N/A	40	41
Lymphoma	62	4,026	3	1	1	N/A	18	0

This data set is not too challenging in terms of finding the four underlying clusters. The number of misclustered subjects by the methods that were compared in Section 4 are all either 0 or 1. Figure 6 displays the order of the splits by the MDP method. First Carcinoid is separated from the rest, then Normal, and finally Colon and SmallCell are separated. We note that this order is consistent with the order of the p-values for the significance of a split in Liu et al. (2008). It is worth mentioning that the order of splits in Liu et al. (2008) is not determined statistically, but provided by biologists.

We also compared the proposed method with other methods in four microarray gene expression data sets: Colon cancer data from Alon et al. (1999), Breast cancer data from West et al. (2001), Prostate and Lymphoma data from Dettling (2004). Table 2 displays the sample size, dimension, and the number of clusters in each data set, and also reports the number of misclustered observations. As in the simulation in Section 4 we fixed the number of clusters for each algorithm in order to make the comparison easier. The model-based Mclust method can be implemented for only the Colon data due to the high dimensionality of the other three data sets. The MDP clustering method shows competitive performance among the compared methods.

For all data examples analyzed here, we tested the significance of a split using the testing procedure proposed in Section 3.2 The p-values are all virtually zero, which indicates strong evidence for the existence of the clusters.

6. Discussion

It is a common belief that analyzing HDLSS data is more restrictive than low dimensional data. In particular, many traditional multivariate methods in textbooks cannot be directly applied due to the singularity of the covariance matrix. In this work we take a positive stand on the HDLSS problem by viewing the excess of variables as a blessing. Specifically, we make good use of the fact that there are enough dimensions in the data so that the orthogonal distance between affine subspaces of clusters becomes a meaningful measure.

The SVD approximation has also been used for K-means clustering by Zha et al. (2001), as a reviewer pointed out. Both our approach and theirs is based on the fact that the PCA directions provide good approximation to the data matrix. However, we found that the two approaches are essentially quite different. They approximate the label vector via pivoted QR decomposition of the first K-PCA direction vectors collectively, whereas we justify the approximation of the label vector based on the individual PCA direction vector. Moreover, they assume that the Gram matrix of the full data is well-approximated by the partitioned within-class Gram matrices. With our notation in the 2-class settings, this means that $\begin{bmatrix}\mathbf{X}'\\ \mathbf{Y}'\end{bmatrix}[\mathbf{X}\,\mathbf{Y}] \approx \begin{bmatrix}\mathbf{X}'\mathbf{X} & 0\\ 0 & \mathbf{Y}'\mathbf{Y}\end{bmatrix}$. This assumption might be reasonable in some situations, especially in lower dimensional applications. We believe that this assumption is not reasonable for most HDLSS data, as their method did not provide a reasonable solution in a small simulation with a simple HDLSS data set that is not shown here.

There are a few future research directions to pursue. For example, instead of a divisive algorithm one can explore an agglomerative approach when constructing the dendrogram. A possible difficulty of this approach might be the computational burden since each step, a merging in this case, involves many calculations of the MDP distance. Nevertheless, comparing the current divisive algorithm with the agglomerative one may shed light on the stability of the MDP linkage method.

Another future direction is the partition-based approach, as a reviewer suggested. The multi-class version of MDP subspace is a natural starting point. One can do an exhaustive search for the c-cluster solution that has the largest distance between c projected piling points. Implementing this idea effectively is not straightforward as there are many possible allocations, and SVD might not work in this case for reducing the search space.

One can allow the underlying distribution to be invariant under any orthogonal transformation (Kelker (1970)). This allows some dependence between the variables. Theorem 13 in Kelker (1970) provides results for the quadratic forms of a spherical distribution, analogous to Cochran's theorem for the normal distribution. If a spherical distribution is assumed for Theorem 3, the χ^2 changes accordingly to the distribution that essentially corresponds to the sum of ν independent squares of standardized variables. Identifying the distribution of the MDP distance in more general settings is proposed as a future work.

Acknowledgement

Ahn's research was partly supported by the NSF Grant DMS-0805758. The authors are grateful to an associate editor and the reviewers for helpful comments.

Appendix

Proof of Theorem 1. In this proof we use 1 and 2 as the class labels for convenience. In the HDLSS geometrical limit, the data vectors from Class 1 and Class 2 form two simplices, denoted $\mathcal{X}$ and $\mathcal{Y}$. Assume that the respective sample sizes are m and n. Since only the relative locations of the data vectors are of interest for our purposes, we can express the data matrices for $\mathcal{X}$ and $\mathcal{Y}$, after scaling by $\sqrt{d}$, as the following:

$$\mathbf{X}' = \begin{bmatrix} \sigma \begin{bmatrix} 1-\frac{1}{m} & -\frac{1}{m} & \cdots & -\frac{1}{m} \\ -\frac{1}{m} & 1-\frac{1}{m} & \cdots & -\frac{1}{m} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{m} & -\frac{1}{m} & \cdots & 1-\frac{1}{m} \end{bmatrix} & \begin{bmatrix} \delta_{\mathrm{X}} & \cdots & \delta_{\mathrm{X}} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \delta_{\mathrm{X}} & \cdots & \delta_{\mathrm{X}} \end{bmatrix} & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \end{bmatrix},$$

$$\mathbf{Y}' = \begin{bmatrix} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} & \begin{bmatrix} -\delta_{\mathrm{Y}} & \cdots & -\delta_{\mathrm{Y}} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ -\delta_{\mathrm{Y}} & \cdots & -\delta_{\mathrm{Y}} \end{bmatrix} & \tau \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix} \end{bmatrix},$$

where

$$\delta_{\mathrm{X}} = \frac{n}{N}\frac{\mu_0}{\sqrt{d-N}} \quad \text{and} \quad \delta_{\mathrm{Y}} = \frac{m}{N}\frac{\mu_0}{\sqrt{d-N}}.$$

Here $\mathbf{X}'$ is partitioned into $m \times m, m \times n$, and $m \times (d-N)$ sub-matrices, and $\mathbf{Y}'$ is partitioned into $n \times (d-N), n \times m$, and $n \times n$ sub-matrices. Note that this formulation ensures the same roles of σ^2, τ^2, and μ_0^2 as in the geometric representation in Section 3.1.

Since the minimum size of a cluster is k, the space for label vectors is

$$\mathcal{L}_k = \{1,2\}^N \setminus \bigcup_{i=0}^{k-1} \{\{1\}^{N-i}, \{2\}^i\} \setminus \bigcup_{i=0}^{k-1} \{\{1\}^i, \{2\}^{N-i}\}.$$

Then for each label vector $\boldsymbol{\ell} \in \mathcal{L}_k$, we consider the corresponding split of the data and calculate the MDP distance. Let $\mathbf{w} = (w_1, \ldots, w_d)'$ be a d-dimensional unit vector for data projection. Let $W_1 = \sum_{i=1}^{m} w_i$, $W_2 = \sum_{i=m+1}^{d-n} w_i$, and $W_3 = \sum_{i=d-n+1}^{d} w_i$. The projections of data vectors for $\mathcal{X}$ are $\sigma(w_j - m^{-1}W_1) + \delta_{\text{x}} W_2, j = 1, \ldots, m$, and the projected data for $\mathcal{Y}$ are $\tau(w_j - n^{-1}W_3) - \delta_{\text{y}} W_2, j = d-n+1, \ldots, d$. Let I_{ij} be the collection of indices of samples that are actually from Class i, but classified into Class j using the label $\boldsymbol{\ell}$ for $i, j = 1, 2$. For example, if all but the last one from Class 1 are classified into Class 1, then $I_{11} = \{1, \ldots, m-1\}$ and $I_{12} = \{m\}$. Furthermore, let J_{2i} denote the shifted index set of I_{2i}, $J_{2i} = d - I_{2i} + 1$ for $i = 1, 2$. Now the piling conditions of the projected data can be written as

$$\sigma(w_i - \frac{1}{m}W_1) + \delta_{\text{x}} W_2 = \tau(w_j - \frac{1}{n}W_3) - \delta_{\text{y}} W_2 \equiv c, \quad i \in I_{11}, j \in J_{21}, \quad \text{(A.1)}$$

$$\sigma(w_i - \frac{1}{m}W_1) + \delta_{\text{x}} W_2 = \tau(w_j - \frac{1}{n}W_3) - \delta_{\text{y}} W_2 \equiv c^*, \quad i \in I_{12}, j \in J_{22}, \quad \text{(A.2)}$$

for some constants c and c^*. These imply that

$$\tau(q - r) = \sigma(p - s), \quad \text{(A.3)}$$

where $w_i = p, i \in I_{11}, w_i = s, i \in I_{12}, w_j = q, j \in J_{21}$, and $w_j = r, j \in J_{22}$. Thus the piling constraints can be simplified as

$$\tau(q - r)(\frac{n_{22}}{n} - \frac{n_{12}}{m}) = (\delta_{\text{x}} + \delta_{\text{y}}) W_2, \quad \text{(A.4)}$$

where $n_{1j} = |I_{1j}|, n_{2j} = |J_{2j}|$. The second step is to maximize the distance between the two piling sites (A.1) and (A.2),

$$\max\{\tau(q - \frac{1}{n}W_3) - \delta_{\text{y}} W_2 - \tau(r - \frac{1}{n}W_3) + \delta_{\text{y}} W_2\}^2 = \max \tau^2 (q - r)^2,$$

subject to the unit length condition

$$n_{11}p^2 + n_{12}s^2 + (d - N)u^2 + n_{21}q^2 + n_{22}r^2 = 1 \quad \text{(A.5)}$$

and (A.4), where $u = w_j$, $j = m+1, \ldots, d-n+1$. The corresponding Lagrangian is

$$\begin{aligned} L &= \tau^2 (q - r)^2 - \lambda_1 \{\tau(q - r)(\frac{n_{22}}{n} - \frac{n_{12}}{m}) - (\delta_{\text{x}} + \delta_{\text{y}}) W_2\} \\ &\quad -\lambda_2 \{n_{11}p^2 + n_{12}s^2 + (d - N)u^2 + n_{21}q^2 + n_{22}r^2 - 1\}, \end{aligned}$$

where λ_1 and λ_2 are positive Lagrangian multipliers. Solving $\partial L/\partial q = 0$ and $\partial L/\partial r = 0$ gives

$$\begin{aligned}\frac{\partial L}{\partial q} &= 2\tau^2(q-r) - \lambda_1\tau(\frac{n_{22}}{n} - \frac{n_{12}}{m}) - 2\lambda_2 n_{21} q = 0,\\ \frac{\partial L}{\partial r} &= -2\tau^2(q-r) + \lambda_1\tau(\frac{n_{22}}{n} - \frac{n_{12}}{m}) - 2\lambda_2 n_{22} r = 0.\end{aligned}$$

This requires that $n_{21}q = -n_{22}r$. Similarly, $\partial L/\partial p = 0$ and $\partial L/\partial s = 0$ yield $n_{11}p = -n_{12}s$. Together with (A.3), we can express p, q, and s in terms of r as

$$q = -\frac{n_{22}}{n_{21}}r,\ p = -\frac{\tau}{\sigma}\frac{n}{m}\frac{n_{12}}{n_{21}}r,\ \text{and}\ s = \frac{\tau}{\sigma}\frac{n}{m}\frac{n_{11}}{n_{21}}r.$$

By plugging these into (A.4), we can also express u in terms of r as

$$u = \frac{\tau}{\mu_0}\frac{n}{n_{21}}\left(\frac{n_{12}}{m} - \frac{n_{22}}{n}\right)\frac{r}{\sqrt{d-N}}.$$

Finally, from (A.5), we get

$$r^2 = \left\{\frac{\tau^2}{\sigma^2}\frac{n^2}{n_{21}^2}\frac{n_{11}n_{12}}{m} + \frac{\tau^2}{\mu_0^2}\frac{n^2}{n_{21}^2}(\frac{n_{12}}{m} - \frac{n_{22}}{n_{21}})^2 + \frac{n_{22}n}{n_{21}}\right\}^{-1},$$

and the maximized distance is

$$D^2_{\text{MDP}}(\boldsymbol{\ell}) = \mu_0^2\left\{\frac{\mu_0^2}{\sigma^2}\frac{n_{11}n_{12}}{m} + \frac{\mu_0^2}{\tau^2}\frac{n_{21}n_{22}}{n} + (\frac{n_{12}}{m} - \frac{n_{22}}{n})^2\right\}^{-1}. \tag{A.6}$$

If $\boldsymbol{\ell} = [\mathbf{1}'_m, -\mathbf{1}'_n]'$ is the true label vector, we have $(n_{11}, n_{12}, n_{21}, n_{22}) = (m, 0, 0, n)$, and the MDP distance becomes μ_0^2. In order to prove that $D^2_{\text{MDP}}(\boldsymbol{\ell}) \le \mu_0^2, \forall \boldsymbol{\ell} \in \mathcal{L}_k$, it suffices to show that the multiplication factor in (A.6) is at most one for all possible choices of label vectors. Let

$$f(x,y) = \frac{\mu_0^2}{\sigma^2}\frac{x(m-x)}{m} + \frac{\mu_0^2}{\tau^2}\frac{(n-y)y}{n} + \left(1 - \frac{x}{m} - \frac{y}{n}\right)^2,$$

where $(x, y) \in \mathcal{S}_k$, the convex hull containing legitimate values of (n_{11}, n_{22}). Then the inverse of the multiplication factor in (A.6) can be viewed as a function f evaluated at $(x, y) = (n_{11}, n_{22})$. Note that the pairs of (n_{11}, n_{22}) that produce a cluster of size less than a preset value k are excluded from $\mathcal{S}_k$. Notice that, as shown in Figure 7(a), for a fixed x (or y), f is a quadratic concave function of y (or x). Furthermore, along the diagonal, $\{(x, y) : y - x = k'\}$ for a fixed k', it can be checked that $f(x, y) = f(x, x + k')$ is also a quadratic concave function of x. Thus, the minimum can occur only at the one of the six corner points of $\mathcal{S}_k$:

$$(0,0), (0, n-k), (k, n), (m, n), (m, k),\ \text{and}\ (m-k, 0),$$

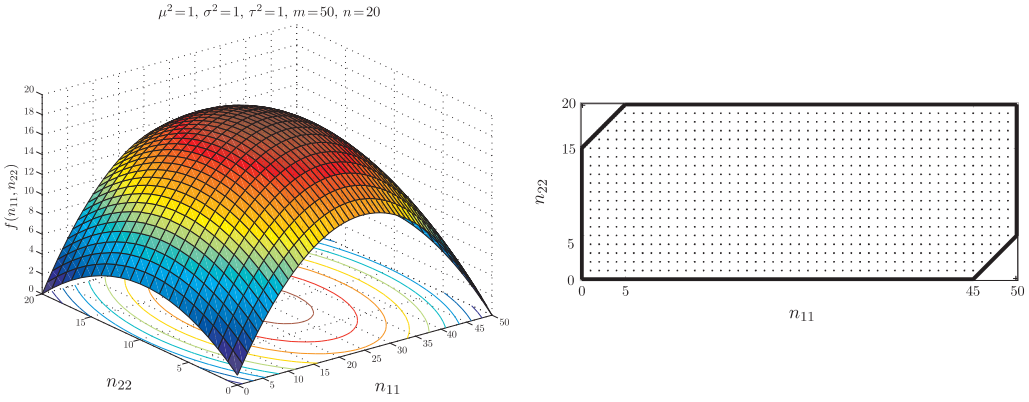

Figure 7. (a) Inverse of multiplication factor in (A.6). (b) the convex hull $\mathcal{S}_k$ containing legitimate values of (n_{11}, n_{22}). Minimum is achieved at one of the six corner points of $\mathcal{S}_k$.

shown in Figure 7(b).

Both $(0,0)$ and (m,n) yield the factor of one as these correspond to true classification in the sense that all the Class 1 samples are separated from the Class 2 samples. Both $(0,n-k)$ and (m,k) imply that k samples from Class 2 are separated from the rest of the samples and yield the multiplication factor $kn^{-1}\{(n-k)\mu_0^2\tau^{-2}+kn^{-1}\}$. Similarly, (k,n) and $(m-k,0)$ represent that k samples from Class 1 are classified differently from the rest and yield the factor $km^{-1}\{(m-k)\mu_0^2\sigma^{-2}+km^{-1}\}$. It can be easily checked that if

$$\mu_0^2 > \max\left\{\frac{n+k}{nk}\tau^2, \frac{m+k}{mk}\sigma^2\right\},$$

then $f(x,y)\geq 1, \quad \forall (x,y)\in\mathcal{S}_k$, with the minimum of one achieved at (m,n) (or $(0,0)$), and this completes the proof.

Proof of Theorem 2. Assume the same representation of the data matrices as in the proof of Theorem 1, and let $\mathbf{w}$ be a projection vector as in the proof of Theorem 1. The optimization conditions for MDP are that the projection of the data vectors in $\mathcal{X}$ $(\mathcal{Y})$ is the projection of their mean vector, and that the distance between the mean projections from each class is maximized. The piling constraint requires that

$$w_1=\cdots=w_m, \quad w_{d-n+1}=\cdots=w_d. \tag{A.7}$$

The maximal distance criterion needs the maximization of $D^2_{\text{MDP}}=\{(\delta_{\text{X}}+\delta_{\text{Y}})W_2\}^2$. Thus the MDP optimization is to maximize W_2^2 subject to (A.7). The solution to this problem is $w_j=0, j=1,\ldots,m,d-n+1,\ldots,d$, $w_j=(d-N)^{-1/2}, j=$

$m+1,\ldots,d-n$, since $\|\mathbf{w}\|^2=1$. Note that this solution indicates that D^2_{MDP} converges to μ_0^2 in probability as $d\to\infty$.

As for the optimization for the principal component analysis (PCA), let $S_1=\sum_{j=1}^m w_j^2$, $S_2=\sum_{j=m+1}^{d-n} w_j^2$, and $S_3=\sum_{j=d-n+1}^{d} w_j^2$. Then the PCA optimization is to maximize

$$G(\mathbf{w})=\sigma^2(S_1-\frac{1}{m}W_1^2)+\tau^2(S_3-\frac{1}{n}W_3^2)+(m\delta_{\text{X}}^2+n\delta_{\text{Y}}^2)W_2^2.$$

The Lagrangian of this problem is

$$L=\sigma^2(S_1-\frac{1}{m}W_1^2)+\tau^2(S_3-\frac{1}{n}W_3^2)+(m\delta_{\text{X}}^2+n\delta_{\text{Y}}^2)W_2^2-\lambda(S_1+S_2+S_3-1),$$

where $\lambda>0$ is the Lagrange multiplier. Solving $\partial L/\partial w_j=0$, $j=1,\ldots,d$, gives

$$\begin{aligned}(\sigma^2-\lambda)w_j-\frac{\sigma^2}{m}W_1&=0,\quad j=1,\ldots,m,\\(\tau^2-\lambda)w_j-\frac{\tau^2}{n}W_3&=0,\quad j=d-n+1,\ldots,d,\\W_1=W_3&=0,\\\{(d-N)(m\delta_{\text{X}}^2+n\delta_{\text{Y}}^2)-\lambda\}W_2&=0.\end{aligned}$$

If $\lambda=(d-N)(m\delta_{\text{X}}^2+n\delta_{\text{Y}}^2)$, then $w_j=0$, $j=1,\ldots,m,d-n+1,\ldots,d$, and the objective function $G(\mathbf{w})$ is maximized at $w_j=(d-N)^{-1/2}, j=m+1,\ldots,d-n$, which is the solution from the MDP optimization. This solution yields the maximum of $G=(m\delta_{\text{X}}^2+n\delta_{\text{Y}}^2)(d-N)=(1/m+1/n)^{-1}\mu_0^2$. If $\lambda=\sigma^2$ (or τ^2), then it is straightforward to see that the maximum value of G is σ^2 (or τ^2). Thus, under the assumption in the theorem that $\mu_0^2>(1/m+1/n)\max(\sigma^2,\tau^2)$, the MDP solution is the same as PCA.

Note that since $\mathbf{v}_{\text{MDP}}$ produces dichotomous projections, the projected data $\mathbf{Z}'\mathbf{v}_{\text{MDP}}$ can be considered as a label vector, denoted by $\tilde{\boldsymbol{\ell}}$. Because we have established $\mathbf{u}_1=\mathbf{v}_{\text{MDP}}$ above, we have $\mathbf{v}_1=s_1^{-1}\mathbf{Z}'\mathbf{u}_1=s_1^{-1}\mathbf{Z}'\mathbf{v}_{\text{MDP}}=s_1^{-1}\tilde{\boldsymbol{\ell}}$, where s_1 is the first singular value. In other words, $\mathbf{v}_1$ is a dichotomous label vector.

Proof of Theorem 3. It suffices to show that $\mathbf{w}'\mathbf{P}\mathbf{w}$ has a $c^2\chi^2_{N-2}$ distribution with $c^2=(\sigma^2/m+\tau^2/n)$. Note that Cochran's theorem is not readily applicable since the projection operator $\mathbf{P}$ depends on the data. Instead we directly show that the distribution of the quadratic form depends on $\mathbf{C}$ only through its rank, $N-2$. There exists a $d\times d$ orthogonal matrix $\mathbf{Q}$ such that $\mathbf{P}=\mathbf{Q}'\mathbf{\Lambda}\mathbf{Q}$ where $\mathbf{\Lambda}$ is a diagonal matrix. Note that $\mathbf{P}$ is symmetric and idempotent, and thus all the diagonal entries of $\mathbf{\Lambda}$ are either zero or one and the multiplicity of one is $\text{rank}(\mathbf{P})=N-2$. We complete the proof by noting that $\mathbf{w}'\mathbf{P}\mathbf{w}=\mathbf{w}'\mathbf{Q}'\mathbf{Q}\mathbf{P}\mathbf{Q}'\mathbf{Q}\mathbf{w}=(\mathbf{Q}\mathbf{w})'\mathbf{\Lambda}(\mathbf{Q}\mathbf{w})$, and $\mathbf{Q}\mathbf{w}$ is $\mathcal{N}_d(\mathbf{0},c^2\mathbf{I}_d)$.

References

Ahn, J. and Marron, J. S. (2010), The maximal data piling direction for discrimination. *Biometrika*, **97**, 254-259.

Ahn, J., Marron, J. S., Muller, K. E. and Chi, Y.-Y. (2007). High Dimension, Low Sample Size Geometric Representation Holds Under Mild Conditions. *Biometrika*, **94**, 760-766.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Amer. Sci.* **96**, 6745-6750.

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1999). When is "Nearest Neighbor" meaningful? *Proc. Int. Conf. on Database Theory*, 217-235.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. and Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Nat. Amer. Sci.* **98**, 13790-13795.

Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459-466.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20**, 3583-3593.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97**, 611-631.

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

Hinneburg, A., Aggarwal, C. C. and Keim, D. A. (2000). What is the nearest neighbor in high dimensional spaces? *Proceedings of the 26th VLDB Conference, Cairo, Egypt.*

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Stat.* **37**, 4104-4130.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization, *Sankhyā A* **32**, 419-438.

Liu, L., Hawkins, M., Ghosh, S. and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proc. Nat. Amer. Sci.* **100**, 13167–13172.

Liu, Y., Hayes, D. N., Nobel, A. and Marron, J. S. (2008). Statistical significance of clustering for high-dimensional, low-sample size data, *J. Amer. Statist. Assoc.* **58**, 236-244.

Loewenstein, Y., Portugaly, E., Fromer, M. and Linial, M. (2008). Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics* **24**, 41-49.

Marron, J. S., Todd, M. J. and Ahn, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.* **102**, 1267-1271.

Pan, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics* **22**, 795-801.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8**, 1145-1164.

Steinbach, M., Ertoz, L. and Kumar, V. (2003). Challenges of clustering high dimensional data. In *New Vistas in Statistical Physics. Applications in Econophysics, Bioinformatics, and Pattern Recognition*, (Edited by L. T. Wille), Springer-Verlag.

Wall, M. E., Dyck P. A. and Brettin, T. S. (2001). SVDMAN-singular value decomposition analysis of microarray data. *Bioinformatics* **17**, 566-568.

Wall, M. E., Rechtsteiner, A. and Rocha, L. M. (2003). Singular value decomposition and principal component analysis, In *A Practical Approach to Microarray Data Analysis*, (Edited by D. P. Berrar, W. Dubitzky and M. Granzow), 91-109. Kluwer: Norwell, MA.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64**, 440-448.

Ward, J. H. (1963). Hierarchical Grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58**, 236-244.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R. and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Nat. Amer. Sci.* **98**, 11462-11467.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105**, 713-726.

Zha, H., He, X., Ding, G., Simon, H. and Gu, M. (2001). Spectral relaxation for k-means clustering. *Neu. Info. Proc. Sys.* **14**, 1057-1064.

Department of Statistics, University of Georgia, Athens, GA 30602-1952, USA.

E-mail: jyahn@stat.uga.edu

Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA.

E-mail: mhlee@stat.colostate.edu

Department of Applied Statistics, Konkuk University, Seoul, South Korea 143-701.

E-mail: youngjoo@konkuk.ac.kr

(Received July 2010; accepted April 2011)