Journal of Classification 22:151-183 (2005)
DOI: 10.1007/s00357-005-0012-9

Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method

Gábor J. Székely

Bowling Green State University, Bowling Green, OH

Maria L. Rizzo

Ohio University, Athens, OH

Abstract: We propose a hierarchical clustering method that minimizes a joint between-within measure of distance between clusters. This method extends Ward's minimum variance method, by defining a cluster distance and objective function in terms of Euclidean distance, or any power of Euclidean distance in the interval (0,2]. Ward's method is obtained as the special case when the power is 2. The ability of the proposed extension to identify clusters with nearly equal centers is an important advantage over geometric or cluster center methods. The between-within distance statistic determines a clustering method that is ultrametric and space-dilating; and for powers strictly less than 2, determines a consistent test of homogeneity and a consistent clustering procedure. The clustering procedure is applied to three problems: classification of tumors by microarray gene expression data, classification of dermatology diseases by clinical and histopathological attributes, and classification of simulated multivariate normal data.

Keywords: Cluster analysis; Hierarchical classification; Ward's minimum variance method.

The authors gratefully acknowledge the constructive suggestions of the editor and four anonymous referees.

This research was partially supported by NSA Grant MDA904-02-1-0091.

Authors' Addresses: Gábor J. Székely, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, email: gabors@bgnet.bgsu.edu; Maria L. Rizzo, Department of Mathematics, Ohio University, Athens, OH 45701, email: rizzo@math.ohiou.edu.

1. Introduction

In this paper we consider the problem of describing the hierarchical structure of multivariate data, where the number of underlying populations is unknown. The objective is to classify the observations (objects) into two or more disjoint, exhaustive clusters corresponding to the distinct populations sampled, and produce a hierarchical structure that lends some insight into a possible nested structure in the data.

We propose a hierarchical clustering procedure based on *joint between-within* cluster distances. As DuBien and Warde (1979) wrote, "Intuitively, the concept of cluster encompasses the duality of homogeneity within clusters and heterogeneity between clusters." However, many standard clustering procedures seek only to minimize within cluster distances, or to maximize between cluster distances. For example, in the single linkage (or nearest neighbor) method, the distance between two groups is equal to the minimum of the distances between all pairs of objects consisting of one object from each group; distances between objects within groups are not considered. Group average method joins clusters with minimum average distance between pairs of objects in different clusters. We introduce a joint between-within e-distance between clusters, and propose an agglomerative hierarchical clustering algorithm that joins clusters at minimum e-distance. The e-distance measures both the heterogeneity between groups and homogeneity within groups. The advantages of this type of criteria are noted by Murtagh (1985, p. 62) who states in reference to Ward's method (Ward 1963), "the two properties of cluster homogeneity and cluster separability are incorporated in the cluster criterion. For summarizing data, it is unlikely that more suitable criteria could be devised."

Our proposed method extends Ward's minimum variance method. Ward's method minimizes the increase in total within-cluster sum of squared error. This increase is proportional to the squared Euclidean distance between cluster centers. In contrast to Ward's method, our cluster distance is based on Euclidean distance, rather than squared Euclidean distance. More generally, we define in the following section an objective function and cluster distance in terms of any power α of Euclidean distance in the interval (0,2] (exponents outside this interval are not applied, for reasons explained in Remark 2 below). Ward's minimum variance method is obtained as the special case when $\alpha = 2$.

Ward's method is one of the geometric or cluster center methods. Methods based on distance between cluster centers will not be effective for clusters with equal means. However, if we choose exponent $\alpha < 2$, the resulting method is not based on distance between cluster centers. Our proposed cluster distance is an empirical distance between the distributions of the sampled populations, and for $0 < \alpha < 2$ cluster distances are always positive for distinct nonempty clusters.

Thus, the special case $\alpha = 2$ differs greatly from every other choice of α in $(0, 2)$. If clusters are characterized by their means, then $\alpha = 2$ is a good choice. If clusters are characterized by their distributions, then $0 < \alpha < 2$ may be a better choice.

Before proceeding to the general case $0 < \alpha \leq 2$, we focus on the simplest choice within this interval, $\alpha = 1$. The clustering method based on e-distance ($\alpha = 1$) presented in Section 2 has several desirable properties:

- statistical consistency (2.2),
- Lance–Williams form (2.3),
- ultrametricity and reducibility (2.4),
- space-dilation (2.4),
- computational tractability (2.4).

In Section 2.5 our cluster distance function is generalized, providing an extension of the Ward minimum variance method. In Section 3, the e-clustering method, denoted $\mathcal{E}$, is applied to three problems: classification of erythemato-squamous diseases in dermatology (3.1), classification of cancer tumors by gene expression data (3.2), and classification of simulated multivariate normal data (3.3). The solutions of the dermatology and cancer examples are compared with the group average method and Ward's method, and solutions of the simulated normal problems are compared with six standard methods.

The following notation and definitions are used. The objects to be clustered are realizations of $\mathbb{R}^d$ valued random variables, represented by an $n \times d$ matrix $\mathbf{X} = (x_{ik})$, where x_{ik} denotes the k^{th} variable or feature of the i^{th} object. The i^{th} object is the d-dimensional vector $\mathbf{x}_i$. The dissimilarity between objects i and j is $\|\mathbf{x}_i - \mathbf{x}_j\|^\alpha$, where $\|\cdot\|$ is the Euclidean norm, and $\alpha = 1$ for our proposed method. The notation $\mathbf{X} \stackrel{\mathcal{D}}{=} \mathbf{Y}$ indicates that the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are identically distributed. The expected value of $\mathbf{X}$ is denoted $E[\mathbf{X}]$.

A clustering is defined to be a partition $\{C_1, \ldots, C_g\}$ of the set of objects into g disjoint non-empty classes or clusters. The number of classes g is not known *a priori*. The centroid of a cluster C is denoted $\overline{C}$, and the sample mean vector of a random sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is denoted $\overline{\mathbf{X}}$.

The following notation is introduced and defined in Section 2:

- $e(C_i, C_j)$ is the e-distance between clusters C_i and C_j based on Euclidean distances between objects, defined in (1),
- $e^{(\alpha)}(C_i, C_j)$ is the $e^{(\alpha)}$-distance between clusters C_i and C_j based on the α power of Euclidean distances between objects, defined in (11),
- $\mathcal{E}^{(\alpha)}(\mathbf{X}, \mathbf{Y})$ is a distance function between random vectors $\mathbf{X}$ and $\mathbf{Y}$, defined in (4),
- $\mathcal{E}(\mathbf{X}, \mathbf{Y}) = \mathcal{E}^{(1)}(\mathbf{X}, \mathbf{Y})$.

2. Hierarchical e-clustering Algorithm

2.1 Cluster e-distance and Objective Function

Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_{n_1}\}$ and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{n_2}\}$ be nonempty subsets of $\mathbb{R}^d$. Define the between-within, or e-distance $e(A, B)$, between A and B as

$$e(A,B) = \frac{n_1 n_2}{n_1 + n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|\mathbf{a}_i - \mathbf{b}_j\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|\mathbf{a}_i - \mathbf{a}_j\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|\mathbf{b}_i - \mathbf{b}_j\|\right). \quad (1)$$

Remark 1. Many standard clustering methods base cluster distance on a type of distance between two summary statistics, such as cluster centers. The e-distance above cannot be expressed in this way. When (1) is applied to singletons $\{\mathbf{a}\}, \{\mathbf{b}\}$, e-distance is proportional to $\|\mathbf{a} - \mathbf{b}\|$. However, whenever the cluster size is greater than 1, $e(A, B)$ cannot be expressed as Euclidean distance, or any power of Euclidean distance between one pair of summary statistics.

Clearly $e(A, B)$ measures jointly the between and within distances between A and B, and small values of e-distance between clusters correspond to homogeneous clusters. Although it is perhaps not immediately obvious, e is non-negative. The non-negativity of cluster distance $e(A, B)$ is a special case of the following Theorem:

Theorem 1. *Suppose $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{R}^d$ are independent and identically distributed (iid) with distribution F, $\boldsymbol{Y}, \boldsymbol{Y}' \in \mathbb{R}^d$ are iid with distribution G, $E\|\boldsymbol{X}\| < \infty$ and $E\|\boldsymbol{Y}\| < \infty$. Then*

$$2E\|\boldsymbol{X} - \boldsymbol{Y}\| - E\|\boldsymbol{X} - \boldsymbol{X}'\| - E\|\boldsymbol{Y} - \boldsymbol{Y}'\| \geq 0, \quad (2)$$

and equality holds if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

Below we present an elementary proof for the special case $d = 1$. This simple proof does not work for $d > 1$; the general proof for $d > 1$ is more technical.

Proof. If $d = 1$ we have

$$2\int_{\mathbb{R}}[F(t) - G(t)]^2 dt$$

$$
\begin{aligned}
&= 2\int_{\mathbb{R}} [F(t)(1-G(t)) + (1-F(t))G(t) - F(t)(1-F(t)) \\
&\qquad - G(t)(1-G(t))]dt \\
&= 2\int_{\mathbb{R}} [P(X \le t < Y) + P(Y \le t < X) \\
&\qquad - P(X \le t < X') - P(Y \le t < Y')]dt \\
&= 2E|X-Y| - E|X-X'| - E|Y-Y'|.
\end{aligned}
$$

To complete the proof for $d > 1$, apply Theorem 2 below, with $\alpha = 1$. □

We develop a hierarchical clustering algorithm that merges the pair of clusters with minimum e-distance at each level. The motivation for this choice of cluster distance is that it is an empirical measure of distance between the sampled distributions. If we apply Theorem 1 to finite populations, the motivation for the e-distance criterion is obvious from the relation (3) below.

Corollary 1. *For all finite nonempty sets $A, B \subset \mathbb{R}^d$, $e(A,B) \ge 0$ and equality holds if and only if $A = B$.*

Proof. Suppose that $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_{n_1}\}$ and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{n_2}\}$ are finite nonempty subsets of $\mathbb{R}^d$. Let $\mathbf{X}, \mathbf{X}'$ be independent and uniformly distributed on the set A, and let $\mathbf{Y}, \mathbf{Y}'$ be independent and uniformly distributed on the set B. Then $E\|\mathbf{X}-\mathbf{Y}\| = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|\mathbf{a}_i - \mathbf{b}_j\|/(n_1 n_2)$, $E\|\mathbf{X}-\mathbf{X}'\| = \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|\mathbf{a}_i - \mathbf{a}_j\|/n_1^2$, and $E\|\mathbf{Y}-\mathbf{Y}'\| = \sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|\mathbf{b}_i - \mathbf{b}_j\|/n_2^2$. Hence

$$
\frac{n_1 n_2}{n_1 + n_2}\,[\,2E\|\mathbf{X}-\mathbf{Y}\| - E\|\mathbf{X}-\mathbf{X}'\| - E\|\mathbf{Y}-\mathbf{Y}'\|\,] = e(A,B). \qquad (3)
$$

By Theorem 1, $e(A,B) \ge 0$. If $e(A,B) = 0$, then by Theorem 1 $\mathbf{X}$ and $\mathbf{Y}$ are identically distributed, which implies that $A = B$. □

The hierarchical e-clustering algorithm applied to n objects has n singleton clusters initially. Since small values of e-distance correspond to homogeneous groups, at each step we compute the e-distance between all pairs of clusters and select the pair with minimum e-distance as the optimal pair to merge. After merging the optimal pair of clusters, the e-distances between clusters are updated. The height $h(k)$ of the corresponding node in the dendrogram is the e-distance between the two clusters merged at step k. The implementation essentially follows the general hierarchical algorithm outlined by Anderberg (1973, pp. 132–136) and Hartigan (1975, pp. 217–218).

For random vectors $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}^d$, and constant α such that $E\|\mathbf{X}\|^\alpha < \infty$ and $E\|\mathbf{Y}\|^\alpha < \infty$, define the real valued function

$$\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 2E\|\mathbf{X}-\mathbf{Y}\|^{\alpha} - E\|\mathbf{X}-\mathbf{X}'\|^{\alpha} - E\|\mathbf{Y}-\mathbf{Y}'\|^{\alpha}, \quad (4)$$

where $\mathbf{X}, \mathbf{X}'$ are iid and $\mathbf{Y}, \mathbf{Y}'$ are iid. Whenever $\mathcal{E}$ appears without the superscript (α), the exponent is $\alpha = 1$. Thus, Theorem 1 states that $\mathcal{E}(\mathbf{X},\mathbf{Y})$ is always nonnegative, and $\mathcal{E}(\mathbf{X},\mathbf{Y}) = 0$ if and only if $\mathbf{X} \overset{\mathcal{D}}{=} \mathbf{Y}$.

Theorem 2. *Let $\mathbf{X}$, $\mathbf{X}'$ be iid random vectors in $\mathbb{R}^d$ and let $\mathbf{Y}$, $\mathbf{Y}'$ be iid random vectors in $\mathbb{R}^d$ independent of $\mathbf{X}$. If α is a constant such that $E\|\mathbf{X}\|^{\alpha} < \infty$ and $E\|\mathbf{Y}\|^{\alpha} < \infty$, then the following statements hold.*

(i) If $0 < \alpha \leq 2$, then $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) \geq 0$.

(ii) If $0 < \alpha < 2$, then $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$ if and only if $\mathbf{X} \overset{\mathcal{D}}{=} \mathbf{Y}$.

(iii) If $\alpha = 2$, then $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$ if and only if $E[\mathbf{X}] = E[\mathbf{Y}]$.

The proof of Theorem 2 is given in the Appendix.

Remark 2. For $\alpha > 2$ the inequality $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) \geq 0$ cannot always hold. For example, let $X = 0$ and $Y = \pm 1$ with probability 1/2. Then $\mathcal{E}^{(\alpha)}(X,Y) = 2 - 2^{\alpha}/2$, which is negative if $\alpha > 2$. For $\alpha < 0$ if $\mathbf{X}$ takes any given value with positive probability, then $E\|\mathbf{X}-\mathbf{X}'\|^{\alpha} = \infty$. Thus it makes sense to restrict our investigation to $0 < \alpha < 2$. In this interval $\alpha = 1$ is the only integer; this is the simplest case, which we have applied for clustering, and it corresponds to our Theorem 1.

2.2 Statistical Consistency

Let A=$\{\mathbf{X}_1,\ldots,\mathbf{X}_{n_1}\}$ and B=$\{\mathbf{Y}_1,\ldots,\mathbf{Y}_{n_2}\}$ be independent random samples of $\mathbb{R}^d$ valued random vectors $\mathbf{X}$ and $\mathbf{Y}$ respectively. Suppose $E\|\mathbf{X}\| < \infty$, $E\|\mathbf{Y}\| < \infty$. Define distances $\mu_{AB} = E\|\mathbf{X}-\mathbf{Y}\|$, $\mu_A = E\|\mathbf{X}_1-\mathbf{X}_2\|$, $\mu_B = E\|\mathbf{Y}_1-\mathbf{Y}_2\|$; and define $e(A,B)$ by (1). Now $e(A,B)$ is a random variable with expected value

$$\begin{aligned} E[e(A,B)] =& \frac{n_1 n_2}{n_1+n_2}\left(2\mu_{AB} - \frac{n_1-1}{n_1}\,\mu_A - \frac{n_2-1}{n_2}\,\mu_B\right) \\ =& \frac{n_1 n_2}{n_1+n_2}\,(2\mu_{AB}-\mu_A-\mu_B) + \frac{n_2\mu_A}{n_1+n_2} + \frac{n_1\mu_B}{n_1+n_2}. \end{aligned} \quad (5)$$

If $\mathbf{X}$ and $\mathbf{Y}$ are identically distributed, $\mu_{AB} = \mu_A = \mu_B$, implying $2\mu_{AB} - \mu_A - \mu_B = 0$ and

$$E[e(A,B)] = \frac{n_2\mu_A + n_1\mu_B}{n_1+n_2} = \mu_{AB} = E\|\mathbf{X}-\mathbf{Y}\|,$$

for all positive integers n_1 and n_2.

If $\mathbf{X}$ and $\mathbf{Y}$ are not identically distributed, $2\mu_{AB} - \mu_A - \mu_B$ equals a positive constant by Theorem 1. Hence if $\mathbf{X} \overset{\mathcal{D}}{\neq} \mathbf{Y}$, and $n = n_1 + n_2$, the expected distance $E[e(A, B)]$ is asymptotically a positive constant times n. As the number of objects to be clustered tends to infinity, if $\mathbf{X} \overset{\mathcal{D}}{=} \mathbf{Y}$ (A and B are homogeneous), the expected distance between A and B tends to a positive constant, and otherwise $E[e(A, B)]$ tends to infinity.

Kaufman and Rousseeuw (1990, p. 241) define dissimilarities between clusters to be *statistically consistent* if dissimilarities tend to a 'meaningful limit' as the sample size tends to infinity. In that sense, the unweighted distance $e(A, B)^* := e(A, B)(n_1 + n_2)/(n_1 n_2)$ is consistent with limiting expected distance $2\mu_{AB} - \mu_A - \mu_B \geq 0$, which is zero if and only if A and B are homogeneous. This asymptotic distance $2\mu_{AB} - \mu_A - \mu_B$ in terms of mean between-cluster and mean within-cluster distances is indeed a meaningful limit with a nice interpretation.

2.3 Lance–Williams Algorithms

An infinite family of agglomerative hierarchical clustering algorithms is represented by the following recursive formula, given by Lance and Williams (1967). Suppose that d_{ij}, d_{ik}, and d_{jk} are the pairwise distances between clusters C_i, C_j, and C_k, and let $d_{(ij)k}$ denote the distance between the new cluster $C_i \cup C_j$ and C_k. An algorithm belongs to this family if $d_{(ij)k}$ can be computed recursively by

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|, \tag{6}$$

where $\alpha_i, \alpha_j, \beta$, and γ are parameters, which may depend on cluster sizes, that together with the cluster distance function d_{ij} determine the clustering algorithm. Several standard clustering algorithms satisfy (6). A table of parameters for standard methods is given by several authors. See e.g. Cormack (1971), Gordon (1999, p. 79), or Milligan (1979).

The e-clustering algorithm can be represented by the Lance–Williams formula, as well. The following recursive formula holds for disjoint clusters C_i, C_j, and C_k size n_1, n_2, and n_3 respectively:

$$\begin{aligned} e(C_i \cup C_j, C_k) = {} & \frac{n_1 + n_3}{n_1 + n_2 + n_3}\, e(C_i, C_k) \\ & + \frac{n_2 + n_3}{n_1 + n_2 + n_3}\, e(C_j, C_k) - \frac{n_3}{n_1 + n_2 + n_3}\, e(C_i, C_j). \end{aligned} \tag{7}$$

Formula (7) is derived in the Appendix. From formula (7) above, if $e_{ij} :=$

$e(C_i, C_j)$ is given by (1), e-distance can be computed recursively by

$$\begin{aligned} e_{(ij)k} &:= e(C_i \cup C_j, C_k) \\ &= \alpha_i\, e(C_i, C_k) + \alpha_j\, e(C_j, C_k) + \beta\, e(C_i, C_j) \\ &= \alpha_i e_{ik} + \alpha_j e_{jk} + \beta e_{ij} + \gamma |e_{ik} - e_{jk}|, \end{aligned} \tag{8}$$

where

$$\begin{aligned} \alpha_i &= \frac{n_i + n_3}{n_1 + n_2 + n_3}, \\ \beta &= \frac{-n_3}{n_1 + n_2 + n_3}, \\ \gamma &= 0. \end{aligned} \tag{9}$$

Thus, equation (8) is a Lance-Williams recursive formula for e-distance, with parameters (9).

It is interesting to note that the parameters of the Lance-Williams recurrence formula for the $\mathcal{E}$ method are identical to the parameters of Ward's minimum variance method. In Ward's minimum variance method, (6) holds with parameters (9) and cluster distance function

$$d_{ij} = d(C_i, C_j) = \frac{n_1 n_2}{n_1 + n_2} \|\overline{C_i} - \overline{C_j}\|^2, \tag{10}$$

where n_1, n_2 are the cluster sizes of C_i and C_j respectively. A derivation of Ward's parameters is presented in Kaufman and Rousseeuw (1990, pp. 232–233) and Späth (1980, p. 179). In Section 2.5 we show that $\mathcal{E}$ and Ward's method are different special cases of a family of methods that minimize $e^{(\alpha)}$-distance functions (11), $0 < \alpha \leq 2$.

2.4 Other Properties of $\mathcal{E}$

The e-clustering algorithm has several other good properties, including ultrametricity and reducibility, space dilation, and computational tractability.

2.4.1 Ultrametricity and Reducibility

The e-clustering algorithm described above has the property that e-distances between merging clusters are monotone non-decreasing (actually strictly increasing). The monotone non-decreasing property is necessary for the hierarchical tree or dendrogram to have no reversals or crossovers. See Morgan and Ray (1995) for examples of crossovers in dendrograms. The importance of monotone hierarchy or *reducibility* property for fast clustering algorithms is discussed by Murtagh (1985, Sec. 3.5). A monotone hierarchy determines a

distance between objects that satisfies the ultrametric property introduced by Hartigan (1967) and Johnson (1967). A distance is *ultrametric* if distances satisfy

$$d_{ij} \leq \max(d_{ik}, d_{jk})$$

for all i, j, k. The ultrametric distance $d(x_i, x_j)$ is the height at which the pair (x_i, x_j) first belong to the same cluster. Milligan (1979) proved that the constraints

$$\begin{aligned} (i) \quad & \alpha_i + \alpha_j + \beta \geq 1, \\ (ii) \quad & \min(\alpha_i, \alpha_j) \geq 0, \\ (iii) \quad & \gamma \geq 0 \end{aligned}$$

are sufficient conditions that distances between merging clusters $d_{(ij)k}$ increase monotonically. Clearly the parameters (9) of the recursive formula (8) for e-distance satisfy these constraints, with $\alpha_i + \alpha_j + \beta = 1$. Therefore e-clustering determines an ultrametric distance between objects. From the respective parameters of the Lance–Williams formula, single linkage, complete linkage, average linkage, and Ward's method have the monotonic/ultrametric property, but the median and centroid methods do not.

2.4.2 Space-dilating Property

A clustering algorithm may tend to distort space. A *space-conserving* algorithm preserves spatial relationships of the original distances, in the following sense:

$$\min(d_{ik}, d_{jk}) \leq d_{(ij)k} \leq \max(d_{ik}, d_{jk}),$$

for merging clusters i, j. This property is discussed by several authors including Chen and Van Ness (1996), DuBien and Warde (1979), Everitt, Landau, and Leese (2001, p.74), and Lance and Williams (1967). Space-conserving algorithms include centroid and group average methods. A space-distorting algorithm can be *space-dilating* or *space-contracting*. Everitt, et al. (2001, p. 74) explain "Space-conserving methods can be thought of as 'averaging' the distances to clusters merged, while space-dilating (contracting) methods move the merged clusters further from (closer to) each other." Mathematical definitions are given by DuBien and Warde (1979) and Chen and Van Ness (1996). Intuitively, all other things being equal, we should prefer space-conserving or space-dilating methods in most applications. DuBien and Warde (1979) studied the two parameter (β, γ) family of the Lance–Williams model, called the flexible strategy, and categorized algorithms into five types in the (β, γ) plane. The single linkage algorithm is space-contracting, evident in its well known propensity for chaining, which produces elongated clusters. Complete linkage

is an example of a space-dilating algorithm. DuBien and Warde (1979) recommend space-conserving and space-dilating algorithms in this family, noting that "space-dilating (β, γ) algorithms should assist in picking up small distances between clusters of data points."

The e-clustering algorithm is space-dilating. The space-dilating property of $\mathcal{E}$ follows directly from Theorem 3.3 of Chen and Van Ness (1996), who give necessary and sufficient conditions for a space-dilating algorithm in terms of the Lance–Williams parameters.

2.4.3 Computational Complexity

The hierarchical e-clustering algorithm requires as input either the raw data or the pairwise dissimilarities. If the dissimilarities (Euclidean distances) are given, the raw data is not required. An efficient method of implementing e-clustering is to store the matrix of e-distances and update e-distances using the Lance-Williams recursive formula defined by (8) and (9). This method is outlined in Anderberg (1973, Section 6.2). Clustering n objects by this method has $O(n^2)$ storage and $O(n^2)$ computational complexity.

2.5 Extension of Ward's Minimum Variance Method

Ward (1963) suggested a general hierarchical clustering procedure where the criterion for selecting the optimal pair of clusters to merge at each step is based on the optimal value of an objective function. The objective function could be any function that reflects the investigator's purpose. Many clustering procedures are encompassed by this very general class. The objective function that Ward used to illustrate the procedure was error sum of squares, and this example is known as "Ward's method". In this section, Ward's method is extended by defining a class of objective functions based on powers $0 < \alpha \leq 2$ of Euclidean distance that contains Ward's objective function as a special case $\alpha = 2$.

When the minimum variance criterion is applied using the Lance-Williams formula, the objective is to minimize the increase in total within cluster variance after merging. This increase (10) is a weighted squared distance between cluster centers. In order that cluster distances can be computed recursively, the initial cluster distance between objects is determined by applying the objective function to singletons. The only choice for initial distance between individual objects that is consistent with this objective function is (proportional to) squared Euclidean distance.

In order to apply the Lance-Williams recurrence formula of Ward's method with any other initial distance in any meaningful way, clearly the definition of "cluster distance" (the objective function) must change. If we lose

the square on the distance, then we lose the "sum of squares" to optimize. We have seen in Section 2.2 that e-distance defined in (1) is a definition of cluster distance that is different from the minimum variance distance, but corresponds to the same Lance-Williams form as Ward's method.

Consider the class of clustering methods with objective functions based on Euclidean distance, and the Lance-Williams form given by (6) and (9). Within this class, consider those methods where the objective function can be expressed as a linear function of α powers of Euclidean distance between objects. Notice that e-distance (1) is one such objective function, with $\alpha = 1$. Ward's error sum of squares is also such a function, with $\alpha = 2$. Hence this class contains at least two different methods (with different properties), $\mathcal{E}$ and Ward's minimum variance method.

In this section we define an objective function and cluster distance formula for each $\alpha \in (0, 2]$ that has the properties:

- The initial distance between singletons $\{\mathbf{a}\}$ and $\{\mathbf{b}\}$ is proportional to $\|\mathbf{a} - \mathbf{b}\|^\alpha$.
- Cluster distances can be updated by a Lance-Williams formula with the same parameters as Ward's minimum variance method.
- If $\alpha = 2$, cluster distances are a weighted squared distance between cluster means.
- If $0 < \alpha < 2$, cluster distances are not squared Euclidean distance and the objective function does not measure variance. Thus the method is not minimum variance. If $\alpha = 1$, the objective function to be optimized is given in terms of cluster distances $e(A, B)$ in equation (1). For all $0 < \alpha < 2$, the distance between distinct clusters with equal means is positive.
- For all $0 < \alpha < 2$ we have statistical consistency, which does not hold for $\alpha = 2$.

We have defined the e-distance in equation (1) in terms of Euclidean distance. More generally, for each α in the interval $(0, 2]$ define the distance $e^{(\alpha)}(A, B)$ between $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_{n_1}\}$ and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{n_2}\}$ as

$$e^{(\alpha)}(A, B) = \frac{n_1 n_2}{n_1 + n_2} \left(\frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|\mathbf{a}_i - \mathbf{b}_j\|^\alpha - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|\mathbf{a}_i - \mathbf{a}_j\|^\alpha - \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \|\mathbf{b}_i - \mathbf{b}_j\|^\alpha \right). \quad (11)$$

Notice that if (11) is applied to a pair of singletons $\{\mathbf{a}\}, \{\mathbf{b}\}$, $e^{(\alpha)}(\{\mathbf{a}\}, \{\mathbf{b}\})$ is proportional to $\|\mathbf{a} - \mathbf{b}\|^\alpha$. When cluster size is greater than 1 and $0 < \alpha < 2$, the

distance formula (11) is not proportional to Euclidean distance (or any power of Euclidean distance) between cluster centers. When the exponent is $\alpha = 2$, however, we will show that $e^{(2)}(A, B)$ is proportional to $\|\bar{\mathbf{a}} - \bar{\mathbf{b}}\|^2$.

For each $\alpha \in (0, 2]$, a hierarchical clustering method based on the cluster distance (11) can be defined analogous to e-clustering. Define the $\mathcal{E}^{(\alpha)}$ clustering method by the cluster distance and objective function (11), where at each step the pair of clusters with minimum $e^{(\alpha)}$-distance are merged. One can derive the recursive formula for updating cluster distances

$$e^{(\alpha)}(C_i \cup C_j, C_k) = \frac{n_1 + n_3}{n_1 + n_2 + n_3} e^{(\alpha)}(C_i, C_k) + \frac{n_2 + n_3}{n_1 + n_2 + n_3} e^{(\alpha)}(C_j, C_k) - \frac{n_3}{n_1 + n_2 + n_3} e^{(\alpha)}(C_i, C_j) \tag{12}$$

analogous to equation (7) by replacing the exponent 1 of Euclidean distance with exponent α, throughout the proof of Proposition 1.

Thus all $\mathcal{E}^{(\alpha)}$ methods, $0 < \alpha \leq 2$, have the same Lance-Williams parameters as Ward's minimum variance method, but with different $e^{(\alpha)}$-distance objective functions. In the special case $\alpha = 2$, initial $e^{(2)}$-distances are proportional to squared Euclidean distances, hence we obtain Ward's minimum variance method. Thus, we should expect that $e^{(2)}(A, B)$ is a weighted squared distance between the centroids of A and B. To see this, notice that if $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_{n_1}\}$ and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{n_2}\}$, then

$$\sum_{i=1}^{n_1} \|\mathbf{a}_i - \mathbf{b}_j\|^2 = \sum_{i=1}^{n_1} \|\mathbf{a}_i - \bar{\mathbf{a}} + \bar{\mathbf{a}} - \mathbf{b}_j\|^2 = n_1 S_1 + n_1 \|\bar{\mathbf{a}} - \mathbf{b}_j\|^2,$$

where $S_1 = (1/n_1) \sum_{i=1}^{n_1} \|\mathbf{a}_i - \bar{\mathbf{a}}\|^2$. Hence

$$\begin{aligned} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \|\mathbf{a}_i - \mathbf{b}_j\|^2 &= n_1 n_2 S_1 + n_1 \sum_{j=1}^{n_2} \|\bar{\mathbf{a}} - \mathbf{b}_j\|^2 \\ &= n_1 n_2 S_1 + n_1 \sum_{j=1}^{n_2} \|\bar{\mathbf{a}} - \bar{\mathbf{b}} + \bar{\mathbf{b}} - \mathbf{b}_j\|^2 \\ &= n_1 n_2 (S_1 + S_2 + \|\bar{\mathbf{a}} - \bar{\mathbf{b}}\|^2), \end{aligned}$$

where $S_2 = (1/n_2) \sum_{i=1}^{n_2} \|\mathbf{b}_i - \bar{\mathbf{b}}\|^2$. Similarly we have $\sum_{j=1}^{n_1} \sum_{i=1}^{n_1} \|\mathbf{a}_i -$

$\mathbf{a}_j\|^2 = 2n_1^2 S_1$, and $\sum_{j=1}^{n_2}\sum_{i=1}^{n_2}\|\mathbf{b}_i - \mathbf{b}_j\|^2 = 2n_2^2 S_2$, so that

$$e^{(2)}(A,B) = \frac{n_1 n_2}{n_1+n_2}\left[\frac{2}{n_1 n_2}(n_1 n_2(S_1 + S_2 + \|\bar{\mathbf{a}} - \bar{\mathbf{b}}\|^2) - \frac{1}{n_1^2}2n_1^2 S_1 - \frac{1}{n_2^2}2n_2^2 S_2\right] = \frac{2n_1 n_2}{n_1+n_2}\|\bar{\mathbf{a}} - \bar{\mathbf{b}}\|^2. \quad (13)$$

Thus $e^{(2)}(A,B)$ is a weighted squared Euclidean distance between cluster centers, equal to twice the cluster distance (10) that is minimized in Ward's method.

If $0 < \alpha < 2$, the simplifications above cannot be made, and the $e^{(\alpha)}$-distance is not a squared Euclidean distance; the $\mathcal{E}^{(\alpha)}$ method is not minimum variance. From Theorem 2, for all $0 < \alpha < 2$, the distance $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y})$ estimated by $e^{(\alpha)}(A,B)$ is always nonnegative, and $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$ if and only if the distributions of $\mathbf{X}$ and $\mathbf{Y}$ are identical.

The case $\alpha = 2$ is fundamentally different from all other $0 < \alpha < 2$. When $\alpha = 2$, $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$ holds under the weaker condition $E[\mathbf{X}] = E[\mathbf{Y}]$.

Remark 3. The last paragraph is related to a classical theorem of Christiaan Huygens, the famous 17th century Dutch natural scientist and mathematician. According to Benzécri (1992, pp. 34-35), if G denotes the center of gravity (mean) of the cloud of points, "Huygens' formula ... and its corollary (it is with respect to G that the inertia of the cloud is least) are true only because we have used in the definition of inertia the square of the distance and not the distance simply."

The $\mathcal{E}^{(\alpha)}$ methods are statistically consistent for all $0 < \alpha < 2$. That is, if A and B are samples from the distributions of $\mathbf{X}$ and $\mathbf{Y}$ respectively, then $\mathbf{X} \overset{\mathcal{D}}{\neq} \mathbf{Y}$ implies that as the total number of objects $n = n_1 + n_2$ to be clustered tends to infinity, $E[e^{(\alpha)}(A,B)]$ tends to infinity; if $\mathbf{X} \overset{\mathcal{D}}{=} \mathbf{Y}$, then $E[e^{(\alpha)}(A,B)]$ tends to a positive constant. For a proof, apply the steps in Section 2.2 used to obtain equation (5), replacing e-distance with $e^{(\alpha)}$-distance, and $\|\cdot\|$ with $\|\cdot\|^\alpha$. In the case of Ward's method statistical consistency does not hold, because when (5) is derived for $e^{(2)}$-distance, $E[e^{(2)}(A,B)]$ is a positive constant (does not tend to infinity as $n \to \infty$) if $E[\mathbf{X}] = E[\mathbf{Y}]$.

Thus, $\mathcal{E} = \mathcal{E}^{(1)}$ may be more effective than $\mathcal{E}^{(2)}$ (Ward) in certain clustering problems where clusters have nearly equal means. This type of problem is illustrated in Example 3, Problem 1 with simulated multivariate normal data.

A Lance-Williams method is determined by the parameters of the recursive equation (6), the cluster distance function d_{ij} in (6), and the algorithm for choosing the optimal pair of clusters to merge at each step. For example, with Lance-Williams parameters (8) and e-distance, at least two different clustering methods are possible. The method that merges clusters with minimum

e-distance at each step is our proposed method $\mathcal{E}$ implemented in this paper. Alternately, we can merge the pair of clusters that maximize the total e-distance

$$e(C_1, \ldots, C_g) := \sum_{1 \le i < j \le g} e(C_i, C_j)$$

between clusters $\{C_1, \ldots, C_g\}$ at each step of the hierarchy. We have implemented both methods, and obtain approximately equivalent results, but the latter method is not ultrametric, and is computationally more difficult.

Although it is possible to define an ad hoc clustering procedure by varying the distance in a Lance-Williams formula, or the criterion for selecting the optimal pair of clusters to merge, such a procedure may have no theoretical foundation. In contrast, both of our e-distance based methods rest upon a theoretical foundation guaranteeing desirable statistical properties of consistency and separation of clusters with equal means. We have shown that by applying e-distance ($\alpha = 1$) rather than squared Euclidean distance between cluster centers ($\alpha = 2$), we obtain a method that applies to a more general class of clustering problems.

3. Applications

Application of the e-clustering algorithm is illustrated in the following three examples. In the first application, the goal is to determine the type of erythemato-squamous disease, a group of diseases in dermatology that share clinical features of erythema and scaling. The second application is classification of human cancer tumors based on gene expression data. The third application clusters simulated multivariate normal data.

Solutions by two alternate clustering procedures are presented for comparison with each application: Ward's method, and the group average linkage method (unweighted pair group method). These methods were selected for comparison because both the group average and Ward's method are considered to perform well in practice. Everitt (1979) wrote "(1) no single method is best in every situation (2) the mathematically respectable single linkage is, in most cases, the least successful for the data used and (3) group average clustering and a method due to Ward (Ward 1963), do fairly well overall." A study of four ultrametric algorithms by Milligan and Isaac (1980) using constructed data sets found that group average method ranked first, complete linkage second, Ward's method third, and single linkage last. Hartigan (1985) found that complete linkage "is the worst of all standard methods for high density clustering," and that group average performs better for this type of problem.

Description of the data and references are given below. A measure of agreement between the known partition and the partition determined by a clustering algorithm is the proportion of pairs of objects in agreement, known as

the Rand Index (Rand 1971). Hubert and Arabie (1985) suggest a null model in which the partitions are selected at random given the fixed cluster sizes. Using the correction for chance proposed by Hubert and Arabie, the Rand Index R is adjusted so that the statistic takes the value 0 under the null model, and maximum value 1 when the partitions are identical. For details see Hubert and Arabie (1985) or Jain and Dubes (1988).

3.1 Diagnosis of Erythemato-Squamous Diseases in Dermatology

The dermatology data analyzed in this application is publicly available from the UCI Machine Learning Repository (Blake and Merz 1998) at ftp.ics.uci.edu. The data was analyzed by Güvenir, Demiröz, and Ilter (1998), and contributed by Güvenir. The erythemato-squamous diseases are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis and pityriasis rubra pilaris. According to Güvenir, et al. (1998), since this group of diseases all share clinical features of erythema and scaling, diagnosis is difficult. Not all samples show the typical histopathological features of the disease. Another difficulty is that a disease may show histopathological features of another disease initially, but have characteristic features at the following stages.

The data consists of 366 objects with 34 attributes. Of the 12 clinical attributes and 22 histopathological attributes, all except two take values in $\{0, 1, 2, 3\}$, where 0 indicates the feature was not present and 3 is the largest amount possible. The attribute family history takes values 0 or 1, and the age of the patient takes positive integer values. The clinical and histopathological attributes are summarized in Table 1. In the following analysis, the attributes have been standardized to zero mean and unit standard deviation. There are eight missing values for age. The distances between pairs of objects with missing values for age were calculated by shortening the pair of vectors to exclude age, and multiplying the Euclidean distance between the shortened pair by $\sqrt{34/33}$.

Agreement between each clustering solution and the recorded diagnosis is compared using the Rand R and adjusted Rand statistic R'. A high value of R' indicates strong agreement, with a maximum possible value of 1. For the six group classification solutions, the adjusted Rand statistics for $\mathcal{E}$, Ward's method, and group average method are $0.92, 0.74$, and 0.63 respectively. Thus the e-clustering solution has the highest agreement with the assumed classes, and is quite different from either the Ward or group average solutions. Improved agreement for Ward's ($R' = .86$) method was obtained with five clusters. The clusters of each of the three solutions are presented in Tables 2, 3 and 4. Note that $\mathcal{E}$ correctly classifies all cases of psoriasis, chronic dermatitis, and pityriasis rubra pilaris, and all but one of the lichen planus cases. All but one of the misclassifications are between seboreic dermatitis and pityriasis rosea. Ward's

Table 1. Attributes of Erythematos Squamous Disease Data

	Clinical Attributes		Histopathological Attributes
1.	erythema	12.	melanin incontinence
2.	scaling	13.	eosinophils in the infiltrate
3.	definite borders	14.	PNL infiltrate
4.	itching	15.	fibrosis of the papillary dermis
5.	koebner phenomenon	16.	exocytosis
6.	polygonal papules	17.	acanthosis
7.	follicular papules	18.	hyperkeratosis
8.	oral mucosal involvement	19.	parakeratosis
9.	knee and elbow involvement	20.	clubbing of the rete ridges
10.	scalp involvement	21.	elongation of the rete ridges
11.	family history	22.	thinning of the suprapapillary epidermis
34.	age	23.	pongiform pustule
		24.	munro microabcess
		25.	focal hypergranulosis
		26.	disappearance of the granular layer
		27.	vaculolization and damage of basal layer
		28.	spongiosis
		29.	saw-tooth appearance of retes
		30.	follicular horn plug
		31.	perifollicular parakeratosis
		32.	inflammatory mononuclear infiltrate
		33.	band-like infiltrate

method is less successful at recovering the clusters, grouping all seboreic dermatitis and pityriasis rosea into a single cluster, while splitting psoriasis into two clusters. The group average method does not reliably distinguish between seboreic dermatitis, pityriasis rosea, or chronic dermatitis, grouping all three diseases into a single cluster. These results suggest that $\mathcal{E}$ provides a reliable differential diagnosis of erythemato-squamous diseases, which performs best among the three methods compared on this data.

Remark 4. The dermatology data was standardized so that the age feature (whole years) would not dominate the distance calculations. To check whether standardization of the 32 ordinal variables influenced the results, we repeated this example with and without standardization, omitting age and family history from the analysis. For all three methods ($\mathcal{E}$, Ward, group average) the clustering results for the 32 ordinal variables were identical for the raw data and the standardized data.

3.2 Classification of Human Tumors Based on Gene Expression Data

The successful treatment of cancer depends in part on accurate classification of tumors. Microarrays are the focus of much scientific research to study

Table 2. Classification of Erythemato-Squamous Diseases by $\mathcal{E}$-clustering Method.[1]

Class	1	2	3	4	5	6	Cases
1. psoriasis	112						112
2. seboreic dermatitis		46		15			61
3. lichen planus			71	1			72
4. pityriasis rosea				47			49
5. chronic dermatitis					52		52
6. pityriasis rubra pilaris						20	20
Total	112	46	72	62	52	20	366

[1] Rand .9743, C. Rand .9195

Table 3. Classification of Erythemato-Squamous Diseases by Ward's Minimum Variance Method.[1]

Class	1	2	3	4	5	6	Cases
1. psoriasis	112						112
2. seboreic dermatitis		61					61
3. lichen planus		1	71				72
4. pityriasis rosea		49					49
5. chronic dermatitis					52		52
6. pityriasis rubra pilaris						20	20
Total	112	111	71	0	52	20	366

[1] Rand .9525, C. Rand .8629

Table 4. Classification of Erythemato-Squamous Diseases by Group Average Method.[1]

Class	1	2	3	4	5	6	Cases
1. psoriasis	108	4					112
2. seboreic dermatitis		61					61
3. lichen planus			72				72
4. pityriasis rosea		49					49
5. chronic dermatitis		52					52
6. pityriasis rubra pilaris						20	20
Total	108	166	72	0	0	20	366

[1] Rand .8534, C. Rand .6329

the variation among tumors. Comparison of gene expression levels of normal and diseased tissue can be used to help identify tumors and appropriate treatment. We applied hierarchical e-clustering to the NCI60 microarray data discussed in Chapter 14 of Hastie, Tibshirani, and Friedman (2001). The data file is

available at http://www-stat-class.stanford.edu/~ tibs/ElemStatLearn/datasets. The raw data are expression levels from cDNA microarrays, in 60 cancer cell lines used in the screen for anti-cancer drugs by the National Cancer Institute. An extensive analysis of the NCI60 data appears in Ross, Sherf, Eisen, Perou, Rees, Spellman, Iyer, Jeffrey, Van De Rijn, Waltham, Pergamenschikov, Lee, Lashkari, Shalon, Myers, Weinstein, Botstein, and Brown (2000). Hierarchical cluster analysis of the NCI60 data is also presented in Dudoit, Fridlyand, and Speed (2000), and Tibshirani, Hastie, Eisen, Ross, Botstein, and Brown (1999).

The data is an array of 6830 gene expression measurements for 64 human cancer samples. The gene expression levels in the NCI60 data are relative to a fixed common reference sample. In the raw data, x_{ij} is $\log_2(Cy5/Cy3)$ (the fluorescence ratio) for gene j in sample i. The data we analyzed, from Hastie, et al. (2001), has been centered to row median zero and column median zero (see the file "nci.info" at URL above). To classify tumors, the genes are regarded as variables. The samples include nine types of cancers: breast (7), central nervous system (CNS) (5), colon (7), leukemia (6), melanoma (8), non-small-cell-lung-carcinoma (NSCLC) (9), ovarian (6), prostate (2), renal (9), and unknown cancer (1). The unknown cancer sample is omitted from this analysis. The two prostate cancer samples are also omitted from this analysis, because of the small class size. We have removed from the analysis 1884 variables with more than two missing values. The resulting data array is a 61×4946 array of real numbers representing gene expression levels for 4946 genes in 61 cancer samples. Note that two of the cell lines are replicated in the data. The samples labeled 'K562B-repro' are replicated leukemia samples, and the samples labeled 'MCF7A-repro' are replicated breast cancer samples. In this analysis, replicated samples are treated as distinct samples, so the clustering solutions should place each pair of replicates into a common class. For further details about the method of data collection and other studies on the NCI60 data, refer to Ross, et al. (2000), Dudoit, et al. (2000), or Tibshirani, et al. (1999).

Hierarchical clustering solutions for three methods, $\mathcal{E}$, Ward's method, and group average method, are presented in Tables 5, 6, and 7. The adjusted Rand statistics for the eight group solution were .4098, .2891, and .0910 for $\mathcal{E}$, Ward, and group average methods respectively. For solutions with best agreement $R' = .5039$ ($\mathcal{E}$, 7 groups), $R' = .3998$ (Ward, 6 groups), and $R' = .1247$ (group average, 3 groups).

Both $\mathcal{E}$ and Ward's method appear to be more successful at recovering the correct classes than the group average method. The corresponding dendrograms for these three solutions are given in Figures 1, 2, and 3. The hierarchical structure is similar but not identical for the $\mathcal{E}$ method and Ward's method. The replicated breast cancer and leukemia samples are each correctly grouped into a common cluster.

Our results are consistent with the analysis of Ross, et al. (2000), that

Table 5. Classification of Cancer Samples by $\mathcal{E}$-clustering Method

	8 Group Solution[1]								Best Agreement[2]								
Class[3]	B	C	K	L	M	N	O	R	B	C	K	L	M	N	O	R	Cases
B	4				2	3			4				2	3			9
C		7								7							7
K			5				3				8						8
L				6		3						6		3			9
M					7	1							7	1			8
N						5								5			5
O		1		4		1				1		4		1			6
R						2		7						2		7	9
Total	4	8	5	10	9	15	3	7	4	8	8	10	9	15	0	7	61

[1] Rand .8612, C. Rand .4098
[2] Rand .8869, C. Rand .5039
[3] B:BREAST, C:COLON, K:LEUKEMIA, L:NSCLC, M:MELANOMA, N:CNS, O:OVARIAN, R:RENAL

Table 6. Classification of Cancer Samples by Ward's Method

	8 Group Solution[1]								Best Agreement[2]								
Class[3]	B	C	K	L	M	N	O	R	B	C	K	L	M	N	O	R	Cases
B	4				2	3			4				2	3			9
C		7								7							7
K			3				3	2			8						8
L			1	7		1					1	7		1			9
M				1	7							1	7				8
N				1		4						1		4			5
O		1		5						1		5					6
R				7		2						7		2			9
Total	4	8	4	21	9	10	3	2	4	8	9	21	9	10	0	0	61

[1] Rand .7918, C. Rand .2891
[2] Rand .8398, C. Rand .3998
[3] B:BREAST, C:COLON, K:LEUKEMIA, L:NSCLC, M:MELANOMA, N:CNS, O:OVARIAN, R:RENAL

cell lines derived from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into independent terminal branches specific to their respective organ types with few exceptions. Cell lines derived from non-small-lung carcinoma and breast tumours were distributed in multiple different terminal branches suggesting that their gene expression patterns were more heterogeneous.

Table 7. Classification of Cancer Samples by Group Average Method

	8 Group Solution[1]								Best Agreement[2]								
Class[3]	B	C	K	L	M	N	O	R	B	C	K	L	M	N	O	R	Cases
B	4			3		2			6			3					9
C	1			6					1			6					7
K		1	5		2						8						8
L				7		1	1		2			7					9
M				8								8					8
N				3		2			2			3					5
O				6								6					6
R				8				1				9					9
Total	5	1	5	41	2	5	1	1	11	0	8	42	0	0	0	0	61

[1] Rand .5164, C. Rand .0910
[2] Rand .5399, C. Rand .1247
[3] B:BREAST, C:COLON, K:LEUKEMIA, L:NSCLC, M:MELANOMA, N:CNS, O:OVARIAN, R:RENAL

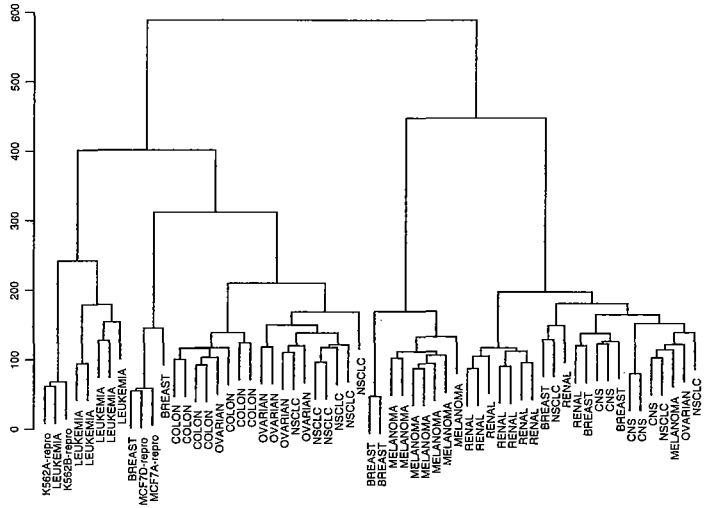

Figure 1. Cancers Clustered by *e*-distance Method

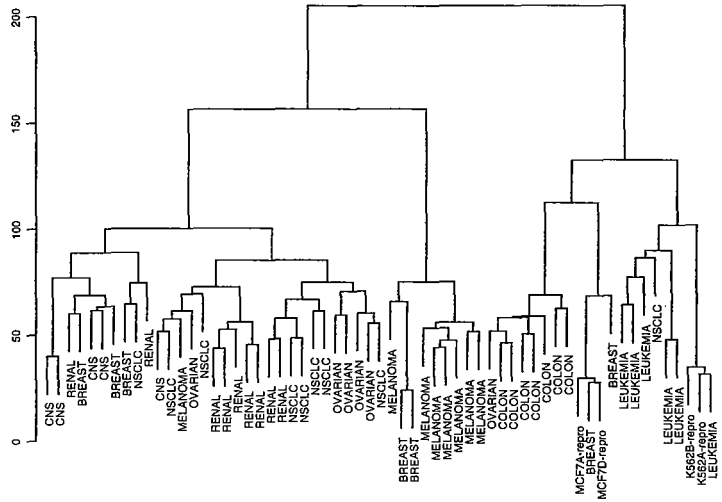

Figure 2. Cancers Clustered by Ward's Method

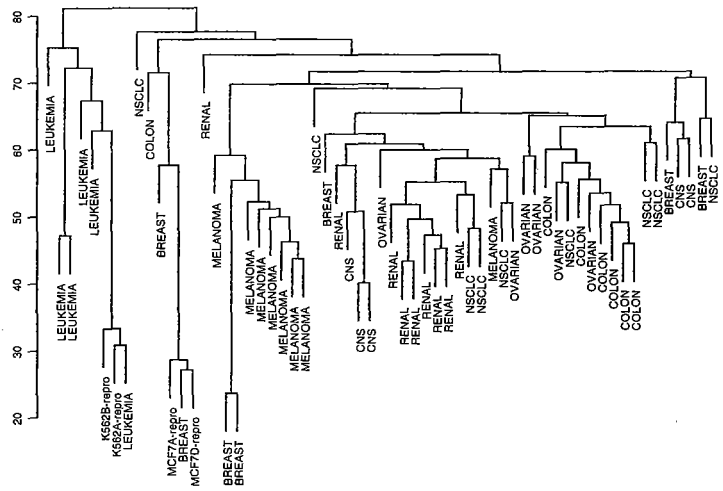

Figure 3. Cancers Clustered by Group Average Method

Their discussion also indicates that it is possible that one of the patients diagnosed with breast cancer had a co-existing melanoma. This possibility appears to be reflected in all solutions, where the melanoma cluster also contains at least one breast cancer sample.

3.3 Simulated Multivariate Normal Data

Theorems 1 and 2 show that $\mathcal{E}$ should be able to separate clusters with nearly equal means, a property not shared by the geometric or cluster center type methods such as centroid, median, or Ward's methods. However, if clusters differ only in location (they have equal distributions after translation), then there is not an obvious theoretical advantage of either $\mathcal{E}$ or Ward's method, because there are no differences to detect other than means; one might expect comparable results from both methods.

To illustrate, we present results for two problems from the Machine Learning Benchmark Problems (Leisch and Dimitriadou 2004). Both problems appeared in Breiman (1996). In each problem, there are two classes with equal probability, and 100 observations.

Problem 1: Class 1 is d-dimensional multivariate normal with zero mean and covariance $4I$, where I denotes the d-dimensional identity matrix. Class 2 is multivariate normal with mean $(a, a, \ldots, a)$ and unit covariance, where $a = 1/\sqrt{d}$.

Problem 2: Class 1 and Class 2 are d-dimensional multivariate normal with unit covariance. The mean of Class 1 is $(a, a, \ldots, a)$ and the mean of Class 2 is $(-a, -a, \ldots, -a)$, where $a = 2/\sqrt{d}$.

For each d, 2000 samples were clustered by $\mathcal{E}$, Ward, complete, average, single, centroid, and median methods. These latter methods are described e.g. in Everitt, et al. (2001) or Gordon (1999).

Figure 4 compares the average corrected Rand index R' of the $\mathcal{E}$ and Ward methods for the first problem with $d = 2 : 20, 25, 30, 35, 40, 50$. The graphs for other methods (and other measures of agreement) are essentially the same, all showing that $\mathcal{E}$ is the superior method for this problem. Table 8 compares average diagonal, Kappa, Rand and corrected Rand indices of all seven clustering methods for a subset of the d values. The results of the simulation clearly demonstrate that of the seven methods, only $\mathcal{E}$ could reliably classify data in Problem 1.

In the second problem, the clusters are spherical and distributions differ only in location. Results are shown in Table 9 for $d = 5, 10, 20, 50$. In low dimension, four methods ($\mathcal{E}$, Ward, complete linkage and group average) perform well. As dimension increases, it appears that both $\mathcal{E}$ and Ward's methods are more effective than complete linkage and group average methods. Figure 5 compares the adjusted Rand index R' of $\mathcal{E}$, Ward's, complete linkage, and group

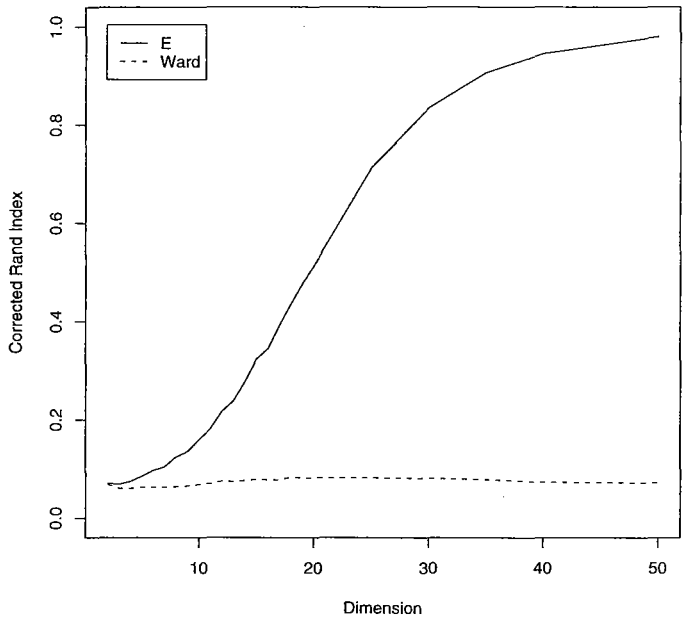

Figure 4. Example 3 (Problem 1). Comparing $\mathcal{E}$ and Ward's Method by Average Corrected Rand Index from 2000 Simulated Samples, 100 Cases with Equal Probability from Class 1: $N_d(0, 4I)$, and Class 2: $N_d((a, \ldots, a), I)$, where $a = 1/\sqrt{d}$.

average methods for $d = 2 : 20, 25, 30, 35, 40, 50$. Overall, for the second problem, we have approximately the same results for $\mathcal{E}$ and Ward's methods, both methods were superior to complete linkage; and all of these performed better than group average, single linkage, centroid and median methods.

4. Summary

The proposed e-distance method $\mathcal{E}$ and Ward's minimum variance method are two different special cases of an infinite family $\{\mathcal{E}^{(\alpha)} : 0 < \alpha \leq 2\}$ of hierarchical clustering methods with identical Lance-Williams parameters, but different objective functions. For each α, the $e^{(\alpha)}$-distance and objective function is a linear function of the α powers of Euclidean distance between all pairs of objects. The $\mathcal{E}$ method we propose is the special case $\alpha = 1$, which corresponds to Euclidean initial distances. Ward's method minimizes $e^{(2)}$-distance, which is a weighted squared Euclidean distance between cluster centers. All $\mathcal{E}^{(\alpha)}$ methods are ultrametric and space-dilating, but Ward's method differs from other $\mathcal{E}^{(\alpha)}$ methods, because $\alpha = 2$ is the only case where $e^{(\alpha)}$-distance can be expressed in terms of a power of Euclidean distance between cluster centers. For all other α in $(0, 2)$, the $e^{(\alpha)}$-distance is positive for all pairs of distinct clusters.

Table 8. Measures of Agreement for Simulated Multivariate Normal Data: Example 3, Problem 1[1]

Method	d	Diag.	Kappa	Rand	C. Rand
$\mathcal{E}$	5	0.6347	0.2693	0.5417	0.0850
Ward	5	0.6172	0.2345	0.5300	0.0625
Complete	5	0.5875	0.1750	0.5175	0.0396
Average	5	0.5168	0.0336	0.4958	0.0010
Median	5	0.5124	0.0248	0.4954	0.0004
Single	5	0.5104	0.0208	0.4952	0.0000
Centroid	5	0.5102	0.0205	0.4952	0.0000
$\mathcal{E}$	10	0.6854	0.3708	0.5784	0.1590
Ward	10	0.6246	0.2491	0.5323	0.0683
Complete	10	0.6076	0.2151	0.5280	0.0607
Average	10	0.5132	0.0264	0.4954	0.0004
Median	10	0.5104	0.0208	0.4952	0.0000
Single	10	0.5101	0.0202	0.4952	0.0000
Centroid	10	0.5100	0.0200	0.4952	0.0000
$\mathcal{E}$	20	0.8452	0.6904	0.7560	0.5128
Ward	20	0.6359	0.2717	0.5384	0.0814
Complete	20	0.6328	0.2657	0.5448	0.0945
Average	20	0.5114	0.0228	0.4952	0.0001
Median	20	0.5101	0.0202	0.4952	0.0000
Single	20	0.5100	0.0201	0.4952	0.0000
Centroid	20	0.5100	0.0200	0.4952	0.0000
$\mathcal{E}$	50	0.9949	0.9898	0.9899	0.9798
Ward	50	0.6281	0.2561	0.5331	0.0716
Complete	50	0.6629	0.3258	0.5687	0.1419
Average	50	0.5104	0.0207	0.4952	0.0000
Median	50	0.5100	0.0200	0.4952	0.0000
Single	50	0.5100	0.0200	0.4952	0.0000
Centroid	50	0.5100	0.0200	0.4952	0.0000

[1] Mean index of agreement from simulation of 2000 samples for each d, of 100 cases with equal prob. from d-variate normal:
Class 1: $N_d(0, 4I)$
Class 2: $N_d((a, a, \ldots, a), I)$, where $a = 1/\sqrt{d}$.

Table 9. Measures of Agreement for Simulated Multivariate Normal Data: Example 3, Problem 2[1]

Method	d	Diag.	Kappa	Rand	C. Rand
$\mathcal{E}$	5	0.9545	0.9091	0.9138	0.8276
Ward	5	0.9548	0.9096	0.9143	0.8287
Complete	5	0.9404	0.8808	0.8917	0.7834
Average	5	0.9223	0.8446	0.8869	0.7746
Median	5	0.5711	0.1423	0.5426	0.0932
Single	5	0.5103	0.0206	0.4952	0.0000
Centroid	5	0.5116	0.0232	0.4961	0.0019
$\mathcal{E}$	10	0.9476	0.8953	0.9014	0.8029
Ward	10	0.9474	0.8949	0.9010	0.8021
Complete	10	0.9287	0.8575	0.8732	0.7465
Average	10	0.8099	0.6198	0.7776	0.5584
Median	10	0.5139	0.0278	0.4958	0.0011
Single	10	0.5101	0.0202	0.4952	0.0000
Centroid	10	0.5101	0.0201	0.4952	0.0000
$\mathcal{E}$	20	0.9324	0.8649	0.8748	0.7496
Ward	20	0.9325	0.8651	0.8750	0.7500
Complete	20	0.8980	0.7960	0.8284	0.6571
Average	20	0.6168	0.2337	0.5926	0.1924
Median	20	0.5102	0.0205	0.4952	0.0000
Single	20	0.5101	0.0201	0.4952	0.0000
Centroid	20	0.5100	0.0200	0.4952	0.0000
$\mathcal{E}$	50	0.9012	0.8024	0.8231	0.6463
Ward	50	0.9015	0.8030	0.8235	0.6471
Complete	50	0.8082	0.6165	0.7149	0.4307
Average	50	0.5146	0.0291	0.4970	0.0035
Median	50	0.5100	0.0200	0.4952	0.0000
Single	50	0.5100	0.0200	0.4952	0.0000
Centroid	50	0.5100	0.0200	0.4952	0.0000

[1] Mean index of agreement from simulation of 2000 samples for each d, of 100 cases with equal prob. from d-variate normal:
Class 1: $N_d((a, a, \ldots, a), I)$
Class 2: $N_d((-a, -a, \ldots, -a), I)$, where $a = 2/\sqrt{d}$.

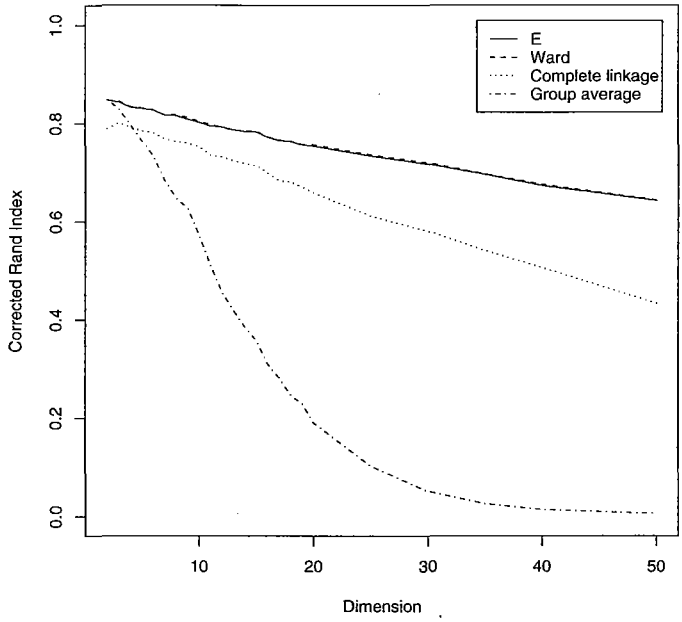

Figure 5. Example 3 (Problem 2). Comparing $\mathcal{E}$, Ward's Method, Complete Linkage, And Group Average Method, By Average Corrected Rand Index from 2000 Simulated Samples, 100 Cases with Equal Probability from Class 1: $N_d(a, \ldots, a, I)$, and Class 2: $N_d((-a, \ldots, -a), I)$, where $a = 2/\sqrt{d}$.

If clusters are characterized by their means, $\alpha = 2$ may be a good choice within this $\mathcal{E}^{(\alpha)}$ family of methods, but if clusters are characterized by their distributions, then $\mathcal{E}$ may be more effective. The $\mathcal{E}$ method, which minimizes e-distance, is statistically consistent (see Section 2.2) in the sense of Kaufman and Rousseeuw (1990, p. 241). Consistency does not hold for Ward's method ($\alpha = 2$), because cluster distance is zero when groups have equal means, while the underlying populations could have different distributions.

The ability of $\mathcal{E}$ to identify clusters with nearly equal centers is potentially an important practical advantage over geometric or cluster center methods such as centroid, median, or Ward's minimum variance methods. Our empirical results suggest that the proposed $\mathcal{E}$ method is a practical and effective approach to recover the underlying hierarchical structure in three quite different types of applications. The applications included high dimensional data, and data with attributes on different scales. In our example clustering simulated normal data with different covariance but nearly equal means, $\mathcal{E}$ clearly outperformed all six standard methods compared. Our second simulation problem involved spherical clusters of equal size. We found that empirical performance of $\mathcal{E}$ was essentially equivalent to Ward's method in this problem. Our empirical results suggest that

the theoretical properties of $\mathcal{E}$ are indeed an advantage for certain clustering problems, without sacrificing the good properties of Ward's minimum variance method for separating spherical clusters.

We have defined an objective function that extends the Ward minimum variance method, and provided the theoretical justification for considering other powers than the squared Euclidean distance. The extension of Ward's method via e-distance and the Lance-Williams recursive formula, illustrated here for $\alpha = 1$, provides a method that theoretically is applicable for a more general class of clustering problems than Ward's minimum variance method.

APPENDIX

Notation and Definitions In Lemma 1 and the proof of Theorem 2, $(\mathbf{x}, \mathbf{y})$ denotes the inner product of $\mathbf{x}$ and $\mathbf{y}$. If $f(\cdot)$ is a complex valued function, $\overline{f(\cdot)}$ denotes the complex conjugate of $f(\cdot)$. The characteristic function of a random vector $\mathbf{X} \in \mathbb{R}^d$ is

$$\hat{f}(\mathbf{t}) = E[\exp(i(\mathbf{X}, \mathbf{t}))] = E[\cos(\mathbf{X}, \mathbf{t}) + i \sin(\mathbf{X}, \mathbf{t})],$$

where i denotes the complex unit. The function $\Gamma(\cdot)$ is the complete gamma function, defined

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t}\, dt, \quad r \neq 0, -1, -2, \ldots$$

We use the following lemma in the proof of Theorem 2.

Lemma 1. *If* $0 < \alpha < 2$, *for all* $\mathbf{x} \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \frac{1 - \cos(\mathbf{t}, \mathbf{x})}{\|\mathbf{t}\|^{d+\alpha}}\, d\mathbf{t} = C(d, \alpha)\|\mathbf{x}\|^\alpha,$$

where $\mathbf{t} \in \mathbb{R}^d$, *and* $C(d, \alpha) > 0$ *is a constant depending only on* d *and* α. *(The integrals at* 0 *and* ∞ *are meant in the principal value sense:* $\lim_{\varepsilon \to 0} \int_{\mathbb{R}^d \setminus \{\varepsilon B + \varepsilon^{-1}\overline{B}\}}$, *where* B *is the unit ball (centered at 0) in* $\mathbb{R}^d$ *and* $\overline{B}$ *is the complement of* B.*)*

Proof. Introduce

$$A = \int_{\mathbb{R}^{d-1}} \frac{dz_2\, dz_3\, \ldots dz_d}{(1 + z_2^2 + z_3^2 + \cdots + z_d^2)^{\frac{d+\alpha}{2}}}.$$

Then by the formulas 3.3.2.1, p.585, 2.2.4.24 p.298 and 2.5.3.13 p.387 of Prudnikov, Brychkov and Marichev (1986), we have

$$A = \frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})} \int_0^\infty \frac{x^{d-2}dx}{(1+x^2)^{\frac{d+\alpha}{2}}} = \frac{\pi^{\frac{d-1}{2}}\,\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{d+\alpha}{2})},$$

and

$$\frac{d}{da}\left(\int_0^\infty \frac{1-\cos ax}{x^{1+\alpha}}\,dx\right) = a^{\alpha-1}\int_0^\infty \frac{\sin x}{x^\alpha}\,dx$$
$$= a^{\alpha-1}\frac{\sqrt{\pi}\,\Gamma(1-\frac{\alpha}{2})}{2^\alpha\,\Gamma(\frac{\alpha+1}{2})}.$$

Introduce the new variables $s_1 := z_1$, $s_k := s_1 z_k$ for $k \geq 2$. Then

$$C(d,\alpha) = A \times \int_{-\infty}^\infty \frac{1-\cos z_1}{|z_1|^{1+\alpha}}\,dz_1$$
$$= \frac{\pi^{\frac{d-1}{2}}\,\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{d+\alpha}{2})} \times \frac{2\sqrt{\pi}\,\Gamma(1-\frac{\alpha}{2})}{\alpha 2^\alpha\,\Gamma(\frac{\alpha+1}{2})}$$
$$= \frac{2\pi^{\frac{d}{2}}\,\Gamma(1-\frac{\alpha}{2})}{\alpha 2^\alpha \Gamma(\frac{d+\alpha}{2})} > 0,$$

and this was to be proved. □

PROOF OF THEOREM 2. Suppose $\mathbf{X} \in \mathbb{R}^d$, with characteristic function $\hat{f}$, $\mathbf{Y} \in \mathbb{R}^d$ with characteristic function $\hat{g}$, and $\mathbf{X}$, $\mathbf{Y}$ are independent. Suppose $\alpha \in \mathbb{R}$ such that $E\|\mathbf{X}\|^\alpha < \infty$ and $E\|\mathbf{Y}\|^\alpha < \infty$. Let $\mathbf{X}'$ and $\mathbf{Y}'$ be random vectors in $\mathbb{R}^d$, such that $\mathbf{X}'$ is independent of $\mathbf{X}$ with characteristic function $\hat{f}$, and $\mathbf{Y}'$ is independent of $\mathbf{Y}$ with characteristic function $\hat{g}$. (Thus $\mathbf{X}, \mathbf{X}'$ are iid and $\mathbf{Y}, \mathbf{Y}'$ are iid.) We need to prove that the function

$$\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 2E\|\mathbf{X}-\mathbf{Y}\|^\alpha - E\|\mathbf{X}-\mathbf{X}'\|^\alpha - E\|\mathbf{Y}-\mathbf{Y}'\|^\alpha$$

satisfies:

(i) If $0 < \alpha \leq 2$, then $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) \geq 0$.

(ii) If $0 < \alpha < 2$, then $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are identically distributed.

(iii) If $\alpha = 2$, then $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$ if and only if $E[\mathbf{X}] = E[\mathbf{Y}]$.

First we derive an expression for $|\hat{f}(\mathbf{t})-\hat{g}(\mathbf{t})|^2$ in terms of the differences $\mathbf{X}-\mathbf{X}'$, $\mathbf{Y}-\mathbf{Y}'$, and $\mathbf{X}-\mathbf{Y}$. By definition,

$$|\hat{f}(\mathbf{t})-\hat{g}(\mathbf{t})|^2 = |\hat{f}(\mathbf{t})-\hat{g}(\mathbf{t})| \cdot \overline{|\hat{f}(\mathbf{t})-\hat{g}(\mathbf{t})|}$$
$$= \hat{f}(\mathbf{t})\overline{\hat{f}(\mathbf{t})} + \hat{g}(\mathbf{t})\overline{\hat{g}(\mathbf{t})} - \hat{f}(\mathbf{t})\overline{\hat{g}(\mathbf{t})} - \overline{\hat{f}(\mathbf{t})}\hat{g}(\mathbf{t}). \qquad (14)$$

Independence of $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}'$ implies that

$$\begin{aligned}\hat{f}(\mathbf{t})\overline{\hat{g}(\mathbf{t})} &= E[\cos(\mathbf{X},\mathbf{t}) + i\sin(\mathbf{X},\mathbf{t})]E[\cos(\mathbf{Y},\mathbf{t}) - i\sin(\mathbf{Y},\mathbf{t})] \\ &= E[\cos(\mathbf{X},\mathbf{t})\cos(\mathbf{Y},\mathbf{t}) + \sin(\mathbf{X},\mathbf{t})\sin(\mathbf{Y},\mathbf{t})] \\ &= E[\cos(\mathbf{X}-\mathbf{Y},\mathbf{t})], \qquad (15)\end{aligned}$$

$$\overline{\hat{f}(\mathbf{t})}\hat{g}(\mathbf{t}) = E[\cos(\mathbf{Y}-\mathbf{X},\mathbf{t})], \qquad (16)$$

$$\hat{f}(\mathbf{t})\overline{\hat{f}(\mathbf{t})} = E[\cos(\mathbf{X}-\mathbf{X}',\mathbf{t})], \qquad (17)$$

$$\hat{g}(\mathbf{t})\overline{\hat{g}(\mathbf{t})} = E[\cos(\mathbf{Y}-\mathbf{Y}',\mathbf{t})]. \qquad (18)$$

Substituting (15)–(18) into (14) we have

$$\begin{aligned}|\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})|^2 &= E[\cos(\mathbf{X}-\mathbf{X}',\mathbf{t}) + \cos(\mathbf{Y}-\mathbf{Y}',\mathbf{t}) - 2\cos(\mathbf{X}-\mathbf{Y},\mathbf{t})] \\ &= E[2(1-\cos(\mathbf{X}-\mathbf{Y},\mathbf{t}) - (1-\cos(\mathbf{X}-\mathbf{X}',\mathbf{t}) \\ &\qquad - (1-\cos(\mathbf{Y}-\mathbf{Y}',\mathbf{t}))],\end{aligned}$$

thus

$$\begin{aligned}&\int_{\mathbb{R}^d} \frac{|\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})|^2}{\|\mathbf{t}\|^{d+\alpha}}\, d\mathbf{t} \\ &\quad = E\left[\int_{\mathbb{R}^d} \frac{2[1-\cos(\mathbf{t},\mathbf{X}-\mathbf{Y})] - [1-\cos(\mathbf{t},\mathbf{X}-\mathbf{X}')] - [1-\cos(\mathbf{t},\mathbf{Y}-\mathbf{Y}')]}{\|\mathbf{t}\|^{d+\alpha}}\, d\mathbf{t}\right].\end{aligned}$$

Apply Lemma 1, to get

$$\begin{aligned}&\int_{\mathbb{R}^d} \frac{|\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})|^2}{\|\mathbf{t}\|^{d+\alpha}}\, d\mathbf{t} \\ &\quad = E[\,2C(d,\alpha)\|\mathbf{X}-\mathbf{Y}\|^\alpha - C(d,\alpha)\|\mathbf{X}-\mathbf{X}'\|^\alpha - C(d,\alpha)\|\mathbf{Y}-\mathbf{Y}'\|^\alpha\,] \\ &\quad = C(d,\alpha)\,\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}). \qquad (19)\end{aligned}$$

Clearly the left hand side of (19) is non-negative, and the positive constant $C(d,\alpha)$ exists for all $0 < \alpha < 2$, so (i) is proved for $0 < \alpha < 2$. Equation (19) is zero if and only if $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$. Therefore, $\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = 0$ is a necessary and sufficient condition that the characteristic functions, and hence the distributions, of $\mathbf{X}$ and $\mathbf{Y}$ are identical, which proves (ii).

Finally, for $\alpha = 2$ we have

$$\mathcal{E}^{(\alpha)}(\mathbf{X},\mathbf{Y}) = \mathcal{E}^{(2)}(\mathbf{X},\mathbf{Y}) = 2\|E[\mathbf{X}] - E[\mathbf{Y}]\|^2,$$

which proves (iii) and the case $\alpha = 2$ in (i). □

Proposition 1 (Recursive Formula for e-distance). *Suppose A, B, and C are disjoint nonempty finite subsets of $\mathbb{R}^d$ such that*

$$e(A, B) \leq \min(e(A, C), e(B, C)).$$

Then a recursive formula for $e(A \cup B, C)$ is given by

$$e(A \cup B, C) = \frac{n_1 + n_3}{n_1 + n_2 + n_3} e(A, C) + \frac{n_2 + n_3}{n_1 + n_2 + n_3} e(B, C) - \frac{n_3}{n_1 + n_2 + n_3} e(A, B). \quad (20)$$

Proof. Suppose that $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_{n_1}\}$, $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{n_2}\}$ and $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_{n_3}\}$ are disjoint, non-empty subsets of $\mathbb{R}^d$ (distinct clusters). Define the constants δ_{11}, δ_{22}, and δ_{12} by

$$\delta_{11} = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|\mathbf{a_i} - \mathbf{a_j}\|,$$

$$\delta_{22} = \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \|\mathbf{b_i} - \mathbf{b_j}\|,$$

$$\delta_{12} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|\mathbf{a_i} - \mathbf{b_j}\|.$$

By definition,

$$\begin{aligned} e(A, B) =& \frac{n_1 n_2}{n_1 + n_2} \left(\frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|\mathbf{a_i} - \mathbf{b_j}\| \right. \\ & \left. - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|\mathbf{a_i} - \mathbf{a_j}\| - \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \|\mathbf{b_i} - \mathbf{b_j}\| \right) \\ =& \frac{n_1 n_2}{n_1 + n_2} (2\delta_{12} - \delta_{11} - \delta_{22}). \end{aligned}$$

Similarly, if

$$\delta_{33} = \frac{1}{n_3^2} \sum_{i=1}^{n_3} \sum_{j=1}^{n_3} \|\mathbf{c_i} - \mathbf{c_j}\|,$$

$$\delta_{13} = \frac{1}{n_1 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_3} \|\mathbf{a_i} - \mathbf{c_j}\|,$$

$$\delta_{23} = \frac{1}{n_2 n_3} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} \|\mathbf{b_i} - \mathbf{c_j}\|,$$

we have

$$e(A,C) = \frac{n_1 n_3}{n_1 + n_3}(2\delta_{13} - \delta_{11} - \delta_{33})$$

and

$$e(B,C) = \frac{n_2 n_3}{n_2 + n_3}(2\delta_{23} - \delta_{22} - \delta_{33}).$$

Consider the cluster $A \cup B$ formed by merging clusters A and B. Denote $A \cup B$ by subscript k, and define the corresponding constants

$$\delta_{k3} = \frac{1}{(n_1 + n_2)n_3} \sum_{j=1}^{n_3} \left(\sum_{i=1}^{n_1} \|\mathbf{a_i} - \mathbf{c_j}\| + \sum_{i=1}^{n_2} \|\mathbf{b_i} - \mathbf{c_j}\| \right),$$

$$\delta_{kk} = \frac{1}{(n_1 + n_2)^2} \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|\mathbf{a_i} - \mathbf{a_j}\| + 2 \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|\mathbf{a_i} - \mathbf{b_j}\| + \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \|\mathbf{b_i} - \mathbf{b_j}\| \right),$$

so that in terms of the original constants we have

$$\delta_{k3} = \frac{n_1 n_3 \delta_{13} + n_2 n_3 \delta_{23}}{(n_1 + n_2)n_3},$$

$$\delta_{kk} = \frac{n_1^2 \delta_{11} + n_2^2 \delta_{22} + 2 n_1 n_2 \delta_{12}}{(n_1 + n_2)^2}.$$

Therefore, the e-distance between the new cluster $A \cup B$ and disjoint cluster C is given by

$$\begin{aligned} e(A \cup B, C) &= \frac{(n_1 + n_2)n_3}{n_1 + n_2 + n_3}[2\delta_{k3} - \delta_{kk} - \delta_{33}] \\ &= \frac{(n_1 + n_2)n_3}{n_1 + n_2 + n_3} \left[\frac{2 n_1 n_3 \delta_{13} + 2 n_2 n_3 \delta_{23}}{(n_1 + n_2)n_3} \right. \\ &\qquad \left. - \frac{n_1^2 \delta_{11} + n_2^2 \delta_{22} + 2 n_1 n_2 \delta_{12}}{(n_1 + n_2)^2} - \delta_{33} \right]. \end{aligned}$$

Simplify

$$\begin{aligned} &- \frac{(n_1 + n_2)n_3}{n_1 + n_2 + n_3} \left[\frac{n_1^2 \delta_{11} + n_2^2 \delta_{22} + 2 n_1 n_2 \delta_{12}}{(n_1 + n_2)^2} \right] \\ &\qquad = \frac{1}{n_1 + n_2 + n_3} [-n_3 e(A,B) - n_1 n_3 \delta_{11} - n_2 n_3 \delta_{22}] \end{aligned}$$

so that

$$\begin{aligned}(n_1+n_2+n_3)\, e(A\cup B, C) &= 2n_1n_3\delta_{13} + 2n_2n_3\delta_{23} \\ &\quad - n_3e(A,B) - n_1n_3\delta_{11} - n_2n_3\delta_{22} - n_1n_3\delta_{33} - n_2n_3\delta_{33} \\ &= n_1n_3[2\delta_{13} - \delta_{11} - \delta_{33}] + n_2n_3[2\delta_{23} - \delta_{22} - \delta_{33}] - n_3e(A,B) \\ &= (n_1+n_3)e(A,C) + (n_2+n_3)e(B,C) - n_3e(A,B).\end{aligned}$$

Therefore the distance between new cluster $A \cup B$ and C is given in terms of $e(A,C)$, $e(B,C)$ and $e(A,B)$ by recursive formula (20). □

References

ANDERBERG, M.R. (1973), *Cluster Analysis for Applications,* New York: Academic Press.

BENZÉCRI, J.-P. (1992), *Correspondence Analysis Handbook*, New York: Marcel Dekker, Inc.

BLAKE, C.L., and MERZ, C.J. (1998), UCI Repository of Machine Learning Databases (http://www.ics.uci.edu/MLRepository.html), Irvine, CA: University of California, Department of Information and Computer Science.

BREIMAN, L. (1996). "Bias, Variance, and Arcing Classifiers", Technical Report 460, Statistics Department, University of California, Berkeley, CA, USA, http://128.32.135.2/tech-reports/

CHEN, Z., and VAN NESS, J.W. (1996), "Space-Conserving and Agglomerative Algorithms", *Journal of Classification, 13,* 157–168.

CORMACK, R.M. (1971), "A Review of Classification," *Journal of the Royal Statistical Society. Series A, 134(3),* 321–367.

DU BIEN, J.L., and WARDE, W.D. (1979), "A Mathematical Comparison of the Members of an Infinite Family of Agglomerative Clustering Algorithms," *The Canadian Journal of Statistics, 7, No. 1,* 29–38.

DUDOIT, S., FRIDLYAND, J., and SPEED, T. (2000), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," Technical Report #576, Department of Statistics, University of California, Berkeley.

EVERITT, B.S., LANDAU, S., and LEESE, M. (2001), *Cluster Analysis* (4th ed.), New York: Oxford University Press, Inc.

EVERITT, B.S. (1979), "Unresolved Problems in Cluster Analysis," *Biometrics, 35(1),* 169–181.

GORDON, A.D. (1999), *Classification* (2nd ed.), Boca Raton: Chapman and Hall.

GÜVENIR, H.A., DEMIRÖZ, G., and ILTER, N. (1998),"Learning Differential Diagnosis of Erythemato-Squamous Diseases using Voting Feature Intervals," *Artificial Intelligence in Medicine, 13(3),* 147–165.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* New York: Springer.

HARTIGAN, J.A. (1967), "Representation of Similarity Matrices by Trees," *Journal of the American Statistical Association, 62(320),* 1140–1158.

HARTIGAN, J.A. (1975), *Clustering Algorithms*, New York: Wiley.

HARTIGAN, J.A. (1985), "Statistical Theory in Clustering," *Journal of Classification, 2,* 67–76.

HUBERT, L., and ARABIE, P. (1985), "Comparing Partitions," *Journal of Classification, 2,* 193–218.

JAIN, A.K., and DUBES, R.C. (1988), *Algorithms for Clustering Data*, New Jersey: Prentice-Hall.

JOHNSON, S.C. (1967), "Hierarchical Clustering Schemes," *Psychometrika, 32(3),* 241–254.

KAUFMAN, L., AND ROUSSEEUW, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley.

LANCE, G.N., and WILLIAMS, W.T. (1967), "A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems," *Computer Journal, 9,* 373–380.

LEISCH, F. and DIMITRIADOU, E. (2004), Original Data Sets from Various Sources. mlbench: Machine Learning Benchmark Problems. R package version 1.0-0.

MILLIGAN, G.W. (1979), "Ultrametric Hierarchical Clustering Algorithms," *Psychometrika, 44(3),* 343–346.

MILLIGAN, G.W., and ISAAC, P.D. (1980), "The Validation of Four Ultrametric Clustering Algorithms," *Pattern Recognition, 12,* 41–50.

MORGAN, B.J.T., and RAY, A.P.G. (1995), "Non-uniqueness and Inversions in Cluster Analysis," *Applied Statistics, 44(1),* 117–134.

MURTAGH, F. (1985), *Multidimensional Clustering Algorithms.* In COMPSTAT Lectures, Vol. 4, (Eds. J.M. Chambers, J. Gordesch, A. Klas, L. Lebart, and P.P. Sint), Vienna:Physica-Verlag.

PRUDNIKOV, A.P., BRYCHKOV, A., and MARICHEV, O.I. (1986), *Integrals and Series,* New York: Gordon and Breach Science Publishers.

RAND, W.M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association, 66,* 846–850.

ROSS, D.T., SHERF, U., EISEN, M.B., PEROU, C.M., REES, C., SPELLMAN, P., IYER, V., JEFFREY, S.S., VAN DE RIJN, M., WALTHAM, M., PERGAMENSCHIKOV, A., LEE, J.C.F., LASHKARI, D., SHALON, D., MYERS, T.G., WEINSTEIN, J.N., BOTSTEIN, D., and BROWN, P.O. (2000), "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics, 24,* 227–235.

SPÄTH, H. (1980), *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Chichester: Ellis Horwood Limited.

TIBSHIRANI, R., HASTIE, T., EISEN, M., ROSS, D., BOTSTEIN, D. and BROWN, P. (1999), "Clustering Methods for the Analysis of DNA Microarray Data," Technical Report, Stanford University.

WARD, J.H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association, 48,* 236-244.