British Journal of Mathematical and Statistical Psychology (2006), 59, 1–34
© 2006 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

K-means clustering: A half-century synthesis

Douglas Steinley
University of Missouri-Columbia, USA

This paper synthesizes the results, methodology, and research conducted concerning the *K*-means clustering method over the last fifty years. The *K*-means method is first introduced, various formulations of the minimum variance loss function and alternative loss functions within the same class are outlined, and different methods of choosing the number of clusters and initialization, variable preprocessing, and data reduction schemes are discussed. Theoretic statistical results are provided and various extensions of *K*-means using different metrics or modifications of the original algorithm are given, leading to a unifying treatment of *K*-means and some of its extensions. Finally, several future studies are outlined that could enhance the understanding of numerous subtleties affecting the performance of the *K*-means method.

1. Introduction

The partitioning of objects is of great interest in many fields, including statistics, psychology, and pattern recognition. If approached by any type of complete enumeration strategy, however, the sheer magnitude of the problem is overwhelming. The number of partitions of N objects into K disjoint and non-empty subsets can be calculated with a Stirling number of the second kind (see Weisstein, 2003, p. 2865):

$$\frac{1}{N!}\sum_{i=1}^{N}(-1)^{N-i}\binom{N}{i}N^K, \tag{1}$$

which in turn can be approximated by $K^N/K!$ (see Kaufman & Rousseeuw, 1990, p. 115). Thus, for example, if 25 objects are to be grouped into 4 clusters, there are approximately 4.69×10^{13} different partitions. For small values of N (between 20 and 30), Hubert, Arabie, and Meulman (2001) and van Os (2000) have used dynamic programming to find optimal partitions. However, as N grows, a brute-force complete enumeration of all the possible partitions with an associated evaluation of some objective (loss) criterion is unrealistic. Because of these computational

*Correspondence should be addressed to Douglas Steinley, Department of Psychological Sciences, University of Missouri-Columbia, 210 McAlester Hall, Columbia, MO 65211, USA (e-mail: steinleyd@missouri.edu).

DOI:10.1348/000711005X48266

difficulties, it is of obvious value to design methods that provide 'good' (and hopefully optimal) partitions within a reasonable amount of computation time.

Cormack (1971) suggested that clusters should be externally isolated and internally cohesive, implying a certain degree of homogeneity within clusters and heterogeneity between clusters. Historically, many researchers attempted to operationalize this definition by minimizing within-group variation (Cox, 1957; Engelman & Hartigan, 1969; Fisher, 1958; Thorndike, 1953). Following these early attempts at maximizing within-group homogeneity, Sebestyen (1962) and MacQueen (1967) independently developed the *K*-means method as a strategy that attempts to find optimal partitions. Since this development, *K*-means has become extremely popular, earning a place in several textbooks on multivariate methods (Johnson & Wichern, 2002, pp. 695–700; Lattin, Carroll, & Green, 2003, pp. 288–297; Timm, 2002, pp. 530–531), cluster analysis (Anderberg, 1973, pp. 162–163; Gordon, 1999, pp. 41–49; Hartigan, 1975, pp. 80–112), statistical learning (Hastie, Tibshirani, & Friedman, 2001, pp. 461–465), and pattern recognition (Duda, Hart, & Stork, 2001, pp. 526–528) textbooks. Unfortunately, these depictions are usually brief and overlook several features, including statistical properties, equivalence to other methods, extensions, cautionary notes, and applications.

More than 20 years ago, when referring to the cluster validation literature, Blashfield, Aldenderfer, and Morey (1982) noted that their review was not exhaustive because 'the task of searching through the voluminous literature on clustering is simply too great' (p. 168). Although electronic databases have made the task of finding and obtaining relevant articles much easier, the same problem holds true even now when considering the coalescence of information pertaining to the most popular non-hierarchical clustering technique, *K*-means clustering. This paper attempts to elucidate these oft-forgotten areas by defining the method, stating several asymptotic theorems, examining related techniques, summarizing research on its performance, and suggesting directions for future explorations with the *K*-means method by collating the literature from more than a dozen different fields of study including, but not limited to, psychology, statistics, mathematics, pattern recognition, and engineering. Although several methods and algorithms are described, the most novel and clever implementations are discussed in detail to provide a building block for future research.

2. The *K*-means method

The *K*-means method is designed to partition two-way, two-mode data (that is, N objects each having measurements on P variables) into K classes ($C_1, C_2, \ldots, C_K$), where C_k is the set of n_k objects in cluster k, and K is given. If $\mathbf{X}_{N\times P} = \{x_{ij}\}_{N\times P}$ denotes the $N \times P$ data matrix, the *K*-means method constructs these partitions so that the squared Euclidean distance between the row vector for any object and the centroid vector of its respective cluster is at least as small as the distances to the centroids of the remaining clusters. The centroid of cluster C_k is a point in P-dimensional space found by averaging the values on each variable over the objects within the cluster. For instance, the centroid value for the jth variable in cluster C_k is

$$\bar{x}_j^{(k)} = \frac{1}{n_k}\sum_{i\in C_k} x_{ij}, \tag{2}$$

and the complete centroid vector for cluster C_k is given by

$$\bar{\mathbf{x}}^{(k)} = (\bar{x}_1^{(k)}, \bar{x}_2^{(k)}, \ldots, \bar{x}_P^{(k)})'. \tag{3}$$

According to Gentle (2002, p. 239), finding these clusters is a 'computationally intensive task' that is 'rather complicated'. Using the notation just introduced, a typical *K*-means algorithm would operate by the following iterative procedure:

(1) *K* initial seeds are defined by *P*-dimensional vectors $(s_1^{(k)}, \ldots, s_P^{(k)})$, for $1 \leq k \leq K$, and the squared Euclidean distance, $d^2(i, k)$, between the *i*th object and the *k*th seed vector is obtained:

$$d^2(i,k) = \sum_{j=1}^{P} (x_{ij} - s_j^{(k)})^2. \tag{4}$$

Objects are allocated to the cluster where (4) is minimum.

(2) After initial object allocation, cluster centroids are obtained for each cluster as described by (3), then objects are compared to each centroid (using $d^2(i, k)$) and moved to the cluster whose centroid is closest.

(3) New centroids are calculated with the updated cluster membership (by calculating the centroids after all objects have been assigned).

(4) Steps 2 and 3 are repeated until no objects can be moved between clusters.

When attempting to find a 'good' partitioning of an object through the iterative method just described, it is of interest to note that we are also attempting to minimize a particular loss criterion, the error sum of squares (SSE):

$$SSE = \sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_j^{(k)})^2. \tag{5}$$

Späth (1980, p. 72) noted that at times, but probably rarely in practice, *SSE* (also referred to as 'squared error distortion' in the pattern recognition literature; Gersho & Gray, 1992) may be further minimized by single object reallocation from one cluster to another. After the initial *K*-means algorithm is performed, a final inspection is made between all points and centroids. If there is an object within C_k such that

$$\frac{n_k}{n_k - 1} d^2(i,k) > \frac{n_{k^*}}{n_{k^*} + 1} d^2(i,k^*), \tag{6}$$

then move the *i*th object from C_k to cluster C_{k^*}, and *SSE* is reduced (see Späth, 1980, p. 72).

The *K*-means iterative relocation algorithm can be formulated using $\mathbf{X}$ and two additional matrices: (a) a membership matrix, $\mathbf{M} = \{m_{ik}\}_{N \times K}$, where m_{ik} equals unity if object *i* belongs to cluster *k*, and zero otherwise, and (b) a cluster representation matrix,

$$\mathbf{R}_{K \times P} = \begin{bmatrix} \mathbf{r}'_1 \\ \mathbf{r}'_2 \\ \vdots \\ \mathbf{r}'_K \end{bmatrix},$$

where the kth row, $\mathbf{r}_k = \{r_{k1}, \ldots, r_{kP}\}$, is a vector most representative of the elements of cluster k. For K-means, each row is represented by the centroid vector of means for cluster k on the P variables. This strategy allows (5) to be rewritten as a function of $\mathbf{M}$ and $\mathbf{R}$,

$$F(\mathbf{R}, \mathbf{M}) = \text{tr}[(\mathbf{X} - \mathbf{MR})'(\mathbf{X} - \mathbf{MR})], \tag{7}$$

which can be estimated by a least squares algorithm procedure that alternates between minimizing (7) with respect to $\mathbf{M}$ given the current estimate of $\mathbf{R}$ and then fixing $\mathbf{M}$ and recomputing $\mathbf{R}$ based on the current estimate of cluster membership, until there is no change in (7). Additionally, it is worthwhile to note that $(\mathbf{X} - \mathbf{MR})'(\mathbf{X} - \mathbf{MR})$ is equivalent to the within sum-of-squares and cross-products matrix, $\mathbf{W}$, making $(\mathbf{X} - \mathbf{MR})'(\mathbf{X} - \mathbf{MR})$ equivalent to $\sum_{k=1}^{K} \mathbf{W}_k$, where $\mathbf{W}_k$ is the within-clusters sums-of-squares-and-cross-products matrix for the kth cluster. This equivalence leads to alternatively representing $\mathbf{W}_k$ as (Hubert *et al.*, 2001, p. 19; Späth, 1985, p. 20)

$$\mathbf{W}_k = \frac{1}{2n_k} \sum_{i \in C_k} \sum_{i^* \in C_k} (\mathbf{x}_i - \mathbf{x}_{i^*})(\mathbf{x}_i - \mathbf{x}_{i^*})' \quad \forall i, i^* = 1, \ldots, n_k, \tag{8}$$

where

$$\mathbf{X}_{N \times P} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix}$$

and the ith row is $\mathbf{x}_i = \{x_{i1}, \ldots, x_{iP}\}$, making the equivalence of (5) and (7) apparent by the relationship of the trace of $\mathbf{W}$ and the sum of the traces of $\mathbf{W}_k$,

$$\text{tr}(\mathbf{W}) = \sum_{k=1}^{K} \text{tr}(\mathbf{W}_k). \tag{9}$$

Furthermore, by expanding (7) the middle two terms cancel, and it is seen that the optimal value of $\mathbf{R}$ is equal to $(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X}$, obtaining the formulation in Gordon and Henderson (1977),

$$SSE = \text{tr}(\mathbf{X}'(\mathbf{I} - \mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}')\mathbf{X}), \tag{10}$$

allowing the problem to be viewed as trying to find an optimal projection of the columns of $\mathbf{X}$; however, to date, no substantive work has been pursued in the conceptual direction provided by (10).

Although tr($\mathbf{W}$) is the most popular function of $\mathbf{W}$ to minimize, several alternatives have been suggested in the literature. $\mathbf{W}$ has two matrix counterparts: the total and between sums-of-squares-and-cross-products matrices, $\mathbf{T}$ and $\mathbf{B}$, respectively, and all are related through the relationship $\mathbf{T} = \mathbf{B} + \mathbf{W}$ (Anderson, 1958; Johnson & Wichern, 2002; Mardia, Kent, & Bibby, 1979; Morrison, 1976). As stated above, minimizing tr($\mathbf{W}$) (Edwards & Cavalli-Sforza, 1965) is equivalent to minimizing (5). However, Friedman and Rubin (1967) attempted to adapt classical multivariate statistics to the clustering problem, such as minimizing $|\mathbf{W}|$ and $|\mathbf{W}|/|\mathbf{T}|$, Wilks's lambda statistic, or maximizing the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ and tr($\mathbf{W}^{-1}\mathbf{B}$), Hotelling's trace criterion; whereas

Windham (1987) suggested minimizing the sum of the determinants of within-cluster sums-of-squares matrices

$$\sum_{k=1}^{K} |\mathbf{W}_k|^{1/P}. \tag{11}$$

Two alternatives to (11) are minimizing

$$\prod_{k=1}^{K} |W_k|^{n_k} \tag{12}$$

and

$$\sum_{k=1}^{K} (n_k - 1)|W_k|^{1/P}, \tag{13}$$

provided by Scott and Symons (1971) and Maronna and Jacovkis (1974), respectively. Additionally, it has been noted (Scott & Symons, 1971) that the optimization of tr(**W**) and |**W**| tends to result in clusters of approximately the same size and shape. Thus, Symons (1981) suggested replacing |**W**| with

$$N\log|\mathbf{W}| - 2\sum_{k=1}^{K} n_k \log\ n_k, \tag{14}$$

and (12) with

$$\sum_{k=1}^{K} (n_k \log|\mathbf{W}_k| - 2n_k \log\ n_k), \tag{15}$$

to avoid restrictions on sample size and cluster shape; Banfield and Raftery (1993) proposed minimizing

$$\sum_{k=1}^{K} n_k \log|\mathbf{W}_k/n_k| \tag{16}$$

or

$$\sum_{k=1}^{K} n_k \log \mathrm{tr}\,(\mathbf{W}_k/n_k) \tag{17}$$

for the same reasons. As with *K*-means, locally optimal solutions for all of these criteria can be found by starting with an arbitrary partition and relocating points until the criterion cannot be reduced (Krzanowski & Marriott, 1995, p. 80).

McRae (1971) implemented the three alternative methods proposed by Friedman and Rubin (1967) into a *K*-means-like algorithm. Bartko, Strauss, and Carpenter (1971) evaluated max(ln(|**T**|/|**W**|)) for clustering psychiatric data and indicated that this procedure exhibited poor recovery performance of substantially meaningful clusters. A few studies of these methods have been undertaken, each finding that the classical *K*-means criterion outperforms |**W**| (Bayne, Beauchamp, Begovich, & Kane, 1980; Blashfield, 1977; Mezzich, 1978), but none has used a broad range of simulated data with known structure as found in studies evaluating the performance of (5) (see Milligan, 1980, 1981, 1985, 1996; Milligan &

Cooper, 1987; Steinley, 2003, 2004a) to evaluate the methods based on more complex functions of **W**, such as (11)–(17), making their conclusions tentative at best.

3. Important considerations

3.1. Local optima

Early researchers were aware that the *K*-means algorithm does not necessarily provide a global optimum, and depending on the starting values used, the algorithm terminates at a local optimum that is never verifiably globally optimal (MacQueen, 1967; Hartigan & Wong, 1979). To avoid local optima, some (Falkenauer & Marchand, 2001; Hartigan, 1975, Chapter 4) suggest performing the *K*-means method several times, with different starting values, accepting the best solution (in terms of *SSE*); however, it has been shown that the number of local optima for data sets of moderate size can run into the thousands (Steinley, 2003), indicating that the results from studies using a small number of random restarts (see Makarenkov & Legendre, 2001, p. 262) may be misleading. Thus, the presence of local optima is a much more pervasive problem than previously thought. Falkenauer and Marchand (2001) contended that the exhibited non-robustness of the *K*-means algorithm makes it very undesirable; however, it has been found that the *K*-means algorithm usually exhibits good cluster recovery properties (Dimitriadou, Dolničar, & Weingessel, 2002; Steinley, 2003).

3.2. Methods of initialization

Because there could be numerous local optima for a data set, the choice of starting values for the *K*-means algorithm is all the more crucial, and several alternatives have been proposed in an attempt to avoid locally optimal solutions. MacQueen (1967) advocated choosing K data points as the initial cluster seeds (a strategy currently implemented in SPSS, 2003, p. 1314); however, this procedure suffers from the influence of the initial ordering of the data. Closely related is the method of randomly choosing K data points as the initial cluster seeds (McRae, 1971); Forgy (1965) and Steinley (2003) randomly partitioned the data units into K mutually exclusive partitions, calculated the partition means, and used these as the initial centroids. Steinley found this method to outperform several other methods currently employed in commercial software packages.

Various deterministic methods have also been proposed. For example, Astrahan (1970) first defined a distance, d_1, and then for each data point, computed the number of data points within d_1 (referred to as the density); the one with the highest density is chosen as the first cluster seed. The remaining $K - 1$ seeds are chosen by decreasing density, as long as they are at least another specified distance, d_2, from the seeds that have already been chosen. Ball and Hall (1967, pp. 72–74) suggest a similar method that is currently implemented in the PROC FASTCLUS procedure in SAS (2004, pp. 1377–1428). Disregarding d_1 and setting the first observation equal to the first cluster seed, the second seed is chosen as the next observation that is at least some distance, d_2, from the first seed. Each subsequent observation is checked, and the remaining observations that are the farthest apart are determined and used as the other $K - 1$ seeds.

Milligan's (1980) idea of using rational starts or starting with the results obtained from a hierarchical agglomerative procedure (such as Ward's method; Ward, 1963) has met with increasing support in the literature (Arabie & Hubert, 1992, 1994; Huberty, DiStefano, & Kamphaus, 1997; Milligan, 1980, 1996; Milligan & Sokol, 1980;

Punj & Stewart, 1983; Waller, Kaiser, Illian, & Manry, 1998). Some researchers may wish to provide their own starting seeds based on previous experience with the subject area (for example, see Belbin, 1987), but as Hartigan (1975) cautioned, 'The number of clusters K should not be decided in advance, but the algorithm should be run with several different values of K' (p. 100), making it impossible to provide predetermined starting seeds.

Bradley and Fayyad (1998) suggested a bootstrap-like algorithm for determining the initial seeds in K-means clustering. Initially, S samples of size n^* ($n^* \ll N$) – sampled with replacement – are taken and each sample is independently clustered into K groups; the sampled data are then combined and reclustered via the K-means algorithm using membership from the previous clusterings to initialize the algorithm. The cluster means from this step act as the 'refined' initial seeds for the clustering of the entire data set. Cerioli and Zani (2001) proposed a density search method based on quadrant counts arising from multivariate histograms to (a) help determine the number of clusters and (b) restrict initial centroids to be chosen from areas of high density; Faber (1994) suggested choosing the starting points uniformly randomly to give preference to points in dense regions.

Hajnal and Loosveldt (2000) suggested that the initialization method implemented by SAS is superior to using a random seed as starting value for K-means. However, in a much broader study examining the performance of the default functions for several commercial software packages (i.e. SAS, SYSTAT, and SPSS), Steinley (2003) found that multiple random restarts (numbering in the thousands) outperformed approaches using Ward's (1963) method as a starting seed (Milligan & Sokol, 1980) and the default options provided by the commercial software.

3.3. Methods to estimate K

One of the hardest problems in cluster analysis is determining the value of K. Milligan and Cooper (1985) conducted an extensive Monte Carlo investigation of 30 methods used in hierarchical cluster analysis to determine the correct number of clusters; however, although Milligan and Cooper mentioned that many methods could be extended to non-hierarchical partitioning methods, none were tested within that context. Additionally, many of these methods require the clusters to be hierarchically nested (making them inappropriate for K-means clustering) or are not appropriate to be evaluated when the clustering is done in a hierarchical manner (such as Marriot's, 1971, method discussed below). This section considers three kinds of methods: algorithmic methods, graphical methods, and formulaic methods, each of which is closely related to K-means clustering.

3.3.1. Algorithmic methods

Algorithmic methods determine K through decisions within the algorithm itself. Like classical K-means, the number of clusters is provided by the user, but the algorithm has an opportunity to modify the user-provided value.

MacQueen (1967) extended his original algorithm by creating a variation of K-means that estimated K by defining a coarsening parameter, φ, and a refining parameter, ψ. As in the traditional K-means algorithm, an initial value of K and cluster seeds are chosen. After calculating the pairwise distances between the initial seeds, if two seeds are closer than φ, they are combined, and the remaining data points are assigned to the cluster with the

closest centroid, and cluster centroids are recomputed. When cluster centroids are updated, pairwise distances are recalculated and the necessary mergers, based on φ, are performed. However, for each data point, if the distance to the nearest centroid is greater than ψ, that data point becomes its own cluster. Unlike the *K*-means procedure, this process ends after one pass through the data. Anderberg (1973, p. 169) noted that there are no clear guidelines for choosing φ and ψ, but if φ is chosen to be fairly small and ψ fairly large, it might be a useful method to detect outliers.

Wishart (1969) proposed a similar method to that of MacQueen (1967) using the user-defined parameters ψ^*, $n_{\min}$, and $K_{\max}$. The procedure begins with an initial partition of the data, such as the random assignment procedure used by Forgy (1965) and Steinley (2003), and the centroids are computed on the basis of the initial partition. The distance from each object to each centroid is calculated, and if the smallest distance exceeds ψ^*, the observation is set aside and the cluster centroids updated; otherwise, the observation is assigned to the cluster with the nearest centroid. If at any time, because of updating the cluster centroids, an object that has been set aside is closer to a cluster centroid than ψ^*, the object is assigned to that cluster. After this procedure converges, if the number of objects within any given cluster is less than $n_{\min}$, these are removed from the analysis and the previous steps repeated. When both of these steps converge, the most similar clusters are merged until there are at most $K_{\max}$ clusters.

Ball and Hall (1965) developed a more elaborate method, ISODATA, for clustering data based on the nearest centroid method. First, the data are subjected to the *K*-means algorithm (as first described), and any clusters that contain fewer than $n_{\min}$ observations are discarded from the analysis. As in MacQueen's (1967) method, clusters are merged if the distance between their centroids is less than φ' (user-defined); while a cluster is split if the product of θ (user-defined) and the standard deviation of any variable (across the entire data set) is exceeded by the within-cluster standard deviation for that variable, and the assignment of data units depends on whether they are above or below the mean of the splitting variable. Anderberg (1973, p. 172) elaborated on this method and cautioned that ISODATA rarely finds elongated clusters; Dubes and Jain (1976) indicated that ISODATA shows no significant improvement over the regular *K*-means algorithm.

Belbin (1987) warned against these techniques, asserting that allowing groups to split and merge within the algorithm may result in combining two groups that have little or nothing in common, leading to the proposal of creating an inter-group distance matrix that can be subjected to hierarchical clustering techniques to determine whether groups should be merged or split.

A classic criterion for determining model complexity is Akaike's information criterion (AIC) (Akaike, 1974); this has been consistently found, however, to overestimate the number of parameters, and it has been suggested that a penalized measure may be more appropriate, creating the Bayesian information criterion (BIC) (Kass & Wasserman, 1995; Schwarz, 1978), while some related methods based on entropy are provided in Celeux and Soromenho (1996). Pelleg and Moore (2000) included the BIC in their *X*-means algorithm to determine the number of clusters, by forming a range for *K*, $[K_{\min}, K_{\max}]$. At each level of *K*, a BIC score is computed and *X*-means chooses *K* to be the value that maximizes BIC. Bischof, Leonards, and Selb (1999) developed a method based on minimum description length (MDL), where the description length is a measure of model fit. Starting with a large value of *K*, the MDL algorithm removes clusters whenever description length can be reduced, and *K*-means is used at each step to optimize the model fit to the data; the process is continued until a stopping criterion is met. Hamerly and Elkan (2002) created an algorithm, *G*-means, to

determine K by choosing a small value for K, running the K-means algorithm, and testing whether the data assigned to each cluster centre are from a Gaussian distribution. If a cluster fails the multivariate normality test, it is split, K-means is run again, and all $K + 1$ clusters retested. The procedure is repeated until no cluster fails a test for multivariate normality. Interestingly, Hamerly and Elkan (2002) created distortion parameters for testing multivariate normality instead of using the well-known, established tests developed by Mardia (1970, 1974, 1975).

3.3.2. Graphical methods

Some authors (Gierl & Schwanenberg, 1998; Gower, 1975; Thorndike, 1953) advocate plotting the objective criterion (5) against varying values of K; a 'flattening' of the curve indicates the correct value of K. At first glance, this is a valid method for determining K due to the monotonic decreasing relationship between (5) and the value of K (i.e. as K increases, (5) decreases); however, the method is highly subjective and prone to the same criticisms as the scree plot for determining the number of factors or the number of components in data reduction, or the stress plot for determining the number of dimensions in scaling. Davidson (2002) went so far as to postulate that this monotonic decreasing relationship results in the algorithm being inconsistent and undesirable.

3.3.3 Formulaic methods

Formulaic methods require the computation of an equation across a range of K and choosing the value either minimizing or maximizing the criterion. Makarenkov and Legendre (2001) used the Calinski and Harabasz (1974) statistic (the method Milligan & Cooper, 1985, recommended above all others) to determine the optimal number of groups in the context of K-means clustering,

$$\left\{\frac{\text{tr}\,(\mathbf{B})}{K-1}\right\} \Big/ \left\{\frac{\text{tr}\,(\mathbf{W})}{N-K}\right\}, \tag{18}$$

where K is chosen to maximize (18). Marriott (1971) recommended minimizing the criterion $K^2|\mathbf{W}|$ to determine the correct value of K, ranking 20th in the Milligan and Cooper (1985) study; this method is best used in conjunction with an algorithm whose objective function relies on $|\mathbf{W}|$ or (11)–(16). Davies and Bouldin (1979) proposed a technique that starts by assigning two observations to each cluster (i.e. $K = N/2$). The within-cluster error sum of squares, SSE_k, ($k = 1, \ldots, K$), is computed for each cluster, leading to a similarity measure between clusters k and k^* calculated by

$$S_{kk^*} = \frac{SSE_k + SSE_{k^*}}{d^2\left(\bar{\mathbf{x}}^{(k)}, \bar{\mathbf{x}}^{(k^*)}\right)}, \tag{19}$$

where the denominator is the distance between the centroids of the two clusters. Then an overall separability score can be written as

$$\bar{S} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{S}_k, \tag{20}$$

where $\mathcal{S}_k$ is the maximum of S_{kk^*} over $k \neq k^*$. This process is repeated for a range of K from 2 to $N/2$, and the 'correct' value of K is chosen to be the value minimizing (20). The Davies and Bouldin (1979) index ranked 10th in the Milligan and Cooper (1985)

study, but it might exhibit better recovery rates when used in conjunction with (9) or (17). Ray and Turi (2000) proposed a very similar method based on an index calculated by

$$\frac{N^{-1}\sum_{k=1}^{K} SSE_k}{\min[d^2(\bar{x}^{(k)}, \bar{x}^{(k^*)})]} \quad \text{for } k \neq k^*, \tag{21}$$

where the numerator is the average within-cluster sum of squares and the denominator is the minimum distance between any two clusters. *K*-means is performed for values in the range $[1, K_{\max}]$, where $K_{\max}$ is the maximum number of clusters (supplied by the user), and the 'correct' number of clusters is the value of K that minimizes (21). Jain and Moreau (1987) embedded (20) (or similar techniques) in a bootstrapping environment to estimate the number of clusters; Wong (1985) provided a similar procedure based on a test statistic developed in Wong and Lane (1983). Kaufman and Rousseeuw (1990, p. 85) recommended choosing K to maximize

$$\left(\sum_{k=1}^{N} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}\right) \Big/ N, \tag{22}$$

where $a(i)$ is the average distance of object i to all other members of its cluster, and $b(i)$ is the minimum distance of object i to any member of a different cluster. Alternatively, Krzanowksi and Lai (1988) recommended maximizing

$$\left|\frac{DIFF(K)}{DIFF(K+1)}\right|, \tag{23}$$

where $DIFF(K) = (K-1)^{2/P}\mathrm{tr}(\mathbf{W}^{K-1}) - K^{2/P}\mathrm{tr}(\mathbf{W}^{K})$. Hubert and Levin (1976) proposed the minimization of

$$\frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)}, \tag{24}$$

where d_w is the sum of the within-cluster distances. The measure in (24) exhibited excellent recovery of the true number of clusters in Milligan (1981), ranking 3rd among all methods in Milligan and Cooper (1985); this index may be particularly relevant when used with (9) because (24) is directly related to the compactness of the clusters – exactly what *K*-means is trying to optimize.

Recently, the choice of K has reemerged as an important research issue in the statistical literature as well. Tibshirani, Walther, and Hastie (2001) proposed the gap statistic to determine K where

$$Gap(K) = \frac{1}{B}\sum_{b}\log(SSW_b(K)) - \log(SSW(K)), \tag{25}$$

and B is a set of uniform data sets, each with the same range as the original data, and $SSW_b(K)$ is the within-cluster sum of squares for the bth uniform data set. Then K is chosen as the smallest value such that $Gap(K) \geq Gap(K+1) - s_{K+1}$ (where s_k is an estimate of the standard deviation of $\log(SSW_b(K))$). Tibshirani *et al.* (2001) found favourable results when comparing (25) to other methods designed to determine K.

However, the generalization of the results provided by Milligan and Cooper (1985) and others (such as Tibshirani *et al.*, 2001) has to be taken in the context of the Monte Carlo procedure implemented and the data generation process used. Like the optimization

procedures above, methods to estimate K based on $\mathrm{tr}(\mathbf{W})$ and $|\mathbf{W}|$ may perform more appropriately when the clusters are of equal size and shape. Even more problematic, however, is that methods to determine the number of clusters based on these criteria may produce an artificial number of clusters when presented with more complex data structures.

3.4. What variables should be used?

3.4.1. Variable standardization

The only comprehensive study examining variable standardization was conducted by Milligan and Cooper (1988); it investigated eight different methods of standardization under several error conditions. Milligan and Cooper concluded that standardizing by the range (instead of the usual z-score) was the most effective method; Dillon, Mulani, and Frederick (1989) indicated variable standardization (z-scores specifically) can result in misleading conclusions when true group structure is present; Späth (1985, p. 23) noted that a z-score is as arbitrary as any other type of scaling. In contrast, Vesanto (2001) argued that z-scores are more interpretable and should be chosen in lieu of standardizing by the range, subsequently creating a measure of 'variable quality' based on the z-score. Vesanto (2001), however, failed to carry out a comprehensive study, and his assertion is not as strongly supported as Milligan and Cooper's results, as they investigated several agglomerative hierarchical techniques.

Schaffer and Green (1996) studied the effects of variable standardization on K-means clustering across ten real data sets and found the data should not be standardized in any of the ways studied by Milligan and Cooper (1988), supporting Stoddard's (1979) conjecture that standardization based on an entire sample variance is likely to eliminate valuable between-cluster variation. However, a study by Steinley (2004a) mirrored the study of Milligan and Cooper, except that the focus was the K-means procedure. In this study, Steinley (2004a) found results parallel to those of Milligan and Cooper and recommended standardization by the range, indicating that the results of Schaffer and Green (1996) may be unreliable because they were trying to make inferences from real data sets while the true underlying structure was unknown.

3.4.2. Variable selection

Fowlkes, Gnanadesikan, and Kettenring (1988) proposed a forward selection technique for identifying a subset of the most 'meaningful' variables in cluster analysis, applying their results to complete linkage hierarchical clustering and noting that the method can be extended to the K-means algorithm. Fletcher and Satz (1985) contended that the distribution of variables must be skewed to support the presence of an underlying group structure; Bajgier and Aggarwal (1991) suggested that testing for negative kurtosis is the most powerful way to determine if univariate mixtures (especially balanced mixtures) are present – Donoghue (1995) extended this notion to multivariate mixtures by creating two indices designed to indicate when a distribution is bimodal or multimodal.

Carmone, Kara, and Maxwell (1999) proposed a variable selection technique dubbed the heuristic identification of noisy random variables (HINoV), based on the Hubert and Arabie (1985) adjusted Rand index (ARI) (see Steinley, 2004b, for a comprehensive review of the ARI). Brusco and Cradit (2001) noticed several limitations with HINoV and proposed a more fundamentally sound variable selection heuristic for K-means clustering (VS-KM) by combining aspects of previous variable selection techniques

(i.e. Carmone *et al.*, 1999; Fowlkes *et al.*, 1988). The method used a simple forward selection method based on partitions (denoted by $\mathcal{P}$) of K clusters created from single variables, $\mathcal{P}_j$, and pairs of variables, $\mathcal{P}_{jj^*}(j, j^* = 1, \ldots, P)$, using the K-means algorithm. A brief description of the algorithm follows.

To select the first pair of variables, two matrices are created: (a) $\mathbf{R}^* = \{r_{jj^*}\}_{P\times P}$, where r_{jj^*} is the ARI between $\mathcal{P}_j$ and $\mathcal{P}_{j^*}$ and (b) $\mathbf{Q} = \{q_{jj^*}\}_{P\times P}$, where q_{jj^*} is $\mathrm{tr}(\mathbf{B})/\mathrm{tr}(\mathbf{T})$ calculated with respect to $\mathcal{P}_{jj^*}$. Three user-defined values, T, $R_{\min}$, and R_{fac}, are required for VS-KM; Brusco and Cradit (2001) recommend choosing $T = 0.25$, $R_{\min} \in [.03, .07]$, and $R_{\mathrm{fac}} \in [.3, .7]$.

(1) If any $r_{jj^*} > T$, choose j and j^* to maximize $\{q_{jj^*}|(r_{jj^*} > T)\}$. If all $r_{jj^*} > T$, choose j and j^* to maximize q_{jj^*} (i.e. not conditioned on r_{jj^*}).
(2) Let S be the set including j and j^*, S^c be its complement, and $\eta = r_{jj^*}$.
(3) Create a partition based on S, $\mathcal{P}_S$.
(4) Compute the ARI, $R_{j'}$, between $\mathcal{P}_S$ and $\mathcal{P}_{j'}$ for all $j' \in S^c$.
(5) Let $\lambda = \max(R_{j'})$. If $\lambda < R_{\min}$, or $\lambda < \eta \times R_{\mathrm{fac}}$, go to step 6; otherwise, letting j'_λ denote the variable that corresponds to $\max(R_{j'})$, set $\eta = \lambda$ and $S = S \cup \{j'_\lambda\}$. If $S^c = \{\emptyset\}$, go to step 6; otherwise go to step 3.
(6) Variables in S are selected for inclusion in a final K-means analysis; while variables in S^c are discarded.

Brusco and Cradit (2001) showed that VS-KM outperformed HINoV in two different studies using both simulated and real data. Although very promising, VS-KM has not been evaluated in conjunction with other variable selection methods. Recently, Brusco (2004) has provided an analogous procedure for variable selection when using K-means clustering to cluster binary data.

3.4.3. Variable weighting

Instead of variable selection, several authors have proposed variable weighting procedures that give differential weights to the P variables, and the weighted squared Euclidean distance,

$$d_w^2(i, i^*) = \sum_{j=1}^{P} w_j^2 (x_{ij} - x_{i^*j})^2. \tag{26}$$

DeSarbo, Carroll, Clark, and Green (1984) introduced the SYNCLUS model as a way of generalising K-means to include a variable weighting mechanism; De Soete (1986, 1988) developed similar methods for optimal variable weighting of ultrametric and additive trees. As an alternative to these, Makarenkov and Legendre (2001) extended De Soete's (1986, 1988) method to K-means clustering and, like DeSarbo *et al.* (1984), based their procedure on (26), minimizing

$$\sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i\in C_k} \frac{w_j^2 (x_{ij} - \bar{x}_j^{(k)})^2}{n_k}, \tag{27}$$

noting that the weights can be set equal to $1/P$, or repeatedly randomly assigned different values (under the constraint that their sum is unity); the solution returning the smallest value of (27) is chosen. Makarenkov and Legendre (2001) used the Polak–Ribiere optimization procedure (see Press, Flannery, Teukolsky, & Vetterling,

1986, p. 303) to minimize (27) and noted that the procedure seemed particularly useful only when 'noisy' variables (i.e. variables without cluster information) existed, but recommended using equal weights (i.e. classic *K*-means) when data are error-perturbed or contain outliers.

Green, Carmone, and Kim (1990) studied the SYNCLUS method for optimal variable weighting in *K*-means clustering. For two empirical data sets, the authors found mixed performance results for the SYNCLUS method; Gnanadesikan, Kettenring, and Tsao (1995) noted that both SYNCLUS and DeSoete's (1986, 1988) method performed poorly. It is unclear whether the poor performance of DeSoete's algorithm carries over to the modification of Makarenkov and Legendre (2001) when compared to other variable weighting mechanisms. Gnanadesikan *et al.* (1995) note that more effective alternatives may be based on previous work by Art, Gnanadesikan, and Kettenring (1982) based on the decomposition

$$\mathbf{T} = \mathbf{W}^* + \mathbf{B}^*, \tag{28}$$

where $\mathbf{W}^*$ is an approximation of $\mathbf{W}$ calculated by using the conceptualization provided in (8) but summing over 'probable' within-cluster pairs. A weights matrix, $\mathbf{Y} = \{y_{jj^*}\}_{P\times P}$, can be incorporated into the calculation of (4) by

$$d^2(i,k) = (\mathbf{x}_i - \bar{\mathbf{x}}^{(k)})\mathbf{Y}(\mathbf{x}_i - \bar{\mathbf{x}}^{(k)})', \tag{29}$$

where possible choices for $\mathbf{Y}$ are

$$\mathbf{Y} = (\mathbf{W}^*)^{-1}, \tag{30}$$

$$\mathbf{Y} = \{\text{diag}(\mathbf{W}^*)\}^{-1}, \tag{31}$$

and

$$\mathbf{Y} = \text{diag}(\mathbf{B}^*)\left\{\text{diag}(\mathbf{W}^*)\right\}^{-1}. \tag{32}$$

The difficult part is determining what pairs to include in the calculation of $\mathbf{W}^*$ – if too many are used, bias will occur; if too few are used, efficiency will decrease. Some methods are discussed in Art *et al.* (1982) and Gnanadesikan, Harvey, and Kettenring (1993), but no method has been shown to be exceptional and preferable to all others (note that this procedure is implemented in the SAS system as PROC ACECLUS).

3.4.4. Data reduction

Barker (1976) found promising results by using principal component analysis to transform a data set into component scores and then performing *K*-means clustering on the component scores that corresponded to eigenvalues greater than one – a technique misused frequently in the earlier development of cluster analysis (see Everitt, 1977; Everitt, Gourlay, & Kendall, 1971). This procedure has been gaining support in the pattern recognition literature (Ben-Hur, Horn, Siegelmann, & Vapnik, 2001; De Backer & Scheunders, 1999), despite several warnings in the statistical, classification, and psychology literature. Chang (1983) asserted that principal component analysis is generally not appropriate for selecting weighted combinations of variables for input into clustering algorithms (the result was demonstrated by attempting to separate two multivariate normal distributions with only their principal component scores); the same procedure (also called 'tandem analysis' or 'tandem

clustering') has been strongly denigrated (Arabie & Hubert, 1994; De Soete & Carroll, 1994) because the first few 'important' components or factors of $\mathbf{X}$ are not guaranteed to define a subspace that is informative with respect to the true, underlying structure in the data, and in some cases the components may even mask the true structure. Kiers (1997) and Vichi and Kiers (2001) suggested the alteration of the true structure may occur because standardized principal components alter the distances between objects; however, when correcting for this, Vichi and Kiers (2001) noted that recovery is still poor and principal components are probably not related to the true cluster structure underlying the data. An equally troublesome procedure is using factor analysis to reduce the dimension of the data and then performing a cluster analysis, such as *K*-means, on the rotated, standardized factor scores, and this has been cautioned against in two empirical studies (Green & Krieger, 1995; Schaffer & Green, 1998).

Methods have been developed that combine *K*-means with multidimensional scaling to represent the 'final' clusters in a low-dimensional space, which is equivalent to a constrained multidimensional scaling problem requiring all N points to occupy K locations in the reduced space (Van Buuren & Heiser, 1989; Heiser & Groenen, 1997). Stute and Zhu (1995) proposed reducing the dimensionality of the data by embedding projection pursuit (Friedman, 1987) within the *K*-means algorithm; De Soete and Carroll (1994) provided a method where each cluster is represented by a point in low-dimensional space, selected to be the centroid to which the observations in the full space are closest. In this latter instance, the objective function minimized is the sum of squared distances between the full-dimensional data and the low-dimensional centroids. Vichi and Kiers (2001) suggested that it is more reasonable to minimize the sum of squared distance between the low-dimensional data and the low-dimensional centroids (achieved by projecting all data points into a subspace) and proposed the factorial *K*-means algorithm by minimizing

$$\mathrm{tr}\,(\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}) - \mathrm{tr}\,(\mathbf{A}'\mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X}\mathbf{A}), \tag{33}$$

where $\mathbf{A} = \{a_{kf}\}_{K\times F}$ is a columnwise orthonormal matrix, and the entries represent the coefficients of the linear combinations of the observed variables in the reduced space of dimensionality F. Vichi and Kiers (2001) found that (33) outperformed preprocessing the data with principal component analysis and De Soete and Carroll's (1994) method of subspace projection.

3.5. Detecting influential observations

Cheng and Milligan (1996a) developed an approach to detect influential observations in *K*-means clustering based on methods presented in Cheng and Milligan (1996b). This method is based on the standard jackknife procedure; influential observations are identified by performing the cluster analysis with all data points, removing a specific point, redoing the cluster analysis, and checking to see whether the removal of the point changes the clustering of the data. Cheng and Milligan (1996b) derived influence measures based on the ARI and found very promising results in determining whether points were influential.

Barnett and Lewis (1994) noted that single-deletion diagnostics can suffer from problems of masking and swamping when multiple outliers are present. Cerioli (1998) recognized this potential disadvantage and implemented a forward search method to detect outliers:

(1) Perform an initial K-means clustering, specifying K' clusters (where $K' \gg K$) and extract the centroids belonging to the K largest clusters (according to the number of observations).
(2) To obtain initial values, assign the K objects closest to the K extracted means in step 1 as the initial centroids.
(3) Let $s = 1$, $m = K + s$, and centroid vectors for the K clusters be $\bar{\mathbf{x}}_{s-1}^{(1)}$, $\bar{\mathbf{x}}_{s-1}^{(2)}, \ldots, \bar{\mathbf{x}}_{s-1}^{(K)}$. Compute

$$d_i = \min[d^2(\mathbf{x}_i, \bar{\mathbf{x}}_{s-1}^{(k)})] \quad \forall k = 1, \ldots, K. \tag{34}$$

(4) Arrange the observations in ascending order according to the d_i values. Let ζ_m be the set of m observations with the smallest d_i.
(5) Perform K-means on items in ζ_m, using $\bar{\mathbf{x}}_{s-1}^{(1)}, \bar{\mathbf{x}}_{s-1}^{(2)}, \ldots, \bar{\mathbf{x}}_{s-1}^{(K)}$ as initial centroids. Denote the centroids of the converged solution by $\bar{\mathbf{x}}_s^{(1)}, \bar{\mathbf{x}}_s^{(2)}, \ldots, \bar{\mathbf{x}}_s^{(K)}$. Recompute d_i, replacing the kth centroid, $\bar{\mathbf{x}}_{s-1}^{(k)}$, with $\bar{\mathbf{x}}_s^{(k)}$.
(6) Stop if $s = n - K$. Otherwise, set $s = s + 1$ and go to step 4.

Cerioli (1998) noted that outliers can be detected by simple graphical displays of the cluster cohesion measure,

$$D(s) = \max\{\max[d(x_i, x_{i'})]\} \quad \forall k = 1, \ldots, K, \forall i, i' \in k. \tag{35}$$

Equation (35) returns small values as long as ζ_m remains outlier-free. Plotting (35) by m is informative because dramatic changes in the measure are visually discernible. Cerioli (1998) also found promising results as long as K was chosen correctly; however, the major disadvantage to this method is the subjectivity required in interpreting the final plot provided by (35).

Cuesta-Albertos, Gordaliza, and Matran (1997) developed trimmed K-means as a method for 'trimming' outliers from a data set, proving consistency for absolutely continuous multivariate distributions with unique trimmed K-means. Garcia-Escudero, Gordaliza, and Matran (1999a, 1999b) extended these results by obtaining asymptotic normality and formulating an asymptotic central limit theorem for trimmed K-means, respectively. Although several 'nice' properties underlie trimmed K-means, Cuesta-Albertos *et al.* (1997) indicated that the procedure is intended for spherical, well-separated clusters of the same size – the simplest problem in cluster analysis. However, in practice, the shape of the clusters is unknown, requiring those who implement trimmed K-means to proceed with caution.

4. Theoretical results concerning *K*-means

4.1. Univariate K-means

Univariate sample K-means is encountered when the goal is to divide an $N \times 1$ vector, $\mathbf{x}$, into K groups. Elias (1970) showed that the population partition is optimal when the within-cluster sums of squares are equal. Based on work in Wong (1982a) on population K-means clusters, Wong (1984) derived two double asymptotic theorems ($K \to \infty$ and $N \to \infty$, indicating that the length of the cluster intervals approaches zero while the size of each cluster approaches infinity) for univariate sample K-means clusters: (a) the sample cluster within-sums of squares are equal; and (b) the size of the cluster intervals is inversely proportional to the cube root of the underlying density at the midpoints (which are sufficiently close to the mean in large samples) of the intervals. However, Wong noted

that the generalization of these univariate results to the multivariate situation is not straightforward because the configuration of the optimal population *K*-means partition in several dimensions is still unclear. To date, this generalization remains unsolved.

4.2. General K-means

4.2.1. Convergence

Pollard (1981) showed that as $N \rightarrow \infty$, the estimated cluster centres, $(\bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)}, \ldots, \bar{\mathbf{x}}^{(K)})$ will converge to the unique vector minimizing (5) (i.e. the true cluster centres), $(\bar{\boldsymbol{\mu}}^{(1)}, \bar{\boldsymbol{\mu}}^{(2)}, \ldots, \bar{\boldsymbol{\mu}}^{(K)})$, where K is assumed to be known; Gaenssler (1988) extended these results to when K is unknown. Pollard (1982) provided a central limit theorem for when the vector of means will be asymptotically normally distributed given that the N objects are independently sampled from a fixed distribution, extending Hartigan's (1978) result for one dimension; Selim and Ismail (1984) verify that the *K*-means algorithm will converge to at least a local optimum.

4.2.2. Admissibility

Fisher and Van Ness (1971) indicated the following forms of admissibility are of most interest when evaluating clustering procedures (only the types relevant to partitioning methods are provided):

(1) *Monotonicity*. A monotone transformation to the dissimilarity measure does not change the clustering.
(2) *Cluster omission*. Removing an entire cluster does not change the clustering of the remaining objects if the technique is reapplied.
(3) *Point proportion*. Duplication of one or more points does not affect the clustering solution.
(4) *Convexity*. The convex hulls of the clusters do not intersect.
(5) *Image admissibility*. No other pattern exists with the same value of K and the same number of objects per cluster that is 'uniformly better' than the one produced by the clustering technique.
(6) *K-group admissibility*. A clustering with K groups exists where all between-cluster distances are larger than all within-cluster distances, and a *K*-group admissible technique will find such clusters.

K-means satisfies condition 4, it is not known whether it satisfies condition 2, and it fails the other four conditions. According to some authors (Dubes & Jain, 1976; Fisher & Van Ness, 1971; Van Ness, 1973), the ideal minimum variance partitioning algorithm should satisfy all conditions except 1 and 3. In combination with basic decision theory, use of such an inadmissible procedure should be cautioned against. Unfortunately, basing decisions solely on mathematical elegance can severely retard the development of methods in an applied setting, as seen by decisions to choose single-link clustering over complete-link clustering solely on the grounds of mathematical tractability (Hubert, 2002).

4.2.3. Invariance

Friedman and Rubin (1967) criticized $\mathrm{tr}(\mathbf{W})$ because, although it is invariant under orthogonal transformations, it is not invariant under non-singular linear transformations and implicitly assumes Euclidean distance; however, their proposed methods are invariant to both kinds of transformations and make no implicit metric assumptions.

Späth (1985, p. 21) counters that in practical applications, invariance is not a property to be overly concerned with and the focus should be on the effects of changing the scale of measurement. This conjecture is followed by a short proof indicating that scale transformations are invariant only if they involve a change of sign; an arbitrary scale transformation or a general linear transformation can result in varying optimal partitions with respect to (5). Thus, it is apparent that the composition of the clusters is highly dependent upon the scale of the data.

4.3. Equivalence of K-means with other methods

4.3.1. Mixture models

Let the population density

$$f = \sum_{k=1}^{K} \alpha_k f_k \tag{36}$$

be a mixture of components, f_k, in proportions α_k. This can also be viewed as a model for K clusters. The f_k are assumed to be of the same parametric family. Then, let $p(k|x) = \alpha_k f_k(x)/\sum \alpha_k f_k(x)$ represent the posterior probability that object x belongs to cluster k. The mixture model can be formulated as

$$f(x) = \sum \alpha_k f_k(x, \theta_k), \tag{37}$$

where θ_k are relevant parameters for the distribution f_k. Given $x_1, \ldots, x_N$, the maximum likelihood estimates (MLEs) for α_k and θ_k satisfy

$$\sum_{i=1}^{N} p(k|x_i) \frac{d}{d\theta_k} \{ f_k(x_i, \theta_k) \} = 0 \tag{38}$$

and

$$\alpha_k = \sum_{i=1}^{N} p(k|x_i)/N. \tag{39}$$

This estimation proceeds in steps. First, given $p(k|x_i)$, estimate θ_k (weighting x_i by its probability of belonging to the kth cluster). Second, given the new θ_k, estimate $p(k|x_i)$. Repeat this process until the estimates do not change (or only change by a minimal, preset amount). This estimation procedure was first used for normal mixtures by Rao (1948). Until the advent of the EM algorithm (Dempster, Laird, & Rubin, 1977), fitting complex mixture models remained difficult. Now the EM algorithm is the standard method used to fit these models (Bartholomew & Knott, 1999; McLachlan & Basford, 1988; McLachlan & Peel, 2000).

After establishing the estimation procedure, the only choice to be made is the parametric family for f_k. Usually the multivariate normal distribution is chosen (e.g. Bartholomew & Knott, 1999; Day, 1969; Dick & Bowden, 1973; Hartigan, 1975, 1985; McLachlan & Basford, 1988; McLachlan & Peel, 2000; Wolfe, 1970). When

$$f_k(x) = (2\pi)^{-p/2} \exp\left\{ -\frac{1}{2}(x - \mu_k)'(x - \mu_k) \right\}, \tag{40}$$

the above estimation procedure reduces to the K-means algorithm (i.e. $\mathbf{\Sigma}_k = \sigma^2 \mathbf{I}$, $k = 1, \ldots, K$), making K-means equivalent to a finite mixture model assuming that the

clusters arise from a multivariate normal distribution with a covariance matrix proportional to the identity. Relaxing the restrictions on $\mathbf{\Sigma}_k$ makes maximizing (40) equivalent to several of the other criteria presented; for example, $\mathbf{\Sigma}_k = \mathbf{\Sigma}$ results in the $|\mathbf{W}|$ criterion and $\mathbf{\Sigma}_k = \sigma_k^2\mathbf{I}$ is equivalent to (17).

De Backer and Scheunders (1999) noted that formulating *K*-means as a finite mixture model and performing the EM algorithm can result in identifiability problems of the within-cluster covariance matrices; this can be overcome, however, by requiring that $P + 1$ observations are assigned to each cluster (Everitt, 1979). Because the EM algorithm can get stuck in local minima, several issues remain pertinent: (a) initialization of the algorithm; (b) determining the number of clusters; and (c) modelling noise and outliers. For a recent review of these topics and fitting distributions more complex than (40) with finite mixture models in general, see Fraley and Raftery (2002).

4.3.2. The maximum F test

Bock (1985, 1996) showed that minimizing (9) is the same as maximizing

$$\frac{\text{tr}\,(\mathbf{B})}{\text{tr}\,(\mathbf{W})}, \tag{41}$$

and, if multiplied by $(n - K)/(K - 1)$, this is just the usual *F*-ratio statistic obtained when testing the equality of means, μ_i, of $N_p(\mu_i, \sigma^2\mathbf{I}_p)$ with $(K - 1)P$ and $(N - K)P$ degrees of freedom (see Bock, 1996). Furthermore, note that (18) also corresponds to this value. Therefore, if the most significant *F*-ratio possible is found, it will correspond to (41), giving the test its name – *the maximum F test*.

4.3.3. Principal points

Principal points are the set of points in *p*-dimensional space that best approximate a given distribution (Flury, 1990); *K* principal points, $\xi_k \in R^p (1 \leq k \leq K)$, are defined as

$$E_F\{d^2(X|\xi_1, \ldots, \xi_K)\} = \min E_F\{d^2(X|y_1, \ldots, y_K)\}, \tag{42}$$

where *F* is the distribution function and $y_1, \ldots, y_K$ are a set of points (Mizuta, 1998). Thus, $\xi_1, \ldots, \xi_K$ are the *K* principal points of random variable *X* minimizing the expected distance of *X* to the nearest ξ_k. According to Mizuta (1998), the *K*-means algorithm can be used to estimate the principal points of a given theoretic distribution, but this conjecture has yet to be investigated. In fact, Tarpey, Li, and Flury (1995) indicated that principal points are special cases of self-consistent points – the set of points to which the *K*-means algorithm converges (not necessarily the global minimum).

4.3.4. Mathematical programming

The *K*-means algorithm can be restated as a non-convex mathematical program where a locally optimal solution is not necessarily the global optimum (Cooper, 1973; Selim & Ismail, 1984). Fisher (1958) provided a dynamic programming algorithm for optimally minimizing (5) in a one-dimensional case, and Jensen (1969) extended Fisher's (1958) formulation to multiple dimensions; however, it was found to be computationally infeasible due to the large number of data points ($N \times P$) which must be evaluated and stored. Vinod (1969) stated the generalized strong property as a formulation for minimizing (5); however, the basis of this property is that 'the optimal partition must have

the property that the hulls of the subsets S_k [i.e. clusters] are convex and non-overlapping' (Diehr, 1973, p. 6). Diehr (1973) proceeded to provide a stronger statement that 'the optimal partition must have the property that the subsets can be defined by hyperplanes which are perpendicular bisectors of the lines joining the group means' (p. 7); however, this is only a necessary, but not sufficient, condition for minimizing (5). Rao (1971) provided an integer programming formulation for this optimization problem, and Diehr provided a quadratic programming formulation. Dodge and Gafner (1994) suggested combining Jensen's (1969) dynamic programming method with integer programming to reduce the complexity of the search for global optima and were able to cluster 62 objects into three clusters in 23 minutes using the L_1-norm (similar procedures have not been developed to optimize the criterion provided in (5)); Gordon and Henderson (1977) noted that the problem of minimizing (5) is an unconstrained minimization problem. In a related problem, an algorithm was developed that provided optimal solutions for minimal sum-of-squares clustering in the context of divisive hierarchical methods (Hansen, Jaumard, & Mladenovic, 1998). Because current clustering problems are concerned with objects numbering in the hundreds of thousands (e.g. fMRI data), these formulations remain tractable only for small problems (Körkel, 1986), and the following statement made by Diehr (1973, p. 17) may still be applicable today: 'Researchers must keep in mind, however, that in most cases the goals of clustering do not justify the computational time to locate or verify the optimal solution'.

5. Formulations related to *K*-means

5.1. Other metric spaces

Hartigan (1975, pp. 92–93) noted that distance measures other than squared Euclidean can be used within the *K*-means framework. As a result, the *K*-means has been changed, enhanced, and extended several times in hopes of developing an algorithm with better partitioning capabilities. In general, *K*-means-like algorithms have been developed for only four metric spaces: L_1, L_2, L_∞, and L_0 (the standard Minkowski power metric where $p \rightarrow \infty$). The first three are able to handle continuous data, while the last was developed for categorical data. Their popularity stems from the simple fact that these four metric spaces have calculable cluster centres (the median, mean, midrange, and mode, respectively), while other Minkowski metrics require an iterative computation of cluster centres (see Späth, 1985, p. 63). This section discusses the L_1, L_∞, and L_0 norms in the context of iterative relocation algorithms.

5.1.1. K-medians

The *K*-medians procedure, as a robustified version of the *K*-means procedure less influenced by outliers, has been proposed several times in the literature (Kaufman & Rousseeuw, 1990, 1987; Massart, Plastria, & Kaufman, 1983; Späth, 1985, p. 71; Vinod, 1969). The *K*-medians algorithm proceeds in the same manner as *K*-means except that (5) is replaced by

$$\sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i \in C_k} |x_{ij} - med_j^{(k)}|, \tag{43}$$

where $med_j^{(k)}$ is the median for the *j*th variable in the *k*th cluster. Jhun and Jin (2000) altered this formulation by stipulating that any single-member cluster is amalgamated with its nearest cluster, claiming it is even better at recovery of elongated clusters and more robust in the presence of outliers.

Garcia-Escudero and Gordaliza (1999) contend that *K*-medians is only insensitive to outliers in special cases and, in general, (43) can be as dramatically affected by outliers as *K*-means, showing that *K*-medians does not exhibit the same robustness properties as the median does in location theory (see Huber, 1981). In fact, although the median may be a robust centralization measure for one random variable, it is unlikely that the 'joint' selection of two medians will be robust for two random variables (Cuesta-Albertos *et al.*, 1997).

5.1.2. K-midranges

The *K*-midranges method (Carroll & Chaturvedi, 1998; Späth, 1985, p. 71) follows the same algorithmic procedure as the *K*-means method, the only difference being that *K*-midranges minimizes a loss function based on the L_∞-norm (also referred to as the 'max metric' and 'dominance metric'; Borg & Groenen, 1997, p. 281), replacing (5) by

$$\sum_{j=1}^{P}\sum_{k=1}^{K}\sum_{i\in C_k}|x_{ij}-mid_j^{(k)}|, \tag{44}$$

where $mid_j^{(k)}$ is the midrange of the *j*th variable on the *k*th cluster, calculated by

$$mid_j^{(k)}=\frac{1}{2}(\max{}^{(k)}x_j+\min{}^{(k)}x_j), \tag{45}$$

the average of the two most extreme values for the *j*th variable in the *k*th cluster. First, initial midranges are chosen and objects in $\mathbf{X}$ are assigned to the cluster with the midrange which is closest to it. Midranges are then recalculated and the procedure repeated until (44) can no longer be reduced. Carroll and Chaturvedi (1998) noted that this method may be best suited to isolating outlying clusters; Späth (1985, pp. 71–72) noted that this procedure is extremely sensitive to outliers and should not be used in practical applications. Like the *K*-means algorithm, *K*-midranges can only guarantee a locally optimal solution.

5.1.3. K-modes

Initially for categorical data, Carroll, Green, and Schaffer (1986) used correspondence analysis to derive spatial coordinates that were then clustered via the *K*-means algorithm. However, Arabie and Hubert (1994) expressed concern with methods that process the data before cluster analysis is performed because the 'true' structure in the data may be corrupted. *K*-modes clustering (Chaturvedi, Green, & Carroll, 2001; Huang, 1998; Huang & Ng, 2003) is a *K*-means-like algorithm designed to derive clusters from the original, unprocessed categorical data. The *K*-modes algorithm replaces (5) with the L_0-norm-based loss function, which is the limiting case of the L_p metric as $p \to 0$, minimizing the distance between each observation and the mode of its parent cluster. It has been suggested that *K*-modes has fewer problems with locally optimal solutions than its counterpart, latent class analysis (Chaturvedi *et al.*, 2001).

5.2. Algorithmic variations

5.2.1. Transfer algorithm

Banfield and Bassill (1977) discussed a general class of non-hierarchical classification techniques that rely on swapping objects between classes. The process proceeds in two phases: (a) a general transfer phase, and (b) a swapping phase. For the *K*-means criterion the procedure starts by performing the standard *K*-means algorithm, then each possible

swapping of two objects is considered, and the objects are swapped if *SSE* is reduced. These two phases are carried out until convergence is reached. To the author's knowledge, this procedure has yet to be analysed in any type of simulation study.

5.2.2. K-harmonic means

Zhang, Hsu, and Dayal (1999) proposed replacing (5) by taking the sum over all data points of the harmonic mean of the squared distance from a data point to all the centres, mathematically represented with the loss function

$$SSHM = \sum_{i=1}^{n} \frac{K}{\sum_{k=1}^{K} 1/(d^2(i,k))}. \tag{46}$$

Because the distance of all x_i from all $\bar{\mathbf{x}}^{(k)}$ is an integral part of the loss function, improvement of (46) can only be gained by changing the values of $\bar{\mathbf{x}}^{(k)}$. Zhang (2000) proposed altering (46) to

$$SSHM_p = \sum_{i=1}^{n} \frac{K}{\sum_{k=1}^{K} 1/(d^p(i,k))}, \tag{47}$$

because taking the Euclidean distance to the pth power allows the data points far from cluster centres to be given increased importance (thus, (46) is a special case of (47) when $p = 2$). The means are updated by taking partial derivatives with respect to the cluster centres, setting them equal to zero, and solving.

Zhang *et al.* (1999) and Zhang (2000) claimed (46) and (47) are essentially insensitive to starting configurations and have a faster rate of convergence than *K*-means when they are started far from the optimal solution. Hartigan (1975, p. 95) notes, however, that for an optimal 2-partition, 'it has many local maxima, and so setting derivatives equal to zero is useless'. To date, an objective evaluation has yet to be conducted to determine if (46) or (47) suffer from the same pitfalls.

5.2.3. Symmetry based K-means

To find 'better' partitions, Su and Chou (2001) added additional steps, as well as adopting a non-metric distance measure, to the *K*-means algorithm. The modified algorithm is as follows:

(1) Initialization. Randomly assign K data points to initialize the cluster centroids $(\bar{\mathbf{x}}^{(k)}, k = 1, \ldots, K)$.
(2) Coarse-tuning. Perform the *K*-means algorithm as described above.
(3) Fine-tuning. After convergence is reached in step 2, the cluster centroid, $\bar{\mathbf{x}}^{k'}$, is found by

$$\bar{\mathbf{x}}^{k'} = \min\ d_s(\mathbf{x}, \bar{\mathbf{x}}^{(k)}), \quad k = 1, \ldots, K, \tag{48}$$

where

$$d_s(\mathbf{x}_i, \bar{\mathbf{x}}) = \min \frac{\|(\mathbf{x}_j - \bar{\mathbf{x}}) + (\mathbf{x}_i - \bar{\mathbf{x}})\|}{\|\mathbf{x}_j - \bar{\mathbf{x}}\| + \|\mathbf{x}_i - \bar{\mathbf{x}}\|}, \quad \forall i,j = 1, \ldots, N, \text{and } i \neq j. \tag{49}$$

(4) If (49) is smaller than a pre-specified value, θ, then assign the vector $\mathbf{x}$ to the cluster specified in (48). If (49) is not smaller than θ, then assign $\mathbf{x}$ according to the cluster which minimizes (4). (Su and Chou set $\theta = 1.8$, giving no theoretical support for this choice of θ.)

(5) Update the cluster means according to the new assignment.
(6) Repeat for a pre-specified number of iterations (no recommendation is given).

Through four examples, the authors indicate that this modified version of *K*-means performs better, but no independent validation studies have been conducted to support their claim.

5.3. Fuzzy K-means

Fuzzy set theory was introduced by Zadeh (1965), prompting Gitman and Levine (1970) to use it to maximally separate clusters. Naturally, a fuzzy version of *K*-means was created (Bezdek, 1974; Dunn, 1974) (also referred to as fuzzy *c*-means), and Bezdek (1980) provided a convergence theorem for this conceptualization. Fuzzy *K*-means can be viewed as a method that does not absolutely assign objects to a cluster (i.e. objects have some probability of arising from different clusters), and the loss function can be formulated as

$$FKM = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{r} d^{2}(i,k), \tag{50}$$

where w_{ik} is a weight indicating the proportion of data point i that belongs to cluster k, and all w_{ik} sum to unity. The r parameter controls the 'fuzziness' of the solution and is greater than or equal to one – the larger its value, the more 'fuzzy' the solution. In addition to this formulation, one of the most common examples of this type of classification is the finite mixture model, which assigns a vector of probabilities, α, to each object (recalling that the sum across α is unity) – providing a 'fuzzy' (or soft) classification of objects. However, in almost all large applications, fuzzy partitions cannot be utilized effectively, and the object is assigned to the group associated with the greatest value of α, defeating the purpose of performing a fuzzy *K*-means (Späth, 1985, p. 13).

5.4. Constrained K-means

Gordon (1973) first introduced the notion of using constraints in cluster analysis; DeSarbo and Mahajan (1984) elucidated the technique by provided an extensive list of constraints applicable to different classification methods, the most relevant for *K*-means clustering being: (a) two objects, x_i and x_j (where x_i and x_j are specific points indicated by the user), must belong to the same cluster, C_k, (b) x_i and x_j must belong to different clusters, (c) C_k has a maximum number of objects, n_k, which can belong to it, and (d) C_k has a minimum value for n_k. Recently, Wagstaff, Cardie, Rogers, and Schroedl (2001) integrated constraints (a) and (b) into a standard *K*-means algorithm and showed enhanced performance, while Gordon and Moore (2000) were able to identify filaments (i.e. line segments) by placing constraints on the relationships between the cluster centroids.

Another purpose of constraining *K*-means clustering is to 'speed up' the algorithm so that large databases can be processed more efficiently. Ng and Han (1994) used simulated annealing to direct the clustering to a small, subset of the data. After the appropriate subsets are chosen, the *K*-means algorithm can be implemented and the results extrapolated to the full data set. Pelleg and Moore (1999) developed a fast, constrained clustering method based on *kd*-trees with sufficient statistics for the clusters stored in the nodes (see Kanungo *et al.*, 2002, for a comprehensive discussion),

resulting in an algorithm that operates as a function of the number of centroids instead of the number of data points. This procedure does have the advantage of using the entire data matrix instead of subsampled parts of the data, and Pelleg and Moore (1999) contended that this formulation of the problem can make the *K*-means algorithm a realistic choice for data sets with billions of objects; unfortunately, they also note that when performed on data embedded in eight-dimensional space or higher, the method performs very poorly. Carmignani, Genco, Lombardo, and Tortorici (1988) reformulated the *K*-means algorithm in terms of parallel computing to improve the speed and efficiency of the traditional algorithm, and Maitra (2001) developed a multi-stage clustering algorithm (within which *K*-means can be embedded) to handle massive data sets.

5.5. Global K-means

Global *K*-means (Likas, Vlassis, & Verbeek, 2003) is a method using the standard *K*-means algorithm in progressive (from $K = 1$ to $K = K_{\max}$) stages to come to the 'optimal clustering'. For $K = 1$, the cluster centroid is taken to be the grand mean of the data, the optimal solution for this trivial situation. Then for $K = 2$, the centroid in the first step is retained and a *K*-means algorithm is performed N times, using each of the remaining data points as a candidate for the second cluster centre. After all N passes have been made, the point which minimizes (5) in conjunction with the first cluster seed (i.e. the grand mean) is retained. In general, for the Kth step, the $K - 1$ centroids from the previous step are retained and the remaining points are searched to determine which will minimize (5) – this point becomes the next cluster centroid. Although no references are given in Likas *et al.* (2003), global *K*-means is very similar to a method described by Hartigan (1975, p. 102), differing only in how new clusters are chosen at each level of *K*. Likas *et al.* (2003, pp. 451–452) boldly claim: 'Since for $K = 1$ the optimal solution is known, we can iteratively apply the above procedure to find the optimal solutions for all *K*-clustering problems $k = 1, \ldots, K$. In addition to effectiveness, this method is deterministic and does not depend on any initial conditions or empirically adjustable parameters'.

5.6. Anticlustering

To create *K* groups which are as similar to each other as possible, Späth (1986a, 1986b) recommended maximizing (5) instead of minimizing, or equivalently minimizing

$$\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2, \tag{51}$$

which provides better results than the more intuitive method of randomly assigning objects to groups.

5.7. Hybrid methods

5.7.1. K-means and Ward's method

The simplest hybrid method involving *K*-means clustering is combining it with Ward's (1963) method. As suggested above, many researchers have advocated using Ward's method to find the initial seeds for a *K*-means cluster analysis; Milligan and Sokol (1980) provided a program which does so.

5.7.2. K-means and single linkage

Wong (1982b) combined *K*-means with single-linkage (Sneath, 1957) hierarchical clustering. The *K*-means method is used to obtain a uniformly consistent estimate (in the form of a histogram) of the density over a *K*-partition of the sample space. The resulting partition is used to construct a distance matrix for the *K* clusters that is subjected to the single-linkage methods. Wong (1982b) noted that this method provides 'consistent' estimates of high-density clusters and, based on theoretical results of the two methods, suggested choosing *K* to be in the neighbourhood N^3; this seems problematic, however, because it depends directly on sample size instead of various functions of the data, such as cluster 'spread'.

6. Unified treatment of *K*-means and related methods

6.1. Bilinear clustering model

The bilinear clustering model was first introduced by Mirkin (1990) and has been used several times (Carroll & Chaturvedi, 1998; Chaturvedi, Carroll, Green, & Rotondo, 1997; Mirkin, 1996, 1998) to build a framework for clustering methods. The model is quite simple:

$$\mathbf{X} = \mathbf{MR} + \mathbf{E}, \tag{52}$$

where $\mathbf{E}$ is a matrix of errors. Equation (52) turns out to be a special case of the CANDCLUS model (Carroll & Chaturvedi, 1995), and all of the procedures involving different metric spaces (i.e. *K*-medians, *K*-modes, and *K*-midranges) can be encapsulated by altering $\mathbf{R}$. Overlapping clustering methods can be represented by relaxing the restrictions placed on $\mathbf{M}$ (i.e. for each row, allowing more than one column to take on the value of unity). Obviously, $\mathbf{M}$ can be altered with various forms of $\mathbf{R}$ to obtain different overlapping clustering criteria for different metrics (extensively discussed in Chaturvedi *et al.*, 1997). The formulation in (52) allows an easily alterable expression for several models, the expansion of classical clustering techniques to newer methods, and facilitates a unified framework for conceptualizing different models.

6.2. Integration of methods via updating techniques

Although (52) provides a succinct way to view the clustering paradigm, a different type of framework allows insight into the inner workings of several of the methods presented. If $\mathbf{R}$ is thought of as a general representative centroid matrix for the *K* clusters, then each row, $\mathbf{r}_k$, can be considered the centroid for cluster k. Also, let w_i represent the weight of observation $\mathbf{x}_i$, m_{ik} be the value in $\mathbf{M}$ for the ith observation with respect to the kth cluster, and $x^{(k)}_{(1)}, x^{(k)}_{(2)}, \ldots, x^{(k)}_{(n_k)}$ be the ordered observations in cluster k. The recomputing of $\mathbf{r}_k$ based on membership and weights can be represented as

$$\mathbf{r}_k = \frac{\sum_{i=1}^{N} m_{ik} w_i \mathbf{x}_i}{\sum_{i=1}^{N} m_{ik} w_i}. \tag{53}$$

For *K*-means, $w_i = 1$, reducing (53) to

$$r_k = \frac{\sum_{i=1}^{N} m_{ik} \mathbf{x}_i}{\sum_{i=1}^{N} m_{ik}}, \tag{54}$$

just the average of the terms belonging to cluster k. Table 1 gives the appropriate m_{ik} and w_i terms for (5), (39), (42), (43), and (46).

Table 1. Formulations for representative vectors, $\mathbf{r}_k$

Method (equation)		m_{ik}	w_i
K-means (5)		$\begin{cases} 1 & \text{if } x_i \in k \\ 0 & \text{otherwise} \end{cases}$	1
K-medians (43)	if n_k odd	$\begin{cases} 1 & \text{if } x_i \in k \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 1 & \text{if } x_i \equiv x^{(x)}_{((n_k+1)/2)} \\ 0 & \text{otherwise} \end{cases}$
	if n_k even	$\begin{cases} 1 & \text{if } x_i \in k \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 1 & \text{if } x_i \equiv x^{(x)}_{(n_k/2)} \text{ or } x^{(x)}_{((n_k+2)/2)} \\ 0 & \text{otherwise} \end{cases}$
K-midranges (44)		$\begin{cases} 1 & \text{if } x_i \in k \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 1 & \text{if } x_i \equiv x^{(x)}_{(n_k)} \text{ or } x^{(x)}_{(1)} \\ 0 & \text{otherwise} \end{cases}$
Mixture model (37)	$\alpha(k\|x_i)$		1
Harmonic K_p-means (47)	$\dfrac{1}{d^{p+2}(i,k)\left(\sum_{k^*=1}^{K} \frac{1}{d^p(i,k^*)}\right)^2}$		$\dfrac{\sum_{k^*=1}^{K} \frac{1}{d^{p+2}(i,k^*)}}{\left(\sum_{k^*=1}^{K} \frac{1}{d^p(i,k^*)}\right)^2}$

By examining the values in Table 1, it is easy to compare the different algorithms directly. For *K*-means each object is weighted equally and every object in the kth cluster contributes directly to calculating its centroid; however, for *K*-medians and *K*-midranges, although cluster membership is still relevant, only the middle region of the cluster or the outer boundaries, respectively affect the calculation of the centroid. For mixture models and *K*-harmonic means, all objects in the data set contribute a certain amount to the calculation for all centroids. For mixture models, all objects are equally weighted, but for *K*-harmonic means, objects are weighted by how close they are to any centroid.

7. Conclusion

This paper reviews a vast collection of the literature on *K*-means cluster analysis. It is clear there are many links between several different fields. Since *K*-means is the simplest version of finite mixture models, it seems that several of the methods developed under the auspices of this procedure may have some degree of transferability to other modelling methods. Although a plethora of work has been pursued in this area, it is clear that much is left to be done. The problem of initialization methods seems to be one of the key focal areas for future research. Due to the pervasive nature of locally optimal solutions, it is necessary to develop intelligent starting seeds so as not to be forced to rely on several thousand random restarts – which becomes a severe problem when analysing extremely large data sets (common to many fields). Another unique factor encountered when analysing large data sets is the issue of data reduction. Noting the

current misgivings present in the literature about several data reduction techniques performed *a priori* to cluster analysis, a concerted and methodological effort must be made to develop techniques which are able to reduce the number of variables while preserving the underlying, true cluster structure of the data. The question of variable standardization also does not have a clear-cut solution. Given that the groups are unknown, a rich area of research would be to determine effective methods for standardizing variables. The problem of drastically different scales is exacerbated by the nature of the minimum variance objective function giving a disproportionate amount of weight to variables on a wider (and larger) scale. The recent development of methods to determine the number of clusters gives promise that the issue is currently being tackled by the research community. Hopefully, this review will provide a useful reference source for researchers desiring to pursue methodological improvements in the area of *K*-means clustering.

Acknowledgements

I am grateful for the constructive comments of two anonymous reviewers and the Editor, which led to substantial improvements in the paper. The author was partially supported by Office of Naval Research Grant no. 000014-02-1-0877.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Arabie, P., & Hubert, L. (1992). Combinatorial data analysis. *Annual Review of Psychology*, *43*, 169–203.

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160–189). Oxford: Blackwell.

Art, D., Gnanadesikan, R., & Kettenring, J. R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica*, *21A*, 75–99.

Astrahan, M. M. (1970). *Speech analysis by clustering, or the hyperphome method*. Stanford Artificial Intelligence Project Memorandum AIM-124. Stanford, CA: Stanford University.

Bajgier, S. M., & Aggarwal, L. K. (1991). Powers of goodness-of-fit tests in detecting balanced mixed normal distributions. *Educational and Psychological Measurement*, *51*, 253–269.

Ball, G. H., & Hall, D. J. (1965). *ISODATA: A novel method for data analysis and pattern classification*. Menlo Park, CA: Stanford Research Institute.

Ball, G. H., & Hall, D. J. (1967). *PROMENADE – an on-line pattern recognition system*. Research Report RADC-TR-67-310. Menlo Park, CA: Stanford Research Institute.

Banfield, C. F., & Bassill, L. C. (1977). Algorithm AS 113. A transfer algorithm for non-hierarchical classification. *Applied Statistics*, *26*, 206–210.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*, 803–821.

Barker, D. (1976). Hierarchic and non-hierarchic grouping methods: An empirical comparisons of two techniques. *Geografiska Annaler*, *5*, 42–58.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.

Bartko, J. J., Strauss, J. S., & Carpenter, W. T., Jr. (1971). An evaluation of taxometric techniques for psychiatric data. *Classification Society Bulletin*, *2*, 2–28.

Bayne, C. K., Beauchamp, J. J., Begovich, C. L., & Kane, V. E. (1980). Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition*, *12*, 51–62.

Belbin, L. (1987). The use of non-hierarchical allocation methods for clustering large sets of data. *Australian Computer Journal*, *19*, 32–41.

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. *Journal of Machine Learning Research*, *2*, 125–137.

Bezdek, J. C. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, *3*, 58–73.

Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*, 1–8.

Bischof, H., Leonards, A., & Selb, A. (1999). MDL principle for robust vector quantisation. *Pattern Analysis and Applications*, *2*, 59–72.

Blashfield, R. (1977). The equivalence of three statistical packages for performing hierarchical cluster analysis. *Psychometrika*, *42*, 429–431.

Blashfield, R. K., Aldenderfer, M. S., & Morey, L. C. (1982). Validating a cluster analytic solution. In H. Hudson (Ed.), *Classifying social data* (pp. 167–176). San Francisco: Jossey-Bass.

Bock, H. H. (1985). On some significance tests in cluster analysis. *Journal of Classification*, *2*, 77–108.

Bock, H. H. (1996). Probability models and hypothesis testing in partitioning cluster analysis. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 377–453). River Edge, NJ: World Scientific.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for *k*-means clustering. In J. Shavlik (Ed.), *Machine learning*: *Proceedings of the fifteenth International Conference* (pp. 91–99). San Francisco: Morgan Kaufmann.

Brusco, M. J. (2004). Clustering binary data in the presence of masking variables. *Psychological Methods*, *9*, 510–523.

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for *K*-means clustering. *Psychometrika*, *66*, 249–270.

Calinski, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*, 1–27.

Carmignani, M., Genco, A., Lombardo, A., & Tortorici, A. (1988). Quadratic and linear systolic solutions for cluster analysis. In E. Diday (Ed.), *Data analysis and informatics V* (pp. 373–380). Amsterdam: North-Holland.

Carmone, F. J., Kara, A., & Maxwell, S. (1999). HINoV: A new model to improve market segment definition by identifying noisy variables. *Journal of Marketing Research*, *36*, 501–509.

Carroll, J. D., & Chaturvedi, A. (1995). A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In R. D. Luce, M. D'Zmura, D. Hoffman, *et al.* (Eds.), *Geometric representations of perceptual phenomena* (pp. 295–318). Mahwah, NJ: Erlbaum.

Carroll, J. D., & Chaturvedi, A. (1998). *K*-midranges clustering. In A. Rizzi, M. Vichi, & H. H. Bock (Eds.), *Advances in data science and classification* (pp. 3–14). Berlin: Springer.

Carroll, J. D., Green, P. E., & Schaffer, C. M. (1986). Interpoint distance comparisons in correspondence analysis. *Journal of Marketing Research*, *22*, 271–281.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*, 195–212.

Cerioli, A. (1998). A new method for detecting influential observations in nonhierarchical cluster analysis. In A. Rizzi, M. Vichi, & H. H. Bock (Eds.), *Advances in data science and classification* (pp. 15–20). Berlin: Springer.

Cerioli, A., & Zani, S. (2001). Exploratory methods for detecting high density regions in cluster analysis. In S. Borra, R. Rocci, M. Vichi & M. Schader (Eds.), *Advances in classification and data analysis* (pp. 11–18). Berlin: Springer.

Chang, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, *32*, 267–275.

Chaturvedi, A. D., Carroll, J. D., Green, P., & Rotondo, J. A. (1997). A feature based approach to market segmentation via overlapping *K*-centroids clustering. *Journal of Marketing Research*, *34*, 370–377.

Chaturvedi, A. D., Green, P. E., & Carroll, J. D. (2001). *K*-modes clustering. *Journal of Classification*, *18*, 35–55.

Cheng, R., & Milligan, G. W. (1996a). *K*-means clustering methods with influence detection. *Educational and Psychological Measurement*, *56*, 833–838.

Cheng, R., & Milligan, G. W. (1996b). Measuring the influence of individual data points in cluster analysis. *Journal of Classification*, *13*, 315–335.

Cooper, L. (1973). M-dimensional location models: Application to cluster analysis. *Journal of Regional Science*, *13*, 41–54.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society A*, *134*, 321–367.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, *52*, 543–547.

Cuesta-Albertos, J. A., Gordaliza, A., & Matran, C. (1997). Trimmed *K*-means: An attempt to robustify quantizers. *Annals of Statistics*, *25*, 553–576.

Davidson, I. (2002). *Understanding K-means non-hierarchical clustering* (Tech. Rep. 02-2). Albany: State University of New York.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*, 224–227.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, *56*, 463–474.

De Backer, S., & Scheunders, P. (1999). A competitive elliptical clustering algorithm. *Pattern Recognition Letters*, *20*, 1141–1147.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering. *Quality and Quantity*, *20*, 169–180.

De Soete, G. (1988). OVWTRE: A program for optimal variable weighting for ultrametric and additive tree fitting. *Journal of Classification*, *5*, 101–104.

De Soete, G., & Carroll, J. D. (1994). *K*-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, *et al.* (Eds.), *New approaches in classification and data analysis* (pp. 212–219). Berlin: Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the E-M algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

DeSarbo, W. S., Carroll, J. D., Clark, L. A., & Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika*, *49*, 57–78.

DeSarbo, W. S., & Mahajan, V. (1984). Constrained classification: The use of a priori information in cluster analysis. *Psychometrika*, *49*, 187–215.

Dick, N. P., & Bowden, D. C. (1973). Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics*, *29*, 781–790.

Diehr, G. (1973, April). *Minimum variance partitions and mathematical programming*. Paper presented at the National Meetings of The Classification Society, Atlanta, Georgia.

Dillon, W. R., Mulani, N., & Frederick, D. G. (1989). On the use of component scores in the presence of group structure. *Journal of Consumer Research*, *16*, 106–112.

Dimitriadou, E., Dolničar, S., & Weingessel, A. (2002). An examination of indices for determining the number of clusters in binary data sets. *Psychometrika*, *67*, 137–160.

Dodge, Y., & Gafner, T. (1994). Complexity relaxation of dynamic programming for cluster analysis. In E. Diday, Y. Lechevallier, M. Schader, *et al.* (Eds.), *New approaches in classification and data analysis* (pp. 220–227). Berlin: Springer.

Donoghue, J. R. (1995). Univariate screening measures for cluster analysis. *Multivariate Behavioral Research*, *30*, 385–427.

Dubes, R., & Jain, A. K. (1976). Clustering techniques: The user's dilemma. *Pattern Recognition*, *8*, 247–260.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern recognition* (2nd ed.). New York: Wiley.

Dunn, J. C. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. *Journal of Cybernetics*, *3*, 32–57.

Edwards, A. W. F., & Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics*, *21*, 362–375.

Elias, P. (1970). Bounds on the performance of optimum quantizers. *IEEE Transactions on Information Theory*, *16*, 172–184.

Engelman, L., & Hartigan, J. A. (1969). Percentage points of a test of clusters. *Journal of the American Statistical Association*, *64*, 1647–1648.

Everitt, B. S. (1977). Cluster analysis. In C. A. O'Muircheartaigh & C. Payne (Eds.), *Exploring data structures* (Vol. 1, pp. 63–88). New York: Wiley.

Everitt, B. S. (1979). Unresolved problems in cluster analysis. *Biometrics*, *35*, 169–181.

Everitt, B. S., Gourlay, A. J., & Kendall, R. E. (1971). An attempt at validation of traditional psychiatric syndromes by cluster analysis. *British Journal of Psychiatry*, *119*, 399–412.

Faber, V. (1994). Clustering and the continuous *K*-means algorithm. *Los Alamos Science*, *22*, 138–144.

Falkenauer, E., & Marchand, A. (2001, June). *Using K-Means? Consider ArrayMiner*. Paper presented at the 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Las Vegas, Nevada.

Fisher, L., & Van Ness, J. W. (1971). Admissible clustering procedures. *Biometrika*, *58*, 91–104.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, *53*, 789–798.

Fletcher, J. M., & Satz, P. (1985). Cluster analysis and the search for learning disabilities subtypes. In B. P. Rourke (Ed.), *Neuropsychology of learning disabilities: Essentials of subtypes analysis* (pp. 40–64). New York: Guilford.

Flury, B. A. (1990). Principal points. *Biometrika*, *77*, 33–41.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, *21*, 768–769.

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification*, *5*, 205–228.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–631.

Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, *62*, 1159–1178.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, *82*, 249–266.

Gaenssler, P. (1988). On a modification of the *K*-means clustering procedure. In E. Diday (Ed.), *Data analysis and informatics V* (pp. 365–371). Amsterdam: North-Holland.

Garcia-Escudero, L. A., & Gordaliza, A. (1999). Robustness of properties of *K*-means and trimmed *K*-means. *Journal of the American Statistical Association*, *94*, 956–969.

Garcia-Escudero, L. A., Gordaliza, A., & Matran, C. (1999a). Asymptotics for trimmed *K*-means and associated tolerance zones. *Journal of Statistical Planning and Inference*, *77*, 247–262.

Garcia-Escudero, L. A., Gordaliza, A., & Matran, C. (1999b). A central limit theorem for multivariate generalized trimmed *K*-means. *Annals of Statistics*, *27*, 1061–1079.

Gentle, J. E. (2002). *Elements of computational statistics*. New York: Springer.

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Boston: Kluwer Academic.

Gierl, H., & Schwanenberg, S. (1998). A comparison of traditional segmentation methods with segmentation based upon artificial neural networks by means of conjoint data from a Monte

Carlo simulation. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 386–392). Berlin: Springer.

Gitman, I., & Levine, M. D. (1970). An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique. *IEEE Transactions on Computers*, *19*, 583–593.

Gnanadesikan, R., Harvey, J. W., & Kettenring, J. R. (1993). Mahalanobis metrics for cluster analysis. *Sankhyā A*, *55*, 494–505.

Gnanadesikan, R., Kettenring, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, *12*, 113–136.

Gordon, A. D. (1973). Classification in the presence of constraints. *Biometrics*, *29*, 821–827.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gordon, A. D., & Henderson, J. J. (1977). An algorithm for Euclidean sum of squares classification. *Biometrics*, *33*, 355–362.

Gordon, G., & Moore, A. (2000). Learning filaments. In P. Langley (Ed.), *Proceedings of the seventeenth International Conference on Machine Learning* (pp. 335–342). San Franscisco: Morgan Kaufman.

Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, *40*, 33–51.

Green, P. E., Carmone, F. J., & Kim, J. (1990). A preliminary study of optimal variable weighting in *K*-means clustering. *Journal of Classification*, *7*, 271–285.

Green, P. E., & Krieger, A. M. (1995). Alternative approaches to cluster-based market segmentation. *Journal of the Market Research Society*, *37*, 221–239.

Hajnal, I., & Loosveldt, G. (2000). The effects of initial values and the covariance structure on the recovery of some clustering methods. In H. A. L. Kiers, J. -P. Rasson, P. J. F. Groenen, & M. Schader (Eds.), *Data analysis, classification, and related methods* (pp. 47–52). Berlin: Springer.

Hamerly, G., & Elkan, C. (2002). *Learning the K in K-means* (Tech. Rep. CS2002-0716). La Jolla, CA: University of California at San Diego.

Hansen, P., Jaumard, B., & Mladenovic, N. (1998). Minimum sum of squares clustering in a low dimensional space. *Journal of Classification*, *15*, 37–55.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria. *Annals of Statistics*, *6*, 117–131.

Hartigan, J. A. (1985). Statistical theory in clustering. *Journal of Classification*, *2*, 63–76.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A *k*-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Huang, Z. (1998). Extensions to the *K*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*, 283–304.

Huang, Z., & Ng, M. K. (2003). A note on *K*-modes clustering. *Journal of Classification*, *20*, 257–261.

Heiser, W. J., & Groenen, P. J. F. (1997). Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika*, *62*, 63–83.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Hubert, L. J. (2002, June). *John van Ryzin's life and work*. Invited talk presented at the North American Classification Society, Madison, WI.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Hubert, L. J., Arabie, P., & Meulman, J. (2001). *Combinatorial data analysis: Optimization by dynamic programming*. Philadelphia: SIAM.

Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, *83*, 1072–1080.

Huberty, C. J., DiStefano, C., & Kamphaus, R. W. (1997). Behavioral clustering of school children. *Multivariate Behavioral Research*, *32*, 105–134.

Jain, A. K., & Moreau, J. V. (1987). Bootstrap technique in cluster analysis. *Pattern Recognition*, *20*, 547–568.

Jensen, R. E. (1969). A dynamic programming algorithm for cluster analysis. *Operations Research*, *17*, 1034–1057.

Jhun, M., & Jin, S. (2000). On a modified K-spatial medians clustering. *Journal of the Korean Statistical Society*, *29*, 247–260.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient *K*-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 881–892.

Kaufman, L., & Rousseeuw, P. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical data analysis based on the L_1-norm and related methods* (pp. 405–416). Amsterdam: Elsevier.

Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*, 928–934.

Kiers, H. A. L. (1997). Discrimination by means of components that are orthogonal in the data space. *Journal of Chemometrics*, *11*, 533–545.

Körkel, M. (1986). Clustering algorithms for the within-class scatter criterion with a restricted number of elements per cluster. In W. Gaul & M. Schader (Eds.), *Classification as a tool for research* (pp. 241–247). Amsterdam: North-Holland.

Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, *44*, 23–34.

Krzanowski, W. J., & Marriott, F. H. C. (1995). *Multivariate analysis II: Classification, covariance structures and repeated measures*. London: Arnold.

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Brooks/Cole.

Likas, A., Vlassis, N., & Verbeek, J. (2003). The global *K*-means clustering algorithm. *Pattern Recognition*, *36*, 451–461.

MacQueen, J. (1967). Some methods of classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Berkeley, CA: University of California Press.

Maitra, R. (2001). Clustering massive datasets with applications in software metrics and tomography. *Technometrics*, *43*, 336–346.

Makarenkov, V., & Legendre, P. (2001). Optimal variable weighting for ultrametric and additive trees and *K*-means partitioning: Methods and software. *Journal of Classification*, *18*, 245–271.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519–530.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. *Sankhyā B*, *36*, 115–128.

Mardia, K. V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test. *Applied Statistics*, *24*, 163–171.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press.

Maronna, R., & Jacovkis, P. M. (1974). Multivariate clustering procedures with variable metrics. *Biometrics*, *30*, 499–505.

Marriott, F. H. C. (1971). Practical problems in a method of cluster analysis. *Biometrics*, *27*, 501–514.

Massart, D. L., Plastria, F., & Kaufman, L. (1983). Non-hierarchical clustering with MASLOC. *Pattern Recognition*, *16*, 507–516.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

McRae, D. J. (1971). MIKCA: A FORTRAN IV iterative *K*-means cluster analysis program. *Behavioral Science*, *16*, 423–424.

Mezzich, J. E. (1978). Evaluating clustering methods for psychiatric diagnosis. *Biological Psychiatry*, *13*, 265–281.

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

Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, *46*, 187–199.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika*, *50*, 123–127.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analysis. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 341–375). River Edge, NJ: World Scientific.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*, 159–179.

Milligan, G. W., & Cooper, M. C. (1987). Methodological review: Clustering methods. *Applied Psychological Measurement*, *11*, 329–354.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*, 181–204.

Milligan, G. W., & Sokol, L. M. (1980). A two-stage clustering algorithm with robust recovery characteristics. *Educational and Psychological Measurement*, *40*, 755–759.

Mirkin, B. G. (1990). A sequential fitting procedure for linear data analysis models. *Journal of Classification*, *7*, 167–195.

Mirkin, B. G. (1996). *Mathematical classification and clustering*. Dordrecht: Kluwer.

Mirkin, B. G. (1998). Mathematical classification and clustering: From how to what and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 172–181). Berlin: Springer.

Mizuta, M. (1998). Two principal points of symmetric distributions. In A. Rizzi, M. Vichi, & H. H. Bock (Eds.), *Advances in data science and classification* (pp. 171–176). Berlin: Springer-Verlag.

Morrison, D. F. (1976). *Multivariate statistical methods* (2nd ed). New York: McGraw-Hill.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In J. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th international conference on very large databases* (pp. 144–155). San Francisco, CA: Morgan Kaufmann.

Pelleg, D., & Moore, A. (1999). Accelerating exact *K*-means algorithms with geometric reasoning. In S. Chaudhuri & D. Madigan (Eds.), *Proceedings of the fifth international conference on knowledge discovery in databases* (pp. 277–281). Menlo Park, CA: AAAI Press.

Pelleg, D., & Moore, A. (2000). *X*-means: Extending *K*-means with efficient estimation of the number of clusters. In *Proceedings of the seventeenth international conference on machine learning* (pp. 727–734). San Francisco: Morgan Kaufmann.

Pollard, D. (1981). Strong consistency of *K*-means clustering. *Annals of Statistics*, *9*, 135–140.

Pollard, D. (1982). A central limit theorem for *K*-means clustering. *Annals of Probability*, *10*, 919–926.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes, the art of scientific computing*. Cambridge: Cambridge University Press.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, *20*, 134–148.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Association B*, *10*, 159–193.

Rao, M. R. (1971). Cluster analysis and mathematical programming. *Journal of the American Statistical Association*, *66*, 622–626.

Ray, S., & Turi, R. H. (2000). Determination of the number of clusters in *K*-means clustering and application in colour image segmentation. In N. R. Pal, A. K. De & J. Das (Eds.), *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques* (pp. 137–143). New Delhi: Narosa.

SAS. (2004). The FASTCLUS procedure. In *SAS/STAT 9.1 user's guide, Volume 2*. Cary, NC: SAS Institute.

Schaffer, C. M., & Green, P. E. (1996). An empirical comparison of variable standardization methods in cluster analysis. *Multivariate Behavioral Research*, *31*, 149–167.

Schaffer, C. M., & Green, P. E. (1998). Cluster-based market segmentation: Some further comparisons of alternative approaches. *Journal of the Market Research Society*, *40*, 155–163.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Scott, A. J., & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, *27*, 387–398.

Sebestyen, G. S. (1962). *Decision making processes in pattern recognition*. New York: Macmillan.

Selim, S. Z., & Ismail, M. A. (1984). *K*-means type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 81–87.

Sneath, P. H. A. (1957). The application of computers in taxonomy. *Journal of General Microbiology*, *17*, 201–226.

Späth, H. (1980). *Cluster analysis algorithms for data reduction and classification of objects*. New York: Wiley.

Späth, H. (1985). *Cluster dissection and analysis: Theory, FORTRAN programs, examples*. New York: Wiley.

Späth, H. (1986a). Anti-clustering: Maximizing the variance criterion. *Control and Cybernetics*, *15*, 213–218.

Späth, H. (1986b). Maximizing partitioning cluster criteria for quantitative data. *Studien zur Klassifikation*, *17*, 221–228.

SPSS (2003). *SPSS 12.0 command syntax reference*. Chicago: SPSS.

Steinley, D. (2003). *K*-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*, 294–304.

Steinley, D. (2004a). Standardizing variables in *K*-means clustering. In D. Banks, L. House, F. R. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering, and data mining applications* (pp. 53–60). New York: Springer.

Steinley, D. (2004b). Properties of the Hubert–Arabie adjusted Rand index. *Psychological Methods*, *9*, 386–396.

Stoddard, A. M. (1979). Standardization of measures prior to cluster analysis. *Biometrics*, *35*, 765–773.

Stute, W., & Zhu, L. X. (1995). Asymptotics of *K*-means clustering based on projection pursuit. *Sankhyā*, *57*, 462–471.

Su, M.-C., & Chou, C.-H. (2001). *A modified version of the K-means algorithm with a distance based on cluster symmetry*. Paper presented at the IEEE Conference on Computing.

Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics*, *37*, 35–43.

Tarpey, T., Li, L., & Flury, B. D. (1995). Principal points and self-consistent points of elliptical distributions. *Annals of Statistics*, *23*, 103–112.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*, 267–276.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B*, *63*, 411–423.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.

Van Buuren, J., & Heiser, W. J. (1989). Clustering *N* objects into *K* groups under optimal scaling of variables. *Psychometrika*, *54*, 699–706.

Van Ness, J. W. (1973). Admissible clustering procedures II. *Biometrika*, *60*, 422–424.

van Os, B. J. (2000). *Dynamic programming for partitioning in multivariate data analysis*. Leiden: Leiden University Press.

Vesanto, J. (2001). Importance of individual variables in the *k*-means algorithm. In D. Cheung, G. J. Williams, & Q. Li (Eds.), *Proceedings of the Pacific–Asia conference in knowledge discovery and data mining* (pp. 513–518). Berlin: Springer.

Vichi, M., & Kiers, H. A. L. (2001). Factorial *K*-means analysis for two-way data. *Computational Statistics and Data Analysis*, *37*, 49–64.

Vinod, H. D. (1969). Integer programming and the theory of groups. *Journal of the American Statistical Association*, *64*, 506–519.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained *K*-means clustering with background knowledge. In *Proceedings of the eighteenth international conference on machine learning* (pp. 577–584). San Francisco, CA: Morgan Kaufmann.

Waller, N. G., Kaiser, H. A., Illian, J. B., & Manry, M. (1998). A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. *Psychometrika*, *63*, 5–22.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Weisstein, E. W. (2003). *CRC concise encyclopedia of mathematics*. Boca Raton, FL: Chapman & Hall/CRC.

Windham, M. P. (1987). Parameter modification for clustering criteria. *Journal of Classification*, *4*, 191–214.

Wishart, D. (1969). *FORTRAN II programs for 8 methods of cluster analysis (CLUSTAN I)*. Computing Contributions, 38th State Geological Survey. Lawrence, KS: University of Kansas.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, *5*, 329–350.

Wong, M. A. (1982a). Asymptotic properties of univariate population *K*-means clusters. *Classification Society Bulletin*, *5*, 44–50.

Wong, M. A. (1982b). A hybrid clustering algorithm for identifying high density clusters. *Journal of the American Statistical Association*, *77*, 841–847.

Wong, M. A. (1984). Asymptotic properties of univariate sample *K*-means clusters. *Journal of Classification*, *1*, 255–270.

Wong, M. A. (1985). A bootstrap testing procedure for investigating the number of subpopulations. *Journal of Statistical Computation and Simulation*, *22*, 99–112.

Wong, M. A., & Lane, T. (1983). A *k*th nearest neighbor clustering procedure. *Journal of the Royal Statistical Society B*, *45*, 362–368.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.

Zhang, B. (2000). Generalized *K*-harmonic means – boosting in unsupervised learning. *Hewlett Packard Technical Report, HPL-2000-137*.

Zhang, B., Hsu, M., & Dayal, U. (1999). *K-harmonic means – a data clustering algorithm* (Hewlett Packard Technical Report, HPL-1999-124). Palo Alto, CA: Hewlett Packard Laboratories.

Received 23 September 2004; revised version received 11 April 2005