PSYCHOMETRIKA—VOL. 72, NO. 1, 93–106
MARCH 2007
DOI: 10.1007/s11336-003-1272-1

VALIDATING CLUSTERS WITH THE LOWER BOUND FOR SUM-OF-SQUARES ERROR

DOUGLAS STEINLEY

UNIVERSITY OF MISSOURI-COLUMBIA

Given that a minor condition holds (e.g., the number of variables is greater than the number of clusters), a nontrivial lower bound for the sum-of-squares error criterion in K-means clustering is derived. By calculating the lower bound for several different situations, a method is developed to determine the adequacy of cluster solution based on the observed sum-of-squares error as compared to the minimum sum-of-squares error.

Key words: K-means, cluster analysis.

1. Introduction

The classification of objects is of great interest in many fields, including psychology. 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 nonempty subsets can be calculated with a Stirling number of the second kind (e.g., 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 $K^N/K!$ (e.g., see Kaufman & Rousseeuw, 1990, p. 115). Thus, for example, if 25 objects are to be grouped into four 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 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 the partitions should be externally isolated and internally cohesive, implying a certain degree of homogeneity within partitions and heterogeneity between partitions (referred to by Cattell & Coulter (1966) as *homostats*). Historically, many researchers attempted to operationalize this definition by minimizing within-group variation (Cox, 1957; Fisher, 1958; Thorndike, 1953; Engelman & Hartigan, 1969). Following these early attempts of maximizing within-group homogeneity, MacQueen (1967) 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 multivariate (Johnson & Wichern, 2002, pp. 695–700; Timm, 2002, pp. 530–531; Lattin, Carroll, & Green, 2003, pp. 288–297), cluster analysis (Anderberg, 1973, pp. 162–163; Gordon, 1999, pp. 41–49; Hartigan, 1975, pp. 80–112) and pattern recognition (Duda, Hart, & Stork, 2001, pp. 526–528) textbooks.

The author was partially supported by the Office of Naval Research Grant #N00014-06-0106.

Requests for reprints should be sent to Douglas Steinley, Department of Psychological Sciences, University of Missouri-Columbia, 210 McAlester Hall, Columbia, MO 65211, USA. E-mail: steinleyd@missouri.edu.

2. The K-Means Method

2.1. Algebraic Representation

The K-means method is designed to partition two-way, two-mode data (i.e., 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

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

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 $1 \le k \le K)$.
2. Based on the initial seeds, the squared Euclidean distance $(d^2(i,k))$ between the ith object and the kth seed vector is obtained:

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

   Objects are allocated to clusters with the minimum squared Euclidean distance to its defining seed.
3. Once all objects have been initially allocated, cluster centroids are calculated as in (3) and replace the initial seeds.
4. Objects are compared to each centroid (using $d^2(i,k)$) and allocated to the cluster whose centroid is closest.
5. New centroids are calculated with the updated cluster membership (by calculating the centroids after all objects have been assigned—the method is not affected by the sequence of the data units (Anderberg, 1973, p. 162)).
6. Steps 4 and 5 are repeated until no objects can be reallocated to different 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):

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

Späth (1980, p. 72) notes that at times, but probably rarely in practice, the 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 ith object from C_k to cluster C_{k^*}, and the SSE is reduced (see Späth, 1980, p. 72). Furthermore, it is important to note that each object only contributes to the centroid of the cluster to which the object currently belongs. A common, misguided practice is to allow every object to contribute to every cluster when the $d^2(i, k)$ are computed, leading to wildly divergent results that are often difficult to interpret—not to mention the occasional nonconvergent implementation of the K-means algorithm.

As noted in Steinley (in press), the K-means algorithm is subject to locally optimal solutions and it has been recommended that several thousand random initializations are used, choosing the solution with the lowest associated value of SSE as the final solution (see Steinley, 2003). However, a notable exception to this has been illustrated in the literature. The global optimum can be achieved for very large object sets if it is assumed that the N objects are ordering along a continuum (see Späth, 1980, pp. 61–64). However, for the purposes of the current research, there is no assumption of an order constraint.

2.2. *Matrix Representation*

This section reviews several different techniques for representing (5) using matrices, resulting in a combination of representations that leads to the lower bound of (5) given K. The K-means iterative relocation algorithm described above can be formulated using $\mathbf{X}$ and two additional matrices:

(a). an $N \times K$ membership matrix, $\mathbf{M} = \{m_{ik}\}$, where entry m_{ik} equals unity if object i belongs to cluster k; zero otherwise; and

(b). a $K \times P$ cluster representation matrix, $\mathbf{R} = \{r_{kp}\}$, where $\mathbf{R}$ can be represented as a stack of row vectors

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

where each row, $\mathbf{r}_k'$, is a 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 an alternating least squares algorithm procedure that alternates between minimizing (7) with respect to $\mathbf{M}$ given the current estimate of $\mathbf{R}$ and minimizing (7) with respect to $\mathbf{R}$ given the current cluster membership. Furthermore, by expanding (7) the middle two terms cancel, and it is seen that $\mathbf{R}$ is equal to $(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X}$, obtaining the formulation in Gordon and Henderson (1977),

$$SSE = \text{tr}(\mathbf{X}'(\mathbf{I} - \mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}')\mathbf{X}), \tag{8}$$

which allows the problem to be viewed as trying to find a particular projection of the columns of $\mathbf{X}$; however, to date, the conceptual direction provided by (8) has been mostly ignored.

2.3. Combining Existing Representations to Derive the Lower Bound

By combining this representation with several of those presented above, it will be possible to derive a nontrivial lower bound, under certain conditions, for (5). Representing $\mathbf{X}$ as a stack of row vectors and partitioning them into their respective clusters,

$$\mathbf{X} = \left[\begin{array}{c} \mathbf{X_1} = \begin{matrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_{n1}' \end{matrix} \\ \hline \mathbf{X_2} = \begin{matrix} \mathbf{x}_{n1+1}' \\ \vdots \\ \mathbf{x}_{n2}' \end{matrix} \\ \hline \begin{matrix} \vdots \\ \vdots \end{matrix} \\ \hline \mathbf{X_K} = \begin{matrix} \mathbf{x}_{(n-nk+1)}' \\ \vdots \\ \mathbf{x}_{nK}' \end{matrix} \end{array}\right],$$

allows the collection of observations for the kth cluster to be represented as the submatrix denoted $\mathbf{X}_k$ of size $n_k \times P$. Recalling the form of (7) and breaking it into K components, allows (5) to be rewritten as

$$SSE = \sum_{k=1}^{K} \text{tr}((\mathbf{X}_k - \mathbf{j}_k\mathbf{r}_k')'(\mathbf{X}_k - \mathbf{j}_k\mathbf{r}_k')), \tag{9}$$

where $\mathbf{j}_k$ is an $n_k \times 1$ vector of ones. Then using the projection notion of (8), (9) can be represented as

$$SSE = \sum_{k=1}^{K} \text{tr}(((\mathbf{I}_k - (\mathbf{j}_k\mathbf{j}_k')/n_k)\mathbf{X}_k)'((\mathbf{I}_k - (\mathbf{j}_k\mathbf{j}_k')/n_k)\mathbf{X}_k)), \tag{10}$$

where $\mathbf{I}_k$ is the $n_k \times n_k$ identity matrix corresponding to the kth cluster. Carrying out the transpose allows us to rewrite (10) as

$$SSE = \sum_{k=1}^{K} \text{tr}(\mathbf{X}_k'(\mathbf{I}_k - (\mathbf{j}_k\mathbf{j}_k')/n_k)'(\mathbf{I}_k - (\mathbf{j}_k\mathbf{j}_k')/n_k)\mathbf{X}_k), \tag{11}$$

and using the properties of traces and projection matrices results in a further reduction of (11) to

$$SSE = \sum_{k=1}^{K} \text{tr}((\mathbf{I}_k - (\mathbf{j}_k\mathbf{j}_k')/n_k)\mathbf{X}_k\mathbf{X}_k'). \tag{12}$$

Expanding the terms in (12), splitting the two unit vectors, and distributing the summation sign reformulates the problem as

$$SSE = \sum_{k=1}^{K} \text{tr}(\mathbf{X}_k'\mathbf{X}_k) - \sum_{k=1}^{K} \text{tr}\left((\mathbf{j}_k'/\sqrt{n_k})\mathbf{X}_k\mathbf{X}_k'(\mathbf{j}_k/\sqrt{n_k})\right). \tag{13}$$

Realizing that $\mathbf{X}'\mathbf{X}$ is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \dots & \mathbf{X}_1'\mathbf{X}_K \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \dots & \mathbf{X}_2'\mathbf{X}_K \\ \vdots & & \ddots & \\ \mathbf{X}_K'\mathbf{X}_1 & \ddots & & \mathbf{X}_K'\mathbf{X}_K \end{bmatrix},$$

leads to the equality

$$\sum_{k=1}^{K} \operatorname{tr}\left(\mathbf{X}_k'\mathbf{X}_k\right) = \operatorname{tr}(\mathbf{X}'\mathbf{X}). \tag{14}$$

Additionally, recalling the form of $\mathbf{M}$ and noting that $\mathbf{M}'\mathbf{M}$ is the $K \times K$ diagonal matrix

$$\mathbf{M}'\mathbf{M} = \begin{bmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_K \end{bmatrix},$$

leads to the realization that an $N \times K$ orthonormal matrix, $\mathbf{O}$, can be formed by dividing the kth column of $\mathbf{M}$ by n_k, allowing (13) to be restated as

$$SSE = \operatorname{tr}(\mathbf{X}'\mathbf{X}) - \operatorname{tr}(\mathbf{O}'\mathbf{X}\mathbf{X}'\mathbf{O}). \tag{15}$$

The first term on the right-hand side of (15) is fixed given the data, allowing the minimization of SSE to be changed to a problem of maximizing the second term on the right-hand side of (15). Given that the data are fixed and letting $\mathcal{O}$ represent the set of all $N \times K$ orthonormal matrices, the problem becomes finding the specific orthonormal matrix, $\mathbf{O}^*$, that maximizes the second term of (15). Using the following theorem of Fan (1949) provides the solution.

Theorem. *Let* $\mathbf{H}$ *be an* $N \times N$ *Hermitian matrix with eigenvalues* $\lambda^{\mathbf{H}} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, *then*

$$\lambda_1 + \cdots + \lambda_K = \max_{\mathcal{O}} \operatorname{tr}(\mathbf{O}'\mathbf{H}\mathbf{O}).$$

Realizing that all symmetric real valued matrices (e.g., $\mathbf{X}\mathbf{X}'$) are special cases of Hermitian matrices, allows the minimum of (15) to be calculated as

$$\psi = \min SSE = \operatorname{tr}(\mathbf{X}'\mathbf{X}) - \sum_{i=1}^{K} \lambda_i^{\mathbf{X}\mathbf{X}'}. \tag{16}$$

If $P \leq K$, ψ will be trivial (i.e., equal to zero), making the derivation in (16) informative only in situations where $K < P$.[1]

It is clear that $\mathbf{M}$ is a subset of $\mathcal{O}$; however, it is likely that $\mathbf{O}^*$ is not of the same form as $\mathbf{M}$, i.e., the values in $\mathbf{O}^*$ are not going to consist of zeros and ones, but rather continuous values. Thus, it may be the case that perfectly clustered data will not achieve the lower bound stated in (16). The following section develops an index to capture specific properties of the lower bound and explores the relationship between the theoretical lower bound and the observed lower bound in a cluster analytic situation.

[1] We are grateful to one of the reviewers for indicating that the same lower bound has previously been derived (see Zha, Ding, Gu, He, & Simon, 2001). The main difference between the two accounts is that the present derivation is based on the historical development of K-means clustering and includes the explicit reference to Fan's (1949) theorem; however, all mathematical results remain the same.

3. An Index Based on the Lower Bound

Although the result provided in (16) is theoretically interesting, in its current form, it lacks direct applicability. Since ψ is fixed given the data, an obvious extension is to measure the distance between the observed sum-of-squares error, ϑ, resulting from a particular clustering of the data. However, given different scales of data sets, it would be impossible to compare the differences across data sets. Thus, a normalized index, ξ, is created by dividing the above difference by the corrected sum-of-squares total (τ), represented by

$$\xi = \frac{\vartheta - \psi}{\tau}. \tag{17}$$

When $\psi = 0$ and $\vartheta = \tau$, ξ has an upper bound of unity, and when $\vartheta = \psi$, a lower bound of zero. Clearly, the closer ξ is to zero, the more compact the clustering of the data. The following section relates ξ to cluster recovery, as measured by Hubert and Arabie's (1985) adjusted Rand index (ARI_{HA}), and develops a test to determine the quality of the clustering solution based on the value of the above index.

4. Assessing the Efficacy of Cluster Solutions Using ξ

To assess the ability of ξ to approximate cluster quality, data were generated from several conditions based on Steinley (2003, 2004a). The seven factors for this study were: (a) the type of overlap among the clusters; (b) the number of clusters; (c) the number of variables; (d) the probability of overlap; (e) the type of distributional family; (f) the size of the data set; and (g) the number of incorrect object assignments. All seven factors are described below. Furthermore, consistent with other studies, three replications were made for each condition (see Brusco & Cradit, 2001; Milligan & Cooper, 1985; Steinley, 2003).

4.1. Factor Description

Type of Overlap: Operationalized by Steinley and Henson (2005), the type of overlap between clusters can be of two different kinds: marginal or joint. Marginal overlap is defined by allowing clusters to overlap on some dimensions (i.e., variables), but not on all dimensions simultaneously. The most familiar cluster generation method using this type of overlap is Milligan's (1985) method where clusters were not allowed to overlap on the first dimension, but were allowed to overlap on all the others. This restriction prevents clusters from overlapping in the joint P-variate space (i.e., on all dimensions simultaneously). Steinley and Henson (2005) relaxed this condition and allowed clusters to overlap on all dimensions, resulting in a more realistic technique for generating clusters.

Number of Clusters: The number of clusters ranged from $K = 4, \ldots, 8$.

Number of Variables: For ψ to have any meaning, the number of variables must be greater than the number of clusters. So, the number of variables ranged from $P = (K + 1), \ldots, 15$.

Probability of Overlap: The probability that two clusters overlapped (either marginal or joint overlap) ranged from $O = 0.10, 0.20, 0.30, 0.40$.

Distributional Family: The clusters were drawn from five different distributional families: normal with equal variances, normal with unequal variances, triangular distributions, uniform distributions, and a mixed distribution. The normal with equal variances generates variables with a covariance matrix proportional to the identity. The normal with unequal variances initially generates data with a diagonal covariance matrix (with no restrictions on the values of the variances), but through arbitrary rotations different correlation structures are achieved.

TABLE 1.
$\mathcal{T}$ for Co-occurrence between two partitions, $\mathcal{P}$ and $\mathcal{Q}$.

		$\mathcal{Q}$				
	Group	q_1	q_2	...	q_C	Totals
	p_1	t_{11}	t_{12}	...	t_{1C}	t_{1+}
	p_2	t_{21}	t_{22}	...	t_{2C}	t_{2+}
$\mathcal{P}$	⋮	⋮	⋮		⋮	⋮
	p_R	t_{R1}	t_{R2}	...	t_{RC}	t_{R+}
	Totals	t_{+1}	t_{+2}	...	t_{+C}	$t_{++} = N$

The triangular distribution generates data from a discrete triangular distribution that results in skewed data. The uniform distribution generates data from a continuous uniform distribution, while the mixed condition randomly selects one of the four aforementioned distributions for each of the variables—resulting in data sets where each variable may rise from a different distribution (see Steinley & Henson (2005) for details concerning the specifics of the data generation process).

Size: The size of the data sets took on the values of $N = 100, 200, 300, 400$, and 500.

Number of Incorrect Object Assignments: To guarantee a wide range of Hubert–Arabie adjusted Rand indices, the values were directly manipulated in the following manner. Assume $\mathcal{P}$ is the true partition of the objects in the generated data. Now, assume $\mathcal{Q}$ is the partitioning of the objects obtained by a clustering algorithm. Then, the cross-classifications of $\mathcal{P}$ and $\mathcal{Q}$ can be represented by the contingency table, $\mathcal{T}$ (see Table 1), where $\mathcal{P}$ contains R classes and $\mathcal{Q}$ contains C classes (for the purposes of this study, it is assumed $R = C$). In $\mathcal{T}$, a generic entry, t_{rc}, represents the number of objects that were classified in the rth class of partition R *and* the cth class of partition C. Perfect agreement between $\mathcal{P}$ and $\mathcal{Q}$ (i.e., $ARI_{HA} = 1.00$) occurs when all entries are on the main diagonal of $\mathcal{T}$. To obtain values of ARI_{HA} less than unity, it is sufficient to move objects from the main diagonal into the body of $\mathcal{T}$. See Steinley (2004b) for a detailed discussion of the manipulation of the partitions and the properties of the ARI_{HA}. For this factor, there were six levels: (a) perfect agreement; and (b) when 5%, 15%, 25%, 35%, and 45% of the objects were randomly assigned to the incorrect partition. This manner of manipulating the data allows the results to be generalizable to several clustering techniques (i.e., the results below are only data dependent and not method dependent).

All factors were completely crossed; however, due to the fact that the value of P for each condition depends on the value of K, the design is not balanced. Thus, instead of 247,500 observations (in this case data sets) there are 202,500 data sets. For each of the data sets, after the objects were randomly assigned to the wrong partition, the solution was compared with the known cluster solution and the ARI_{HA}, ψ, and ξ were computed.

4.2. Results

If ARI_{HA} values are treated as a response variable and the manipulated factors as explanatory variables, a seven-way ANOVA can be conducted (see Brusco & Cradit, 2001; Milligan & Cooper, 1988; Steinley, 2004a; for similar designs). The results for the main effects of the ANOVA are displayed in Table 2.

Consistent with results provided by Steinley (2004b), none of the data dependent manipulated factors had large effects on ARI_{HA} ($\eta^2 < .001$ for all factors). Thus, ARI_{HA} is invariant to properties of the data set and is sensitive only to partition agreement ($\eta^2 = 0.995$).

TABLE 2.
ANOVA with ARI_{HA} as the response.

Source	DF	SS	MS	F	η^2
Type of overlap	1	*	*	0.08	*
#Variables	10	*	*	0.21	*
Prob. overlap	4	*	*	0.18	*
Distribution	4	*	*	0.05	*
Data set size	4	0.09	0.02	143.49	*
#Clusters	4	52.04	13.01	87903.2	*
%Incorrect	5	16550.34	3310.07	2.236E7	0.995
Total	202499	16639.64			

* $\leq .001$.

A similar analysis can be conducted with ξ as the response variable (results presented in Table 3). Similar to ARI_{HA}, the data dependent factors do not have large effects on ξ; whereas the degree of partition agreement has an extremely large effect on the response variable ($\eta^2 = 0.90$). Given that the variances of the two indices (ARI_{HA} and ξ) are almost completely explained by the same variable, it is reasonable to assume that, in a research setting, the unobservable value of ARI_{HA} may be well approximated by the cluster dependent value of ξ.

5. Validating Cluster Structure

Given that, when desired, most clustering procedures will always provide a clustering of the data (whether appropriate or not), this section provides two methods for determining the validity of a cluster solution provided by a K-means cluster analysis. First, a heuristic method based on simple linear regression is given so the researcher can quickly assess the appropriateness of a given partitioning of the data. Second, a more rigorous procedure based on generating appropriate reference distributions is provided. The latter procedure allows for a straightforward method of significance testing for cluster validity.

5.1. Heuristic Method

Given the primary influence of the percentage of observations assigned to the incorrect partition on both ARI_{HA} and ξ, it is instructive to investigate the change in the average value

TABLE 3.
ANOVA with ξ as the response.

Source	DF	SS	MS	F	η^2
Type of overlap	1	64.51	64.51	15083.6	*
#Variables	10	56.12	5.61	1312.28	*
Prob. overlap	4	45.34	11.33	2650.39	*
Distribution	4	111.29	27.82	6505.45	*
Data set size	4	5.08	1.27	297.24	*
#Clusters	4	25.69	6.42	1502.03	*
%Incorrect	5	10816.84	2163.36	505803	0.90
Total	202499	12025.82			

* $\leq .001$.

TABLE 4.
Mean of ARI_{HA} and ξ with respect to incorrect partition assignment.

%Incorrect	Mean ARI_{HA}	Mean ξ
0	1.00	0.13
5	0.88	0.22
15	0.65	0.40
25	0.47	0.55
35	0.31	0.67
45	0.19	0.77

for these indices across the six levels of this factor (see Table 4 for details). It is clear that the ARI_{HA} and ξ are inversely related, as the value of ARI_{HA} decreases as the value of ξ increases ($r = -.95$)

Predicting ARI_{HA} from ξ using a simple linear regression provides

$$ARI_{HA} = \beta_0 + \beta_1 \xi, \tag{18}$$

and when (18) is fit to the data, the estimated equation is

$$\widehat{ARI_{HA}} = 1.09 - 1.12\xi \tag{19}$$

as the predictive equation. The model exhibits an adjusted $R^2 = 0.90$, a respectable value for predictive purposes. Using the heuristic values of ARI_{HA} provided by Steinley (2004b), we assume that values less than 0.65 represent poor cluster recovery. Substituting 0.65 into the left-hand side of (19) and solving for ξ yields a value of $\xi = 0.40$. Now, we can use $\xi = 0.40$ as a threshold value to decide whether or not to accept a given partitioning of the data. Dividing the generated data into two groups, $\mathcal{G}_1$ if $\xi < 0.40$ (i.e., acceptable partitions) and $\mathcal{G}_2$ if $\xi \geq 0.40$ (i.e., unacceptable partitions), the power of the decision rule can be tested. Table 5 provides the descriptive statistics for the two groups.

The difference between the average ARI_{HA} of the two groups is 0.49, and when considering the pooled standard deviation (0.15), the effect size (Cohen's d) is 3.27—clearly an enormous value for d. Given the relative nonimportance of data specific characteristics and the range of data sets in the simulation, the value of 0.40 is generalizable to several data analytic situations. At the very least, by using (19), the researcher is able to take advantage of external knowledge of the data set to get a rough idea of how well the resulting partition approximates the true cluster structure.

5.2. *Nonparametric Method*

Although the heuristic method above can provide a quick, rough idea of the quality of cluster solution, it may be desirable to implement a more rigorous procedure. Similar to Steinley (2004b) and Tibshirani, Walther, and Hastie (2001), it is possible to generate an underlying reference

TABLE 5.
Descriptive statistics for $\mathcal{G}_1$ and $\mathcal{G}_2$.

Group	N	Mean ARI_{HA}	Std
$\mathcal{G}_1$	87,215	.87	0.13
$\mathcal{G}_2$	115,285	.38	0.16

distribution, with the desired properties controlled for, to compare to the observed data set. Since both ARI_{HA} and ξ primarily depend on the percent of incorrect assignments, it is possible to fix several properties of the generated data to reflect the characteristics of the observed data.

Assume the data matrix $\mathbf{X}_{N\times P} = \{x_{ij}\}$ is observed, where N represents the number of observations and P represents the number of variables. Then, after a cluster analysis is performed on $\mathbf{X}$ to obtain K clusters, a corresponding value of ξ, based on the resulting partition ($\mathcal{P}_{\mathbf{X}}$) when the data set is clustered via the K-means procedure, can be computed (denoted as $\xi_{\mathbf{X}}$). The primary goal is to determine whether the $\mathcal{P}_{\mathbf{X}}$ is representative of a "true," underlying cluster structure or merely an artifact of the clustering algorithm, resulting in the hypothesis:

H_o : The observed $\mathcal{P}_{\mathbf{X}}$ adequately fit the data (i.e., it reflects a true, underlying cluster structure).

with the corresponding alternative hypothesis:

H_a : The observed $\mathcal{P}_{\mathbf{X}}$ does not adequately fit the data.

It is assumed that "fit" using cluster structure is measured in terms of the ARI_{HA}. Thus, if the ARI_{HA} is low (i.e., a clustering technique is unable to recover the true cluster structure), then the fit of the partition to the data is poor; whereas, if the ARI_{HA} is high and a clustering technique is able to recover the true cluster structure the fit of the partition to the data is good. Following Steinley (2004b), the adequate fit cut-off used will be 0.65; however, it is important to realize that any value of ARI_{HA} can be substituted into the following procedure. To test H_o, the following algorithm is implemented (afterward, each step is described in detail):

1. Set $A = 0$, $B = 0$; define $\xi_{\mathbf{X}}$ as the *observed* value of ξ, α as the desired significance level, B_s as the number of samples to be drawn from the baseline distribution.
2. Generate a data set with K clusters, N observations, and P variables from an appropriate underlying distribution. In general, before the cluster analysis is performed, the researcher is not aware of the underlying distribution or the appropriate cluster structure of the data. Since these factors do not have an appreciable impact on either ARI_{HA} or ξ, it is sufficient to choose the underlying distribution, the type of cluster overlap, and the degree of cluster overlap randomly. Thus, whatever small effect these unobservable factors (predata analysis) have on the two outcome measures will be accounted for by the randomization method (see the description below for alternative data set generation suggestions).
3. After the data set has been generated, cluster the data set via the K-means procedure and calculate ξ_B^* and $ARI_{HA}^{(B)}$ for this data set. If $\xi_B^* > \xi_{\mathbf{X}}$, then $A = A + 1$.
4. If $B < B_s$, then $B = B + 1$ and return to Step 2.
5. If $A/B_s < \alpha$, reject H_o.

5.2.1. Step 1. This steps only requires two user-defined values, α and B_s. Standard values for α can be employed (say .05 or .01), while larger values of B_s will lead to more accurate results due to reduction of sampling variability. B_s should be no smaller than 100, while sizes of 1000 or more will lead to much more reliable results.

5.2.2. Step 2. The data generation procedure described in the algorithm above allows for the most general reference distribution. However, the data generation process is completely flexible and should be tailored to the specific needs of the researcher. For instance, when generating classical clusters where there is always at least one dimension that clearly defines the cluster structures, then Milligan's (1985) generation procedure could be implemented. On the other

hand, if clusters with a specific skew or kurtosis were desired, then the use of the procedure developed by Waller, Underhill, and Kaiser (1999) would be appropriate.

Furthermore, any preconceived notions about the size and shape of the true cluster structure can be reflected in the reference distribution by either restricting the size of the generated clusters or the distribution from which the variables defining the cluster structure are drawn. In short, any generation procedure that has been used previously can be substituted in this step. Thus, the user is only limited by his/her imagination or creativity, allowing the nonparametric testing procedure to be extremely flexible.

5.2.3. Steps 3–5. This procedure will compute the area to the right of $\xi_{\mathbf{X}}$ in the distribution of ξ^*. If this area is less than α (indicating $\xi_{\mathbf{X}}$ is in the upper-tail of the ξ^* distribution), the test indicates poor model fit and lack of true cluster structure recovery by the K-means method.

Alternatively, the test can be conducted in terms of a cut-off value. If the values of ξ^* are ordered from smallest to largest $(\xi^*_{(1)}, \xi^*_{(2)}, \ldots, \xi^*_{(B_s)})$, then the value $\xi^*_{(\alpha B_s)}$ serves as cut-off value. If $\xi_{\mathbf{X}} > \xi^*_{(\alpha B_s)}$, the null hypothesis is rejected. (*Note*: This technique can be considered analogous to model testing in the exploratory factory analysis setting where a chi-squared test statistic is used.)

5.3. Illustrative Example

This section applies the above methodology to two example data sets. The first is the Fisher (1936) iris data. The sepal length, sepal width, petal length, and petal width are measured in millimeters on 50 iris specimens from each of three species, *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Thus, the final data matrix for the iris data contains 150 observations measured on four variables. Using the true classes, as defined by the species, the computed index is $\xi_{\text{iris}} = 0.13$; however, when the K-means procedure is used to cluster the iris data, the observed index is $\xi_{\text{iris}(K\text{-means})} = .08$ while the degree of cluster recovery is $ARI_{HA}^{(\text{iris})} = 0.71$, indicating good recovery by Steinley's (2004b) guidelines.

The second example data set is a fabricated data set containing an artificial cluster structure. The fabricated data set is generated from a multivariate normal distribution with four variables and 150 observations; however, there is no "true" underlying cluster structure present in the data. The data are then clustered into four clusters, via the K-means procedure, resulting in a computed index of $\xi_{\text{fab}} = 0.51$, and since there is no underlying cluster structure, a value for ARI_{HA} is not defined.

Since each data set has the same number of objects and variables, it is sufficient to generate a single reference distribution for both scenarios. In this example, $\alpha = .05$ and $B_s = 1000$. The reference distribution was generated in the manner described within the algorithm. Specifically, for each of the 1000 data sets, the following decisions were made:

(a) The underlying reference distribution was randomly chosen from the set: {normal with equal variances, normal with unequal variances, uniform, triangular, mixed}, where each distribution had an equal probability of being selected (i.e., each had a probability of .20 of being selected).
(b) The type of cluster overlap (i.e., marginal or joint) was also chosen randomly, with each having a .5 probability of being selected for a given data set.
(c) The amount of cluster overlap for each data set was chosen from a random continuous uniform distribution $\mathcal{U}(0, .45)$.

Thus, for each of the 1000 data sets, there are ten combinations of distribution and type of overlap (steps (a) and (b)) that are randomly assigned an amount of overlap from the interval [0, .45]

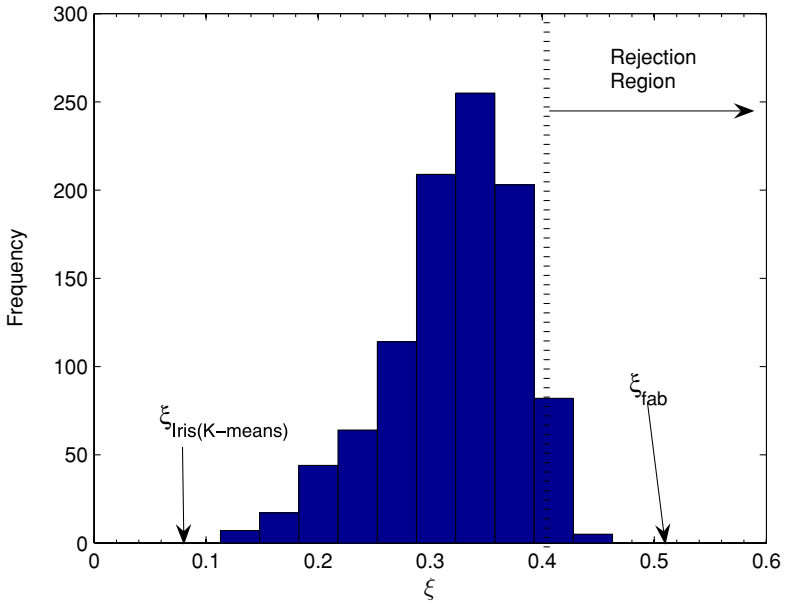

FIGURE 1.
Reference distribution of ξ.

(step (c)). It is important to note that it is possible for each data set to have different overlaps that give rise to the cluster structure.

Figure 1 denotes the reference distribution, with the rejection region representing the top 5% of the observed values for ξ (where the exact cut-off value is $\xi_{\text{cutoff}} = .418$). For the iris data, $A/B_s = 1$, while for the fabricated data $A/B_s = 0$, indicating that ξ_{iris} was always below the ξ^* values for all generated data sets while ξ_{fab} was always above the ξ^* values for all generated data sets. Clearly, the data set with a true cluster structure falls in the acceptance region, while the data set with artificial cluster falls in the rejection region.

6. Conclusion

This paper effectively derives a lower bound for the K-means loss criterion and, based on the lower bound, an index that is invariant to several properties of the data set (i.e., number of clusters, number of variables, distribution of variables, size, etc.) is created. In turn, this index is related to a separate cluster recovery index and a powerful test to determine the quality of a cluster solution is developed. The primary limitation to the proposed index is its ineffectiveness when the number of clusters is larger than the number of variables; however, given the recent increase in data set sizes (for instance, microarray data) this situation should not be too difficult an obstacle to overcome.

Additionally, it is crucial to understand that the effectiveness of the procedure relies heavily on the choice of reference distributions to generate the cluster structure. In the example above, the implementation of the nonparametric testing procedure relied on a very conservative approach that allowed the reference distribution to be drawn from a very wide range of distributions. However, researchers may be tempted to make certain a priori assumptions about the multidimensional structure of the clusters hypothesized to be present within the data. A common assumption would be that the clusters were multivariate normal and spherically shaped. Unfortunately, if the

reference distribution is generated on an assumed cluster structure that, in actuality, is incorrect, the conclusions drawn from the nonparametric procedure will be tenuous at best and wildly misleading at worst. On the other hand, if the correct cluster structure is chosen, the nonparametric test will become more powerful. To gain maximal confidence in the results obtained from the outlined procedure, the researcher should conduct some type of sensitivity analysis. For example, if the researcher chooses a more restrictive range of reference distributions to sample from for the initial test, a follow-up test (using a more liberal sampling scheme for the possibilities of reference distributions) should be conducted to determine the sensitivity of the outcome of the test to the assumptions made upon implementation of the procedure. Finally, these concerns extend beyond the choice of reference distribution to the other choices made by the researcher as well (i.e., overlap, relative cluster density, etc.), and appropriate sensitivity analyses should be conducted to determine the quality of the initial conclusions.

Most importantly, this strategy provides a simple tool to aid researchers in determining if the partitioning provided by a method minimizing the sum of squares error (such as K-means) is truly reflecting the underlying nature of the data. Current research is following two avenues of inquiry:

(a) identifying techniques to transform the continuous values of $\mathbf{O}^*$ to the corresponding discrete values of $\mathbf{M}$, possibly avoiding the iterative nature of the traditional K-means procedure; and
(b) developing a similar eigendecomposition and lower bound for fuzzy (i.e., overlapping) clustering, creating a comparable index and establishing the relationship between the lower bound and the degree of fuzziness in the final cluster solution.

References

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

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

Cattell, R.B., & Coulter, M.A. (1966). Principles of behavioral taxonomy and the mathematical basis of the taxonome computer program. *British Journal of Mathematical and Statistical Psychology, 19*, 237–269.

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

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

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

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

Fan, K. (1949). On a theorem of Weyl concerning eigenvalues of linear transformations. *Proceedings of the National Academy of Sciences of the United States of America, 35*, 652–655.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics, 7*, 179–188.

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

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

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

Golub, G.H., & Van Loan, C.F. (1996). *Matrix computations* (3rd ed.). Baltimore: The Johns Hopkins University Press.

Gordon, A.D. (1999). *Classification* (2nd ed.). New York: Chapman & Hall/CRC.

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

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

Hubert, L., & 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.

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

Kaufman, L., & Rousseeuw, P.J. (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 Science.

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

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

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.

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

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. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5*, 181–204.

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

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

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-Verlag.

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

Steinley, D. (in press). *K*-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*.

Steinley, D., & Henson, R. (2005). OCLUS: An analytic method for generating clusters with known overlap. *Journal of Classification, 22*, 221–250.

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, Series B, 63*, 411–423.

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

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

Waller, N.G., Underhill, J.M., & Kaiser, H.A. (1999). A method for generating simulated plasmodes and artificial test clusters with user-defined shape, size, and orientation. *Multivariate Behavioral Research, 34*, 123–142.

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

Zha, H., Ding, C., Gu, M., He, X., & Simon, H.D. (2001). Spectral relaxation for *K*-means clustering. *Nueral Information Processsing Systems, 14*, 1057–1064.

Manuscript received 23 Nov 2004
Final version received 20 Feb 2006
Published Online Date: 13 Jun 2007