PSYCHOMETRIKA—VOL. 82, NO. 2, 275–294
JUNE 2017
DOI: 10.1007/s11336-017-9561-1

VARIANCE-BASED CLUSTER SELECTION CRITERIA IN A K-MEANS FRAMEWORK FOR ONE-MODE DISSIMILARITY DATA

J. FERNANDO VERA

UNIVERSITY OF GRANADA

RODRIGO MACÍAS

CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS, UNIDAD MONTERREY

One of the main problems in cluster analysis is that of determining the number of groups in the data. In general, the approach taken depends on the cluster method used. For K-means, some of the most widely employed criteria are formulated in terms of the decomposition of the total point scatter, regarding a two-mode data set of N points in p dimensions, which are optimally arranged into K classes. This paper addresses the formulation of criteria to determine the number of clusters, in the general situation in which the available information for clustering is a one-mode $N \times N$ dissimilarity matrix describing the objects. In this framework, p and the coordinates of points are usually unknown, and the application of criteria originally formulated for two-mode data sets is dependent on their possible reformulation in the one-mode situation. The decomposition of the variability of the clustered objects is proposed in terms of the corresponding block-shaped partition of the dissimilarity matrix. Within-block and between-block dispersion values for the partitioned dissimilarity matrix are derived, and variance-based criteria are subsequently formulated in order to determine the number of groups in the data. A Monte Carlo experiment was carried out to study the performance of the proposed criteria. For simulated clustered points in p dimensions, greater efficiency in recovering the number of clusters is obtained when the criteria are calculated from the related Euclidean distances instead of the known two-mode data set, in general, for unequal-sized clusters and for low dimensionality situations. For simulated dissimilarity data sets, the proposed criteria always outperform the results obtained when these criteria are calculated from their original formulation, using dissimilarities instead of distances.

Key words: dissimilarity, cluster analysis, K-means, SYNCLUS, variance-based criterion, number of clusters.

1. Introduction

The K-means algorithm for clustering (Hartigan & Wong, 1979; MacQueen, 1967), is one of the most popular optimization clustering techniques. It produces a partition of the rows of a two-mode $N \times p$ matrix $\mathbf{X}$ into a specified number K of non-overlapping groups, on the basis of their proximities. The rows of $\mathbf{X}$ are in general considered N observations of p variables, assuming that these are measured on a continuous scale, and each observation is classified in the cluster with the nearest mean value, typically accounted for in terms of squared Euclidean distances (see Steinley, 2006 for a review of K-means clustering).

A major problem in cluster analysis, and in K-means in particular, is that of determining the number of clusters. Many approaches to this problem have been suggested, and in general the number of groups chosen depends on the cluster method used (see Everit, Landau, Leese, & Stahl, 2011 for a further review). The most common approach taken is to choose the number of clusters that optimizes a certain criterion. Several methods have been suggested in this direction; commonly adopted criteria in K-means are based on a function of the within-cluster dispersion (Calinski &

Correspondence should be made to J. Fernando Vera, Department of Statistics and O.R., Faculty of Sciences, University of Granada, 18071 Granada, Spain. Email: jfvera@ugr.es

© 2017 The Psychometric Society

Harabasz, 1974, Hartigan, 1975, Krzanowski & Lai, 1985, Tibshirani, Walther & Hastie, 2001, Sugar & James, 2003), calculated in terms of the p-dimensional vectors of observations.

The performance of these dispersion-based criteria has been tested in different situations by simulation, but as with any simulation study, the results obtained are not generalizable. In a real situation, the results obtained depend on the unknown cluster structure and on the algorithm employed to determine group membership, and therefore, the simulation findings are only comparable in the context of the Monte Carlo experiment. One of the most detailed comparative studies is that of Milligan and Cooper (1985), who conducted simulations to compare thirty procedures used in hierarchical cluster analysis, and who recommended the Calinski and Harabasz (1974) index above all others. This index is formulated in terms of the decomposition of the total point scatter of clustered points. The proposed procedure involves selecting the number of clusters that maximizes the ratio of between-cluster to within-cluster dispersion. Although the results of Milligan and Cooper (1985) were obtained in a hierarchical cluster context, the latter index has also been widely employed in K-means clustering (see also Rocci & Vichi, 2008 for a formulation of the index in two-mode multi-partitioning).

More recently, Tibshirani et al. (2001) proposed the Gap statistic, which compares the change in within-cluster dispersion with that expected under an appropriate null probabilistic distribution, usually the uniform distribution in an appropriate rectangle containing the data. Subsequently, Sugar and James (2003) developed a nonparametric method based on a measure of the average distance per dimension between each observation and its closest cluster center. These authors compared their so-called jump method with the Calinski and Harabasz (1974) index, Hartigan's rule (1975), the Krzanowski and Lai test (Krzanowski & Lai, 1985), the Silhouette statistic (Kaufman, & Rousseeuw, 1990), and the Gap statistic (Tibshirani et al. 2001). Empirical results showed their method to be highly successful at selecting the correct number of clusters, for a wide range of practical problems, outperforming the other approaches considered. In the context of intelligent K-means (iK-means), Chiang and Mirkin (2010) showed that Hartigan's HK "rule of thumb" (Hartigan, 1975), in conjunction with an HK adjusted iK-Means algorithm, achieved the best results in terms of number of clusters. These authors compared this criterion with the Calinski and Harabasz criterion, the Gap statistic, the jump statistic and the Silhouette statistic. It was also compared with three other procedures based on the average between-partitions distance index (see Mirkin, 2005) and on the consensus distribution area, i.e., the area under the cumulative distribution function of entries in what is called a consensus matrix, that is, the matrix whose (i, j)-th entry is the proportion of those clustering runs in which the entities (i, j) are in the same cluster (Monti, Tamayo, Mesirov & Golub, 2003).

Underlying the K-means procedure, and many of the criteria proposed to determine the number of clusters, is the hypothesis that the variables in $\mathbf{X}$ are measured on a continuous scale. Nevertheless, in some applications the variables in $\mathbf{X}$ may be a mixture of continuous, ordinal and/or nominal, and some entries will often be missing. Variables of mixed type and missing values may require a new methodology for data clustering. In this situation, a suitable dissimilarity matrix can be generated as the basis for clustering (DeSarbo, Carroll, Clark, & Green, 1984). Furthermore, in many clustering applications, an inter-object dissimilarity matrix $\mathbf{\Delta}$ may arise directly. Examples of such situations can be found in diverse areas. In marketing, clusters of items may be visualized in terms of adjacency data or co-purchase items (Condon, Golden, Lele, Raghavan, & Wasil, 2002); in psychology, clustering can be applied to words for items based on proximity scores determined by patients presenting a specific pathology (Elvevag, & Storms, 2003); in sociology, clustering can help determine groups of social uncertainly sources based on human perceptions (Priem, Love, & Shaffer, 2002); in information retrieval, from the Web or other databases, words or terms may be clustered according to the semantic distances between pairs (Cilibrasi, & Vitanyi, 2007); and in genetic, genomes may be clustered by considering normalized compression distances between gene expression data (Ito, Zeugmann, & Zhu, 2010).

In any situation in which there exists a one-mode set of dissimilarity data, if the dissimilarities are arranged appropriately, an informative partition of the objects will produce a block-shaped partition of the dissimilarity matrix, i.e., a partition that can be identified by drawing horizontal and vertical lines between various of its rows and columns related to the corresponding partition of the objects; moreover, a block-shaped partition of the dissimilarity matrix will induce a partition in the objects such that the objects within a group have a cohesive structure and the groups are well isolated from each other. This relation between any partition of the objects and the related block-shaped partition of the corresponding matrix of Euclidean distances can also be appreciated in terms of clustering criteria in K-means. Thus, minimizing the within-group sums of squares criterion for a two-mode data matrix $\boldsymbol{X}$ is equivalent to minimizing the sum of all the corresponding squared Euclidean distances between two objects from each group (see, for example, Everit et al., 2011, Chapter 5).

When the only available information is a dissimilarity matrix $\boldsymbol{\Delta}$ between objects, a generalized K-means procedure can be formulated in terms of the nonnegative dissimilarities, by minimizing the lack of homogeneity within each cluster (see Everit et al., 2011). To determine the number of clusters in this framework, from the above-mentioned criteria, the Kaufman and Rousseeuw (1990) Silhouette plot method can be directly applied in terms of dissimilarities. In addition, the formulation of a variance-based dependent criterion adapted to a one-mode dissimilarity matrix is also feasible, considering a measure of the total point scatter and splitting it into the within-cluster scatter and the between-cluster scatter.

In this paper, we consider the decomposition of the total dispersion within a dissimilarity matrix (see Heiser, & Groenen, 1997) in order to adapt variance-based criteria—in particular, the Callinski and Harabasz criterion and Hartigan's rule—for direct application on a dissimilarity matrix. A comprehensive Monte Carlo experiment is carried out to study the performance of the proposed variance-based criteria in determining the appropriate number of clusters for a dissimilarity matrix. In addition, the results obtained with the proposed criteria are compared with those given by the application of the classical formulation of the same criteria for artificial two-mode data sets. An important finding of this analysis is that the effectiveness of the variance-based criteria tested increases considerably when they are applied from the Euclidean distances rather than from the original data matrix $\mathbf{X}$. The performance of the proposed criteria is also illustrated for real dissimilarities.

2. A K-Means Clustering Procedure for Dissimilarities

In the application of K-means clustering from a dissimilarity matrix, a generalized K-means algorithm can be defined by characterizing the extent to which observations assigned to the same cluster tend to be close to one another. Let $O = \{o_1, \ldots, o_N\}$ be a set of N objects, and $\boldsymbol{\Delta} = \{\delta_{ij}\}$, $i, j = 1, \ldots, N$, a symmetric matrix of the dissimilarities between them. Assume that each object is allocated to one and only one of K clusters, denoting by $\mathbf{E}$ an indicator matrix of order $N \times K$, whose elements e_{ik} are equal to one if object o_i belongs to cluster k, or zero otherwise. Thus, if we denote by J_k the set of size N_k of objects belonging to cluster k, for $k = 1, \ldots, K$, the hypothesis that the clusters form a partition is expressed as $J_k \bigcap J_l = \emptyset$, for $k \neq l$, and $\bigcup J_k = O$. From $\mathbf{E}$, we can construct a block-shaped partition matrix $P(\boldsymbol{\Delta})$ of blocks $\boldsymbol{\Delta}_{kl}$, where $\delta_{ij} \in \boldsymbol{\Delta}_{\boldsymbol{kl}}$ if $o_i \in J_k$ and $o_j \in J_l$, $\forall i, j = 1, 2, \ldots, N$.

When the information comes from a one-mode dissimilarity matrix $\boldsymbol{\Delta}$, the concepts of lack of homogeneity (minimization) and separation (maximization) can be employed to develop adequacy criteria. The total point scatter, which is not cluster dependent, can be expressed in terms of any classification matrix $\mathbf{E}$ into K clusters as,

$$\tau = \sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{ij} = \sum_{k=1}^{K}\sum_{i=1}^{N} e_{ik}\left(\sum_{j=1}^{N} e_{jk}\delta_{ij} + \sum_{l\neq k}^{K}\sum_{j=1}^{N} e_{jl}\delta_{ij}\right) = \omega(K) + \beta(K), \tag{1}$$

where $\omega(K)$ denotes the within-cluster points scatter and $\beta(K)$ the between-cluster points scatter given by

$$\omega(K) = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jk}\delta_{ij},$$
$$\beta(K) = \sum_{k=1}^{K}\sum_{l\neq k}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}\delta_{ij}.$$

Therefore, since we wish to assign close points to the same cluster, the criterion of minimizing the within-cluster points scatter $\omega(K)$, which is equivalent to maximizing the between-cluster points scatter, characterizes the extent to which observations assigned to the same cluster tend to be close to one another.

Several clustering criteria have been derived from a two-mode $N \times p$ matrix $\mathbf{X}$ using the decomposition of the total points scatter, and of these the K-means procedure is one of the most widely employed. Denoting by $\mathbf{x}_i$ the ith-row of matrix $\mathbf{X}$, $i = 1, \ldots, N$, and by $\bar{\mathbf{x}}_k$ the coordinates in p dimensions of the kth-centroid, $k = 1, \ldots, K$, the K-means method can be formulated in this context by considering dissimilarities as squared Euclidean distances, $\delta_{ij} = d_{ij}^2$, with $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$, and denoting by $d_{ik} = d(\mathbf{x}_i, \bar{\mathbf{x}}_k)$ the Euclidean distance from the ith object to the cluster k. Then, for cluster k, the minimization of the lack of homogeneity is equivalent to minimizing

$$\begin{aligned}\omega_k(K) &= \sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jk}d_{ij}^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jk}\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right)'\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right) + \sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jk}\left(\mathbf{x}_j - \bar{\mathbf{x}}_k\right)'\left(\mathbf{x}_j - \bar{\mathbf{x}}_k\right) \\ &= 2N_k\sum_{i=1}^{N} e_{ik}d_{ik}^2,\end{aligned} \tag{2}$$

from which the minimization of

$$W(K) = \sum_{k=1}^{K}\sum_{i=1}^{N} e_{ik}d_{ik}^2 = \sum_{k=1}^{K}\frac{1}{2N_k}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jk}d_{ij}^2, \tag{3}$$

represents the classical K-means criteria on the basis of the Euclidean distance matrix.

To provide an algorithm for K-means clustering that can be directly applied to Euclidean distances between objects (among other characteristics), DeSarbo et al. (1984) proposed the SYNCLUS (SYNthesized CLUStering) method. In a Cluster-MDS framework, Heiser and Groenen (1997) proposed a K-means-related procedure, the minimal distance method (MD), to partition the objects into K clusters, when the only available information comes from a dissimilarity matrix.

Vera, Macías, and Angulo (2008) adapted this procedure to include constraints, and Vera, Macías and Heiser (2013) employed it for two-mode preference data. In a probabilistic framework, Vera, Macías, and Heiser (2009) and Vera, Macías, and Angulo (2009) have proposed a latent class model for a dissimilarity matrix, under the hypothesis of Gaussian and of lognormal distributions, respectively. Taking into account that K-means can be formulated as a particular limit of the EM algorithm for Gaussian mixtures (see, e.g., Bishop, 2006, Section 9.3 or Press, Teukolsky, Vetterling, & Flannery, 2007, Section 16.1), this latent class model can be viewed as an alternative to K-means when considering a probabilistic distribution for dissimilarities.

The well-known equivalent formulation of the K-means procedure in terms of Euclidean distances can be employed for clustering in a dissimilarity matrix $\mathbf{\Delta}$, by minimizing the lack of homogeneity criterion. Thus, considering a randomly assigned initial classification for the objects, a similar algorithm to that proposed by the SYNCLUS method can be employed in this dissimilarity framework, to minimize (3). From this initial classification, each point is then reassigned to the cluster presenting the closest centroid. To this end, the point-centroid squared dissimilarities are defined as

$$D_{jk}^2 = \frac{1}{N_k}\sum_{i=1}^{N} e_{ik}\delta_{ji}^2 - D_k^2,\ \forall j = 1,\ldots N,\ \forall k = 1,\ldots K, \tag{4}$$

where

$$D_k^2 = \frac{1}{2N_k^2}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jk}\delta_{ij}^2.$$

Thus, o_j is assigned to cluster k for minimum $D_{jk}^2, \forall k = 1,\ldots,K$, and objects are relocated simultaneously, for $j = 1,\ldots,N$, and this is repeated iteratively until the loss function

$$\mathrm{EP} = \sum_{k=1}^{K}\sum_{j=1}^{N} e_{jk}D_{jk}^2 \tag{5}$$

cannot be further minimized, i.e., until no points change from one cluster to another. If $\mathbf{\Delta}$ represents the Euclidean distances for a configuration of points $\boldsymbol{X}$, the minimization of (5) defines a K-means algorithm.

3. Variance-Based Clustering Criteria for a Dissimilarity Matrix

If a subjacent probabilistic distribution is considered for the dissimilarities as in Vera et al. (2009), or in Vera et al. (2009), hypothesis testing combined with bootstrap sampling, or statistical criteria such as the BIC (Schwartz, 1978) information criterion, can be employed to determine the number of clusters. However, in a general deterministic framework as in the situation we are considering here, in which the only information comes from a dissimilarity matrix, variance-based criteria can be formulated to determine the number of clusters in terms of dissimilarities. Given the allocation of the objects into K clusters, the orthogonality of the least squares estimates with their residuals makes it possible to break down the total variance of a dissimilarity matrix in terms of the between-cluster and the within-cluster variability for a given block-shaped partition.

3.1. Analysis of Dispersion for a Dissimilarity Matrix

For a one-mode dissimilarity matrix $\mathbf{\Delta}$, the total dissimilarity scatter τ can be written as,

$$\tau(\mathbf{\Delta}) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(\delta_{ij} - \overline{\delta}\right)^2, \tag{6}$$

where w_{ij} represents weights that can be assigned to each pair of objects, e.g., to deal with missing dissimilarities, and $\overline{\delta}$ represents the overall mean, called the overall Sokal–Michener dissimilarity (Sokal, & Michener, 1958), which is given by,

$$\overline{\delta} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\delta_{ij}.$$

Any partition of the object space $P(O)$ induces a block-shaped partition in the dissimilarities. Thus, given a block-shaped partition $P(\mathbf{\Delta})$, and considering the Sokal–Michener dissimilarity between clusters, which for each block $\mathbf{\Delta}_{kl}$, $k \leq l$, is defined as

$$\overline{\delta}_{kl} = \sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}\frac{w_{ij}\delta_{ij}}{\mathring{w}_{kl}}, \quad \text{where } \mathring{w}_{kl} = \sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}w_{ij},$$

the well-known orthogonality to the residuals of least squares quantities makes the following relation hold

$$\sum_{k\leq l}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}w_{ij}\delta_{ij}^2 = \sum_{k\leq l}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}w_{ij}\left(\delta_{ij} - \overline{\delta}_{kl}\right)^2 + \sum_{k\leq l}\mathring{w}_{kl}\overline{\delta}_{kl}^2. \tag{7}$$

In this context, orthogonality implies that for any block $\mathbf{\Delta}_{kl}$, for $k \leq l$, the weighted cross product of $\overline{\delta}_{kl}$ and $(\delta_{ij} - \overline{\delta}_{kl})$ vanishes, and thus, the equation $\delta_{ij} = (\delta_{ij} - \overline{\delta}_{kl}) + \overline{\delta}_{kl}$ when squared, multiplied by w_{ij}, and summed, yields the latter expression.

The first component on the right-hand side of (7) represents the sum of the square deviations of the dissimilarities with respect to the average of the block to which they belong, while $\sum_{k\leq l}\mathring{w}_{kl}\overline{\delta}_{kl}^2$ represents the total dispersion between the clusters. Thus in general, the total variability can be decomposed as,

$$\sum_{k\leq l}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}w_{ij}(\delta_{ij} - \overline{\delta})^2 = \sum_{k\leq l}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}w_{ij}(\delta_{ij} - \overline{\delta}_{kl})^2 + \sum_{k\leq l}\mathring{w}_{kl}(\overline{\delta}_{kl} - \overline{\delta})^2, \tag{8}$$

where $\overline{\delta}$ is the overall mean of dissimilarities. The above two expressions are equivalent and represent a dispersion analysis on a block-shaped partition from the one-mode dissimilarity matrix. The first component represents the within-block dispersion with $(N(N-1) - K(K+1))/2$ degrees of freedom, whereas the second component represents the between-block dispersion with $K(K+1)/2$ degrees of freedom. The within-block dispersion and the between-block dispersion are denoted by

$$W^*(K) = \sum_{k\le l}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jl}w_{ij}(\delta_{ij} - \bar{\delta}_{kl})^2,$$
$$B^*(K) = \sum_{k\le l} \mathring{w}_{kl}(\bar{\delta}_{kl} - \bar{\delta})^2. \tag{9}$$

3.2. The Calinski–Harabasz Index in Terms of a Dissimilarity Matrix

The Calinski–Harabasz (1974) index (or pseudo-F value) is defined for a continuous data matrix $\mathbf{X}$ as:

$$\mathrm{CH}(K) = \frac{B(K)/(K-1)}{W(K)/(N-K)}, \tag{10}$$

where $W(K)$ and $B(K)$ represent the trace of the within-group dispersion matrix $\mathbf{W}$ and of the between-group dispersion matrix $\mathbf{B}$, respectively, defined by

$$W(K) = tr(\mathbf{W}) = tr\sum_{k=1}^{K}\sum_{i=1}^{N} e_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)' = \sum_{k=1}^{K}\sum_{i=1}^{N} e_{ik}d_{ik}^2, \tag{11}$$

$$B(K) = tr(\mathbf{B}) = tr\sum_{k=1}^{K} N_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})' = \sum_{k=1}^{K} N_k d_{k\cdot}^2 \tag{12}$$

where $d_{k\cdot}$ is the Euclidean distance between $\bar{\mathbf{x}}_k$ and $\bar{\mathbf{x}}$. As shown by Gower and Krzanowski (1999), the total sum of squares $T = tr(\mathbf{X}'\mathbf{X})$ can be written as,

$$T = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = W(K) + B(K), \tag{13}$$

and taking into account (3), $B(K)$ can also be written in terms of pairwise Euclidean distances as,

$$B(K) = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 - \sum_{k=1}^{K}\frac{1}{2N_k}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ik}e_{jk}d_{ij}^2. \tag{14}$$

The value of K that maximizes CH is then chosen as the optimum number of clusters. CH(1) is not defined; even if it were modified by replacing $K-1$ with K, its value at 1 would be 0. Since $\mathrm{CH}(K) > 0$ for $K \ge 1$, the maximum would never occur at $K = 1$.
In the context of a block-shaped partition from a one-mode dissimilarity matrix, the number of distinct blocks is $K(K+1)/2$ and the number of different dissimilarities is $N(N-1)/2$ (without including the diagonal entries). Thus, the optimum number of clusters K is associated with a large value of

$$\mathrm{CH}^*(K) = \frac{B^*(K)/(K(K+1)/2)}{W^*(K)/([N(N-1) - K(K+1)]/2)}. \tag{15}$$

For the case of $K = 1$, the term $B^*(1)$ is equal to zero and $\mathrm{CH}^*(1) = 0$. Therefore, the maximum will never be achieved for $K = 1$.

3.3. A Formulation of the Hartigan Statistic in Terms of Dissimilarities

To determine the number of clusters, Hartigan (1975) proposed the statistic

$$H(K) = \left[\frac{W(K)}{W(K+1)} - 1\right](N - K - 1), \tag{16}$$

where $W(K)$ is the trace of $\mathbf{W}$ (see 11). Starting with $K = 1$, a new cluster is added as long as $H(K)$ is sufficiently large. Instead of using an approximate F-distribution cutoff, Hartigan suggested an empirical rule, of adding a new cluster if $H(K) > 10$. Then, the estimated number of clusters is the smallest $K \geq 1$ such that $H(K) \leq 10$. Unlike the previous criterion, this index is well defined for $K = 1$ and can identify when there is no cluster structure. Thus, considering again the number of different blocks, $K(K+1)/2$, and the number of dissimilarities without including the diagonal entries, $N(N-1)/2$, for a block-shaped dissimilarity matrix, the criterion $H(K)$ can be formulated as

$$H^*(K) = \left[\frac{W^*(K)}{W^*(K+1)} - 1\right](([N(N-1) - K(K+1)]/2) - 1), \tag{17}$$

where W^* is defined on (9). In order to find an empirical rule similar to that proposed by Hartigan, we conducted a simulation study (results not reported here) and found that for any value of K, the ratio in brackets at (16) does not differ significantly when $W^*(K)$ is considered instead of $W(K)$. Nevertheless, the value of the correction factor in a dissimilarity framework, $(([N(N-1) - K(K+1)]/2) - 1)$, is very large for a moderately large value of N, and so the value of $H^*(K)$ varies on a large scale for any K. Thus, a new empirical rule, different from that proposed by Hartigan, is defined to determine the condition for adding a new cluster $(K+1)$. This rule is based on the magnitude of the ratio of the correction factor between the original and the proposed criterion given by

$$o(N, K) = \frac{([N(N-1) - K(K+1)]/2) - 1}{N - K - 1} = \frac{N+K}{2} - \frac{1}{N-K-1}. \tag{18}$$

Therefore, according to Hartigan's rule, and considering the proportion in (18), the estimated number of clusters is the smallest value $K \geq 1$ such that $H^*(K) \leq 10\, o(N, K)$. This rule depends on N and K, but for a value of $N >> K$, $H^*(K)/H(K) \approx N/2$. Therefore, the proposed rule for large values of N (for practical issues, it should be considered at least $N > K(K+1)$) can be written as $H^*(K) \leq 5N$. This rule is formulated experimentally and in terms of a within-cluster variance reduction, without any statistical distribution consideration, and produces good results for general dissimilarity values.

3.4. The Silhouette Statistic for a Dissimilarity Matrix

The Silhouette statistic (Kaufman, & Rousseeuw, 1990) is defined for a given object o_i as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{19}$$

where $a(i)$ is the average distance of object o_i from all other objects in its own cluster and where $b(i)$ is the average distance of object o_i from all objects in the nearest cluster. For each object o_i,

the index $s(i)$ measures the (standardized) difference between $b(i)$ and $a(i)$, and this difference is large if the solution contains at least enough clusters to satisfactorily capture the variability in the objects. In general, according to the values of $a(i)$ and $b(i)$, $s(i)$ can be expressed as

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i) \end{cases} \tag{20}$$

Thus, $s(i) \in [-1, 1]$. When $s(i)$ is close to the value 1, object i is nearer to its own cluster than to a neighboring one and is considered to be well classified. When $s(i)$ is closer to the value -1, the opposite relationship applies and object i is considered to be misclassified. When the index is close to zero, it is not clear whether the object should have been assigned to its current cluster or to a neighboring one. Kaufman and Rousseeuw (1990) suggested choosing the number of clusters that maximizes the average value of $s(i)$ over the entire data set, denoted by SIL.

In the context of a one-mode dissimilarity matrix, and given a block-shaped partition $P(\mathbf{\Delta})$, $a(i)$ and $b(i)$ (both averaged distances) can be expressed as:

$$\widetilde{a}(i) = \frac{\sum_{i,j \in J_l} \delta_{ij}}{\mathring{w}_{ll}}, \quad \widetilde{b}(i) = \min_{t \neq l} \left\{ \frac{\sum_{j \in J_t} \delta_{ij}}{\mathring{w}_{lt}} \right\}, \quad o_i \in J_l.$$

Because (19), which is now defined in terms of $\widetilde{a}(i)$ and $\widetilde{b}(i)$ and denoted by $\widetilde{s}(i)$, does not make sense for $K = 1$, the number of clusters suggested is the value of $\widehat{K} > 1$ such that

$$\widetilde{\mathrm{SIL}}(\widehat{K}) = \max_{K>1}\{\overline{S}(K)\}, \quad \text{where } \overline{S}(K) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{s}(i), \quad \text{for } K > 1.$$

A reasonable clustering must be characterized for a value of $\widetilde{\mathrm{SIL}}(\widehat{K})$ larger than 0.5, and a small value should be interpreted as a substantial absence of grouping structure in the data (Kaufman, & Rousseeuw, 1990).

4. A Comparative Simulation Study

A simulation study was conducted to explore the behavior of the criteria, generating Euclidean distances from artificially grouped rectangular data sets. In addition, grouped dissimilarity data were directly generated from mixtures of normal distributions.

4.1. Experimental Results for Simulated Euclidean Distances

Artificial clustered data sets were generated in accordance with the well-known Milligan algorithm (Milligan, 1985), and taking into account different cluster densities (only the most significant results are shown here, due to space limitations). The data sets were generated considering a structure in three, four, five and six non-overlapping clusters, in 2, 3, 4, 6 and 8 dimensions. The distribution of points across the clusters was performed according to three levels of density. The first level generated an equal number of points in each cluster (or as equal as possible). The second level generated a cluster containing 10% of the total data, while the third level required one cluster containing 60% of the data. The remaining points were distributed as evenly as possible

across the other clusters present in the data. For each combination of factors, 100 data sets each consisting of $N \times p$ observations for $N = 50, 100, 200, 500$ were generated, and the Euclidean distances between the pairs of observations were calculated as dissimilarities. For a given density level and number of clusters and dimensions, the following main steps for the generation process were followed (see Milligan, 1985 for an extensive description).

- Non-overlapping cluster boundaries are randomly generated for the first dimension, assuming a uniform distribution from 10 to 40. The boundary length is taken as three times the standard deviation of the cluster, and the midpoint of this length represents the mean of the cluster in this dimension. A random ordering of the clusters is assumed, and the boundaries of clusters k and l are separated by a random quantity $f(S_k + S_l)$, where S_k is the standard deviation of cluster k and f is a uniform random variable with a range of 0.25–0.75.
- The cluster boundaries are determined in the same way for each of the remaining dimensions. The maximum range of the data is limited to two-thirds of the range of the first dimension.
- A multivariate normal with a centroid given by the midpoints of the boundary lengths is used to generate the within-cluster points. The diagonal of the variance-covariance matrix is given by the standard deviations and the off-diagonal elements are zeros. The only points generated that are accepted are those which lie within the cluster boundaries.

The solutions obtained by the K-means algorithm as described in Sect. 2 were used to calculate the criteria applicable. For the K-means algorithm, 200 iterations and 200 random restarts were employed, for a range of clusters from $K = 2, \ldots, 10$. Although for the SYNCLUS method (DeSarbo et al., 1984), a particular set of K center points was chosen as the initial solution, in the proposed implementation the utilization of random seed points in connection with random restarts produced the best results.

In addition to the use of the CH, Hartigan and Silhouette criteria in their original formulations for a rectangular $N \times p$ data matrix $\mathbf{X}$, in this comparative study, the two following variance-based criteria were considered. The first criterion is the jump method proposed by Sugar and James (2003), which is based on distortion, i.e., the average distance, per dimension, between each observation and its closest cluster center. Formally, this distortion is defined as

$$d_K = \frac{1}{p} \min_{\mathbf{c}_1,\ldots,\mathbf{c}_K} E[(\mathbf{X} - \mathbf{c}_\mathbf{X})^\mathbf{T} \boldsymbol{\Gamma}^{-1} (\mathbf{X} - \mathbf{c}_\mathbf{X})], \tag{21}$$

where $\boldsymbol{\Gamma}$ is the within-cluster covariance matrix, $\mathbf{c}_1, \ldots, \mathbf{c}_K$ is a set of cluster center candidates, and $\mathbf{c}_\mathbf{X}$ is the one closest to $\boldsymbol{X}$. In practice, d_K is estimated by applying the K-means algorithm to the dissimilarity data and by considering the classification obtained for the data matrix $\mathbf{X}$. The jump is defined as $J_K = d_K^{-q} - d_{K-1}^{-q}$, assuming that $d_0^{-q} \equiv 0$, where $q > 0$ is an adequate value (a typical value is $q = p/2$). Then, the appropriate number of clusters is estimated as the value $K^* = \arg\max_K J_K$, i.e., the value of K associated with the largest jump.

Another approach was proposed by Krzanowski and Lai (1985), who recommended choosing the K value that maximizes the expression

$$\mathrm{KL}(K) = \left| \frac{\mathrm{DIFF}(K)}{\mathrm{DIFF}(K+1)} \right|, \tag{22}$$

where $\mathrm{DIFF}(K) = (K-1)^{2/p} W(K-1) - k^{2/p} W(K)$. Although these two additional criteria are also widely employed in K-means, due to the explicit use of the p dimensionality in their formulation, this approach is only applicable to a rectangular data set.

Table 1 shows the results obtained for $N = 50, 500$, and $K = 4, 6$, for the simulated data sets. The criteria to determine the number of clusters, both in their original formulations and in the one proposed here, were applied to the same classification obtained by the proposed K-means algorithm in terms of Euclidean distances. The values in the table are the recovery percentages of the actual number of clusters, taking into account the dimensions in which the data were generated. Because the jump method and the KL statistic are formulated in terms of the dimensionality of the data, the results were only calculated in terms of the original data.

In all the settings considered and for equal-sized clusters (first level of cluster density), the CH index in its classical formulation presents a very consistent performance in recovering the real number of clusters (above 75%), especially for $N \geq 200$, and when considering data generated in a space with a dimension larger than three for $N \leq 100$. Of the remaining criteria analyzed in their original formulations, the Silhouette, the KL and the jump (the latter except for $K = 3$ and $N \leq 100$, and $K = 4$ for $N = 50$) also present a very regular behavior pattern in all dimensions, although slightly less so than the CH index, while the Hartigan rule performs erratically in some settings, especially for low dimensions and $N \geq 100$. In the Euclidean distance space, the adapted CH* index shows a slightly better performance, i.e., slightly higher percentages of success than those observed in its classical formulation, especially for low dimensions, while the performance of the Hartigan adapted rule, H^*, was in general somewhat irregular, i.e., sometimes showing larger percentage values and sometimes lower ones, and its performance worsens as the size of N increases, although in some situations for $N \leq 100$ there is a significant improvement over the original formulation.

Considering the results obtained for the second and third levels of cluster density, i.e., in which one cluster contains 10 and 60% of the total data, the effectiveness of most of the criteria seems to be lower than when the clusters present equal densities. However, the CH index consistently recovers the actual number of groups, particularly in the higher dimensions (with the exception of the third level for $K = 6$, and up to 6 dimensions). Among the other criteria in their original forms, the KL criterion and the Silhouette statistic show a regular behavior pattern in all settings, although their performance in high dimensions is slightly inferior to that obtained by the CH index. In the Euclidean distance space, the CH* index maintains an efficacy similar to that of its original formulation for data generated in a high dimension, although its performance increases with lower dimensions. The H^* rule is in some cases significantly more effective than in its original formulation, but its performance is inconsistent.

In general, the CH* best recovered the actual number of clusters in all settings considered, in comparison with the other adapted criteria, and even surpassed the original formulation in situations in which different cluster densities were considered, in particular for lower dimensions. Accordingly, this criterion is very attractive for application to real data sets, because in real-world experimental situations, the assumption of different cluster densities is usually more realistic.

The irregular performance of the jump method in some of the situations analyzed could be accounted for by the proposed transformation, $p/2$. Indeed, smaller values for this transformation would be expected to give better results for this criterion, but unfortunately in real-world experimental situations, the number of clusters, and therefore the exact transformation value, is unknown. With respect to the Hartigan rule, the inconsistent performance in both spaces seems to confirm the findings reported by Milligan (1985), although the iK-means algorithm of Chian and Mirkin (2010) considerably improves the performance of this index for rectangular data sets.

The results obtained in the simulation experiment suggest that when the information comes from a matrix $\boldsymbol{X}$, application of the CH* adapted index to the derived Euclidean distance matrix $\mathbf{D}$ is a suitable procedure for determining the number of clusters, in particular for lower dimensions.

TABLE 1.

Recovery percentages of the actual number of clusters for the original criteria and for the adapted criteria, according to the clustering solutions obtained by the K-means algorithm applied to the distance matrices derived from the data sets generated by the Milligan algorithm, for $N = 50$, 500, and $K = 4$, 6.

Criteria\Dim	$N = 50$										$N = 500$									
	Original					Adapted					Original					Adapted				
	2	3	4	6	8	2	3	4	6	8	2	3	4	6	8	2	3	4	6	8
Data generated for $K = 4$ clusters																				
Equal sized																				
CH	62	97	99	99	99	86	87	96	98	98	87	93	100	100	100	93	92	97	99	100
HA	20	55	86	99	100	42	59	67	81	90	0	0	0	0	0	3	4	3	6	11
SIL	70	77	88	87	97	70	77	88	87	97	67	78	82	87	96	67	78	82	87	96
KL	84	86	91	95	93	–	–	–	–	–	83	87	96	98	100	–	–	–	–	–
JUMP(p/2)	69	73	63	55	36	–	–	–	–	–	90	90	99	100	100	–	–	–	–	–
10% cluster																				
CH	43	82	92	98	98	71	83	92	98	98	60	97	96	97	97	83	92	97	94	94
HA	12	55	75	99	99	37	65	72	79	79	0	0	0	0	0	0	1	1	6	12
SIL	60	69	79	86	91	60	69	79	86	91	61	70	73	85	89	61	70	73	85	89
KL	65	71	83	86	88	–	–	–	–	–	74	82	89	96	97	–	–	–	–	–
JUMP(p/2)	50	55	58	40	22	–	–	–	–	–	81	93	97	99	98	–	–	–	–	–
60% cluster																				
CH	9	33	71	93	97	59	78	81	93	94	25	60	84	98	99	76	81	76	93	96
HA	1	9	20	62	88	29	41	51	73	75	0	0	0	0	0	0	0	0	0	0
SIL	56	64	82	84	92	56	64	82	84	92	55	63	78	88	95	55	63	78	88	95
KL	34	46	55	75	74	–	–	–	–	–	41	45	43	63	83	–	–	–	–	–
JUMP(p/2)	28	40	31	34	17	–	–	–	–	–	67	76	86	94	98	–	–	–	–	–
Data generated for $K = 6$ clusters																				
Equal sized																				
CH	57	82	98	99	99	74	84	96	97	99	94	99	100	100	100	92	94	97	100	100
HA	69	90	99	99	99	58	66	79	88	98	0	0	0	0	0	0	2	2	9	22
SIL	75	70	85	96	97	75	70	85	96	97	72	72	83	93	98	72	72	83	93	98
KL	84	89	93	98	94	–	–	–	–	–	83	97	98	99	99	–	–	–	–	–
JUMP(p/2)	93	95	100	98	95	–	–	–	–	–	88	95	99	100	100	–	–	–	–	–

TABLE 1.
continued

Criteria\Dim	$N = 50$										$N = 500$									
	Original					Adapted					Original					Adapted				
	2	3	4	6	8	2	3	4	6	8	2	3	4	6	8	2	3	4	6	8
10% cluster																				
CH	46	84	99	100	100	78	86	96	93	99	90	99	100	100	100	88	96	99	100	100
HA	62	94	99	100	100	50	70	84	88	96	0	0	0	0	0	1	4	2	10	17
SIL	66	71	87	93	99	66	71	87	93	99	64	72	89	95	96	64	72	89	95	96
KL	77	81	89	98	96	–	–	–	–	–	81	93	99	100	99	–	–	–	–	–
JUMP	84	88	97	96	96	–	–	–	–	–	80	97	100	100	100	–	–	–	–	–
60% cluster																				
CH	9	11	16	47	83	38	63	71	83	95	21	13	31	88	100	55	68	78	90	97
HA	11	13	20	60	90	17	32	33	62	69	5	5	3	1	2	6	7	3	1	3
SIL	48	61	78	91	96	48	61	78	91	96	42	68	69	84	98	42	68	69	84	98
KL	19	33	51	67	77	–	–	–	–	–	23	29	50	70	85	–	–	–	–	–
JUMP(p/2)	29	38	44	58	30	–	–	–	–	–	43	54	73	92	98	–	–	–	–	–

4.2. Experimental Results for Simulated Dissimilarity Data

To further test the performance of the proposed procedure, clustered nonnegative dissimilarity data were directly generated from a mixture of $K(K+1)/2$ normal distributions, following a methodology similar to that discussed in Vera et al. (2009), but now in terms of dissimilarities. Therefore, in contrast to the previous Monte Carlo experiment, a mixture distribution is assumed here for the direct simulation of the clustered dissimilarities, for which the means vectors and covariance matrices of each normal component in the mixture are calculated from a previous Monte Carlo experiment, using the Milligan algorithm as described above. Thus, from a given partition $\mathbf{E}$ of a data matrix $\mathbf{X}$ into K clusters obtained with the Milligan algorithm, a block-shaped partition of the corresponding matrix of Euclidean distances $\mathbf{D}$ is derived, and the corresponding mean distances μ_{kl}, and variances σ^2_{kl}, for $k \leq l$ are calculated as follows:

$$\mu_{kl}(E) = \frac{\sum_{i<j} e_{ik} e_{jl} d_{ij}}{\sum_{i<j} e_{ik} e_{jl}} \tag{23}$$

$$\sigma^2_{kl}(E) = \frac{\sum_{i<j} e_{ik} e_{jl} \left(d_{ij} - \mu_{kl}\right)^2}{\sum_{i<j} e_{ik} e_{jl}}. \tag{24}$$

Assuming $\delta_{ij} \sim \mathcal{N}(\mu_{kl}, \sigma^2_{kl})$, for $\delta_{ij} \in \mathbf{\Delta}_{kl}$, nonnegative dissimilarities δ_{ij} are directly generated from the mixture of Gaussian components, with means and variances given by (23) and (24), by assuming a probability of δ_{ij} of belonging to each block given by the size of the experiment design. According to $\mathbf{E}$, a block-shaped dissimilarity matrix $\mathbf{\Delta}^*$ of size $N \times N$ is thus generated by means of the corresponding partition block components $\mathbf{\Delta}^*_{kl}$, for $N = 50, 100, 200, 500$, where the dimensions of the block components are determined according to the three density levels as before; diagonal blocks of the same size, one diagonal block containing 10% of the total diagonal block dissimilarities and one diagonal block containing 60% of the total diagonal block dissimilarities, while the remaining diagonal blocks are considered to be of equal sizes.

Thus, 100 dissimilarity matrices $\mathbf{\Delta}^*$ were generated for each of the above N values and three block size designs, for the values of $K = 3, \ldots, 8$. The performance of these adapted criteria was tested over the 7200 dissimilarity matrices and thus generated, and the percentages of recovery of the true numbers of clusters were analyzed for the proposed formulation, as well as by considering the classical formulation of the criteria (when available) in terms of (3) and (14), when dissimilarities are considered instead of Euclidean distances, i.e., formulating

$$\widetilde{W}(K) = \sum_{k=1}^{K} \frac{1}{2N_k} \sum_{i=1}^{N} \sum_{j=1}^{N} e_{ik} e_{jk} \delta^2_{ij}, \tag{25}$$

$$\widetilde{B}(K) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta^2_{ij} - \sum_{k=1}^{K} \frac{1}{2N_k} \sum_{i=1}^{N} \sum_{j=1}^{N} e_{ik} e_{jk} \delta^2_{ij}. \tag{26}$$

Note that $\widetilde{W}(K)$ and $\widetilde{B}(K)$ denote the classical formulation for the within- and between-group dispersion in terms of dissimilarities, and related criteria are denoted accordingly.

Table 2 shows the results for the CH^* and H^* criteria, those for the Silhouette when dissimilarities are used instead of distances, and those for the $\widetilde{\mathrm{CH}}$ and $\widetilde{H}$ criteria, i.e., the results when (25) and (26) are employed in their formulation (only results for $K = 4, 6, 8$ are shown).

TABLE 2.
Recovery percentages of the actual number of clusters for the original criteria and for the adapted criteria, according to the clustering solutions obtained by the K-means algorithm applied to the dissimilarity matrices generated from a mixture of Gaussian blocks, for $N = 50, 100, 200, 500$, and $K = 4, 6, 8$.

Diagonal density	$\widetilde{\text{CH}}$	$\widetilde{H}$	$\widetilde{\text{SIL}}$	CH*	H^*	$\widetilde{\text{CH}}$	$\widetilde{H}$	$\widetilde{\text{SIL}}$	CH*	H^*
			$N = 50$					$N = 100$		
Data generated for 10 distinct blocks, associated with $K = 4$										
Equal size	96	86	77	97	99	100	44	87	97	100
10% one block	94	85	75	89	100	99	41	80	93	100
60% one block	90	81	62	88	96	95	27	65	93	99
			$N = 50$					$N = 100$		
Data generated for 21 distinct blocks, associated with $K = 6$										
Equal size	87	90	81	95	94	98	70	70	97	99
10% one block	85	87	78	93	96	99	71	69	94	100
60% one block	74	82	49	88	76	95	34	49	81	94
			$N = 50$					$N = 100$		
Data generated for 36 distinct blocks, associated with $K = 8$										
Equal size	77	87	61	94	72	90	79	64	88	90
10% one block	73	84	64	87	70	94	89	60	93	98
60% one block	38	60	37	63	36	75	45	41	71	73
			$N = 200$					$N = 500$		
Data generated for 10 distinct blocks, associated with $K = 4$										
Equal size	99	0	79	98	100	99	0	74	98	100
10% one block	95	0	66	91	100	97	0	69	96	100
60% one block	89	0	58	90	100	92	0	67	88	100
			$N = 200$					$N = 500$		
Data generated for 21 distinct blocks, associated with $K = 6$										
Equal size	100	18	76	97	99	100	2	76	92	100
10% one block	100	16	74	96	100	100	0	68	96	100
60% one block	91	16	56	79	98	93	5	49	78	98
			$N = 200$					$N = 500$		
Data generated for 36 distinct blocks, associated with $K = 8$										
Equal size	92	56	67	91	92	97	24	65	93	98
10% one block	97	57	57	94	97	93	22	51	92	93
60% one block	78	46	43	71	75	70	29	37	63	72

As in the previous simulation study, the CH* index very efficiently recovers the actual number of clusters. However, what is remarkable is the performance of the new formulation of the Hartigan criterion (H^*), which in general obtains the best results in all the situations considered for $N > K(K + 1)$. The Silhouette index produces an irregular performance in most of the settings analyzed. Although the alternative formulations of $\widetilde{\text{CH}}$ and $\widetilde{H}$ also obtain good results, especially the $\widetilde{\text{CH}}$ criterion for dissimilarities (the $\widetilde{H}$ showed an irregular performance for $N = 100$, and poor performance for $N > 100$), the formulation proposed in this paper in terms of variance decomposition produced the best results, especially for the H^* index. Therefore, when the only information is a dissimilarity matrix Δ between a set of $N >> K$ objects, the proposed H^* index constitutes a very acceptable procedure for determining the actual number of clusters.

It is well known that the K-means clustering algorithm will perform optimally when the data are generated from normal distributions with equal variances, and it should be noted that K-means clustering is designed to find non-overlapping groups (see Steinley 2006, and Steinley and Brusco, 2007). Therefore, the inclusion of overlap between the clusters may lead to confused results in

TABLE 3.
Recovery percentages of the actual number of clusters for the original criteria and for the adapted criteria, according to the clustering solutions obtained by the K-means algorithm applied to the dissimilarity matrices generated from a mixture of overlapping Gaussian blocks, for $N = 50, 500$, and $K = 4, 6, 8$.

	$N = 50$					$N = 500$				
Diagonal density	$\widetilde{\text{CH}}$	$\widetilde{H}$	$\widetilde{\text{SIL}}$	CH*	H^*	$\widetilde{\text{CH}}$	$\widetilde{H}$	$\widetilde{\text{SIL}}$	CH*	H^*
Average overlap, .01										
Data generated of 10 distinct blocks, associated with $K = 4$										
Equal size	95	87	0	75	100	100	0	0	28	100
Different size	95	64	0	73	99	83	0	0	3	100
Data generated of 21 distinct blocks, associated with $K = 6$										
Equal size	3	15	0	0	4	100	0	0	40	100
Different size	6	17	0	0	3	74	0	0	10	100
Data generated of 36 distinct blocks, associated with $K = 8$										
Equal size	21	23	3	7	7	100	3	0	0	100
Different size	24	31	2	7	6	100	2	0	0	100
Average overlap, .05										
Data generated of 10 distinct blocks, associated with $K = 4$										
Equal size	19	78	0	0	59	0	0	0	0	100
Different size	21	63	0	0	56	0	0	0	0	100
Data generated of 21 distinct blocks, associated with $K = 6$										
Equal size	0	19	0	0	0	0	0	0	0	100
Different size	1	19	0	0	0	0	1	0	0	100
Data generated of 36 distinct blocks, associated with $K = 8$										
Equal size	3	1	6	0	0	0	0	0	0	100
Different size	1	0	4	0	0	0	0	0	0	100

terms of the performance of cluster criteria, mainly due to the expected poor performance for the k-means procedure. Taking this into account, we considered the inclusion of overlap only in determining how the performance of the study criteria withstands data analytic situations that are not ideal.

Accordingly, 100 dissimilarity data sets were generated for the values of $N = 50, 100, 200, 500$, and $K = 3, \ldots, 8$, following the above-described procedure, thus generating dissimilarities for each mixture of normal distributions, using the R package MixSim (see Melnykov, Chen and Maitra, 2012 for further details). In this package, equal- and unequal-sized clusters were considered, and two average degrees of cluster pairwise overlap given by the values of $\bar{\omega} = 0.01$, i.e., a moderate degree of overlap, and $\bar{\omega} = 0.05$, i.e., a high degree of overlap, were employed (see Maitra and Melnykov, 2010 for further details of the meaning of the cluster overlap parameter, in terms of misclassification probabilities).

Table 3 shows the results obtained for 4, 6 and 8 clusters with overlapping data (results for $N = 100, 200$ are not shown). The $\widetilde{\text{CH}}$ and, even more so, the proposed H^* criteria seem to perform well in general for a moderate degree of overlap ($\bar{\omega} = 0.01$) in all situations for $N > 100$, although with up to $K = 7$ clusters for $N = 100$, and up to $K = 4$ clusters for $N = 50$. For a high degree of overlap ($\bar{\omega} = 0.05$), in general the performance decreases, as expected (the number of clusters is usually underestimate, perhaps due to the performance of the K-means procedure in this context), especially for $N < 500$, for both equal- and unequal-sized clusters. Nevertheless, for $N = 500$ the H^* criterion performed well in all situations, while for $N = 200$ it performed well up to $K = 7$, and for $N = 100$ up to $K = 4$ clusters. In the remaining situations, the

TABLE 4.
Clustering results for the famous people data, for $K = 5$.

Cluster	Members
1	Takemitsu, Stockhausen, Kagel, Xenakis, Ligeti
	Kurtag, Martinu, Berg, Britten, Crumb
	Penderecki, Bartok, Beethoven, Mozart, Debussy
	Hindemith, Ravel, Schoenberg, Sibelius, Villa-Lobos
	Boulez, Kodaly, Prokofiev, Schubert
2	Rembrandt, Rubens, Beckmann, Botero, Braque
	Chagall, Duchamp, Escher, Frankenthaler, Giacometti
	Hotere, Kirchner, Kandinsky, Kollwitz, Klimt
	Malevich, Modigliani, Munch, Picasso, Rodin
	Schlemmer, Tinguely, Villafuerte, Vasarely, Warhol
3	Rowling, Hosseini, McCullough
	Friedman, Paolini, Grisham
	Osteen, Gladwell, Trudeau, Levitt, Kidd
	Haddon, Brashares, Guiliano, Maguire
	Snicket, Patterson, Kostova, Kantor, Wiles
4	Pythagoras, Archimedes, Euclid, Thales, Descartes
	Lagrange, Laplace, Leibniz
	Euler, Gauss, Hilbert, Galois, Cauchy
	Dedekind, Poincare, Godel, Ramanujan
	Riemann, Erdos, Thomas Zeugmann, Jan Poland
5	Rolling, Stones, Madonna, Elvis, Depeche
	Mode, Pink, Floyd, Elton, John
	Beatles, Phil, Collins, Toten, Hosen
	McLachan, Prinzen, Aguilera, Queen, Britney
	Spears, Scorpions, Metallica, Blackmore, Mercy
	Cage, Brown, Frey, Warren, Oz Sparks, Roberts, Lewis, Pascal, Newton

performance of these criteria both in the classical and in the proposed formulation was poor, while the Silhouette criterion in general performed poorly with all the data sets tested.

5. Application to Real Data

To illustrate the performance of the adapted criteria when applied to real data, we analyzed a dissimilarity matrix previously examined by Poland and Zeugmann (2006). The data represent web distances, as termed by Cilibrasi and Vitányi, (2004, 2007), which are semantic distances between pairs of words or terms. This distance is calculated by counting how often the terms occur in the web (page counts) using the Google search engine, and therefore, this distance is also called the Google distance, although it is far from being a metric (see details, e.g., in Poland, & Zeugmann, 2006 or Cilibrasi, & Vitányi, 2007). The data set analyzed in this study consists of the Google distances between 125 famous people classified into five groups of 25 individuals: composers, artists, last year's bestselling authors, mathematicians and pop music performers. The K-means algorithm described in Sect. 2 was applied to this set of dissimilarities.

For this data set, the CH^* index reached its optimal value at $K = 2$, while the H^* rule suggested a value of $K = 5$ (this represents the minimum value for which the H^* statistic is less than 625, for $N = 125$). Although the Silhouette statistic also suggested five clusters, the corresponding average value was less than 0.04, which indicates that the choice of the appropriate

number of clusters under this criterion is misleading, as observed by Kaufman and Rousseeuw (1990).

Thus, only the adapted Hartigan criterion H^* reliably identified the known number of groups into which the people are classified. Table 4 shows the clustering results of the people in $K = 5$ groups. Of the 125 names, only 12 were misclassified with respect to the known groups. The largest number of misclassified names was in the least popular groups, i.e., bestselling authors (7 misclustered) and mathematicians (4 misclustered). The seven bestselling authors (Brown, Frey, Warren, Oz, Sparks, Roberts and Lewis), two mathematicians (Pascal and Newton) and one composer (Cage) who were misclassified were assigned to cluster 5, the pop music performers. The other two misclassified mathematicians (Kantor and Wiles) were assigned to Group 3, composed mainly of bestselling authors. On the other hand, all the names in the group of artists were classified in cluster 2. A similar partition into five classes was obtained by Poland and Zeugmann (2006) by applying the spectral clustering method. In this case, of the 125 names, 9 were misclassified, with the mathematicians producing the highest number of names that were misclassified (four names).

The analysis of the cluster structure in the real-world example confirms the validity of the H^* rule to identify the appropriate number of clusters from a dissimilarity matrix, which does not necessarily represent Euclidean distances between clustered objects, with a low level of overlap.

6. Conclusions

This paper proposes a methodology to determine the correct number of clusters in a K-means framework, when the only available information is a dissimilarity matrix between the objects. The decomposition of the total variability in terms of the block-shaped partition derived for the dissimilarity matrix is considered, and the formulation of variance-based criteria in terms of this decomposition is proposed to determine the number of clusters on a dissimilarity matrix. A similar algorithm to that used in the SYNCLUS method (DeSarbo et al., 1984) is considered as a K-means procedure for dissimilarities, adapting the existing cluster selection criteria for two-mode data.

Besides the Kaufman and Rousseeuw (1990) Silhouette plot method, which can be directly applied in terms of dissimilarities, the formulation of the Calinski and Harabasz (1974) criterion and of Hartigan's rule (1975) is proposed in terms of the decomposition of the variability derived from the partition given for $\mathbf{\Delta}$. The formulation of a variance-based criteria in terms of dissimilarities also makes this advisable for two-mode data sets.

A comprehensive simulation study to analyze the performance of these criteria in determining the appropriate number of clusters was carried out in terms of Euclidean distances. The results obtained for the proposed criteria were compared with those obtained using the classical formulation for the original two-mode data sets, in addition to the jump method (Sugar, & James, 2003) and the Krzanowski and Lai test (Krzanowski, & Lai, 1985). Since the proposed methodology is different from the classical one, the results need not be equivalent. An important finding derived from this analysis, perhaps due to the implicit reduction in dimensionality, is that in general, the efficiency in recovering the number of clusters of the variance-based criteria tested increases when they are applied from the Euclidean distances rather than from the original data matrix $\mathbf{X}$, when this is expressed in low dimensionality. The results obtained from the direct application of the proposed procedure to simulated dissimilarity data sets were compared with the number of clusters predicted when the classical CH and Hartigan criteria were calculated by replacing the Euclidean distances by the dissimilarities in their formulation. The performance of the proposed criteria is also illustrated for real dissimilarities.

In terms of two-mode data sets, the results obtained show that the CH* index is highly successful at selecting the correct number of clusters, outperforming the other approaches examined. In terms of the Euclidean distance matrix, in general the adapted CH* method obtained results

similar to those of the usual CH index when applied to the original two-mode data sets. For K-means clustering on a two-mode data matrix, the results obtained suggest that using the adapted CH* index from a derived auxiliary one-mode Euclidean distance matrix may be an appropriate procedure to determine the optimum number of clusters, particularly in low dimensions. In terms of dissimilarities, a considerable improvement in the performance of the proposed formulation for the Hartigan criterion was obtained, and the proposed H^* index was the best procedure to determine the correct number of clusters, even when some degree of overlapping was present, which makes this procedure appropriate for experimental situations.

As is well known, the performance of different methods employed to determine the number of clusters depends on the cluster method used. In this paper, variance-based methods for dissimilarities were formulated in a K-means framework using SYNCLUS, which is a K-means procedure for Euclidean distances. Nevertheless, other cluster procedures in a generalized K-means framework can also be employed, and the performance of the proposed formulation for variance-based criteria in terms of dissimilarity data, in conjunction with this clustering procedure, is currently being investigated by the authors.

Acknowledgments

The authors would like to thank the Editor, the AE, and two anonymous referees for valuable criticisms and suggestions that greatly improved the paper. This work has been partially supported by Grant ECO2013-48413-R (J. Fernando Vera) of the Ministerio de Economía y Competitividad of Spain (co-financed by FEDER), and by Grant CB-2015-02-252996 (Rodrigo Macías) of CONACYT.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

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

Chiang, M. M., & Mirkin, B. (2010). Intelligent choice of the number of cluster in K-means clustering: An experimental study with different cluster spreads. *Journal of Classification*, *27*, 3–40.

Cilibrasi, R., & Vitanyi, P. (2004). Automatic meaning discovery using Google. http://xxx.lanl.gov/abs/cs.CL/0412098.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(3), 370–383.

Condon, E., Golden, B., Lele, S., Raghavan, S., & Wasil, E. (2002). A visualization model based on adjacency data. *Decision Support Systems*, *33*, 349–362.

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

Elvevag, B., & Storms, G. (2003). Scaling and clustering in the study of semantic disruptions in patients with Schizophrenia: A re-evaluation. *Schizophrenia Research*, *63*, 237.

Everit, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis. Wiley series in probability and statistics* (5th ed.). New York: Wiley.

Gower, J. C., & Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(4), 505–519.

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

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

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

Ito, K., Zeugmann, T., & Zhu, Y. (2010). Clustering the normalized compression distance for Influenza virus data. *Lecture Notes in Computer Science*, *6060*, 130–146.

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

Krzanowski, W. J., & Lai, Y. T. (1985). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, *44*, 23–34.

McQueen, J. (1967). Some methods for classification and analysis of multivariate observations, In *Fifth Berkeley Symposium on Mathematical Statistics and Probability* (vol. II, pp. 281–297).

Maitra, R., & Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, *19*(2), 354–376.

Melnykov, V., Chen, W.-C., & Maitra, R. (2012). MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, *51*(12), 1–25.
Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika*, *50*(1), 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.
Mirkin, B. (2005). *Clustering for data mining: A data recovery approach*. Boca Raton, FL: Chapman and Hall.
Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*, 91–118.
Poland, J., & Zeugmann, T. (2006). Clustering the google distance with eigenvectors and semidefinite programming. In *Knowledge Media Technologies, First International Core-to-Core Workshop*, "Diskussionsbeiträge, Institut für Medien und Kommunikationswisschaft" (vol. 21, pp. 61–69). Technische Universität Ilmenau.
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.). New York: Cambridge University Press.
Priem, R. L., Love, L., & Shaffer, M. A. (2002). Executives perceptions of uncertainty sources: A numerical taxonomy and underlying dimensions. *Journal of Management*, *28*, 725–746.
Rocci, R., & Vichi, M. (2008). Two-mode multi-partitioning. *Computational Statistics and Data Analysis*, *52*, 1984–2003.
Schwarz, A. J. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
Sokal, R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.
Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*, 1–34.
Steinley, D., & Brusco, M. J. (2007). Initializing *K*-means batch clustering: A critical evaluation of several techniques. *Journal of Classification*, *24*, 99–121.
Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Asssociation.*, *98*, 750–762.
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.
Vera, J. F., Macías, R., & Angulo, J. M. (2008). Non-stationary spatial covariance structure estimation in oversampled domains by cluster differences scaling with spatial constraints. *Stochastic Environmental Research and Risk Assessment*, *22*, 95–106.
Vera, J. F., Macías, R., & Angulo, J. M. (2009). A latent class MDS model with spatial constraints for non-stationary spatial covariance estimation. *Stochastic Environmental Research and Risk Assessment*, *23*(6), 769–779.
Vera, J. F., Macías, R., & Heiser, W. J. (2009). A latent class multidimensional scaling model for two-way one-mode continuous rating dissimilarity data. *Psychometrika*, *74*(2), 297–315.
Vera, J. F., Macías, R., & Heiser, W. J. (2013). Cluster differences unfolding for two-way two-mode preference rating data. *Journal of Classification*, *30*, 370–396.

Manuscript Received: 24 OCT 2013
Final Version Received: 28 APR 2016
Published Online Date: 13 FEB 2017