Journal of Classification 32:268-284 (2015)
DOI: 10.1007/s00357-015-9178-y

On the Added Value of Bootstrap Analysis for *K*-Means Clustering

Joeri Hofmans

Vrije Universiteit Brussel; KU Leuven

Eva Ceulemans

KU Leuven

Douglas Steinley

University of Missouri

Iven Van Mechelen

KU Leuven

Abstract: Because of its deterministic nature, *K*-means does not yield confidence information about centroids and estimated cluster memberships, although this could be useful for inferential purposes. In this paper we propose to arrive at such information by means of a non-parametric bootstrap procedure, the performance of which is tested in an extensive simulation study. Results show that the coverage of hyper-ellipsoid bootstrap confidence regions for the centroids is in general close to the nominal coverage probability. For the cluster memberships, we found that probabilistic membership information derived from the bootstrap analysis can be used to improve the cluster assignment of individual objects, albeit only in the case of a very large number of clusters. However, in the case of smaller numbers of clusters, the probabilistic membership information still appeared to be useful as it indicates for which objects the cluster assignment resulting from the analysis of the original data is likely to be correct; hence, this information can be used to construct a partial clustering in which the latter objects only are assigned to clusters.

Keywords: *K*-means; Bootstrapping; Clustering.

The research reported in this paper was supported by the Fund for Scientific Research – Flanders (Project G.0146.06) and by the Research Fund of K.U.Leuven (GOA/2005/04 and GOA/2010/02).

Corresponding Author's Address: J. Hofmans, Vrije Universiteit Brussel, Faculty of Psychology and Educational Sciences, Work and Organizational Psychology, Pleinlaan 2, 1050 Brussels, Belgium, email: joeri.hofmans@vub.ac.be.

Published online: 8 July 2015

1. Introduction

Given object by variable data, researchers are often interested in retrieving clusters of objects that behave similarly on the variables under study. For example, when patients are tested on a set of clinical variables, one may wish to identify clusters of patients with similar symptom profiles (which subsequently may be conceived to correspond with certain clinical syndromes). Alternatively, in consumer psychology, researchers may want to identify clusters of consumers with a similar spending pattern, which then allows for a segmentation of the market into different consumer types. Finally, in organizational research, one may want to find groups of employees who score similarly on a range of performance criteria in order to utilize them optimally for the organization.

Several types of clusterings exist, with the simplest type being a partitioning, that is, a subdivision of the group of objects into a set of mutually exclusive clusters, with each object belonging to a single cluster only. *K*-means (MacQueen 1967; Sebestyen 1962) is a well-known method to arrive at such a partitioning, and is frequently used in practice (Kogan 2007; Steinley 2006). For a given data matrix $\boldsymbol{X}_{I\times J}$ with I objects taking scores on J variables, *K*-means searches for a binary partitioning matrix $\boldsymbol{P}_{I\times K}$ (which represents the memberships of the I objects to the K clusters) and a centroids matrix $\boldsymbol{C}_{K\times J}$ (which contains the centroids for the K clusters). The pair of a partition and a centroid matrix are chosen such that the following least squares loss function is minimized:

$$\min_{P,C} \|\mathbf{X} - \mathbf{PC}\|^2. \tag{1}$$

Minimizing loss function (1) implies that the centroid for some cluster equals the vector of means on all variables taken across all objects belonging to that cluster.

Because of its deterministic nature, *K*-means does not yield confidence information about estimated cluster memberships and centroids. Yet, such confidence information could be useful for inferential purposes. Given this limitation, one may consider two possible options:

A first option is to shift to a stochastic method. A popular method of choice in this respect is mixture modeling (McLachlan and Peel 2000). Such an approach yields valuable confidence information. However, there are also some disadvantages associated with mixture modeling. In particular, a standard likelihood-based estimation of mixture models does not directly yield discrete membership information; rather, from the estimated parameters one may derive probabilistic membership information in the form of posterior probabilities (the conversion of which into discrete memberships requiring yet another transformation, based on

Bayes' rule). To obtain directly discrete membership information in the estimation of mixture models, one could adopt a classification-likelihood approach; this approach, however, is not free from problems, as its results have been shown to be asymptotically biased (e.g., Bryant and Williamson 1978).

A second option is to stay within a deterministic framework and to search for a way to obtain confidence information. In the present paper, we explore this second alternative. To arrive at such information, we appeal to a non-parametric bootstrap procedure. Nonparametric bootstrapping implies the creation of a pseudo sampling distribution for an estimate under study by drawing (bootstrap) samples with replacement from the original data (Efron 1979; Efron and Tibshirani 1993). In the past, nonparametric bootstrapping has been successfully applied to various methods of multivariate analysis. Examples include work on canonical variate analysis by Krzanowski (1989), on three-way methods by Kiers (2004), on linear and nonlinear principal component analysis by Timmerman, Kiers, and Smilde (2007) and Linting, Meulman, Groenen, and Van der Kooij (2007), respectively, and on multilevel simultaneous component analysis by Timmerman, Smilde, Kiers, Ceulemans, and Stouten (2009). With regard to *K*-means cluster analysis, a number of authors proposed to use nonparametric bootstrapping to choose the number of clusters, *K* (e.g., Jain and Moreau 1987; Maitra, Melnikov, and Lahiri 2012). In the present paper, we focus on the application of nonparametric bootstrap analysis to *K*-means to obtain confidence information.

To do so, we start by sampling with replacement *I* objects from the original $I \times J$ object-by-variable data matrix $\boldsymbol{X}_{I\times J}$, which then yields a bootstrap sample $\boldsymbol{X}^{b}_{I\times J}$. Next, each of the $b = 1,\ldots,B$ bootstrap samples is subjected to a *K*-means analysis, with some prespecified number of clusters *K*, and for each bootstrap sample a set of (pseudo) parameter estimates –in our case, the centroid matrix ($\boldsymbol{C}^{b}_{K\times J}$)– is saved. Subsequently, the results of the bootstrap analyses are to be integrated across the bootstrap samples (somewhat similar in spirit to an integration across locally optimal solutions – see Steinley 2008). In particular, in our case, the centroid matrices resulting from the bootstrap analyses are combined across bootstrap samples to yield (uni- or multivariate) confidence regions. In addition, the centroid matrix obtained for each bootstrap sample can also be used to construct a partitioning of all *I* objects of the original data set (including those that do not belong to the bootstrap sample) by assigning each object to the cluster with the closest centroid; subsequently, the resulting cluster memberships can be integrated across bootstrap samples by calculating for each object the proportion of bootstrap samples according to which it belongs to each of the clusters. Obviously, the calculation of confidence regions and membership

proportions requires that the centroids of the different bootstrap samples have been properly aligned beforehand. This implies that a solution is needed for the well-known label switching problem with the cluster labels resulting from a *K*-means analysis being only unique up to a permutation (similar as in mixture models: e.g., Jasra, Holmes, and Stephens 2005).

A key question with regard to non-parametric bootstrapping as applied to *K*-means pertains to the quality of the resulting confidence information. Unfortunately, however, previous research does not yield a satisfactory answer to this question. More specifically, up to now, only three studies have addressed this issue. In a first study, Monti, Tamayo, Mesirov, and Golub (2003) applied subsampling (i.e., sampling without replacement a subset of objects from the original dataset) to a *K*-means cluster analysis. However, they did not inspect the subsampling confidence information as such; rather, they combined the partitioning matrices $\boldsymbol{P}^{b}_{I\times K}$ into a dissimilarity matrix, to subsequently subject this to a new cluster analysis. In a second study (Dudoit and Fridlyand 2003), nonparametric bootstrapping was applied to Kaufman and Rousseeuw's (1990) Partitioning Around Medoids or PAM method. This study further also included a very brief evaluation of the quality of the bootstrap confidence results. In particular, Dudoit and Fridlyand (2003) found that the proportion of objects that is assigned to a wrong cluster is lower for a clustering that is obtained by applying Bayes' rule to the nonparametric bootstrap membership proportions than for a clustering on the basis of the original full data set. Moreover, the bootstrap confidence information was shown to be a good indicator of the accuracy of the cluster assignments. In a third study, Jhun (1990) compared the consistency of confidence intervals obtained with nonparametric bootstrapping with those obtained via asymptotical methods (see Pollard 1982), from which he concluded that the bootstrap works better than those asymptotical methods. Yet, the problem with these studies is that in the first study the subsampling output is regarded as an intermediary product, rather than as an informative end-product. In the second study, *K*-medoids instead of *K*-means was studied; furthermore, the bootstrap results were evaluated on the basis of a small number of datasets only, which makes the conclusions of the evaluation tentative. In the third study, a comparison with asymptotical methods is available, but this comparison relies on a very small simulation study. As a consequence, the quality of nonparametric bootstrap results within a *K*-means context remains an open issue.

In the present paper, we will address this issue by means of an extensive simulation study. The remainder of this paper is structured as follows: In Section 2 the research questions are introduced. Section 3 and 4 describe the data generation and data analysis respectively. In Section 5

the results of the simulation study are reported. In Section 6 a number of concluding remarks are given.

2. Research Questions

To clarify the research questions, we first distinguish between three different pairs of matrices: (1) the true centroid matrix $\boldsymbol{C}_{K\times J}^{T}$ and the true partitioning matrix $\boldsymbol{P}_{I\times K}^{T}$, containing the true centroids and true cluster memberships respectively; in a simulation study both matrices are known; (2) the centroid matrix $\boldsymbol{C}_{K\times J}^{D}$ and partitioning matrix $\boldsymbol{P}_{I\times K}^{D}$ that result from a *K*-means analysis applied to the data matrix $\boldsymbol{X}_{I\times J}$; (3) the b^{th} bootstrap centroid matrix $\boldsymbol{C}_{K\times J}^{b}$ ($b = 1,\ldots,B$) that results from a *K*-means analysis of the b^{th} bootstrap sample $\boldsymbol{X}_{I\times J}^{b}$ (obtained by sampling with replacement from the rows of $\boldsymbol{X}_{I\times J}$) along with the partitioning matrix $\boldsymbol{P}_{I\times K}^{b}$ that is obtained by assigning each object from the full data set $\boldsymbol{X}_{I\times J}$ to the cluster with the closest centroid.

Making use of these matrices, two research questions will be answered: 1. Concerning the centroids, we want to examine the quality of the bootstrap centroid confidence intervals that can be derived from the $\boldsymbol{C}_{K\times J}^{b}$ matrices. More specifically, we want to know how often those bootstrap confidence intervals contain the true centroids (i.e., the values in $\boldsymbol{C}_{K\times J}^{T}$). 2. Concerning the cluster memberships, we want to test whether the probabilistic cluster memberships that can be derived from the $\boldsymbol{P}_{I\times K}^{b}$ matrices can be used to construct a deterministic cluster assignment that resembles the true underlying memberships ($\boldsymbol{P}_{I\times K}^{T}$) better than the cluster assignment obtained from the analysis of the original data matrix ($\boldsymbol{P}_{I\times K}^{D}$).

3. Data Generation

The data $\boldsymbol{X}_{I\times J}$ are simulated using Milligan's (1985) well-known cluster generation method. In particular, the data are generated as follows: (1) The number of clusters (2, 4, or 6), the number of observations (100, 200, or 500), the number of variables (4 or 8), and the homogeneity level of cluster size (all clusters of equal size, 10% of the observations in one cluster and the other clusters of equal size, 60% of the observations in one cluster and the other clusters of equal size) are specified. (2) For each variable, lower and upper boundary values are generated for each cluster. In the actual data generation (Step 3), the midpoint between the lower and upper boundaries will act as the mean of the data generating distribution for the cluster in question, and the difference between the two boundaries (which will further be called the boundary length) will be put equal to three times the standard deviation of this distribution. Because cluster

overlap is not allowed on the first variable, the boundary values on this variable are generated in such a way that the different clusters are well separated. In particular, firstly a random ordering of the clusters is specified. Next, the boundary length for each cluster on this variable is drawn from a uniform distribution with support ranging from 10 to 40. The cluster boundaries of neighboring clusters i and j are further separated by the random quantity $f \times (S_i + S_j)$, with S_i and S_j being the standard deviations of the generating distributions of clusters i and j respectively, and f being drawn from a uniform random variable with a support ranging from .25 to .75. For the remaining variables, the boundary lengths are again randomly drawn from a uniform distribution with support ranging from 10 to 40. Given these boundary lengths, the midpoints between the upper and lower boundaries are randomly located on each of the remaining dimensions, subject to the constraint that all boundary points are located within an interval with a length equal to two-thirds of the range of the first variable (which implies that on all dimensions, except for the first, cluster overlap is possible). (3) The data points for each cluster are generated from a multivariate normal distribution with a diagonal variance-covariance matrix, and with mean vector and standard deviations as specified above. Points that fall outside the cluster boundaries are not accepted. (4) Outliers are generated for each cluster, with the number of outliers being proportional to the size of the cluster. Either 20% or 40% of the observations are generated as outliers. These are drawn from a multivariate normal distribution with the same vector of means as in Step 3 and with the same variance-covariance matrix, yet now multiplied by a factor of nine. A point is retained as an outlier if it falls outside the boundaries on at least one dimension of the variable space. At this moment, the true data have been constructed with known matrices $\boldsymbol{P}_{I\times K}^{T}$ and $\boldsymbol{C}_{K\times J}^{T}$, with $\boldsymbol{C}_{K\times J}^{T}$ containing the means of the multivariate normal distributions that have been used to generate the data points. (5) Noise is added to the data points. Noise is generated from a multivariate normal distribution with mean zero and with the same variance-covariance matrix as in Step 3, and subsequently multiplied by .2, .6, and 1 to generate three different levels of noise (for an illustration, see Figure 1). (6) All variables are standardized, thereby yielding the data matrix $\boldsymbol{X}_{I\times J}$. The true centroid matrix $\boldsymbol{C}_{K\times J}^{T}$ is rescaled accordingly using the same linear transformations.

In summary, the data are generated according to a fully crossed factorial design in which 6 factors are systematically manipulated:

1. Number of clusters (i.e., K): 2, 4 and 6
2. Number of observations (i.e., I): 100, 200, and 500
3. Number of variables (i.e., J): 4 and 8

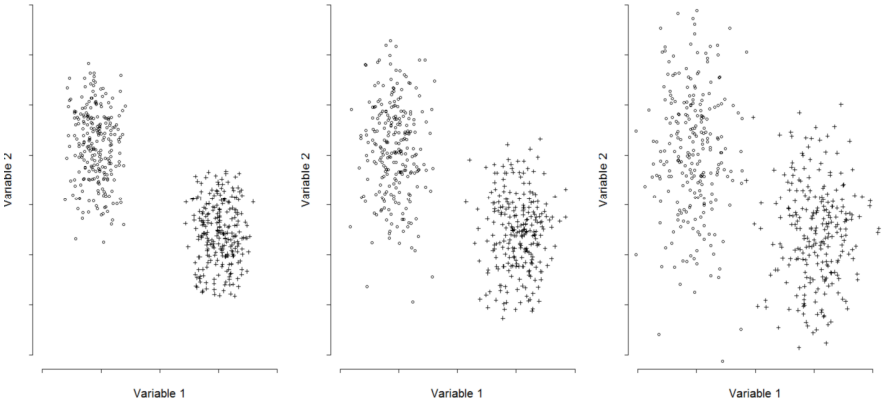


Figure 1: Scatterplots in which the effect of the error manipulation on a two-cluster solution is shown. The only difference between the three scatterplots is that the errors are multiplied by .2, .6, and 1 in the left, middle, and right panel respectively.

4. Homogeneity level of cluster size: equal size, 10% in one cluster and the other clusters of equal size, 60% in one cluster and the other clusters of equal size
5. Percentage of outliers: 20% or 40% of the observations.
6. Noise level: multiplication by .2, .6, and 1.

Within each cell of the design, 10 different datasets were generated, resulting in a total of 3,240 simulated datasets.

In addition, to test the performance of our bootstrap procedure for data sets with a large number of clusters, we also generated data with 50 clusters, 500 observations, 8 variables, and equally sized clusters. The number of outliers (20% or 40%) and the degree of noise (multiplication by .2, .6, and 1) were manipulated, and for each cell of the design, 10 different datasets were generated. This resulted in 60 additional datasets.

4. Data Analysis

Algorithm

In the present study, we adopted the *K*-means algorithm as implemented in MATLAB. This algorithm operates according to the following iterative procedure: (Step 1) *Initialization stage*: Initialize **C** by randomly selecting *K* objects as centroids. (Step 2) *Update of* ***P***: For each object *i* and for each cluster *k*, the fit between the object and the cluster is calculated by:

$$d^2(i,k) = \sum_{j=1}^{J} \left(x_{ij} - c_{kj}\right)^2; \quad (2)$$

Subsequently, each object is allocated to the cluster for which (2) is minimal. If this step results in an empty cluster, a new cluster is created consisting of the object that fits its current cluster worst. (Step 3) *Update of* $\boldsymbol{C}$: New centroids are calculated on the basis of the updated cluster memberships. (Step 4) *Loop*: Steps 2 and 3 are repeated until no further objects can be reallocated. (Step 5) *Online-updates of* $\boldsymbol{P}$ *and* $\boldsymbol{C}$: This step is similar to Step (2), but now the cluster (re)allocation of each object is the one that yields the lowest value for the overall *K*-means loss function (1), taking into account the changes in all centroids implied by a possible reallocation. After each object reallocation, the centroids are immediately updated. This step is repeated until none of the objects is reallocated.

To lower the risk of ending in a local optimum, for each analysis 100 runs of the *K*-means algorithm were performed, each with a different random initialization of the centroid matrix (see Hand and Krzanowski 2005; Steinley 2003). The run that minimized loss function (1) was retained.

Bootstrap Procedure

The bootstrap procedure runs according to the following sequence: (Step 1) The original data $\boldsymbol{X}_{I\times J}$ are analyzed with *K*-means for a given value of *K*, thereby resulting in $\boldsymbol{P}^{D}_{I\times K}$ and $\boldsymbol{C}^{D}_{K\times J}$. (Step 2) From the full data matrix $\boldsymbol{X}_{I\times J}$, *I* objects are drawn with replacement, and this is done 1,000 times, yielding 1,000 bootstrap samples $\boldsymbol{X}^{b}_{I\times J}$. (Step 3) Each of these bootstrap samples is subjected to a *K*-means analysis with the same value for *K*, thereby resulting in a centroid matrix $\boldsymbol{C}^{b}_{K\times J}$ for b = 1 ,…, 1,000. (Step 4) To align the centroids of the 1000 bootstrap samples (and to deal with the label switching problem) the *K* clusters of each bootstrap solution are permuted such that there is maximal correspondence between $\boldsymbol{C}^{b}_{K\times J}$ and $\boldsymbol{C}^{D}_{K\times J}$ (in the least squares sense)[1]. To find the (globally optimal) solution for this alignment problem we rely on the implementation by Munkres (1957) of the so-called Hungarian algorithm (Munkres 1957).

[1] One might argue that the quality of our proposed bootstrap procedure might be affected by the choice of the criterion that is optimized to deal with the alignment issue (including the label switching problem). Furthermore, it is possible that two additional obstacles to cluster alignment also have an impact on this quality, viz., that (1) the bootstrap procedure may lead to the generation of 'false' clusters based on bootstrap-induced replicates rather than true classes (Monti et al. 2003; Möller and Radke 2006), and that (2) it might mask true clusters. Yet, it is exactly the quality of the procedure as a whole, including the way in which the alignment and label switching problems are dealt with, that will be investigated in the present simulation study.

(Step 5) The permuted centroid matrix for the b^{th} bootstrap sample is then used to construct a matrix $\boldsymbol{P}^{b}_{I\times K}$ with cluster membership values for all I objects in the original data matrix, by assigning each object to the cluster with the closest centroid. (Step 6) For the centroids, bootstrap confidence intervals are constructed using two methods. First, we calculated a bias-corrected and accelerated (BC_a) confidence interval for each element of each centroid, resulting in a combined confidence region in the form of a hyperbox. Second, we implemented a method in which we constructed the confidence intervals by drawing hyper ellipsoids around the cluster centroids (see Linting et al. 2007 for a similar procedure). To draw these hyper ellipsoids, we relied on the covariances calculated between each pair $(c^{b}_{kj}, c^{b}_{kj'})$ of elements of the centroid vector for the k^{th} cluster across all bootstrap samples. Regarding the cluster memberships, starting from the matrices $\boldsymbol{P}^{b}_{I\times K}$ we calculated for each object i and for each cluster k the proportion of matrices in which object i belongs to cluster k. This information was stored in the probabilistic membership matrix $\boldsymbol{P}^{A}_{I\times K}$. MATLAB code for performing this bootstrap procedure can be obtained from the first author.

5. Results

First, we report on a manipulation check, in which we study how the recovery of the clustering and the centroids is affected by the manipulated data characteristics. Subsequently, we turn to the actual research questions where we first discuss the quality of the bootstrap centroid confidence intervals. Finally, we evaluate whether the probabilistic cluster memberships resulting from the bootstrap procedure can be used to improve the initial deterministic cluster assignment.

K-means Analysis of the Generated Data Matrices: Impact of Manipulated Data Characteristics

To study the impact of the manipulated data characteristics on the recovery of the clusterings, we compared the partitioning matrices $\boldsymbol{P}^{D}_{I\times K}$ obtained from the analyses of the generated data with the true partitioning matrices $\boldsymbol{P}^{T}_{I\times K}$. This comparison was done by computing the Adjusted Rand Index (ARI), which is a measure of similarity of two partitionings that takes on a value of 1 when there is perfect agreement and a value of 0 when the similarity is at chance level (Hubert and Arabie 1985). In the 2-, 4-, and 6-cluster conditions, the average ARI was .849 (SD = .160). Moreover, recovery increased if the number of clusters increased ($F(2,2916)=886.78$; $p<.001$; $\eta^2_p=.378$), the number of variables increased ($F(1,2916)=528.21$; $p<.001$; $\eta^2_p=.153$), the amount of outliers decreased ($F(1,2916)=474.54$; $p<.001$; $\eta^2_p=.140$), the amount of noise decreased

($F(2,2916)=13.02$; $p<.001$; $\eta^2_p=.009$), and the observations were more equally distributed across clusters ($F(2,2916)=50.84$; $p<.001$; $\eta^2_p=.034$). In the 50-cluster conditions, the ARI was on average .855 (SD = .024), and there was no effect of the amount of noise ($F(2,54)=.31$; $p=.734$; $\eta^2_p=.011$), nor of the number of outliers ($F(1,54)=.085$; $p=.772$; $\eta^2_p=.002$). These effects are in line with effects that are generally found in similar simulation studies (see Steinley 2003, 2006; Steinley and Brusco 2007).

With respect to the recovery of the centroids, the mean squared distance between the true centroids and the centroids derived from the data (with the mean being calculated across clusters and datasets) equaled .041 on average (SD = .11) for the 2-, 4, and 6-cluster conditions. This distance decreased if the number of clusters increased ($F(2,2916)=3.94$; $p=.020$; $\eta^2_p=.003$), the number of variables increased ($F(1,2916)=53.98$; $p<.001$; $\eta^2_p=.018$), the number of observations increased ($F(2,2916)=31.66$; $p<.001$; $\eta^2_p=.021$), the amount of outliers decreased ($F(1, 2916)=12.44$; $p<.001$; $\eta^2_p=.004$), and the observations were more equally distributed across clusters ($F(2, 2916)=41.97$; $p<.001$; $\eta^2_p=.028$). For the 50-cluster conditions, the mean squared distance between the true centroids and the centroids derived from the data (i.e., across clusters and variables) was on average .157 (SD = .026), and was not affected by the amount of noise ($F(2,54)=.63$; $p=.537$; $\eta^2_p=.023$), or the number of outliers ($F(1,54)=.03$; $p=.861$; $\eta^2_p=.001$).

Coverage Probability of Bootstrap Confidence Intervals for Centroids

For the element-wise confidence intervals, the average coverage probability (coverage being calculated for each variable separately) of the 95% bias-corrected and accelerated bootstrap confidence intervals equaled .895 (SD = .095) for the 2-, 4-, and 6-cluster conditions[2]. This coverage probability was affected by the number of variables ($F(1,2916)=210.85$;

[2] We calculated an index of bias of the cluster means by first computing the element-wise squared difference between the data-based centroids and the average of the bootstrapped centroids, and by then averaging these squared distances across all elements of the difference matrix. For the 2-, 4, and 6-cluster conditions, this bias equaled .009 on average (SD = .030) and was affected by the number of observations ($F(2,2916)=103.69$; $p<.001$; $\eta^2_p=.066$), the distribution of the observations across clusters ($F(4, 2916)=50.80$; $p<.001$; $\eta^2_p=.034$), the interaction between the number of clusters and the distribution of the observations across clusters ($F(4, 2916)=74.39$; $p<.001$; $\eta^2_p=.093$), and the interaction between the number of clusters, the number of observations, and the distribution of the observations across clusters ($F(8, 2916)=20.19$; $p<.001$; $\eta^2_p=.052$). In particular, the bias decreased when the number of observations increased and when the observations were more equally distributed across clusters. The effects of the other data characteristics, as well as their interactions were small (i.e., $\eta^2_p<.030$).

$p<.001$; $\eta^2_p=.067$) and the number of observations ($F(2,2916)=79.90$; $p<.001$; $\eta^2_p=.052$); the effects of the other data characteristics, as well as their interactions were small (i.e., $\eta^2_p<.030$). In Figure 2, the average coverage probability is shown as a function of the number of observations and the number of variables. It can be seen that in the conditions with 100 observations and 8 variables, the average coverage is relatively close to the desired 95%, whereas it decreases if the number of observations increases and the number of variables decreases. This pattern can be interpreted as follows: On the one hand all variables yield some information on the clustering; thus, the clusters and the centroids are recovered worse when the number of variables decreases. On the other hand, more observations lead to smaller confidence intervals, irrespective of the number of variables. These small confidence intervals are problematic if there is less information on the clustering (as is the case with less variables), because in that case the clustering results may become inconsistent across the bootstrap samples. For the 50-cluster conditions, the average coverage probability of the 95% bias-corrected and accelerated bootstrap confidence intervals was .869 (*SD* = .029), and this was not affected by the amount of noise ($F(2,54)=.07$; $p=.936$; $\eta^2_p=.002$), or the number of outliers ($F(1,54)=.98$; $p=.326$; $\eta^2_p=.018$).

For the hyper-ellipsoid confidence regions, the average coverage probability for the 2-, 4-, and 6-cluster conditions (coverage again being calculated for each variable separately) equaled .943 (*SD* = .070), and was affected by the number of clusters ($F(2,2916)=201.05$; $p<.001$; $\eta^2_p=.121$), homogeneity of cluster size ($F(2,2916)=59.28$; $p<.001$; $\eta^2_p=.039$), the number of clusters by homogeneity of cluster size interaction ($F(4,2916)=33.26$; $p<.001$; $\eta^2_p=.044$), and the number of variables by number of observations interaction ($F(2,2916)=58.54$; $p<.001$; $\eta^2_p=.039$). The effects of the other data characteristics, as well as their interactions were small (i.e., $\eta^2_p<.030$). The effect of the combination of the number of clusters and homogeneity of cluster size can be seen in Figure 3. In almost all conditions, coverage was close to the desired 95%. However, a substantial decrease in coverage probability showed up in conditions with one large and many small clusters. For the 50-cluster conditions, the coverage probability of the hyper-ellipsoid confidence regions was on average .919 (*SD* = .020), and did not depend on the amount of noise ($F(2,54)=1.14$; $p=.326$; $\eta^2_p=.041$), or the number of outliers ($F(1,54)=2.11$; $p=.153$; $\eta^2_p=.038$).

Potential of probabilistic Cluster Memberships Derived From Bootstrap Analyses

In this section we will address the question whether a better deterministic cluster assignment can be constructed on the basis of the

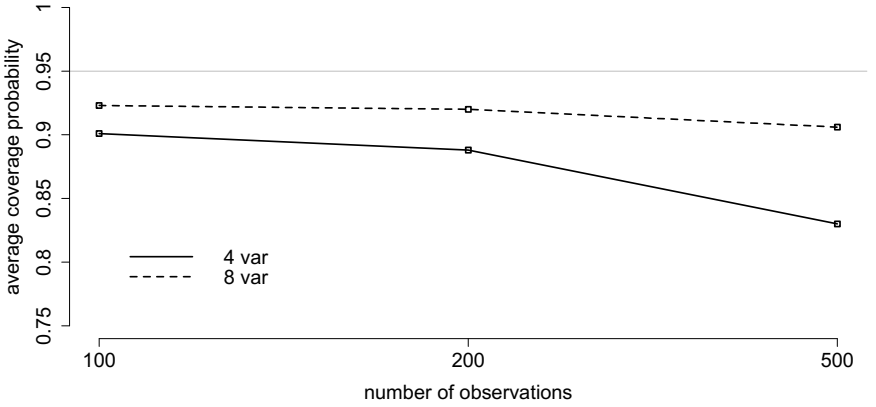


Figure 2: Average coverage probability for the element-wise 95% confidence intervals as a function of the number of observations (on the X-axis) and the number of variables (separate lines). The horizontal grey line shows the nominal 95% coverage probability.

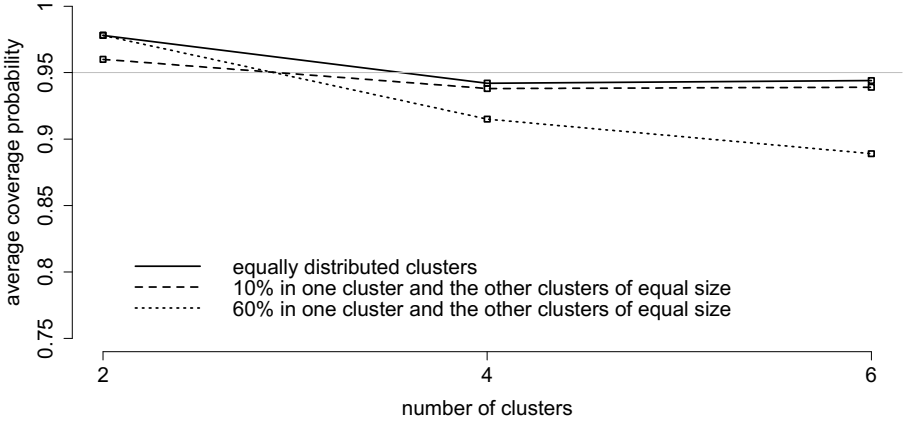


Figure 3: Average coverage probability for the 95% confidence regions as a function of the number of clusters (on the X-axis) and the cluster density (separate lines). The horizontal grey line shows the nominal 95% coverage probability.

probabilistic cluster memberships in $\boldsymbol{P}^{A}_{I\times K}$ derived from the bootstrap analyses. To answer this question, we tested whether there was a better correspondence between the true partitioning $\boldsymbol{P}^{T}_{I\times K}$ and a partitioning based on a dichotomization of the probabilistic cluster membership matrix $\boldsymbol{P}^{A}_{I\times K}$ than between $\boldsymbol{P}^{T}_{I\times K}$ and the partitioning based on the data $\boldsymbol{P}^{D}_{I\times K}$. A straightforward way to dichotomize the probabilistic cluster membership matrix $\boldsymbol{P}^{A}_{I\times K}$ is to apply Bayes' rule, which implies that object *i* is assigned

to the cluster for which $\mathbf{P}_{ik}^{\mathrm{A}} = \max_{\mathrm{k}=1}^{\mathrm{K}} \mathbf{P}_{ik}^{\mathrm{A}}$. We conducted an ANOVA with $\boldsymbol{P}_{I\times K}^{D}$ versus dichotomized $\boldsymbol{P}_{I\times K}^{A}$ as the within-subjects factor, the data characteristics as the between-subjects factors, and recovery of the clustering as measured by the ARI as the dependent variable[3]. For the 2-, 4-, and 6-cluster conditions, this analysis revealed that cluster recovery for the analysis based on the data did not significantly differ from the recovery based on the dichotomized probabilistic cluster memberships derived from the bootstrap analyses ($F(1,3239)=.24$; $p=.622$; $\eta^2_p<.001$). However, for the 50-cluster conditions, cluster recovery was significantly better for the bootstrap-based memberships than for the memberships stemming from the analysis of the data ($F(1,59)=313.12$; $p<.001$; $\eta^2_p=.841$), with average recoveries of .904 (*SD* = .020) and .855 (*SD* = .024) respectively.

Because for the 2-, 4-, and 6-cluster conditions cluster recovery did not improve by dichotomizing the probabilistic membership matrix by means of Bayes' rule, we further examined whether the values in the probabilistic membership matrix $\mathbf{P}_{ik}^{\mathrm{A}}$ were meaningfully related to recovery of the true cluster memberships. The effect was done by applying a restricted version of Bayes' rule in which object *i* is only assigned to a cluster if the maximal $\mathbf{P}_{ik}^{\mathrm{A}}$ value for that object exceeds a prespecified threshold. We further varied the thresholds for each data set in such a way that exactly 95, 90, 85, 80, and 75% of the objects were assigned. To evaluate the performance of this partial assignment rule, the ARI was computed for each threshold separately. From Figure 4, it can be seen that the mean and median ARI increased when less objects were assigned (with all pairwise comparisons being statistically significant except for the one between the 85% and 80% conditions). A similar, yet more extreme, pattern of findings was obtained for the 50-cluster conditions, with mean ARI's of .932 (*SD* = .024), .954 (*SD* = .024), .974 (*SD* = .023), .984 (*SD* = .019), and .991 (*SD* = .014) when assigning 95, 90, 85, 80, and 75% of the objects respectively. Note that the gain in recovery is not due to an improved cluster assignment of the assigned objects, but rather to the omission of objects with a bad cluster assignment. In sum, we conclude that the bootstrap information can be used to identify objects with a better cluster assignment.

[3] Note that the ARI measures what Efron and Tibshirani (1993) refer to as 'prediction error'. Traditionally, the prediction error would be expressed as the percentage of incorrect classifications. However, one important problem with these classification rates is that they do not correct for chance-based correct classification (Steinley 2004). By correcting for correct classification due to chance, the ARI allows for valid comparisons of cluster recovery between solutions with different numbers of clusters (Hubert and Arabie 1985).

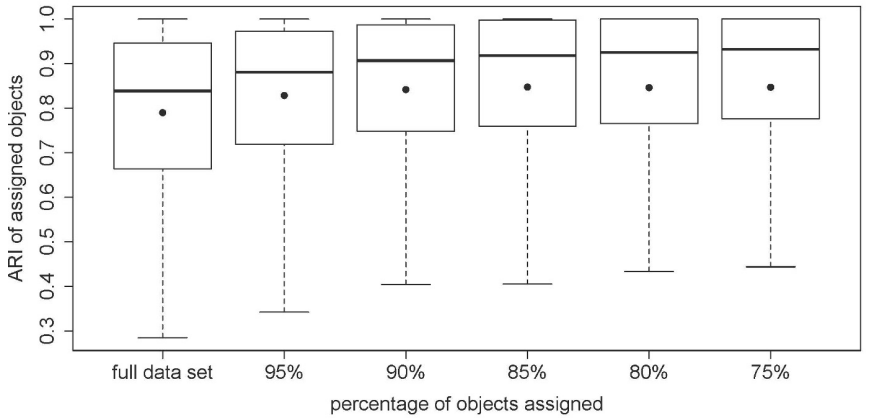


Figure 4: Recovery performance of partial cluster assignment derived from bootstrap analyses as a function of the percentage of assigned objects. The points in the boxes represent the average recovery, whereas the lines represent the median recovery. Note that in the analysis of the full data set, all objects are assigned.

Finally, the effect of the data characteristics on the gain in recovery was tested by performing an ANOVA with the difference in cluster recovery between the data-based assignment and the bootstrap-based 75% partial assignment as dependent variable, and the data characteristics as independent variables. Moreover, we also included the quality of the clustering resulting from the data-based analysis (in terms of the ARI) as an additional predictor. This analysis revealed that the impact of the data characteristics and of the quality of the data-based clustering result on the gain in recovery was small (with for all effects $\eta^2_p < .04$). A similar result was found when the dependent variable was the difference in recovery between the data-based assignment and the bootstrap-based 95% partial assignment (all η^2_p's $\leq .02$).

6. Concluding Remarks

The proposed bootstrap procedure yields confidence information about the centroids in the form of confidence intervals. Regarding the element-wise confidence intervals, the quality of these intervals is reasonably good, although the average coverage probability is lower than the nominal level in most conditions. Furthermore, coverage probability appeared to depend primarily on two data characteristics, that is, the number of variables and the number of observations. Fortunately, these characteristics are manifest, which implies that users can use them to assess whether bootstrap-based confidence intervals can be expected to work for a particular problem at hand. From their part, confidence regions

based on hyper ellipsoids around the centroids appear to perform even better: In almost all conditions the coverage is close to the nominal coverage probability of 95%. Downward deviations from this nominal level show up when there is one big cluster together with many very small clusters. The latter is a latent characteristic and therefore not readily accessible to the user. In this case the effect is less troublesome, though, as coverage is in general good.

Regarding cluster memberships, we derived probabilistic membership information from the bootstrap analysis. This was done in a conceptually more sound way than in other bootstrap-based approaches in which membership information on some object is only derived from the bootstrap samples that include the object in question (e.g., Dudoit and Fridlyand 2003). In contrast, our approach relies on the bootstrap distribution of the centroid estimates, which is subsequently used to derive probabilistic membership information on each object from all bootstrap samples. We further showed that the resulting information can be used to improve the cluster assignment of individual objects, yet only when the number of clusters is very large. However, in case of smaller numbers of clusters, the probabilistic membership information still appeared to be useful as it indicates for which objects the cluster assignment resulting from the analysis of the original data is likely to be correct. This information may then be used to improve the recovery of the true clustering by only retaining the objects that are classified in the same way across bootstrap samples. Otherwise, the fact that the gain in recovery due to removing part of the objects was virtually unaffected by the manipulated characteristics, suggests that the procedure works equally well across a wide range of data sets.

With regard to possible directions for future research, one might first note that our simulation study was confined by the specifics of the data-generating mechanism proposed by Milligan. In future research, it may be worthwhile to investigate other data-generating mechanisms, which may yield more complex clustering structures. Indeed, part of our results could perhaps have been hampered by ceiling effects, and one may wonder whether for more complex data a bootstrap-based procedure could lead to more sizeable gains (as we now already saw in the case of a large number of clusters).

Second, it might also be interesting to evaluate the usefulness of other types of resampling techniques. As mentioned before, a bootstrap procedure may cause the generation of 'false' clusters based on bootstrap-induced replicates (Monti et al. 2003; Möller and Radke 2006); moreover, it may also mask true clusters (e.g., if the obtained clustering comprises a false cluster resulting from bootstrap-induced replicates, at least one of the true clusters will not be retrieved). Obviously, these problems could

hamper the cluster alignment along the bootstrap replicates and, hence, the quality of the confidence and probabilistic membership information resulting from the bootstrap procedure. One possible way to resolve this problem could be to resort to subsampling, that is, to drawing objects without replacement from the original data. The use of this sampling scheme avoids a possible bias due to artificial replicates (Monti et al. 2003). On the downside, however, it requires the specification of an additional parameter that refers to the size of the fraction that is to be drawn from the original data, which is often not an easy task (Möller and Radke 2006). Anyhow, further research is needed to assess the performance of such a subsampling procedure.

A third and final possible direction for future research could pertain to a comparative evaluation of the performance of bootstrap-based extensions of *K*-means with that of stochastic approaches based on mixture models. This comparison may focus on the quality of the inferential information resulting from both types of approaches.

References

BRYANT, P.G., and WILLIAMSON, J.A. (1978), "Asymptotic Behavior of Classification Maximum Likelihood Estimates", *Biometrika, 65*, 273–281.

DAVISON, A.C., and HINKLEY, D.V. (1997), *Bootstrap Methods and Their Application*, Cambridge UK: Cambridge University Press.

DUDOIT, S., and FRIDLYAND, J. (2003), "Bagging to Improve the Accuracy of a Clustering Procedure", *Bioinformatics, 19*, 1090–1099.

EFRON, B. (1979), "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics, 7*, 1–26.

EFRON, B., and TIBSHIRANI, R. (1993), *An Introduction to the Bootstrap*, London: Chapman and Hall.

HAND, D.J., and KRZANOWSKI, W.J. (2005), "Optimising *K*-Means Clustering Results With Standard Software Packages", *Computational Statistics and Data Analysis, 49*, 969–973.

HENNIG, C. (2007), "Cluster-Wise Assessment of Cluster Stability", *Computational Statistics and Data Analysis, 52*, 258–271.

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

JAIN, A.K. (2010), "Data Clustering: 50 Years Beyond *K*-Means", *Pattern Recognition Letters, 31*, 651–666.

JAIN, A.K., and MOREAU, J.V. (1987), “Bootstrap Technique in Cluster Analysis”, *Pattern Recognition*, *20*, 547–568.

JASRA, A., HOLMES, C.C., and STEPHENS, D.A. (2005), "Markow Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling", *Statistical Science, 20*, 50–67.

JHUN, M. (1990), "Bootstrapping *K*-Means Clustering", *Journal of the Japanese Society for Computational Statistics, 3*, 1–14.

KAUFMAN, L., and ROUSSEEUW, P.J. (1990), *Finding Groups in Data*, New York: Wiley.

KIERS, H.A.L. (2004), "Bootstrap Confidence Intervals for Three-way Methods", *Journal of Chemometrics, 18*, 22–36.

KOGAN, J. (2007), *Introduction to Clustering Large and High-Dimensional Data*, Cambridge: Cambridge University Press.

KRZANOWSKI, W.J. (1989), "On Confidence Regions in Canonical Variate Analysis", *Biometrika, 76*, 107–116.

LINTING, M., MEULMAN, J.J., GROENEN, P.J.F., and VAN DER KOOIJ, A.J. (2007), "Stability of Nonlinear Principal Components Analysis: An Empirical Study Using the Balanced Bootstrap", *Psychological Methods, 12*, 359–379.

MACQUEEN, J. (1967), "Some Methods of Classification and Analysis of Multivariate Observations", in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

MAITRA, R., MELNYKOV, V., and LAHIRI, S.N (2012), "Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets", *Journal of the American Statistical Association*, *107*, 378–392.

MCLACHLAN, G.J., and PEEL, D. (2000), *Finite Mixture Models*, New York: Wiley.

MILLIGAN, G.W. (1985), "An Algorithm for Generating Artificial Test Clusters", *Psychometrika, 50*, 123–127.

MÖLLER, U., and RADKE, D. (2006), "Performance of Data Resampling Methods for Robust Class Discovery Based on Clustering", *Intelligent Data Analysis, 10*, 139–162.

MONTI, S., TAMAYO, P., MESIROV, J., and GOLUB, T. (2003), "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data", *Machine Learning, 52*, 91–118.

MUNKRES, J. (1957), "Algorithms for the Assignment and Transportation Problems," *Journal of the Society for Industrial and Applied Mathematics*, *5(1),* 32–38.

POLLARD, D. (1982), "A Central Limit Theorem for *K*-Means Clustering", *Annals of Probability, 10*, 919–926.

SEBESTYEN, G.S. (1962), *Decision Making Processes in Pattern Recognition*, New York: Macmillan.

STEINLEY, D. (2003), "Local Optima in *K*-Means Clustering: What You Don't Know May Hurt You", *Psychological Methods, 8*, 294–304.

STEINLEY, D. (2004), "Properties of the Hubert-Arabia Adjusted Rand Index", *Psychological Methods, 9*, 386–396.

STEINLEY, D. (2006), "*K*-Means Clustering: A Half-Century Synthesis", *British Journal of Mathematical and Statistical Psychology, 59*, 1–34.

STEINLEY, D. (2008), "Stability Analysis in *K*-Means Clustering", *British Journal of Mathematical and Statistical Psychology*, *61*, 255–273.

STEINLEY, D., and BRUSCO, M.J. (2007), "Initializing *K*-Means Batch Clustering: A Critical Evaluation of Several Techniques", *Journal of Classification, 24*, 99–121.

STEINLEY, D., and BRUSCO, M.J. (2011), "Evaluating Mixture Modeling for Clustering: Recommendations and Cautions", *Psychological Methods, 16*, 63–79.

TIMMERMAN, M.E., KIERS, H.A.L., and SMILDE, A.K. (2007), "Estimating Confidence Intervals for Principal Component Loadings: A Comparison Between the Bootstrap and Asymptotic Results", *British Journal of Mathematical and Statistical Psychology, 60*, 295–314.

TIMMERMAN, M.E., KIERS, H.A.L., SMILDE, A.K., CEULEMANS, E., and STOUTEN, J. (2009), "Bootstrap Confidence Intervals In Multi-Level Simultaneous Component Analysis", *British Journal of Mathematical & Statistical Psychology, 62*, 299–318.