Journal for Person-Oriented Research
2021; 7(1): 22-35
Published by the *Scandinavian Society for Person-Oriented Research*
Freely available at https://journals.lub.lu.se/jpor and https://www.person-research.org
https://doi.org/10.17505/jpor.2021.23449

How to Use Model-Based Cluster Analysis Efficiently in Person-Oriented Research

Bence Gergely[1,2] and András Vargha[1,3]

[1] Károli Gáspár University, Budapest, Hungary
[2] University of Amsterdam, Amsterdam, The Netherlands
[3] Eötvös Loránd University, Budapest. Hungary

Corresponding author:
András Vargha
Email: vargha.andras@kre.hu

To cite this article:
Gergely, B., & Vargha, A. (2021). How to use model-based cluster analysis efficiently in person-oriented research. *Journal for Person-Oriented Research, 7*(1), 22-35. https://doi.org/10.17505/jpor.2021.23449

Abstract: Model-based cluster analysis (MBCA) was created to automatize the often subjective model-selection procedure of traditional explorative clustering methods. It is a type of finite mixture modelling, assuming that the data come from a mixture of different subpopulations following given distributions, typically multivariate normal. In that case cluster analysis is the exploration of the underlying mixture structure. In MBCA finding the possible number of clusters and the best clustering model is a statistical model-selection problem, where the models with differing number and type of component distributions are compared. For fitting a certain model MBCA uses a likelihood based Bayesian Information Criterion (BIC) to evaluate its appropriateness and the model with the highest BIC value is accepted as the final solution. The aim of the present study is to investigate the adequacy of automatic model selection in MBCA using BIC, and suggested alternative methods, like the Integrated Completed Likelihood Criterion (ICL), or Baudry's method. An additional aim is to refine these procedures by using so called quality coefficients (QCs), borrowed from methodological advances within the field of exploratory cluster analysis, to help in the choice of an appropriate cluster structure (CLS), and also to compare the efficiency of MBCA in identifying a theoretical CLS with those of various other clustering methods. The analyses are restricted to studying the performance of various procedures of the type described above for two classification situations, typical in person-oriented studies: (1) an example data set characterized by a perfect theoretical CLS with seven types (seven completely homogeneous clusters) was used to generate three data sets with varying degrees of measurement error added to the original values, and (2) three additional data sets based on another perfect theoretical CLS with four types. It was found that the automatic decision rarely led to an optimal solution. However, dropping solutions with irregular BIC curves, and using different QCs as an aid in choosing between different solutions generated by MBCA and by fusing close clusters, optimal solutions were achieved for the two classification situations studied. With this refined procedure the revealed cluster solutions of MBCA often proved to be at least as good as those of different hierarchical and *k*-center clustering methods. MBCA was definitely superior in identifying four-type CLS models. In identifying seven-type CLS models MBCA performed at a similar level as the best of other clustering methods (such as *k*-means) only when the reliability level of the input variables was high or moderate, otherwise it was slightly less efficient.

Keywords: mixture models, person-oriented methods, model-based cluster analysis, integrated completed likelihood criterion, Baudry's method

Introduction

Model-based cluster analysis (MBCA) was created to automatize the demanding model-selection procedure of traditional explorative clustering methods (e. g., hierarchical and *k*-means clustering). With these methods the

standard procedure for selecting the number of clusters is first selecting a set of quality coefficients (QCs) and assessing the given cluster structure (CLS) by evaluating whether most of these QCs reach a predetermined cutoff value (Vargha et al., 2016; Vargha & Bergman, 2019).

The problem with this approach is the researcher's subjectivity. A problem with selecting the optimal number of clusters is the wide variety of rather heuristic approaches. By using different methods researchers may obtain entirely different solutions, thus the replicability of these studies could be questionable. Instead of using a cutting point approach as with the QCs and the MORI coefficients (see Vargha & Bergman, 2019), MBCA offers a basically automatized way of choosing the best solution of clustering.

Model-based cluster analysis

MBCA is a type of finite mixture modelling, assuming that the data come from a mixture of different subpopulations following given distributions, typically multivariate normal. In that case cluster analysis is the exploration of the underlying mixture structure. According to Fraley and Raftery (2002) finding the possible number of clusters and the best clustering model is a statistical model-selection problem, where the models with differing number and type of component distributions are compared. For fitting a certain model MBCA uses an iterative Expectation-Maximization (EM) algorithm initialized by hierarchical model-based clustering, and the BIC criterion to evaluate its appropriateness (Hasnat et al., 2016; Scrucca et al., 2016; Vaithyanathan & Dom 2013). According to Fraley et al. (2012) the larger the value of the BIC[1], the stronger the evidence for the model and the number of clusters.

Once we fitted a mixture model with k components, a probabilistic clustering of the data into k clusters can be obtained by using the posterior probabilities of component memberships. Unlike traditional clusterings, MBCA uses a soft assignment, where each data point (case) in the multi-dimensional space has a probability of belonging to each cluster (fuzzy clustering). By means of these probabilities we can fix cluster belongingness by assigning each case to the cluster of a component distribution that generates the case's expression profile with maximum probability.

Selecting the best cluster solution

Fraley and Raftery (2002) proposed a solution to the problem of choosing the right number of clusters. After conducting the analysis for a set of cluster numbers (i.e., 1-10) and structure types (see, e.g., Scrucca et al. 2016), one can decide the best cluster solution by plotting the BIC curve and searching local maximums of BIC. A decisive first local maximum indicates strong evidence for a model (Fraley & Raftery, 1998). This process allows a certain automatization of finding the best cluster solution, without the need of using heuristic approaches like the elbow-method (Myers, 1996) on one of the QCs (Vargha & Bergman, 2019).

Biernacki et al. (2000) assert that clustering in MBCA is not merely choosing the best approximating mixture model, but rather finding a meaningful taxonomy of the data, and so BIC may not be ideal to determine the number of clusters. They argue that the goal of selecting the number of mixture components for estimating the underlying probability density is well met by BIC (Dasgupta & Raftery, 1998). It may occur, however, that a non-Gaussian cluster will be represented by a mixture of two or more Gaussian distributions, leading to the overestimation of the number of clusters (Baudry et al., 2008; Bertoletti et al., 2015).

To overcome this problem, Biernacki et al. (2000) proposed the integrated completed likelihood (ICL) criterion, which is the BIC penalized through an entropy term which measures cluster overlaps. Here they assumed that every object belongs to one distribution with a probability of 1 and to all others with a probability of 0. This way ICL can be regarded as an adjusted maximized log-likelihood criterion, which takes the special clustering goal of MBCA into account. Therefore, the ICL-based MBCA should be regarded as a cluster-oriented approach whereas the BIC-based MBCA is a component-oriented approach. According to Biernacki et al. (2000), ICL may outperform the traditional BIC-based evaluation when the data consist of non-Gaussian components.

Baudry et al. (2010) suggested another method combining the principles of the BIC and the ICL approach. Their proposed algorithm consists of the following steps. First, one fits a Gaussian mixture model, and using BIC, selects the best number of components. Second, one hierarchically combines mixture components of this solution using the entropy of the conditional membership distributions to decide which two components to merge at each stage. Last, one should inspect the entropy plot and looking for an elbow (see, e.g., Figure 5).

The purpose of the paper

In a perfect classification every object belongs to one of k possible classes in the p-dimensional space of the p input variables, and all objects in the same class have the same value pattern, falling into the same p-dimensional point. This theoretical, 'true' data set is regarded to be error free and can be defined by k discrete points in the p-dimensional space, where these k points, the theoretical centroids, define k different types (Vargha and Bergman, 2019). All objects falling into one such p-dimensional centroid is regarded as a theoretical cluster. In practice, such theoretical clusters do not exist because variables defining them are always sub-

[1] We note that the BIC in this approach has an opposite sign than in its usual formula (see, e.g., Wit, Heuvel, & Romeijn, 2012), that is why one looks for BIC maximum in MBCA and not BIC minimum.

mitted to some degree of measurement error. If these errors are independent and normally distributed, the multivariate distribution of the empirical data set will follow a mixture of *p*-dimensional normal distributions with *k* components, whose centers are the *k* theoretical centroids (Vargha & Bergman, 2019).

If we select a set of variables for submitting to cluster analysis in a certain population, three situations are possible:

A. There is a CLS in the population with well-identifiable types that can be more or less hidden by error.

B. There is a partial CLS in the population, where some clear types exist, but they do not cover the whole population.

C. No real CLS and no real types exist.

The purpose of the present paper was to clarify the usefulness of different MBCA methods (BIC- and ICL-based, and Baudry's method) for automatic model selection in MBCA when situation A holds, and to give usable instructions for determining the optimal number of types in a typical person-oriented scenario. If the automatic interpretation of MBCA does not recover the given structure, an important question is what indicators can help in identifying the true model and number of clusters. An additional aim was to refine these MBCA procedures by analyzing BIC-curves and using QCs to help in the choice of an acceptable cluster structure, and also to compare the efficiency of MBCA in identifying a theoretical CLS with the efficiency of different other clustering methods.

Method

In the following, we will create two concrete theoretical CLSs with some continuous unobservable theoretical traits, one based on a real empirical study, the other being an artificial one. Then we define observable variables (measuring tools) to each trait that include a certain amount of error and see how the theoretical CLS can be identified by MBCA, using either the BIC criterion, the ICL criterion, or Baudry's method.

The selection of theoretical cluster structures

In a sociolinguistic study of Romanian ethnic minority living in Hungary, Vargha and Borbély (2017) applied five sociolinguistic variables (minority language competence, language use in family and church, minority identity, and attitude toward minority language). They explored seven clusters with attractive QCs and MORI indices (see Vargha & Borbély 2017). The explored clusters identified seven types of speakers which coincides with the clearly not linear process of language shift and assimilation from a bilingual minority status to a monolingual Hungarian status. Based on these positive results we chose our artificial theoretical CLS the centroids of this solution[2] (see Figure 1 and Appendix 2).

The second theoretical structure, borrowed from Bergman et al. (2017), was a simpler one, having four artificially defined types with four variables (see Figure 2).

Data generation based on theoretical cluster structures

The applied theoretical and empirical samples were the same as the ones in Vargha and Bergman (2019). The only difference was that Vargha and Bergman (2019) studied the adequacies of conventional hierarchical and *k*-means solutions in terms of several QCs and MORI coefficients based on them, whereas the present study focused on the adequacy of MBCA solutions.

The first theoretical sample (Teo7types) was defined the following way. It consisted of 7 clusters based on 5 variables just as in Figure 1. Each theoretical cluster was errorless, consisting of cases with exactly the same value-pattern corresponding to the CLS of Figure 1. Cluster sizes were borrowed from the study of Vargha and Borbély (2017), multiplying those values by 5 (with resulting cluster sizes 195, 160, 175, 115, 95, 50, and 115, respectively), yielding a substantial sample size of 905.

The second theoretical sample (Teo4types) was defined similarly. It consisted of 4 clusters based on 4 variables just as in Figure 2. Each theoretical cluster was errorless, consisting of cases with exactly the same value-pattern corresponding to the CLS of Figure 2. Cluster sizes were borrowed from Bergman et al. (2017): 160, 40, 160, and 40, respectively, for types A to D, to have a sample of moderate size ($N = 400$).

For both theoretical samples three artificial empirical samples (Emp7type1, Emp7type2, and Emp7type3 for Teo7types, and Emp4type1, Emp4type2, and Emp4type3 for Teo4types) were defined the following way. For each original true variable value in the theoretical data set, a new value was created by adding an independent random $N(0; \sigma_i)$, variable, where σ_i was set to .5, .75 and 1 for the three samples, respectively. Then each data value was rounded to the nearest integer. Data values less than 1 or greater than 5 were set to 1 or 5, respectively. This algorithm yielded five-point integer valued variables in the empirical samples, which is a usual case in psychological practice. The three different σ_i values yielded three levels of measurement reliability. These levels were identified by computing r^2 values (explained variance proportions) between the theoretical and the corresponding empirical variables, used as reliability estimates. In Emp7type1 these fall into the .65-.81 region (mean = .74), in Emp7type2 into the .52-.71 region (mean = .61), and in Emp7type3 into the .40-.57 region (mean = .50). Similarly, in Emp4type1 these fall into

[2] The original scales were modified by appropriate linear transformations to scales whose theoretical minimum and maximum values were 1 and 5.

the .71-.86 region (mean = .78), in Emp4type2 into the .53-.72 region (mean = .65), and in Emp4type3 into the .43-.60 region (mean = .50). These three means in both cases represent three (high, moderate, or low) levels of reliability of the empirical variables in the three empirical samples (Cohen, 1977).

Figure 1

The chosen 1st artificial theoretical (Teo7types) CLS with 7 clusters and 5 variables: Design with seven types (TC1 to TC7) and five variables (V1 to V5).

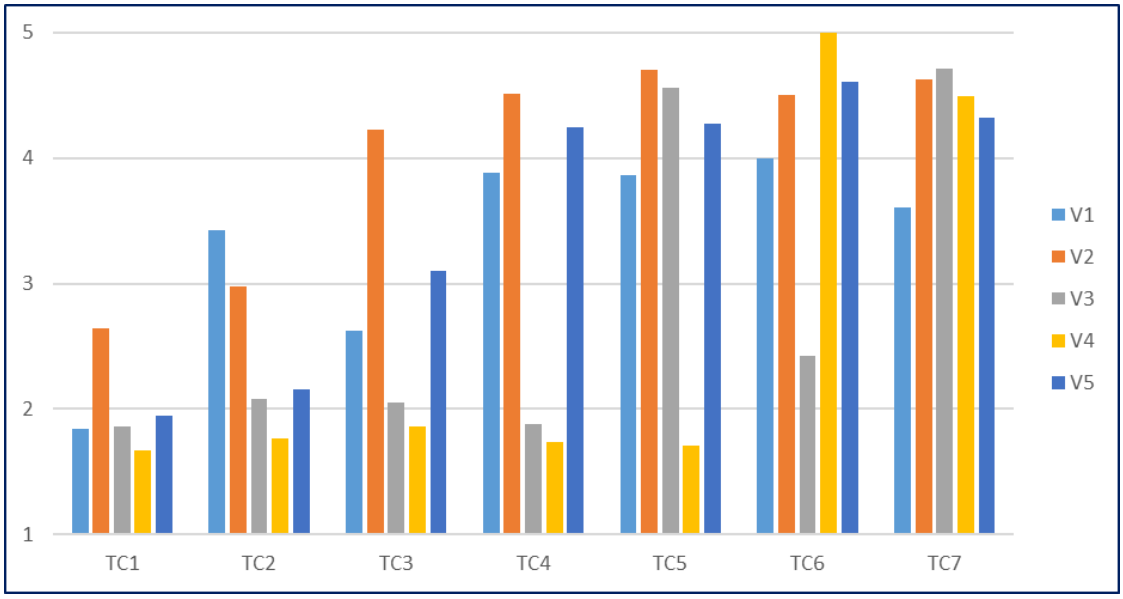
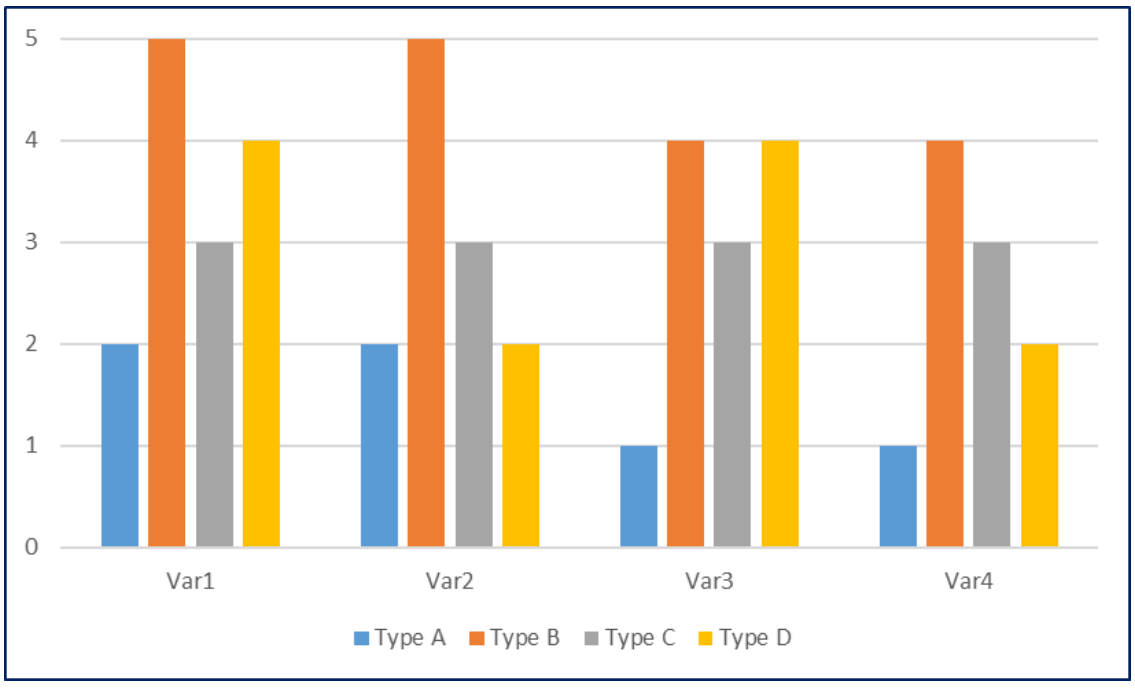


Figure 2

The chosen 2nd artificial theoretical (Teo4types) CLS with 4 clusters and 4 variables: Design with four types (Type A to Type D) and four variables (Var1 to Var4).



The *SD* values varied between 0.99 and 1.24 for the high, between 1.03 and 1.28 for the moderate, and between 1.15 and 1.34 for the low reliability samples, so the variation of the *SD* values can be regarded small. The correlations between the empirical variables were small also within the theoretical types.

The normality of all final variables in each theoretical cluster for both CLS types were checked via sample skewness and kurtosis. For the 7-type situation 24 (22.9%) out of the possible 105 cluster by variable by reliability level combinations either skewness or kurtosis was significant at $p < .001$ level, and for the 4-type situation 14 (23.3%) out of the possible 60 cluster by variable by reliability level combinations either skewness or kurtosis was significant at $p < .001$ level. This means that due to rounding and truncating the resulting variables were definitely nonnormal in a substantial proportion (23%) of cases.

Statistical analyses

We conducted the MBCA analyses with the mclust package in R (Fraley and Raftery 2003), and the post-hoc analyses in ROPstat (Vargha, Torma, and Bergman 2015). We performed MBCA using both the BIC and the ICL criterion, and also with Baudry's method in all six empirical samples. (1) First, we identified the model with the highest BIC and ICL values in the component range $k = 2$ to 10. If this maximum was attained with an irregular curve (due to a sudden jump), the model was excluded and the next best model was selected. (2) Then, saving the promising solution (solutions) we checked whether there are cluster centroids close to each other in a solution, using the Centroid module of ROPstat. If two clusters were close to each other, we combined them. (3) Then, we identified the optimal solution also with Baudry's method. Here we used the cluster number and model from the original MBCA having the highest BIC value, and merged clusters pairwise until only one cluster remained (using the *clustCombi* function of mclust package). (4) Next, we compared the – occasionally improved – empirical cluster solutions with the true theoretical clustering looking for good matches between pairs of centroids of the two CLSs. We computed also the percentage of correctly classified cases to measure the match between the empirical and the theoretical CLSs the following way. An empirical cluster was regarded as a representative of a theoretical cluster (type) that occurred with the highest frequency in that empirical cluster. We summed the frequencies of the represented types over the clusters and divided it by the total sample size. (5) Finally, if there were several competing empirical solutions, we compared and evaluated them using the following QCs according to the suggestions of Vargha et al. (2015, 2016), Vargha and Bergman (2019), and Bergman, Vargha, and Kövi (2017): HC cluster homogeneity coefficients and their average (HCmean), explained error sum of square percentage (EESS%), simplified version of the Silhouette Coefficient (SilCoef), and a generalized Dunn index (GDI24). Their detailed formulas can be found in Appendix 1.

In the mclust package there are 14 available models for more than two groups represented by triplets EII, VII, EEI, VEI, EVI, VVI, EEE, EVE, VEE, EEV, VEV, EVV, and VVV. The first letter in a triplet refers to the cluster size (volume), the second to the shape, and the third to the orientation of the multivariate distribution. *E* always stands for "equal", *V* for "variable" and *I* for "coordinate axes". For example:

- EVI denotes a model in which the volumes of all clusters are equal (*E*), the shapes of the clusters may vary (*V*), and the orientation is the identity (*I*) or coordinate axes;
- EEE means that the clusters have the same size, shape and orientation in the *p*-dimensional space;
- VEI means that the cluster sizes vary, but they have the same shape and the orientation equal to coordinate axes (see in detail in Scrucca et al. 2016).

Results

In the high reliability Emp7type1 sample an EEV (ellipsoidal, equal shape) model with 10 components, denoted by EEV10 (see Figure 3) had the largest BIC value, and using the ICL criterion we had a very similar result. But regarding that the EEV curve has a sudden huge jump from $k = 9$ to $k = 10$, differing strongly from all other curves, we labelled it an irregular one and dropped it. Among the remaining solutions an EII7, an EEI7, and a VII7 model proved to be the best both with regard to the BIC and the ICL criterion (see Figures 3 and 4).

To check the adequacy of these three 7-cluster solutions we compared their centroids with those of the theoretical model (see Figure 1) by means of the Centroid module of ROPstat, and found a high level match for all three solutions. The pairwise ASED differences between corresponding cluster centroids were smaller than 0.016 for all three models. 91.4% of the observations were correctly classified for EII7, 91.7 for EEI7, and 91.4 for VII7, despite the substantial error put on the theoretical scores, the enforced rounding, and the applied restricting 5-point scale.

It should be noted that the EEV10 solution had substantially worse QCs than those of EII7, EEI7, and VII7 (EESS%: 61.8 vs. 75.8-75.9; SilCoef: 0.10 vs. 0.65-0.66; GDI24: 0.03 vs. 1.14-1.20; HCmean: 0.77 vs. 0.48-0.49), showing that it was really inadequate. Baudry's automatic method, starting with this wrong EEV10 solution could not obviously yield an acceptable solution. Baudry's 6-cluster solution (see Figure 5) had only two centroids having an ASED difference less than .10 from the nearest theoretical centers, and its QCs were even worse than those of the EEV10 solution.

In the moderate reliability Emp7type2 sample the best solution was an isolated irregular VEV10 model for both BIC and ICL. Dropping it, two 7-cluster models (EII7 and EEI7) turned out to be the most appropriate ones with BIC.

The ICL-plot showed a totally different picture with a local maximum yielded by the VVI and VVE curves at $k = 3$ (see Figure 6). The QCs of VEV10, VVI3, VVE3 were substantially worse than the QCs of EII7 and EEI7 in terms of HCmean, measuring the average cluster homogeneity (for the former three: 0.98-1.10, for the latter two: 0.67).

Figure 3

The BIC-plot of MBCA curves in the high reliability Emp7type1 sample for k = 1 to 10

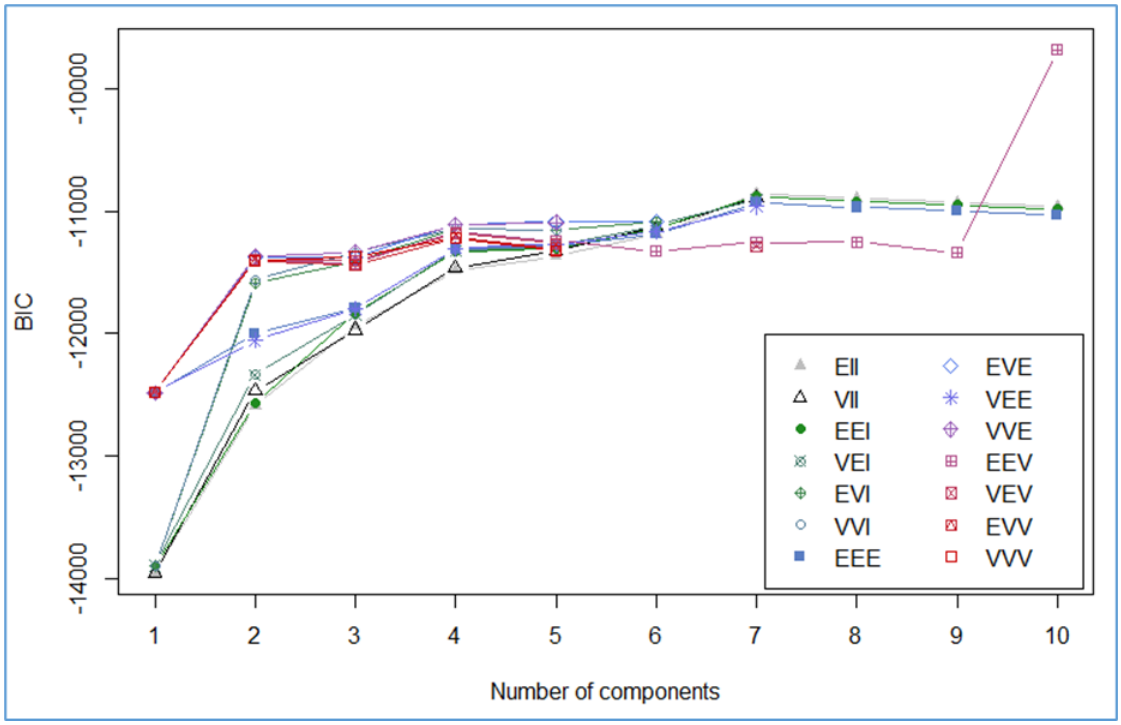


Figure 4

The ICL-plot of the best three MBCA curves in the high reliability Emp7type1 sample after dropping the irregular EEV curve for k = 4 to 10

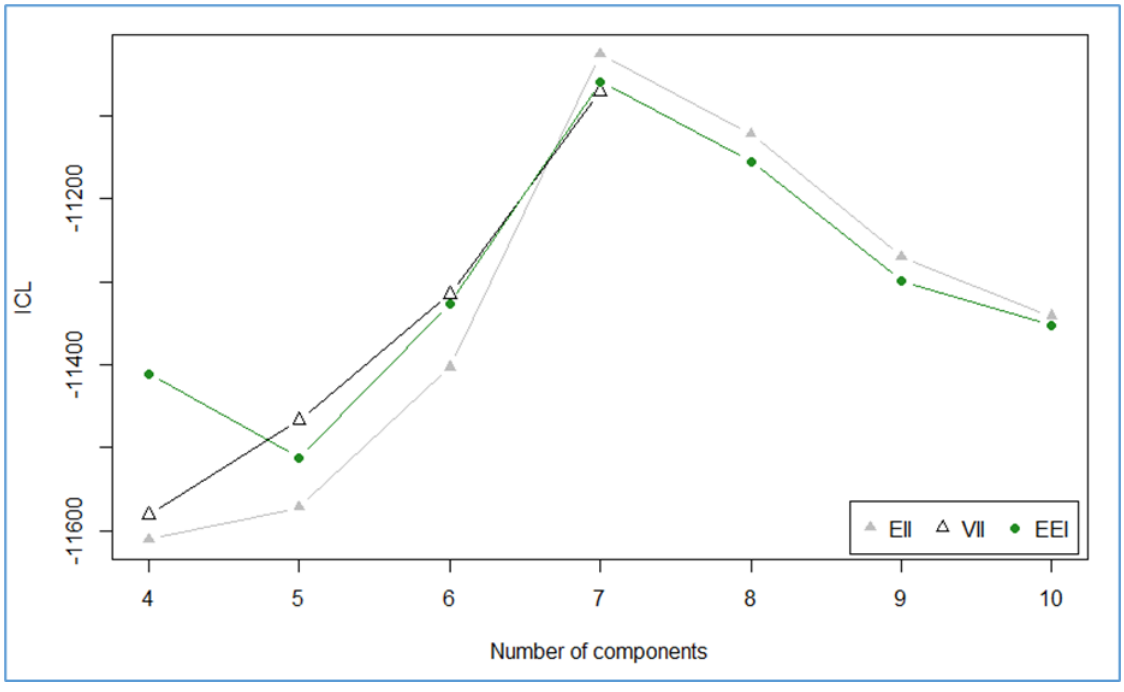


Figure 5
The Entropy plot of Baudry's method in the high reliability Emp7type1 sample

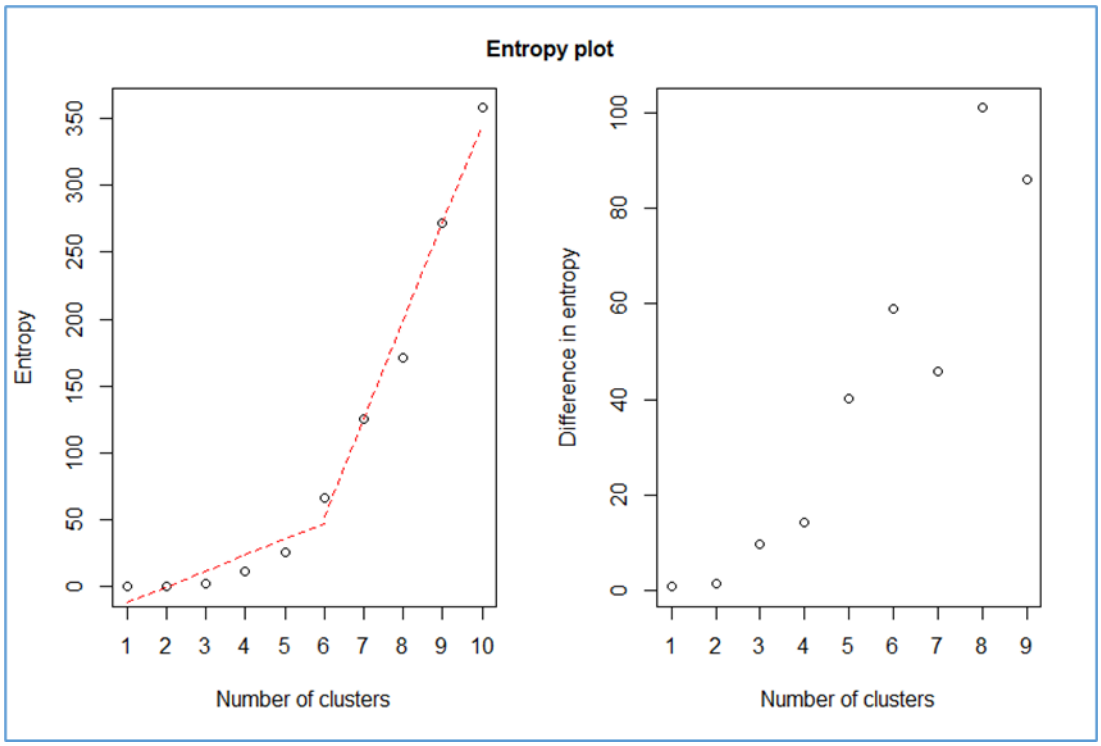
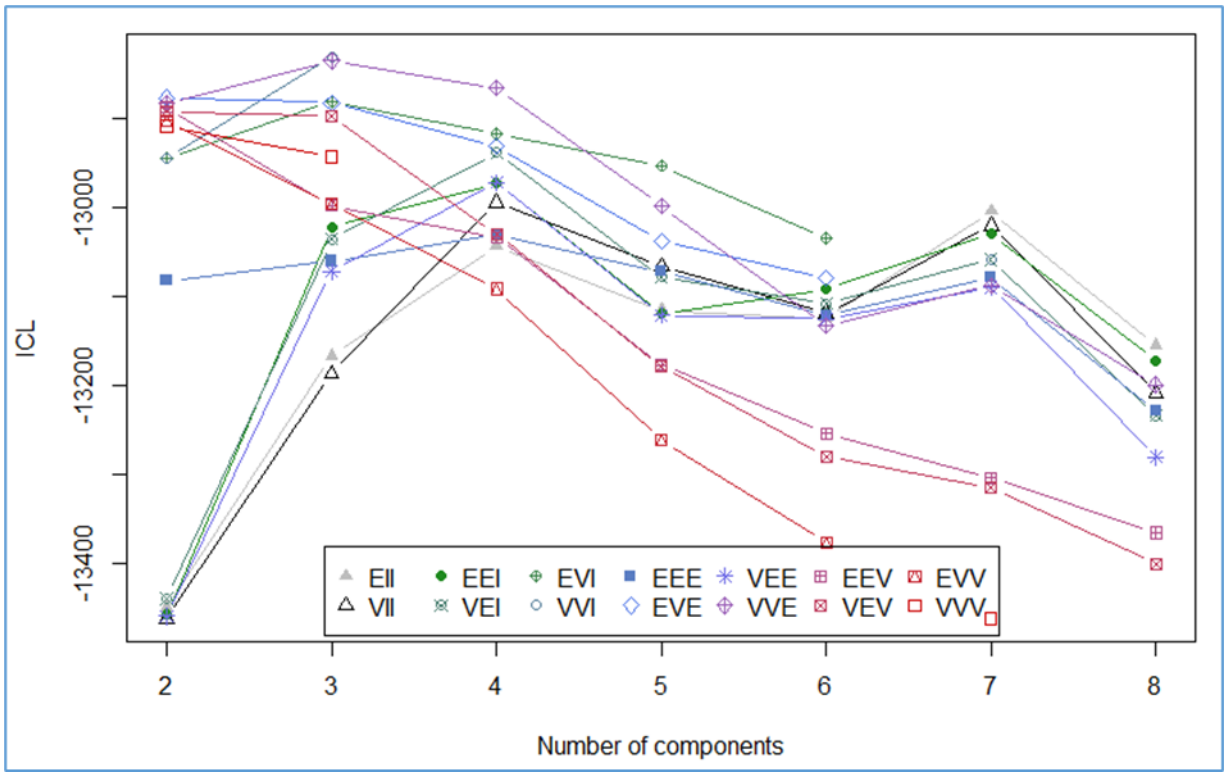


Figure 6
The ICL-plot of MBCA curves in the moderate reliability Emp7type2 sample for k = 2 to 8



We compared the centroids of EII7 and EE7 with the theoretical model as in the high reliability case and found an acceptable match. The pairwise ASED differences between corresponding cluster centroids were never greater than 0.08. Out of the 905 theoretical observations 76.4% were correctly classified for both EII7 and EE7 MBCA solutions, and there were only three theoretical clusters where the percentage of correct identifications did not exceed 85% (they fell between 62.5% and 71.4%).

Baudry's method, starting with the wrong VEV10 solution, could not yield an acceptable final solution. It indicated a 7-cluster solution having only one centroid having an ASED difference less than .10 from the nearest theoretical center, and all of its QCs were substantially worse than those of EII7 and EEI7, especially in terms of GDI24 (.53 vs. .90 and .86).

In the low reliability Emp7type3 sample the automatic decision led to a VEI6 model. Comparing this solution with the theoretical model, two of the six clusters had a good match with two theoretical types (with ASED differences .022 and .046) and two others had an acceptable match (with ASED differences .062 and .101). This shows that if the reliability of the input variables is only at the 50%-level, a partial identification of the original CLS is still possible. The analysis with the ICL criterion yielded a messy picture very similar to the one of Figure 6, having a first high peak at k = 3 on the VVI and VVE curves (see Figure 7). Here again, Baudry's method led to a bad four clusters solution.

In the high reliability Emp4type1 sample the automatic decision with both BIC and ICL led to a wrong model selection with EEV9 and EEV6 yielding decisive local peaks (see Figure 8). After dropping this irregular EEV curve, EII5 and EEI5 seemed to be best models with both BIC and ICL. In the EII5 and EEI5 solutions the first two clusters had similar centroids (with ASED = .45 for both models), so we fused them and obtained a modified 4-cluster solutions from EII5 and EEI5 (denoted by EII5m and EEI5m). The ASED differences between corresponding theoretical and empirical centroids were never greater than 0.02. EII correctly identified 98% of the cases, while EEI correctly identified 97.3%. Baudry's method, starting with the EEV9 model, yielded here a bad 7-cluster solution.

In the moderate reliability Emp4type2 sample the automatic decision led again to a wrong model selection with both BIC and ICL due to a huge jump on the EEV curve at k = 8, similar to the one in Figure 3. After dropping this irregular curve, EII5 and EEI5 seemed to be the best models. In the EEI5 solution one cluster had no element, so it was really a 4-cluster solution, denoted by EEI5(4). In the EII5 solution the first two clusters had similar centroids (with ASED = .47), so we fused them, yielding a 4-cluster modified EII5m solution.

Figure 7

The ICL-plot of MBCA curves in the low reliability Emp7type3 sample for k = 1 to 10

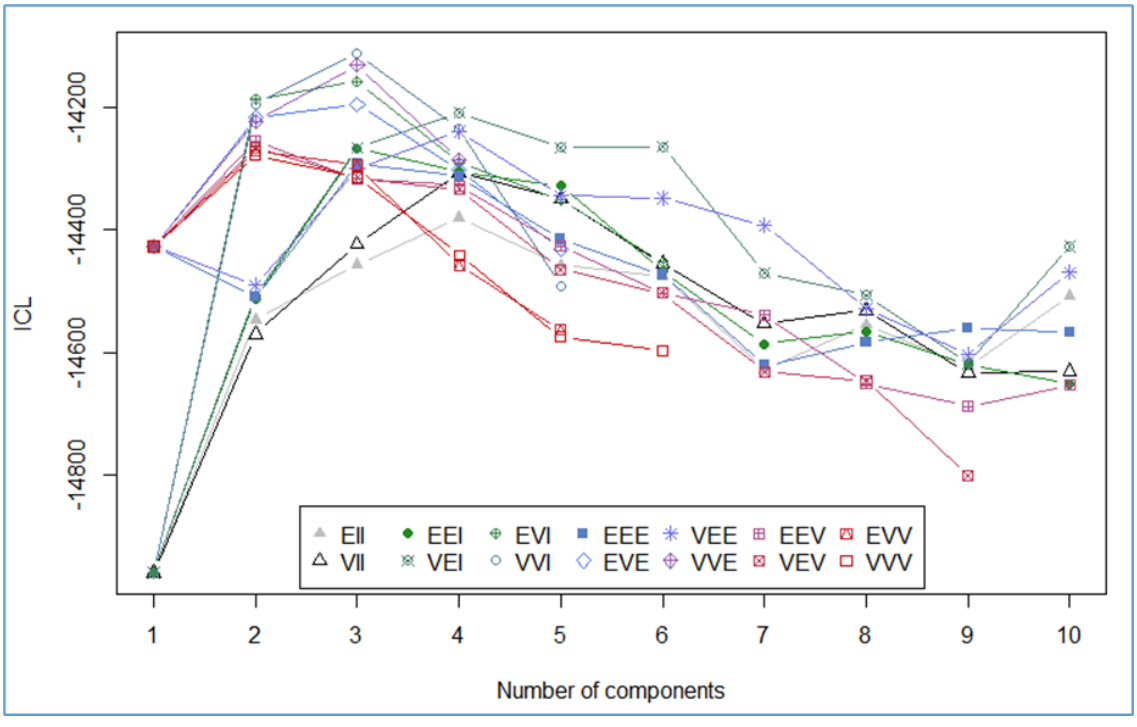


Figure 8

The BIC-plot of MBCA curves in the high reliability Emp4type1 sample for k = 3 to 10

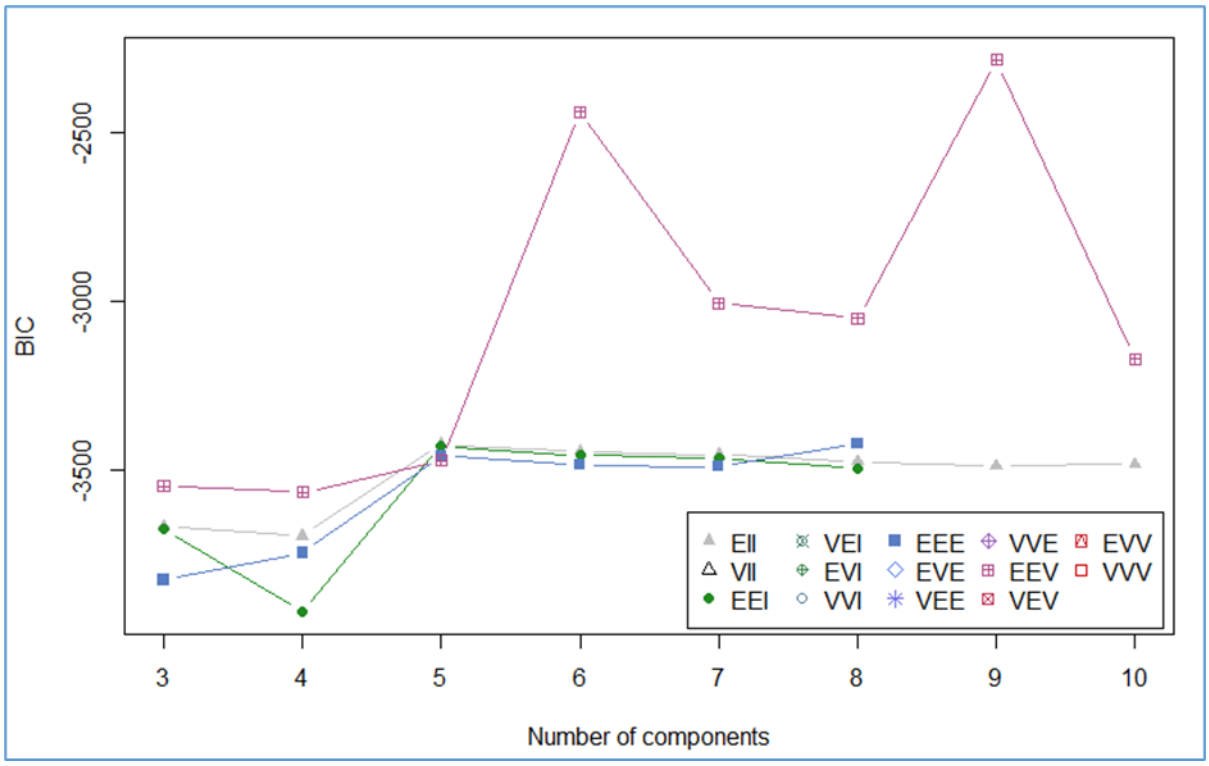
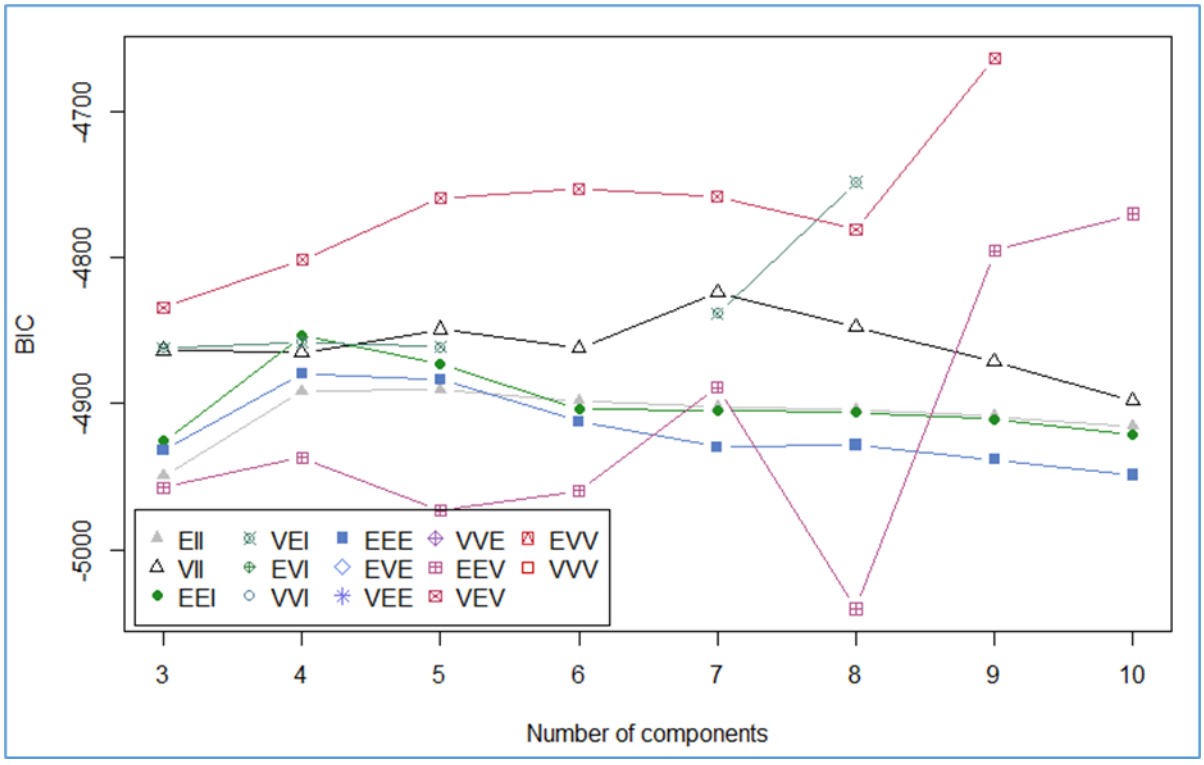


Figure 9

The BIC-plot of MBCA curves in the low reliability Emp4type3 sample for k = 3 to 10



The pairwise ASED differences between the corresponding cluster centroids were never greater than 0.1. Out of the 400 observations 91.8% were correctly classified for EII5m and 93.3% for EEI5m, and there were only one theoretical cluster in both solutions where the percentage of correct identifications did not exceed 90% (81.3% and 85.6%, respectively). Baudry's method, starting with the EEV8 model, yielded here a bad 5-cluster solution, where three clusters fell very close to each other (with pairwise ASED differences less than .40).

In the low reliability Emp4type3 sample, after excluding the irregular curves of VEV9, EEV10, and VEI8 models, the VII7 model had the highest BIC and ICL value (see Figure 9). In the VII7 solution four clusters were close to each other, so we fused them, and this 4-cluster solution (VII7m) happened to be acceptable. The ASED differences were not greater than 0.06 and 83.5% of the theoretical cases were correctly identified.

Baudry's method, starting with the VEV9 model, yielded a 5-cluster solution, where the centroids of three clusters fell very close to each other (with pairwise ASED differences less than .06). When we fused them, the resulting 3-cluster solution had three clusters fairly similar to the theoretical clusters (with pairwise ASED differences not greater than 0.037), but being much weaker than the former VII7m solution in terms of the QCs (EESS%: 33.6 vs. 51.6; SilCoef: 0.48 vs. 0.57; GDI24: 0.68 vs. 0.90; HCmean: 1.33 vs. 0.97).

In Table 1 we summarized the MBCA results of the six empirical samples with BIC and ICL criteria and for Baudry's method. It is astonishing that the automatic decision rarely leads to an optimal solution. Quite often there is local maximum on an irregular curve (see Figures 3, 8, 9), yielding an obviously bad solution. But even if we drop these irregular curves, the next best solutions often have to be submitted to some further modification by fusing close clusters. These modifications were needed in the Emp4type samples, where the starting models had more than four clusters.

It is worth to note that – quite unexpectedly – the ICL-based decisions were never better than the BIC-based ones. In the ICL-plots we found almost always the same irregular curves than on the BIC-plots, and if not, we found local maximums far from the optimal one (see Figures 6 and 7). Baudry's method did not lead to good solutions either.

In Table 2 we summarized some adequacy measures of the best final MBCA solutions. Here, column Type# denotes the number of theoretical clusters (types), and Correct% the proportion of correctly classified cases in the different cluster solutions. The main conclusions that can be drawn based on Table 2 are as follows.

- MBCA can recover the theoretical CLS very well if the reliability level of the set of input variables is high, and quite well if it is moderate.
- MBCA can recover the theoretical CLS at least partially even if the reliability level of the set of input variables is low. For example, in the Emp4type2 sample the two best MBCA solutions could correctly identify the true type of more than 90% of 400 cases.
- The QCs of the best cluster solutions are always excellent for the high reliability measurements (EESS% > 0.75, SilCoef > 0.65, GDI24 > 1.1, HCmean < 0.50), and they are fairly good even for the moderate reliability measurements.

Table 1

Summary of MBCA results on the six empirical samples with BIC and ICL criteria and for Baudry's method. Suffix m after a model name indicates that the cluster solution was modified by fusing some close clusters.

Sample	BIC	ICL	Baudry
Emp7type1	1. Irregular bad EEV10; 2. Good EII7, EEI7, VII7.	Irregular bad EEV10 solution.	Starting with EEV10, bad 6-cluster solution.
Emp7type2	1. Irregular bad VEV10; 2. Good EII7, EEI7.	Bad VVI3, VVE3 solutions.	Starting with VEV10, bad 7-cluster solution.
Emp7type3	Weak VEI6 solution.	Bad VVI3, VVE3 solutions.	Starting with VEI6, bad 4-cluster solution.
Emp4type1	1. Irregular bad EEV9; 2. Good EII5m, EEI5m.	Irregular bad EEV9 solution.	Starting with EEV9, bad 7-cluster solution.
Emp4type2	1. Irregular bad EEV8; 2. Good EII5m, EEI5.	Irregular bad EEV8 solution.	Starting with EEV8, bad 5-cluster solution.
Emp4type3	1. Irregular bad VEV9; 2. Acceptable VII7m.	Irregular bad VEV9 solution.	Starting with VEV9, weak 5-cluster solution.

Table 2

Basic characteristics of the final best MBCA solutions. Suffix m after a model name indicates that the cluster solution was modified by fusing some close clusters.

Type#	Reliability	Model	EESS%	SilCoef	GDI24	HCmean	HC range
7	High	EII7	75.9	0.66	1.15	0.49	0.46-0.52
7	High	EEI7	75.8	0.65	1.14	0.49	0.47-0.52
7	High	VII7	75.8	0.65	1.20	0.49	0.45-0.52
7	Moderate	EII7	66.6	0.54	0.90	0.67	0.51-0.71
7	Moderate	EEI7	66.6	0.54	0.86	0.67	0.51-0.73
7	Low	VEI6	54.6	0.48	0.57	0.91	0.57-1.07
4	High	EII5m	79.0	0.81	1.76	0.42	0.22-0.49
4	High	EII5m	78.9	0.80	1.68	0.43	0.22-0.48
4	Moderate	EII5m	67.8	0.71	1.00	0.65	0.56-0.73
4	Moderate	EEI5(4)	67.1	0.70	0.96	0.66	0.53-0.77
4	Low	VII7m	51.6	0.57	0.90	0.97	0.45-1.33

Table 3

Percentage of correctly classified cases in the solutions of different clustering methods in the six simulated datasets

	Percentage of correctly classified cases					
Sample	HCA Ward	KCA ROPstat	HCA R/diana	KCA R/kmeans	*k*-medoids R/pam	Best MBCA solution
Emp7types1	83.6	91.3	87.3	91.3	85.2	91.7
Emp7types2	67.2	77.0	66.3	77.9	74.0	76.4
Emp7types3	60.8	66.6	55.5	65.7	63.4	60.1
Emp4types1	95.5	94.8	95.3	94.8	95.5	98.0
Emp4types2	81.8	90.0	84.8	90.0	90.8	93.3
Emp4types3	81.0	79.0	73.0	78.5	82.3	83.5

Comparing MBCA with other clustering methods

In order to check the adequacy of MBCA we compared its efficiency in identifying the true theoretical CLSs with different other clustering methods.

First, since we used the same datasets as Vargha and Bergman (2019), we were able to compare the results of our MBCA analyses with these hierarchical (HCA) and *k*-means (KCA) clustering results (see Vargha and Bergman, 2019, Table 6). The former was a Ward-type agglomerative HCA (Ward, 1963), the latter a relocation procedure performed on the result of a Ward's HCA (Vargha et al., 2015). In addition, we performed in all empirical datasets the following cluster analyses (all in R): DIANA divisive HCA[3] (Kaufman and Rousseeuw, 1990), Hartigan-Wong type KCA[4] (Hartigan and Wong, 1979), and *k*-medoids clustering[5] (Park and Jun, 2009). In all these new analyses the cluster number was specified based on the cluster number of the best MBCA solution, although in the case of our special theoretical CLSs and empirical samples there are other efficient ways of determining the optimal cluster number (Vargha and Bergman, 2019). After performing these cluster analyses we computed for each solution the percentage of correctly classified cases and summarized them in Table 3 along with the same data of the best MBCA solution.

[3] With *diana* function of package cluster (see https://www.rdocumentation.org/packages/cluster/versions/2.1.2/topics/diana)

[4] With *kmeans* function of package stats (see https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans)

[5] With *pam* function of package cluster (see https://www.rdocumentation.org/packages/cluster/versions/2.1.2/topics/pam)

Based on Table 3 one can conclude that in the case of Emp7types samples MBCA yielded very similar results as the best of the competing clustering methods (mainly KCA) in the high and the moderate reliability case, and yielded a weaker result only in the low reliability case. In the Emp4types samples MBCA always yielded the best classification. Overall higher percentages in the Emp4types cases indicate that the recovery of the original structure seems to be more efficient in simpler structures, where there are fewer theoretical types.

Discussion

Some studies have reported that the BIC-based approach of MBCA can sometimes overestimate the number of clusters, because it is possible that a cluster consists of several mixture components following a Gaussian distribution (Fraley & Raftery, 1998, 2002; Dasgupta & Raftery, 1998). To improve the efficiency of MBCA, Biernacki et al. (2000) suggested that the ICL criteria could be a better method to find the appropriate number of clusters. Baudry et al. (2007) suggested that starting with the best model of the automatic interpretation of MBCA and hierarchically merging the clusters until only one cluster remains, one can find the appropriate cluster number by evaluating the mean entropy plot and searching for an elbow.

The aim of the present study was to check the adequacy of the automatic interpretation of MBCA with all of the above methods. An additional aim was to refine these MBCA procedures by using QCs to help in the choice of an appropriate cluster structure, and compare the efficiency of MBCA with the efficiency of other clustering methods in the same datasets.

To achieve this goal we created empirical datasets, one based on a theoretical CLS with seven types (Teo7types) and one based on a theoretical CLS with four types (Teo4types). After assigning the value pattern of every participant to the given cluster's centroid (Teo7types and Teo4types) the empirical datasets were created by adding different amounts of random errors to the theoretical models, so the generated empirical samples could be characterized with varying (low, medium, and high) levels of reliability. Due to rounding and truncating the resulting variables were strongly nonnormal in 23% of the possible cluster by variable by reliability level combinations for both CLS types. We performed MBCA with BIC, ICL and Baudry's method in all six datasets.

The most important result of the performed analyses is that the automatic decision rarely leads to an optimal solution. Quite often there is local maximum on an irregular curve yielding a bad solution. Even if one drops the irregular curves, the next best solutions may have to be submitted to some further modifications by fusing close clusters. This occurred actually in our four-type samples. It is also important that the ICL-based decisions were never better than the BIC-based ones, and the automatic Baudry's method did not lead to good solutions either. Consequently it appears that the model-based approach also contains subjective elements in that a variety of different models are produced and the final solution is chosen according to the interpretation of the values of different quality indices, such as EESS%, HCmean, MORI coefficients, etc.

The resulting structures were always better for Teo4types. In the high reliability situations, we obtained excellent solutions for both theoretical models. Quite unexpectedly, the resulting structures were often acceptable even in the low reliability case. In case of competing solutions, the values of the computed quality coefficients (EESS%, SilCoef, GDI24 and HCmean) were consistent with the goodness of the solution in terms of the similarity to the true theoretical model and the reliability level of the variables in the different empirical samples.

Vargha and Bergman (2019) analyzed the same datasets using HCA and KCA. Thus we were able to compare the efficiency of MBCA with these methods. In addition, we performed in all empirical dataset DIANA divisive HCA, Hartigan-Wong type KCA, and *k*-medoids clustering. These results indicated that MBCA yielded definitely better results (in terms of correctly classified cases) in all three four-type datasets and the same quality results in the high and moderate reliability seven-type datasets.

Although these results indicate that a not automatized, sophisticated use of MBCA can be successfully applied in exploring theoretical types in person-oriented research, further investigations are needed to identify those circumstances where MBCA is obviously better than the competing clustering methods.

A limitation of our conclusions is that they are based on analyses with just two examples of theoretical CLSs. We can argue, however, that if a method does not perform well in such clear-cut, ideal situations, there are good reasons to question its general usefulness. This failure refers mainly to the automatized use of MBCA. Our suggested approach of improving MBCA often yielded high quality results as compared with those of other clustering methods. This has to be certainly confirmed by using other types of theoretical CLSs as well.

With the given datasets we wanted to simulate a design typically used in practice. So, in the data generation process, we used linear transformation to create a five-point scale, rounding the obtained continuous data and enforcing them to the 1-5 range. Owing to this specific design of data creation, our results need to be carefully interpreted.

Author contributions

The first author wrote the manuscript with input from the second author. The second author designed the models and the computational framework. Both authors carried out the analyses and wrote together the discussion part.

Declaration of interests

The authors declare that there are no conflicts of interests.

Acknowledgements

The research was funded by the New National Excellence Program (ÚNKP-19-1-I-KRE-4). The preparation of the present study was also supported by a research grant obtained from the Faculty of Humanities, Károli Gáspár University of the Reformed Church (Person- and Family-oriented Health Study, Grant No. 2018/20643B800) and by the National Research, Development and Innovation Office of Hungary (Grant No. K 116965). The authors wish to thank the many useful suggestions of the editor and an anonymous reviewer.

Action editor

Lars R. Bergman served as action editor for this article.

Open access

References

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics, 19*(2), 332-353. https://doi.org/10.1198/jcgs.2010.08111

Bergman, L. R., Vargha, A., & Kövi, Z. (2017). Revitalizing the typological approach: Some methods for finding types. *Journal for Person-Oriented Research, 3*(1), 49-62. https://doi.org/10.17505/jpor.2017.04

Bertoletti, M., Friel, N., & Rastelli, R. (2015). Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron, 73*(2), 177-199. https://doi.org/10.1007/s40300-015-0064-5

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence, 22*(7), 719-725. https://doi.org/10.1109/34.865189

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (rev. ed.).* New York: Academic Press.

Dasgupta, A., & Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American statistical Association, 93*(441), 294-302. https://doi.org/10.2307/2669625

Desgraupes, B.: (2017). Clustering Indices. University Paris Ouest, Lab Modal'X. https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf. (Downloaded: May 28, 2020)

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578-588. http://dx.doi.org/10.1093/comjnl/41.8.578

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association, 97,* 611-631. https://doi.org/10.1198/016214502760047131

Fraley, C., & Raftery, A. E. (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification, 20,* 263-286. https://doi.org/10.1007/s00357-003-0015-3

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation (Vol. 597, p. 1). Technical report.

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*(1), 100-108. https://doi.org/10.2307/2346830

Hasnat, M. A., Alata, O., & Tremeau, A. (2016). Model-based hierarchical clustering with Bregman divergences and Fisher's mixture model: application to depth image analysis. *Statistics and Computing, 26*(4), 861-880. https://doi.org/10.1007/s11222-015-9576-3

Myers, J. H. (1996). *Segmentation and positioning for strategic marketing decisions.* Chicago: American Marketing Association.

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications, 36*(2), 3336-3341. https://doi.org/10.1016/j.eswa.2008.01.039

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20,* 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Scrucca L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016): mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal, 8*(1), 289-317. https://doi.org/10.32614/RJ-2016-021

Vaithyanathan, S., & Dom, B. E. (2013). Model-based hierarchical clustering. arXiv preprint arXiv:1301.3899.

Vargha, A., Torma, B., & Bergman, L. R. (2015). ROPstat: a general statistical package useful for conducting person-oriented analyses. *Journal for Person-Oriented Research, 1*(1-2), 87-98. https://doi.org/10.17505/jpor.2015.09

Vargha, A., Bergman, L. R., & Takács, S. (2016). Performing cluster analysis within a person-oriented context: Some methods for evaluating the quality of cluster solutions. *Journal for Person-Oriented Research, 2*(1-2), 78-86. https://doi.org/10.17505/jpor.2016.08

Vargha, A., & Borbély, A. (2017). Application of modern classification methods in the study of bilingualism. *Glottotheory, 8*(2), 203-216. https://doi.org/10.1515/glot-2017-0013

Vargha, A., & Bergman, L. R. (2019). MORI coefficients as indicators of a "real" cluster structure. *Hungarian Statistical Review,*

2(1), 3-23. https://doi.org/10.35618/hsr2019.01.en003

Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*, 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wit, E., Heuvel, E. V. D., & Romeijn, J. W. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica, 66*(3), 217-236. https://doi.org/10.1111/j.1467-9574.2012.00530.x

Appendix 1

1. HC (Homogeneity Coefficient) of a cluster. This is the average of the pairwise within-cluster distances of its cases. To evaluate a cluster solution, HCmean can be used as a QC. It is the weighted mean of the cluster HC values (the weights are the cluster sizes).
2. EESS% (Explained Error Sum of Square percentage). This is a multivariate generalization of eta-squared used in analysis of variance:

   $$EESS\% = 100*(SStotal - SScluster)/SStotal, \quad (1)$$

   where SStotal is the sum, over the entire sample, of each case's sum of squared deviations of each variable value from the mean for the entire sample in that variable, and SScluster is the sum, over the clusters, of the within-cluster sums of squared deviations of the cases from the variable centroids.
3. SilCoef (a simplified version of the Silhouette Coefficient). This is defined as follows: First, compute SCi for each case i in the sample using formula (2):

   $$SCi = (B - A)/\max(A, B), \quad (2)$$

   where A is the distance from the case to the centroid of the cluster which the case belongs to and B is the minimal distance from the case to the centroid of any other cluster. SilCoef is the average of the SCi values of all cases. A high SilCoef value indicates that, on average, cases are substantially closer to their own cluster centers than to the nearest of the other cluster centers.
4. GDI24 (Generalized Dunn Index). The GDI24 index is a special case of the family of generalized Dunn indices and it can be defined as follows (Desgraupes, 2017):

   $$GDI24 = D/HCmax, \quad (3)$$

   where D is the distance of the two nearest cluster centroids, and HCmax is the HC value of the most heterogeneous cluster.

Appendix 2

We made the analysis script and the simulated data publicly available. See: https://osf.io/ghvrk/?view_only=1be719cafe3c473380641205000cb3eb