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.

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)

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