MORI coefficients as indicators of a 'real' cluster structure*

András Vargha
Professor
Károli Gáspár University
of the Reformed Church,
Eötvös Loránd University,
Hungary
E-mail: vargha.andras@kre.hu

Lars R. Bergman
Professor emeritus
Stockholm University,
Sweden
E-mail: lrb@psychology.su.se

The present study examines some methods for identifying a 'real' cluster structure in situations where objects are classified according to their value profile in a set of variables. The authors investigate the usefulness of various coefficients used mainly in cluster analysis for evaluating the quality of a cluster solution with an emphasis on the newly developed MORI coefficient. Three types of data sets are analysed: a bivariate normal data set and two different data sets where the underlying classification structure is perfect but includes errors of measurement. The findings indicate that the MORI coefficient is a useful tool for detecting the existence of a real classification structure; they also indicate the degree of correspondence between the underlying theoretical classification structure and the empirical cluster structure found in the analyses.

KEYWORDS:
Classification.
Cluster analysis.
MORI.

DOI: 10.35618/hsr2019.01.en003

* The preparation of the present study was 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).

Our study focuses on some methods for identifying a 'real' cluster structure where objects are classified according to their value profile in a set of variables. In the present paper we test the usefulness of various coefficients used mainly in CA (cluster analysis) for evaluating the quality of a cluster solution, with an emphasis on the newly developed MORI coefficient (*Vargha–Bergman–Takács* [2016]; this coefficient is described in a later section). Due to the high complexity of most classification situations, it is extremely difficult to derive conclusions with a high degree of generalizability, and therefore, we chose to address only two classification cases, admittedly idealized but still of considerable interest. First, suppose that the objects to be classified (in our case, mostly persons) are characterized by p quantitative variables. This means that each object is a point in the p-dimensional Euclidean space. The number of objects can be arbitrary. The two classification cases are as follows:

1. A perfect classification structure, where every object belongs to one of k possible classes (p-dimensional sets) 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. Mathematically, this structure can be defined by k discrete points in the p-dimensional space, where these k discrete points, the theoretical centroids, define k different types. The totality/population of all objects in the p-dimensional space having the same type is regarded as a theoretical class. In practice, such a data set does not exist; there may exist only an empirical data set corresponding to the theoretical one. The variables generating the empirical data set are identical to the theoretical variables except that a measurement error is added to each true score. 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 centres are the k theoretical centroids.

2. A 'bad' alternative to Case 1, where there exists only a single point in the p-dimensional space where the objects accumulate. In this case, the data are distributed according to a unimodal p-dimensional distribution. This case will be represented in our study by a multivariate normal distribution. The univariate normal components may or may not be correlated with each other. In each case, no real classification structure exists, in the sense that the structure can be perfectly described by the single centre of a multivariate unimodal distribution.

In Case 1, it is assumed that the data set subjected to empirical analysis reflects a theoretically 'true' data set but with errors of measurement added to the variable values. Two examples of a data set of this type are cluster-analysed. We then study the extent to which the size of QCs (quality coefficients) for solutions with different numbers of clusters indicates the 'true' TCLS (theoretical classification structure) with regard to the number of classes, degree of reproduction of its class centroids, and degree of correspondence between cluster membership and true class membership of the objects. The emphasis is on the usefulness of the MORI coefficient for these purposes. Case 2 is examined by analysing a random sample from a bivariate normal data set with correlated variables.

1. QCs and MORI coefficients

In CA, after a cluster solution has been obtained, clustering QCs are often used to evaluate how good a cluster model is. In the abundant literature on classification analysis, many QCs have been introduced (e.g. *Desgraupes* [2017]). Different QCs focus on different aspects of a cluster structure; thus, it is very important to choose an appropriate set of QCs to evaluate a concrete model. In general, QCs try to measure one or both of two main characteristics of a cluster structure, namely, compactness (cohesion or cluster homogeneity) and separability (see *Vargha–Bergman–Takács* [2016] for details). In the analyses described in the present study, the following QCs were used to evaluate the goodness of an ECLS ([empirical cluster structure]; see *Vargha–Bergman–Takács* [2016] and *Bergman–Vargha–Kövi* [2017] for detailed descriptions):

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). If the input variables are not in standardized form, HCmean is highly dependent on the variances. For this reason, in this case, we suggest dividing HCmean by the average of the variances of the input variables, thus obtaining *HCmeanS*, a standardized form of HCmean.

2. EESS% (*explained error sum of square percentage*). This is a multivariate generalization of eta-squared used in analysis of variance:

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

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. *PB* (*cluster point-biserial correlation*). This is the Pearson correlation, computed on the sample of all pairs of cases, between the binary variable of whether a pair's cases belong to the same cluster (0) or not (1), and the distance between that pair's cases. A well-known formula for PB (e.g. *Glass–Hopkins* [1996]) is:

$$\mathrm{PB} = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}}. \qquad /2/$$

Here, M_0 is the average pairwise within-cluster case distance; M_1 is the average pairwise between-cluster case distance; $n = N(N-1)/2$ is the number of pairs of cases in the total sample of size N; n_0 and n_1 denote the number of pairs of cases that belong to the same and to different clusters, respectively; and, s_{n-1} is the standard deviation of the pairwise differences between cases in the total sample of size n.

4. *CLdelta.* Considering that PB depends primarily on the $M_1 - M_0$ difference, the first component in formula /2/ – being a kind of standardized difference of $M_1 - M_0$ – can also be used as a QC, called CLdelta. It can be explained analogously to the well-known Cohen's delta effect size measure (*Cohen* [1977]). CLdelta indicates the extent to which cases are closer to their own cluster members than to cases from other clusters.

5. A simplified version of the *Silhouette coefficient* (*SC*). This is defined as follows: First, compute SC_i for each case i in the sample using formula /3/:

$$\mathrm{SC}_i = (B - A) / \max(A, B), \qquad /3/$$

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. SC is the average of all the cases' SC_i values. A high SC value indicates that, on average, cases are substantially closer to their own cluster centres than to the nearest of the other cluster centres.

HCmean (or HCmeanS) and EESS% reflect the cohesion of an ECLS, whereas PB, CLdelta, and SC primarily reflect the extent of separability. Based on suggestions in the literature and on our own empirical experience, given a good structure, HCmeanS is expected to be well less than 1, EESS% > 65%, PB > 0.30, CLdelta > 0.80, and SC > 0.50. (See Table 1.)

Table 1

Basic features of the applied clustering quality coefficients

Quality coefficient	Minimum	Maximum	Indication of an acceptable cluster structure
HCmean	0	no limit	< 1 if the input variables are standardized – otherwise it depends on the scales of the input variables
HCmeanS	0	no limit	< 1
EESS%	0	100	> 65%
PB	0	1	> 0.30
CLdelta	0	no limit	> 0.80
SC	0	1	> 0.50

There are plenty of QCs (e.g. *Desgraupes* [2017], where 43 QCs are explained) and one may ask why we chose only these coefficients for our analyses, and not others. In a person-oriented context, it is not uncommon that two distinct types are relatively close to each other. Primarily, for this reason, cluster structures in such situations have to be evaluated primarily by cohesion indices, whose best representatives are EESS% (measuring the explained variance proportion of the input variables through a cluster code variable) and HCmean or HCmeanS (directly measuring the average homogeneity of the clusters). Secondarily, cluster structures have to be evaluated by global separation indices (assessing both compactness and separability), from among which PB is a well-explainable correlation-type measure and CLdelta is a Cohen's delta type coefficient used in two-group comparisons for assessing the effect size. Only in some special cases of person-oriented studies do separation indices carry relevant information. For this reason, in our analyses, we chose one well-known representative, SC, from this class of QCs.

To obtain further evidence for the quality of a cluster solution for real data, it is also important to show that the solution is – significantly and in a measurable way – better than a solution obtained on a random data set of the same size, with the same number of variables and same number of clusters. For this purpose, *Vargha–Bergman–Takács* [2016] developed the MORI coefficient. MORI measures the relative improvement of an ECLS (as measured by a QC) obtained for real data com-

pared to that obtained for the ECLSs resulting from analysing several types of random data sets with the same general properties as the real data set. MORI is computed according to the following formula:

$$\text{MORI} = \frac{\text{QC} - \text{QC}_{\text{rand}}}{\text{QC}_{\text{best}} - \text{QC}_{\text{rand}}}. \qquad /4/$$

In this formula, QC is the quality measure of an ECLS that we would like to evaluate, QC_{rand} is the average of the QCs of the ECLSs resulting from analysing simulated random data, and QC_{best} is the value of QC obtained when the cluster structure is perfect. Specifically, $\text{QC}_{\text{best}} = 100$ for EESS%, 1 for PB and SC, and 0 for HCmean. If QC_{best} is theoretically infinitely large (as in the case of CLdelta), it is suggested that the denominator of /4/ be set equal to QC_{rand}. In this case, MORI measures the improvement in QC relative to the base value of QC_{rand} (*Vargha–Bergman–Takács* [2016]).

In the validation module of the ROPstat statistical package (*Vargha–Torma–Bergman* [2015]), MORI can be computed for all the QCs presented above for each of these four options for the type of random control data set:

1. independent random permutations of the values of the input variables,
2. random data from independent continuous uniform distributions,
3. random data from independent normal distributions,
4. random data from correlated normal distributions, with intercorrelations matching the correlations among the input variables.

For each of the options above, significance levels and confidence intervals can be computed for the MORI coefficients. ROPstat allows at most 100 independent random Rep (replications).

The aim of this study is also to compare the usefulness of these four types of random control data sets for identifying a TCLS.

2. Creation of theoretical and empirical samples

Suppose that each subject in a population of size N can be characterized by a set of quantitative traits (value pattern). A TCLS exists in this population if each subject´s value profile is identical to one of k value patterns of the given traits, called types, and $k < N$. Each type defines one class in the population. In psychology, k is

normally expected to be much smaller than N, say less than 10. Sometimes, we refer to a TCLS as a real or natural classification structure, considering the fact that it truly exists in nature and can potentially be detected. In practice, finding a TCLS is difficult because the data set for such a structure is not observed directly; there exists only an empirical data set that corresponds to the theoretical data set but with errors of measurements added to the variables in the theoretical data set.

Below, we present examples of two concrete TCLSs with some continuous unobservable theoretical traits: one based on a real empirical study and the other a completely artificial example. For each example, we will construct observed variables (measurements of each trait) with an error component added to the original theoretical value (empirical data set). Then, analyses will be performed on this data set to see how well the studied TCLS can be identified by standard k-means CAs and by the QCs and the MORI coefficient introduced above.

The first theoretical sample (Teo7types) was derived from a sociolinguistic classification study of Romanian ethnic minority persons living in Hungary (*Vargha–Borbély* [2017]). Five sociolinguistic variables were included in the value profile (minority language competence, language use in family, language use in church, minority identity, and attitude toward minority language) and a seven-cluster solution was found, with attractive MORI indices (see Table 8; *Vargha–Borbély* [2017]). The clusters identified seven types of speakers in the clearly nonlinear process of language shift and assimilation from a bilingual minority status to a monolingual Hungarian status. Based on these positive results, we chose as our artificial TCLS the centroids of this solution.[1] (See Table 2.)

Table 2

Teo7types data set – centroids and sizes of the theoretical clusters derived from the Romanian sample

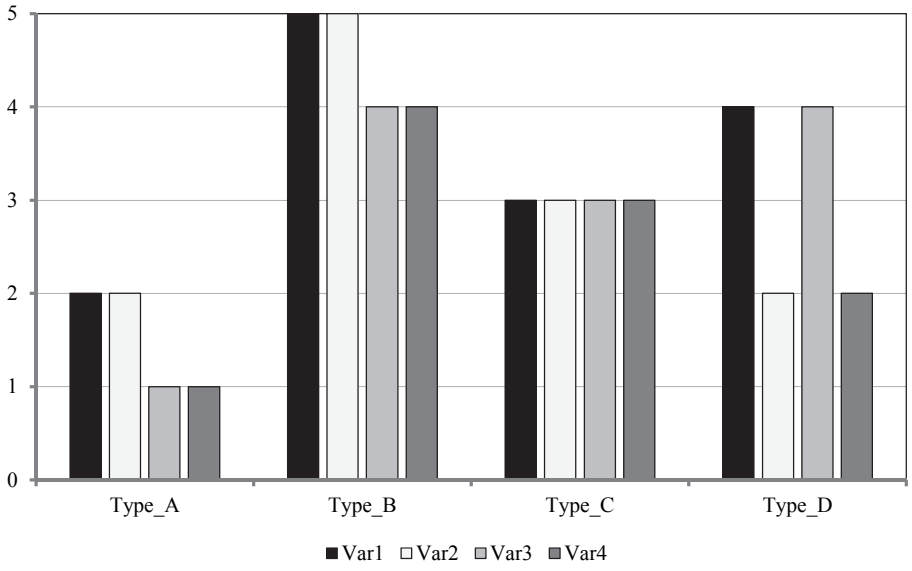
Theoretical cluster	Variable					Cluster size
	V1	V2	V3	V4	V5	
TC1	1.84	2.64	1.86	1.67	1.95	195
TC2	3.43	2.98	2.08	1.77	2.16	160
TC3	2.62	4.23	2.05	1.86	3.10	175
TC4	3.88	4.51	1.88	1.74	4.25	115
TC5	3.86	4.70	4.56	1.71	4.27	95
TC6	4.00	4.50	2.42	5.00	4.61	50
TC7	3.61	4.63	4.71	4.49	4.32	115

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

It consisted of 7 clusters based on 5 variables, as in Table 2. Each theoretical cluster was error-free, consisting of cases with the same value pattern, corresponding to the TCLS in Table 2. The relative cluster sizes were borrowed from *Vargha–Borbély* [2017] but the sizes were multiplied by 5 to have a substantial sample size of 905. (See the last column in Table 2.)

The second, simpler theoretical sample (Teo4types) was borrowed from *Bergman–Vargha–Kövi* [2017]. It is an artificial sample ($N = 400$) consisting of 4 clusters based on 4 variables, as in Figure 1. Each theoretical cluster was error-free, consisting of cases with the same value pattern corresponding to the TCLS of Figure 1. Cluster sizes were also borrowed from *Bergman–Vargha–Kövi* [2017]: 160, 40, 160, and 40, respectively, for types A to D.

Figure 1. Teo4types data set – artificial TCLS with four clusters (types) and four variables



Note. Var: variable.

For each theoretical sample, three artificial empirical samples (Emp7types1, Emp7types2, and Emp7types3 for Teo7types, and Emp4types1, Emp4types2, and Emp4types3 for Teo4types) were constructed in the following way: For each original (true) variable value in the theoretical data set, a new value was created by adding the value of an independent random $N(0; SD_i)$ variable, where SD_i was set to 0.50, 0.75, 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 5-point Likert-scale variables in the empirical samples, which is a common situation in practice. The three different SD_i values yielded three levels of reliability for the derived variables. These levels were assessed by computing r^2 values as explained variance proportions between the theoretical and the corresponding empirical variables, used as reliability estimates. These r^2 values were between 0.65 and 0.81 (mean = 0.74) in Emp7types1, between 0.52 and 0.71 (mean = 0.61) in Emp7types2, and between 0.40 and 0.57 (mean = 0.50) in Emp7types3. Similarly, they were between 0.71 and 0.86 (mean = 0.78) in Emp4types1, between 0.53 and 0.72 (mean = 0.65) in Emp4types2, and between 0.43 and 0.60 (mean = 0.50) in Emp4types3. These r^2 means represent the three levels of reliability of the empirical variables (high, moderate, and low).

3. Results

In this session we will perform a CA within a bivariate random normal data set with $\rho = 0.8$ and will obtain some surprisingly high quality coefficients. Then results of analyses of two empirical data sets, each corresponding to a theoretical data set, will be presented.

3.1. CA of a bivariate normal data set

When a CA is performed on a sample, a certain 'best solution' is usually chosen, often based on the values of one or more QCs and on what appears meaningful and theoretically relevant. Since a CA will always provide a partition of the objects, it may be possible that the obtained solution is merely an artefact of the partitioning procedure. Our primary goal is to find a method that can identify these artefacts. Consider now the bivariate normal data case (Case 2). We generated a random sample of size 1,000 from a bivariate standard normal distribution with $\rho = 0.8$. (See Figure 2.) Based on this random sample, we conducted a hierarchical CA (Ward's method with ASED [average squared Euclidean distance] of the cases), whose 4-cluster solution seemed to be the most promising (first large drop of EESS% occurring from $k = 4$ to 3[2] and clearly fulfilling the criteria mentioned in the previous section for all QCs; see Table 1).

[2] EESS% values in the hierarchical solution from $k = 8$ to $k = 3$: 87.29 ($k = 8$), 85.22 ($k = 7$), 83.02 ($k = 6$), 80.53 ($k = 5$), 76.35 ($k = 4$), 69.50 ($k = 3$), and 50.29 ($k = 2$).

Figure 2. Scatter plot of a sample from a bivariate normal distribution with $\rho = 0.8$
($N = 1{,}000$)

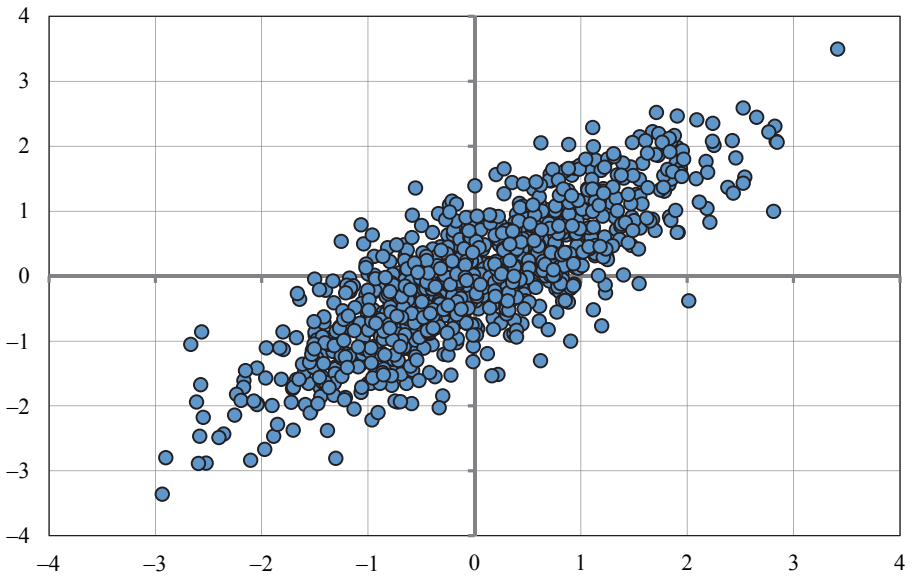


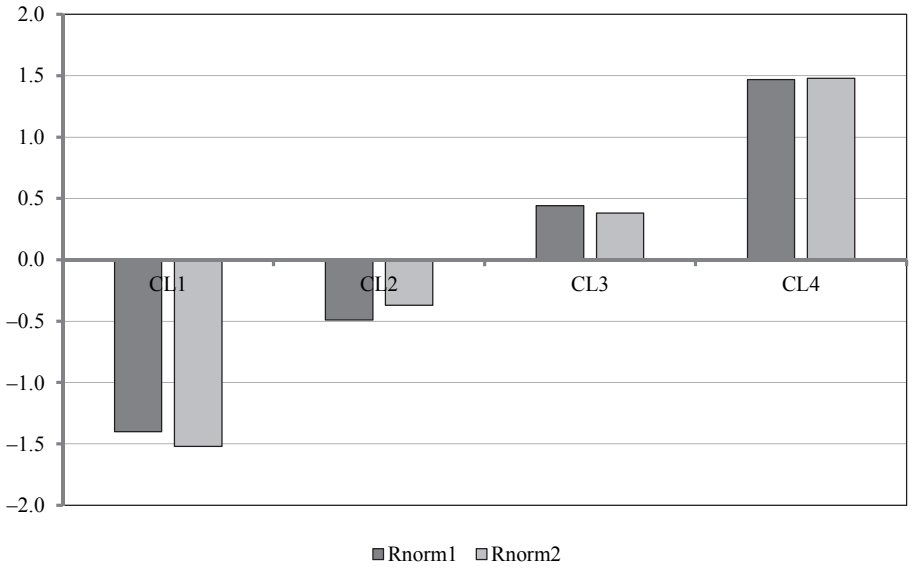
Table 3

Quality coefficients of the relocated 4- and 3-cluster solutions of 1,000 bivariate random normal data points with $\rho = 0.8$

Cluster number	Quality coefficient					
	EESS%	PB	SC	HCmean	CLdelta	HC range
$k = 4$	79.89	0.395	0.706	0.403	0.876	0.33–0.56
$k = 3$	73.09	0.446	0.746	0.545	0.922	0.43–0.66

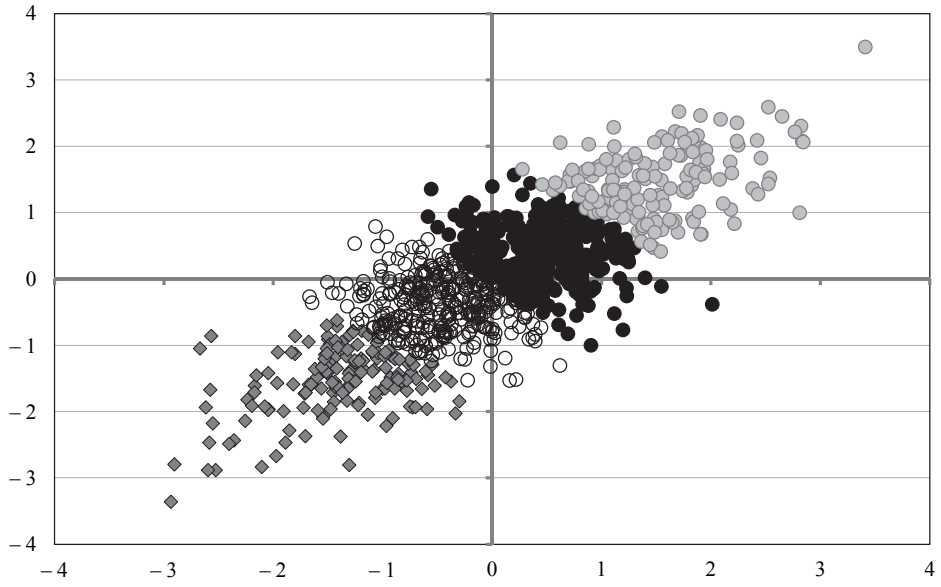
After performing a relocation on the obtained hierarchical solution (*k*-means CA), QCs improved even further (see the $k = 4$ row in Table 3) and indicated a good ECLS in terms of all the QCs (see Table 1). The centroids of the four clusters can be seen in Figure 3 (the two input variables are denoted here by Rnorm1 and Rnorm2). The cluster sizes of CL1 (cluster 1) to CL4 (cluster 4) were 155, 344, 343, and 158, respectively. To illustrate this solution, we created, once again, the scatter plot of Rnorm1 and Rnorm2, but used different shades (and/or forms) for the cases belonging to the different clusters. (See Figure 4.)

Figure 3. The centroids of the 4-cluster k-means solution in a sample of 1,000 bivariate random normal data points with $\rho = 0.8$



Note. CL1 to CL4 are clusters; Rnorm 1 and Rnorm 2 are input variables.

Figure 4. The four clusters of Figure 3 on the scatter plot of Figure 2



Finally, using the MORI coefficient, we validated the 4-cluster *k*-means solution with types (1), (2), and (4) described below (in this case, type (3) is identical to type (1), so it was dropped), using Rep = 100. (See Table 4.) Here, recall that MORI measures the extent to which the QC value for an explored cluster structure is better than the average of the values of that QC (see Formula /4/) obtained for 100 random control data sets of a certain type [(1), (2), or (4)], where *k*-means CAs are also performed with the same number of clusters ($k = 4$) and variables ($p = 2$). In type (1) (random permutation control), the random data file is generated by applying a random permutation on each column of the input variables in the original data file. This transformation leaves the distribution of the individual input variables unchanged but removes all the correlations among them. In type (2) (independent random uniform control), the random data file is created by generating a uniformly distributed random data sample of the same size. In type (4) (correlated random normal control), the random data file is created by generating a normally distributed random data sample of the same size, where the theoretical correlations between the generated random input variables are the same as the correlations between the original input variables.

The MORI values obtained are summarized in Table 4. Based on these values, the following conclusions can be drawn:

1. There is a large inconsistency among the MORI indices for the different types of the random control data set.

2. In the correlated normal type (4), all the MORI values are around zero, as expected.

3. However, if the QCs of the obtained 4-cluster *k*-means solution are compared to the average of the QCs of 100 4-cluster *k*-means solutions of uncorrelated random data sets (see rows 1 and 2), the MORI values of the QCs measuring cohesion (EESS% and HCmean) increase substantially, especially with the random permutation option. This indicates that a strong relationship between the input variables can, in itself, be a factor that creates dense regions in the population, which, in turn, may be identified – certainly often falsely – as homogeneous clusters (see Figure 4) fairly better than those in random data sets of independent variables.

4. The low – sometimes even negative – MORI values of the global separation indices PB and CLdelta seem to be able to indicate for all types of random control data sets that the explored ECLS is not better than the structure obtained from a random data sample.

Table 4

MORI validation indices for five QCs for the k-means 4-cluster solution based on a data set with two correlated random normal variables with $\rho = 0.8$
(Rep = 100)

Type of random control data set	Quality coefficient				
	EESS%	PB	SC	HCmean	CLdelta
(1) Random permutation*	0.44	0.03	0.16	0.44	0.02
(2) Independent uniform distribution*	0.19	–0.26	–0.05	0.19	–0.27
(4) Correlated normal distribution**	0.00	–0.01	0.01	0.00	–0.01

* All positive MORI coefficients were significantly greater than 0 at the 1% significance level.
** No MORI coefficients were significantly different from 0 at the 5% significance level.
Note. Here and in the following tables, Rep: number of replications. Type (3) (independent random normal control) was dropped as it is identical to type (1).

The behaviour of PB and CLdelta is, however, not always typical. It becomes obvious if we study the *k*-means 3-cluster solution[3] based on the bivariate standard normal data set and validate it with the random permutation control (type (1)). In this case, the MORI values for PB (0.10) and CLdelta (0.14) are substantially above 0 at $p < 0.01$, and for EESS% (0.42) and HCmean (0.42), they are so high that they falsely suggest that the 'real' TCLS has been found. (The MORI value for SC is 0.25.) Similar results were obtained by *Vargha* and *Borbély* [2017] too when working with three correlated random normal variates. They showed that, despite the lack of a clear TCLS, the MORI values for EESS% and SC determined by all the types of random control data set, except for those for the correlated normal type, exceeded the 0.35 level substantially, sometimes exceeding even 0.50 (see Table 7; *Vargha–Borbély* [2017]).

Summarizing the internal validation results of the explored 3- and 4-cluster solutions based on the bivariate standard normal data set, we can conclude that only the MORI indices for the correlated normal random control data set could reliably indicate that the explored 3- and 4-cluster solutions, despite exhibiting sufficiently high QC values, are mere artefacts.

Without a clear definition of the concept of a TCLS, in general, we will not be able to assess whether cluster solutions similar to the ones obtained above are simply parsimonious descriptions of the data sets or classifications corresponding to existing theoretical models. Below, this issue is addressed by analysing two empirical data sets that correspond to two different theoretical structures.

[3] In this solution, approximately half of the subjects are included in a central cluster, one-fourth in a generally-low cluster, and one-fourth in a generally-high cluster.

3.2. Analysis of two empirical data sets, each corresponding to a theoretical data set

On each of the six empirical samples, we performed a hierarchical CA with Ward's method, followed by relocations for $k = 2$ to 10 clusters. Then, a validation was performed for each solution, using all four types of random control data with Rep = 100 independent repetitions. MORI values were computed for all the five QCs defined earlier. Finally, at each reliability level, the best solutions were compared to the theoretical cluster structures of the theoretical samples (Teo7types and Teo4types). All these analyses were performed with the ROPstat statistical software (*Vargha–Torma–Bergman* [2015]).

The statistical analyses aimed to address the following three questions:

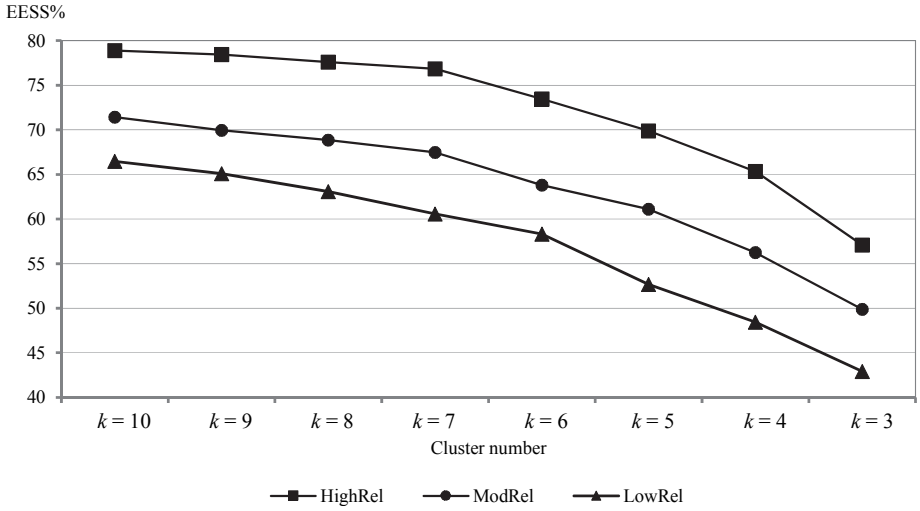
> *Q1*. Were the indicators able to identify the existence of some real cluster structure (i.e. reject the null hypothesis that the cluster structure found could have been generated by some type of random data)?
>
> *Q2*. Which indicators provided the most useful information for identifying the true number of clusters in the TCLSs?
>
> *Q3*. To what extent did the MORI coefficient reflect the degree of correspondence between the ECLSs and TCLSs?

Which indicators indicated that some real cluster structure existed? (*Q1*.) Significance tests were performed for all the QCs and MORI coefficients for the all six empirical samples, and for each of the four options of random control data. All tests produced significant results at the 1% level. Hence, in this limited sense, all coefficients indicated that some real structure existed.

Indicators for the true number of clusters. (*Q2*.) It is often suggested that the analyst should 'choose the number of clusters k after which EESS% first decreases by a large amount' (elbow method, see *Thorndike* [1953], *Milligan–Cooper* [1985], *Myers* [1996]). In the Emp4types samples, this suggestion seemed to work, but in the Emp7types samples, where the number of theoretical types was seven, it did not work. (See Figure 5.) For this reason, we sought a better procedure for determining the true number of clusters, which allows for less subjectivity and ambiguity.

Analysing the MORI values for EESS% in the six empirical samples, we noticed a systematic behaviour. MORI often reached its maximum for the real type of control data set. (See Table 5.) This pattern was most salient in the high- and moderate-reliability samples, with the random permutation option as the type of random control data. (See Figures 6 and 7.) In this situation, the five QCs were less useful.

Figure 5. EESS% in k-means cluster analyses performed in the three Emp7types samples for k = 3 to 10



Note. Here and hereinafter, HighRel: high reliability; ModRel: moderate reliability; LowRel: low reliability.

Table 5

MORI values for EESS% in the high-reliability Emp7types1 sample for the four types of random control data set
(Rep = 100)

Type of random control data set	Cluster number						
	k = 4	*k* = 5	*k* = 6	*k* = 7	*k* = 8	*k* = 9	*k* = 10
(1) Random permutation	0.281	0.325	0.355	**0.408**	0.391	0.387	0.374
(2) Independent uniform distribution	0.368	0.406	0.434	**0.468**	0.450	0.438	0.416
(3) Independent normal distribution	0.475	0.511	0.539	**0.575**	0.566	0.562	0.553
(4) Correlated normal distribution	0.447	0.478	0.499	**0.534**	0.520	0.512	0.497

Note. The highest MORI value in each row is denoted in bold. All the MORI coefficients were significantly greater than 0 at the 1% significance level.

Figure 6. MORI patterns of EESS% for the three Emp7types samples using the random permutation option

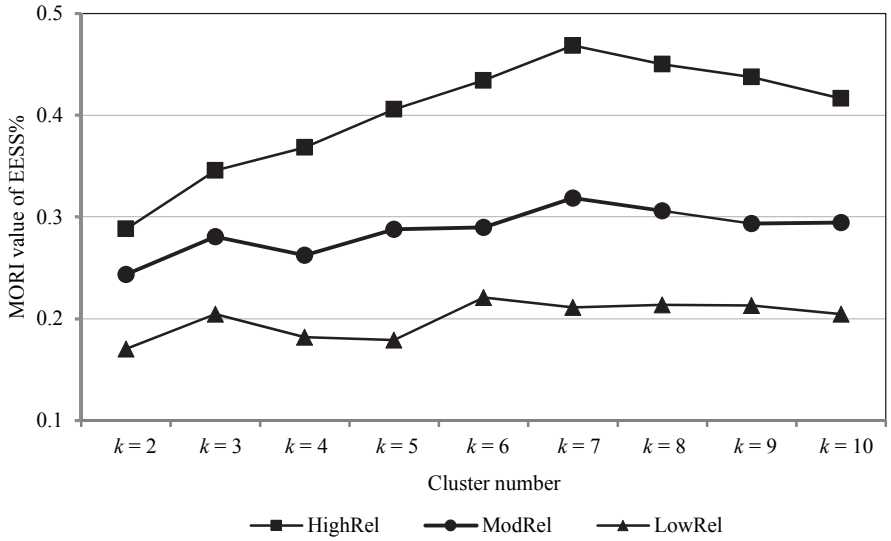
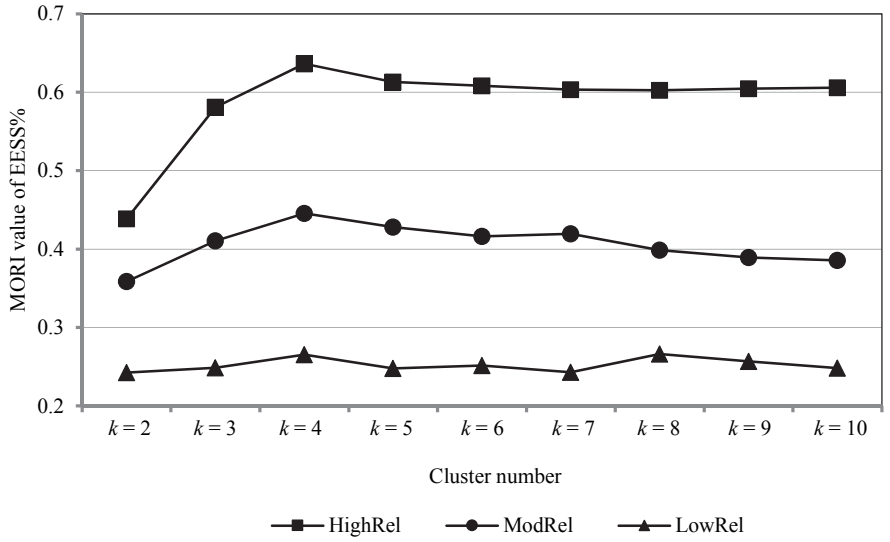


Figure 7. MORI patterns of EESS% for the three Emp4types samples using the random permutation option



Degree of correspondence between ECLSs and TCLSs. (*Q3*.) In Table 6, we summarized some adequacy measures of the best *k*-means and hierarchical cluster

solutions reflecting how well they were able to reproduce the theoretical cluster structure. The main conclusions that can be drawn based on Table 6 are as follows:

1. In four out of six samples, the relocation of the best hierarchical solution improved the quality of ECLS to a substantial extent (in terms of similarity to the TCLS types and in recovery of TCLS class memberships). However, in the Emp4types samples, where the number of theoretical types was smaller (4 compared to 7), in two out of three samples, the hierarchical solution was better. This means that a general preference for a *k*-means analysis instead of a hierarchical one was not justified.

2. The recovery of the simpler Teo4types structure was more efficient than that of Teo7types, which has more theoretical types.

3. Both the hierarchical and *k*-means analysis were able, in our case, to recover the TCLS at least partially, even if the reliability level of the set of input variables was low. For example, in the Emp4types3 sample with a reliability of 0.50, the best cluster solutions were able to correctly identify about 80% of the 400 cases.

4. It can be seen from Figures 6 and 7 that the size of MORI for EESS% indicated well the degree of correspondence between the ECLS structure and TCLS structure reported in Table 6, and, as expected, the size of MORI and the structure correspondence decreased when the reliability of the studied sample decreased.

Table 6

Comparison of the best hierarchical and k-means cluster solutions for the theoretical structure, using two criteria

Sample	Percentage of correctly classified cases (Criterion 1)		Pairwise ASED differences between original and explored cluster centres (Criterion 2)	
	Hierarchical	*k*-means	Hierarchical	*k*-means
Emp7types1	83.6	91.3	0; 0; 0.004; 0.005; 0.013; 0.031; 0.130	0; 0.001; 0.001; 0.002; 0.004; 0.007; 0.022
Emp7types2	67.2	77.0	0.022; 0.022; 0.054; 0.071; 0.084; 0.086; 0.110	0.001; 0.008; 0.019; 0.021; 0.024; 0.055; 0.058
Emp7types3	60.8	66.6	0.007; 0.061; 0.065; 0.076; 0.115; 0.201; 0.202	0.035; 0.045; 0.055; 0.089; 0.104; 0.182; 0.321
Emp4types1	95.5	94.8	0; 0.001; 0.003; 0.013	0; 0; 0.011; 0.086
Emp4types2	81.8	90.0	0.002; 0.012; 0.146; 0.176	0; 0.002; 0.093; 0.139
Emp4types3	81.0	79.0	0.014; 0.033; 0.104; 0.183	0.005; 0.011; 0.176; 0.209

Note. ASED: average squared Euclidean distance.

Replication of findings on a new ECLS sample. A limitation of our findings is that only one sample with errors added to the true scores was produced for each condition. It is possible, but not probable, that, if we had replicated the simulations, the findings might have been different to a moderate extent. For this reason, as a partial check, we performed a replication of all three Emp7types samples with completely new random data and computed the MORI coefficients for all five QCs using the correlated random normal option. This was done for $k = 4$ to 10 clusters. As can be seen in Table 7, the difference between the MORI values of the old and new empirical samples is small (especially for EESS% and HCmean).

Table 7

Absolute differences of the MORI indices for the five quality coefficients between the Emp7types data and the new randomized empirical data (averages and ranges for k = 4 to 10 clusters)

Level of reliability	Quality coefficient				
	EESS%	PB	SC	HCmean	CLdelta
	Average				
HighRel	0.010	0.007	0.026	0.007	0.003
ModRel	0.009	0.010	0.016	0.027	0.025
LowRel	0.013	0.016	0.017	0.019	0.026
	Range				
HighRel	0.002–0.017	0.000–0.023	0.009–0.044	0.003–0.013	0.001–0.005
ModRel	0.002–0.023	0.001–0.019	0.004–0.044	0.010–0.051	0.006–0.038
LowRel	0.000–0.048	0.001–0.036	0.003–0.043	0.010–0.030	0.012–0.065

4. Conclusion

It should be mentioned that, because only examples have been analysed, no generalization beyond them is feasible. However, we believe the examples are rather typical for a number of commonly occurring classification situations. In fact, *the finding that certain procedures did not function well for the purpose of identifying the true cluster structure* in our examples might have a practical value. This strongly suggests these procedures are not generally useful.

Among the studied QCs, the MORI index provided the most useful information regarding the existence and degree of a ‘real’ cluster structure. First, the MORI index

of EESS%, using the correlated normal random control option, was a good indicator that a real TCLS had been found in the empirical analysis, if the average MORI value had reached or exceeded 0.35 around the explored best number of clusters. (See Table 6.) In these situations, the true number of clusters is the value of k for which MORI reaches its maximum, using the random permutation option. (See Figures 6 and 7.) If the MORI value for EESS% around the explored best number of clusters using the correlated normal option is below 0.20, this may indicate the lack of a real TCLS or the low reliability of the variables used in the CA. (See the last row of Table 4.)

Determining the 'true' number of clusters is not a simple task in CA. As indicated by *Franke–Reisinger–Hoppe* [2009], *Milligan–Cooper* [1985] investigated 30 stopping rules to determine the number of clusters in an extensive Monte Carlo study and found that the *Calinksi-Harabasz* index [1974] proved to be one of the best criteria. However, in our study, the Calinksi-Harabasz index indicated $k = 2$ as the best number of clusters for all six empirical samples, obviously falsely (findings not reported in the Results section). We suggest that our stopping rule using MORI is better in many settings.

Franke–Reisinger–Hoppe [2009] also developed a new ICA (index of clustering appropriateness), with the following formula:

$$\text{ICA} = \text{RS}/\text{MSRS}, \qquad /5/$$

where RS is the explained variation in a clustering solution and MSRS the mean SRS across a number of random experiments, with SRS being the simulated explained variation in one of these random experiments. ICA has a similar logic to MORI based on EESS%. (See Formula /4/.) ICA reflects the increased explanatory power of an explored cluster solution relative to the average result based on simulated random uniform data with the same number of clusters and variables. This corresponds to our type (2) random control data set. However, of the four types of random control data sets, type (1) (random permutation) and type (4) (correlated random normal) appear, in our case, to have been more useful. (See Figures 6 and 7, and Table 4.) These two types have complementary features. On the one hand, using the random permutation method, the distributions of the input variables are preserved and, thus, the cluster-forming influence of the intercorrelations of the input variables can be assessed. On the other hand, with the correlated random normal method, the intercorrelations of the input variables are preserved and, thus, the cluster-forming influence of the special multivariate distribution of the input variables (specifically, its deviance from the multivariate normal) can be assessed.

Results from the analyses on the bivariate normal data set showed that a cluster structure of apparently high quality had been found, as shown by all five QCs. Even when the MORI validation indices were computed for these coefficients, the size of the cohesion coefficients indicated a good cluster structure, except when the random correlated normal control option was applied. These findings suggest that the application of the MORI validation procedure using the correlated random normal control option can be useful for deciding whether a clustering structure is 'real' in the sense that it cannot be explained by an underlying data model of a multivariate normal distribution. This was also found in the analysis of the two examples of empirical data sets.

Two limitations of the present study are that only one set of data was produced for each reliability condition and that all variables were 5-point scales. The results of the replication on a new set of random data support that the first limitation is not severe. The 5-point scale type of the variables is common in practice when items are used as variables, but it is possible that, if, instead, truly continuous variables had been analysed, it would have resulted in somewhat different findings.

References

BERGMAN, L. R. – VARGHA, A. – KÖVI, Z. [2017]: Revitalizing the typological approach: some methods for finding types. *Journal for Person-Oriented Research.* Vol. 3. No. 1. pp. 49–62. https://doi.org/10.17505/jpor.2017.04

CALINSKI, T. – HARABASZ, J. [1974]: A dendrite method for cluster analysis. *Communications in Statistics, Theory and Methods*. Vol. 3. No. 1. pp. 1–27.

COHEN, J. [1977]: *Statistical Power Analysis for the Behavioral Sciences*. Revised Edition. Academic Press. New York.

DESGRAUPES, B. [2017]: *Clustering Indices*. University Paris Ouest. Lab Modal'X. https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

FRANKE, N. – REISINGER, H. – HOPPE, D. [2009]: Remaining within-cluster heterogeneity: a meta-analysis of the 'dark side' of clustering methods. *Journal of Marketing Management*. Vol. 25. Nos. 3–4. pp. 273–293. http://dx.doi.org/10.1362/026725709X429755

GLASS, G. V. – HOPKINS, K. D. [1996]: *Statistical Methods in Education and Psychology*. 3rd Edition. Allyn & Bacon. Boston.

MILLIGAN, G. W. – COOPER, M. C. [1985]: An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. Vol. 50. No. 2. pp. 159–179. https://doi.org/10.1007/BF02294245

MYERS, J. H. [1996]: *Segmentation and Positioning for Strategic Marketing Decisions*. American Marketing Association. Chicago.

THORNDIKE, R. L. [1953]: 'Who Belongs in the Family?' *Psychometrika.* Vol. 18. No. 4. pp. 267–276. https://doi.org/10.1007/BF02289263

VARGHA, A. – BERGMAN, L. R. – TAKÁCS, SZ. [2016]: Performing cluster analysis within a person-oriented context: some methods for evaluating the quality of cluster solutions. *Journal for Person-Oriented Research*. Vol. 2. Nos. 1–2. pp. 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*. Vol. 8. No. 2. pp. 203–216. https://doi.org/10.1515/glot-2017-0013

VARGHA, A. – TORMA, B. – BERGMAN, L. R. [2015]: ROPstat: a general statistical package useful for conducting person-oriented analyses. *Journal for Person-Oriented Research*. Vol. 1. Nos. 1–2. pp. 87–98. https://doi.org/10.17505/jpor.2015.09