STATISTICS IN TRANSITION-new series, December 2012
Vol. 13, No. 3, pp. 569—580

AN EMPIRICAL ANALYSIS OF THE EFFECTIVENESS OF WISHART AND MOJENA CRITERIA IN CLUSTER ANALYSIS

Artur Mikulec, Aleksandra Kupis-Fijałkowska

ABSTRACT

Mojena and Wishart criteria are methods of selecting the optimal grouping result of agglomerative cluster analysis methods (hierarchical). Two criteria were proposed by Mojena in the 70's of the 20th century: the upper tail rule and moving average quality control rule, both based on an analysis of the fusion levels of objects in the dendrogram with the aim to determine the cut-off point of it, i.e. to choose the optimal clustering result. The third criterion: tree validation was created by Wishart and evaluates the randomness of the objects clustering in the dendrogram.

The purpose of this paper is to present the results of the empirical analysis of the effectiveness of Mojena and Wishart criteria for the number of clusters selection, in comparison to other applicable criteria in this area, including those proposed by: Baker and Hubert, Calinski and Harabasz, Davies and Bouldin, Hubert and Levine. The empirical analysis has been carried out in *ClustanGraphics 8* Program and selected packages in R environment for the generated data sets.

Key words: upper tail rule, moving average quality control rule, Mojena criteria, Wishart criterion (tree validation), *ClustanGraphics 8*.

1. Introduction

The assessment methods of the grouping result – in the broad sense – are related to the three issues of cluster analysis, respectively: determining the number of clusters, comparing two (or more) classification results and the grouping result quality assessment. The grouping result evaluation stage, i.e. selecting the number of clusters in the analysis which is based on the hierarchical clustering algorithms is one of the last steps here, however, it is extremely important in the classification process. In fact, when the classification sequence $P_0, P_1, ..., P_{n-1}$ is given, on the basis of some formal criteria a decision on the final grouping result selection should be taken.

The article aims to present the results of an empirical effectiveness analysis of the two Mojena criteria [Mojena 1977], both based on the distance analysis of the objects fusions in the graph tree – *best cut significance test (upper tail rule, moving average quality control rule)* and the Wishart criterion [Wishart 2006] which evaluates the randomness of objects clustering in the dendrogram – *tree validation*. The above mentioned criteria have been compared with other procedures for selecting the number of clusters, including the following: Baker and Hubert (BH), Caliński and Harabasz (CH), Davies and Bouldin (DB), and Hubert and Levine (HL). All criteria are based on determining the number of classes and the structure of the grouping result. The article discusses different cases of clusters generated on the basis of the same (identical) covariance matrix of variables.

2. Methods of determining the number of clusters

The cluster analysis literature widely presents and describes in detail many methods that aim to select the number of clusters [see Gan et al., 2007; Gatnar, Walesiak 2009; Mikulec 2012].

The two Mojena criteria and the Wishart criterion, which are the basis of this work, can be found among very few procedures for (determining) the number of clusters (in addition to the following indices: Beale, Duda and Hart, RMSSTD or RS) that are dedicated to the hierarchical classification methods, for example the agglomeration one. Nonetheless, also other procedures mentioned in the introduction can be used as the selection criteria for the number of clusters in the agglomeration methods (see Table 1) – they differ due to the construction of the internal criterion for the grouping result assessment.

Table 1. Chosen methods of determining the number of clusters in the data set [a]

CRITERION	Formula, confidence interval	Criterion of selecting the number of clusters
Baker & Hubert	$BH(u)=\frac{S_+ - S_-}{S_+ + S_-},\ BH(u)\in\langle -1;1\rangle$	$\hat{u}=\arg\max_u[BH(u)]$
Caliński & Harabasz	$CH(u)=\frac{tr(B_u)/(u-1)}{tr(W_u)/(n-u)},\ CH(u)\in R_+$	$\hat{u}=\arg\max_u[CH(u)]$
Davies & Bouldin	$BD(u)=\frac{1}{u}\sum_{q=1}^{u}\max_{r,q\neq r}\left(\frac{S_q+S_r}{d(q,r)}\right)$	$\hat{u}=\arg\min_u[BD(u)]$
Hubert & Lewine	$HL(u)=\frac{D(u)-l_w D_{\min}}{l_w D_{\max}-l_w D_{\min}},\ HL(u)\in(0;1)$	$\hat{u}=\arg\min_u[HL(u)]$

Table 1. Chosen methods of determining the number of clusters in the data set [a] (cont.)

CRITERION	Formula, confidence interval	Criterion of selecting the number of clusters
Upper tail rule (Mojena I)	$\alpha_{x+1} > \overline{\alpha} + k \cdot s_{\alpha}$	classification P_x, where the corresponding step $x : x = 1,...,n-2$ is the first one which satisfies the given inequality
Moving average (Mojena II)	$\alpha_{x+1} > \overline{\alpha}_x + L_x + b_x + k \cdot s_x$ .where: $L_x = \frac{(y-1)b_x}{2}$, $b_x = \frac{6\left[2\sum_{f=x-y+1}^{x} w_f \alpha_f - (y+1)\sum_{f=x-y+1}^{x} \alpha_f\right]}{y(y^2-1)}$, $w_f = w_{f-1} + 1,\ f = (x-y+2),...,x$, $w_{x-y+1} = 1$	classification P_x, where the corresponding step $x : x = y, y+1,...,n-2$ is the first one which satisfies the given inequality
Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)	A comparison of the classification sequence results obtained as a result of agglomeration methods application with the family of trees generated by a random permutation of the data set	H_0 where the structure of grouping objects as a given tree is random (no structure), $H_1 :\sim H_0$

a n – number of objects $(i = 1,...,n)$; m – number of characteristics $(j = 1,...,m)$; u – number of groups $(q, r, s = 1,...,u)$; $K_q - q$ cluster; S_+, S_- – number of the distance pairs, consistent and inconsistent respectively; $tr(B_u)$, $tr(W_u)$ – trace of the covariance matrix, between-groups (B_u) and within-groups (W_u) respectively; $S_q = \sqrt[t]{(1/n_q)\sum_{i \in K_q}\sum_{j=1}^{m}\left|x_{ij}^{q} - z_{qj}\right|^t}$ – measurement of dispersion for objects in the q group (K_q), where for $t = 1$ it is the average distance of objects in the q cluster (K_q) from the center of gravity, i.e. the medoid in the group, and for $t = 2$ it is a standard deviation of distance of objects in the q cluster (K_q) from the center of gravity, i.e. the medoid in the group (for the r group the measurement S_r can be

analogically obtained); $d(q,r) = \sqrt[p]{\sum_{j=1}^{m} \left| z_{qj} - z_{rj} \right|^p}$ – measurement of the distance between the gravity centers, i.e. medoids (z_{qj}, z_{rj}) of the q and r groups, respectively Manhattan distance for $p = 1$ and the Euclidian distance for $p = 2$; $D(u)$ – sum of all within-groups distances; l_w – number of within-groups distances; $D_{\min}, D_{\max}$ – within-groups distance, respectively the smallest and the largest; $\alpha_x = \min_{i<o}[d_{io}]$, $(i, o = 1,...,n-x)$ – measure of dissimilarity (of distances) between the clusters; α_{x+1} – level (distance) of groups fusion in the step $x+1$, $\overline{\alpha}$ – average level (distance) of groups fusion, s_α – standard deviation of the level (distance) of group fusion; constant k, $k \in (2,75;\ 3,5)$; y – number of values of the level (distance) of classes fusion α in a given case, which is used to calculate the moving average; $\overline{\alpha}_x$ – moving average of the α parameter value calculated in the step x; L_x – correction for the delayed "trend" level (distance) of classes fusion calculated in the step x; b_x – „moving" meansquare slope of the trend line for the level (distance) of classes fusion in the step x; s_x – „moving" standard deviation of the α parameter (distance) value.

Source: based on [Mojena 1977; Wishart 2006; Gatnar, Walesiak 2009].

3. Assumptions and the empirical analysis scheme

The empirical analysis of the effectiveness of two Mojena criteria and Wishart criterion against other four criteria, namely Baker and Hubert (BH), Caliński and Harabasz (CH), Davies and Bouldin (DB), and Hubert and Levine (HL), was conducted for:

- 2-5 clusters,
- 2-5 variables,
- clusters with the following structure (100 objects):
  - 2 clusters containing 50 and 50 objects respectively,
  - 3 clusters containing 30, 30 and 40 objects respectively,
  - 4 clusters containing 10, 20, 30 and 40 objects respectively,
  - 5 clusters containing 5, 10, 15, 30 and 40 objects respectively,
- clusters without noisy variables,
- clusters generated on the basis of the same (identical) covariance matrix of variables,
- Euclidian distance measure,
- the three most popular agglomeration methods – the complete linkage, the average linkage and the Ward's method.

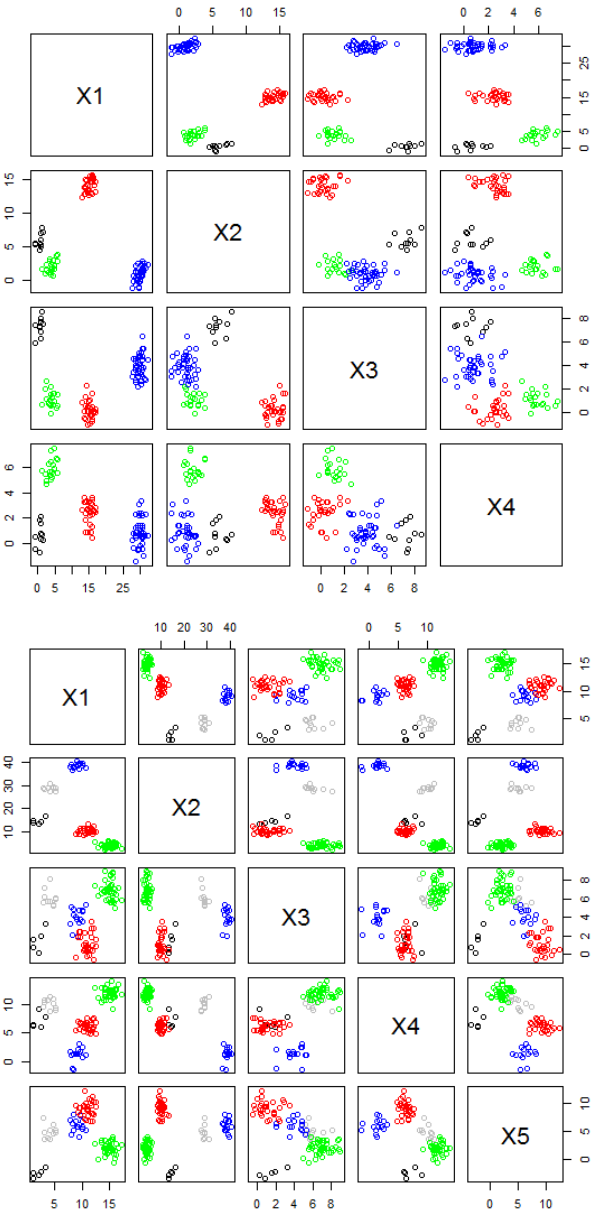
As a result, 16 data sets[1] were analyzed, the authors took into account 4 variants of the number of clusters and 4 variants of the number of variables. Three agglomeration methods were used here, one of the analyzed data sets is presented on the Figure 1.

The calculations were performed on the data sets generated with `cluster.Gen` function of `clusterSim` package [Walesiak, Dudek 2012] working in the R environment and using the *ClustanGraphics 8* software application [Wishart 2006], and the computations scheme looked as follows:

- Step 1, the data sets were generated on the basis of the assumptions (16 sets) for which the correct clusters structure was known (2-5 clusters),
- Step 2, in the *ClustanGraphics 8* application the cluster analysis was carried out using three agglomerative clustering algorithms (48 results obtained), and each dendrogram result with all clustering sets of objects was written to a file,
- Step 3, in the *ClustanGraphics 8* application the correct number of clusters was determined – as a result of grouping, according to two Mojena criteria and the Wishart criterion,
- Step 4, in the R environment (clusterSim) other indices for determining the number of clusters were calculated – Baker and Hubert (BH), Caliński and Harabasz (CH), Davies and Bouldin (DB), Hubert and Levine (HL) for divisions of 2 to 10 clusters, and the optimal result was chosen according to each criterion for selecting the number of classes: BH (max), CH (max), DB (min), HL (min),
- Step 5, with the given result of cluster analysis – the clusters structure indicated by the chosen criteria of determining the number of clusters for each analyzed dataset – adjusted Rand index was calculated in order to assess consistency of objects belonging to the cluster formed on the basis of each criterion, and the factual cluster of the analyzed set of objects (the known structure of the classes),
- Step 6, taking into account all clustering results, i.e. correct in the structure and objects belonging to the cluster – the Rand index values close to the unity, and incorrect in the structure and objects belonging to the cluster – the Rand index values close to zero, the assessment of the effectiveness of the analyzed criteria for selecting the number of clusters was done by averaging the Rand index values for each agglomeration method and each criterion.

[1] The complete characteristic of all analyzed data sets cannot be presented in the article as it is very wide.

Figure 1. The data set generated for 4 clusters and 4 variables, and for 5 clusters and 5 variables [a]



a In the case of sets with the structure 4 clusters vs. 4 variables and 5 clusters vs. 5 variables, the best criteria for determining the number of clusters (independently from the agglomeration method) were the following ones: BH, CH, DB, and the Wishart criterion, all determined the correct number of clusters and the grouping result structure (in 100%). The upper tail rule was effective for

the sets above only when the Ward's method was used. The moving average rule criterion determined incorrect number of clusters and incorrect clustering result structure for a set of clusters vs. 4 variables when the Ward's method was used. HL criterion incorrectly determined the number of clusters and clustering result structure for the sets above.

Source: own study using `R` environment `clusterSim` package.

4. The results of the empirical analysis

The Table 2 below presents, for each analyzed criterion of the number of clusters selection, the number of correct and incorrect indications according to the agglomerative clustering methods for the 16 analyzed data sets, where the clusters were generated on the basis of the same (identical) covariance matrix of variables. The results were summarized to create a basis for the assessment of the validation of indications determining the number of clusters for different criteria.

Table 2. Assessment of the number of clusters according to the number of clusters selection criteria

Method	The indication			
	correct		incorrect	
Baker and Hubert (BH)				
Average linkage	10	62.50%	6	37.50%
Complete linkage	11	68.75%	5	31.25%
Ward's	12	75.00%	4	25.00%
Caliński and Harabasz (CH)				
Average linkage	13	81.25%	3	18.75%
Complete linkage	12	75.00%	4	25.00%
Ward's	12	75.00%	4	25.00%
Davies and Bouldin (DB)				
Average linkage	6	37.50%	10	**62.50%**
Complete linkage	10	62.50%	6	37.50%
Ward's	12	75.00%	4	25.00%

Table 2. Assessment of the number of clusters according to the number of clusters selection criteria (cont.)

Method	The indication			
	correct		incorrect	
Hubert and Levine (HL)				
Average linkage	1	6.25%	15	**93.75%**
Complete linkage	3	18.75%	13	**81.25%**
Ward's	0	0.00%	16	**100.00%**
The upper tail rule (Mojena I)				
Average linkage	1	6.25%	15	**93.75%**
Complete linkage	1	6.25%	15	**93.75%**
Ward's	7	43.75%	9	**56.25%**
The moving average rule (Mojena II)				
Average linkage	13	81.25%	3	18.75%
Complete linkage	13	81.25%	3	18.75%
Ward's	3	18.75%	13	**81.25%**
Tree validation – the criterion on the randomness of the objects clustering in the dendrogram (Wishart)				
Average linkage	10	62.50%	6	37.50%
Complete linkage	12	75.00%	4	25.00%
Ward's	14	87.50%	2	12.50%

Source: own computations.

Baker and Hubert criterion – independently of the agglomerative clustering method – in about 2/3 to 3/4 cases indicated the correct number of classes.

More correct results on the indications of the number of clusters were obtained on the basis of Harabasz and Caliński criterion, its accuracy level of the correct number of clusters selection, independently of the agglomerative clustering method, was at least 75.00%.

Davies and Bouldin index more often indicated the correct number of clusters for the analyzed data sets when the complete linkage and the Ward's methods were applied, whereas for the average linkage method the validation level of indications did not exceed 37.50%.

The next two criteria of Hubert and Levine, and the upper tail (Mojena I criterion) did not give the correct solutions for the number of clusters selection in the analyzed data sets, where the clusters were generated on the basis of the same (identical) covariance matrix of variables.

The level of incorrect clusters indications in the case of Hubert and Levine – independently of the agglomerative clustering method – exceeded 81.25%, and in the case of the upper tail rule misclassification rate was equal to 56.25% (the Ward's method) and 93.75% (average linkage and complete linkage methods).

On the basis of the correct indications results of clusters number for Mojena II criterion (moving average rule) it can be stated that in the case of average linkage and complete linkage methods in most cases (81.25%) it is able to determine the correct number of clusters for the analyzed data sets. It is specific that the moving average rule completely failed in the agglomerative clustering with the Ward's method.

The last of the analyzed criteria, the tree validation rule, which evaluates the randomness of the objects clustering in the dendrogram, indicated correctly the number of classes for the average linkage agglomeration method in about 2/3 cases, whereas in the case of complete linkage and the Ward's methods percentage of the correct number of clusters indications, where this decision rule was also used, it was equal respectively to 75.00% and 87.50%. Therefore, next to the Baker and Hubert and Harabasz and Caliński criteria, it should be recognized as the most stable and effective one.

It should be emphasized that the correctness (frequency) of the number of clusters indications by the different criteria is the main, however not the sufficient premise for the "quality" assessment, i.e. relevance of the considered criterion for selecting the number of clusters. The most important is the consistency of the grouping result in the term of objects belonging to the respective cluster, and therefore consistency of the clustering result with the known clustering structure in the generated data sets, for which the assessment based on the adjusted Rand index was taken.

Therefore, Table 3 presents the consistency assessment of the clustering result with the known clustering structure, which is based on the average of adjusted Rand index values with respect to each of the agglomeration methods and each criterion.

The averaging operation was made with taking into account clusters structure for all clustering results, i.e. the correct and incorrect number of clusters indications for each criterion, that are related to all 16 data sets (see Table 2).

Thus, in the effectiveness assessment two aspects should be included – the number of "good" and "bad" solutions (clustering results) indicated by different

criteria for selecting the number of clusters and the consistency of each grouping result with the known structure of the classes.

Table 3. Consistency of the grouping result according to the number of clusters selection criteria

METHOD	Criterion	Average value of adjusted Rand's index
Average linkage	Baker and Hubert (BH)	0.904
	Caliński and Harabasz (CH)	0.919
	Davies and Bouldin (DB)	0.871
	Hubert and Levine (HL)	0.713
	The upper tail rule (Mojena I)	**0.459**
	The moving average rule (Mojena II)	**0.945**
	Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)	**0.946**
Complete linkage	Baker and Hubert (BH)	0.888
	Caliński and Harabasz (CH)	0.887
	Davies and Bouldin (DB)	0.859
	Hubert and Levine (HL)	0.778
	The upper tail rule (Mojena I)	**0.354**
	The moving average rule (Mojena II)	**0.905**
	Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)	**0.890**
Ward's	Baker and Hubert (BH)	0.894
	Caliński and Harabasz (CH)	0.894
	Davies and Bouldin (DB)	0.905
	Hubert and Levine (HL)	0.622
	The upper tail rule (Mojena I)	0.771
	The moving average rule (Mojena II)	**0.602**
	Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)	**0.943**

Source: own calculations.

5. Conclusions

Taking into account the results of the analyzes (see Table 2, Table 3) it can be stated that from the selected procedures aiming to determine the number of clusters, the most effective for the agglomeration methods are two of three criteria which are dedicated to the agglomerative clustering algorithms.

In the case of using medium linkage rule, high correctness in the indications of the number of clusters and the highest average consistency of the clustering result with the known clustering structure were obtained for the Wishart rule – the criterion of the randomness of the objects clustering in the dendrogram (0.946). Also, Mojena II criterion of the moving average gave high correctness and consistency (0.945). Moreover, in the cluster analysis using the complete linkage method high correctness of the grouping result and the highest average result consistency with the known classes structure was guaranteed by the same criterion (Mojena II, 0.905). The Wishart criterion on the randomness of the objects clustering in the dendrogram was characterized by a comparable high correctness and consistency (0.890). However, in the classification using the Ward's method, only Wishart criterion – tree validation (0.943) gave high correctness for the number of clusters indication and the highest average grouping result consistency with the known classes structure.

The conducted study shows that Mojena I criterion of the upper tail is absolutely ineffective in correct determination of the number of clusters, and recognition of the structure of objects belonging to the clusters. The grouping result consistence for the analyzed data sets and the mentioned criterion in the case of the medium linkage, complete linkage and the Ward's methods were equal to (0.459), (0.354) and (0.602) respectively.

Among other methods of selection of the number of clusters in a given data set, the following two criteria should be mentioned: Baker and Hubert (BH) and Caliński and Harabasz (CH), both characterized by high correctness of the number of clusters indications and consistency of the grouping result with the known class structure. Moreover, Hubert and Levin criterion usually erroneously determines the number of clusters.

REFERENCES

GAN, G., MA C., WU, J., Data clustering: theory, algorithms, and applications, SIAM, Philadelphia 2007.

GATNAR, E., WALESIAK, M. (ed.), Statystyczna analiza danych z wykorzystaniem programu R, Wydawnictwo PWN, Warsaw 2009.

MIKULEC, A., Metody oceny wyniku grupowania w analizie skupień, [in:] Jajuga K., Walesiak M. (ed.), Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2012.

MOJENA, R., Hierarchical grouping methods and stopping rules: an evaluation, „Computer Journal" 1977, vol. 20 (4), p. 359-363.

WISHART, D., Clustangraphics primer: a guide to cluster analysis, (the 4th edition), Edinburgh 2006.

www.clustan.com