ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 141, 1997

*Agnieszka Rossa**

A MONTE CARLO INVESTIGATION OF TWO DISTANCE MEASURES BETWEEN STATISTICAL POPULATIONS AND THEIR APPLICATION TO CLUSTER ANALYSIS

Abstract. The paper deals with a simulation study of one of the well-known hierarchical cluster analysis methods applied to classifying the statistical populations. In particular, the problem of clustering the univariate normal populations is studied. Two measures of the distance between statistical populations are considered: the Mahalanobis distance measure which is defined for normally distributed populations under assumption that the covariance matrices are equal and the Kullback-Leibler divergence (the so called Generalized Mahalanobis Distance) the use of which is extended on populations of any distribution.

The simulation study is concerned with the set of 15 univariate normal populations, variances of which are chanched during successive steps. The aim is to study robustness of the nearest neighbour method to departure from the variance equality assumption when the Mahalanobis distance formula is applied. The differences between two cluster families, obtained for the same set of populations but with the different distance matrices applied, are studied. The distance between both final cluster sets is measured by means of the Marczewski-Steinhaus distance.

Key words: hierarchical cluster analysis methods, robustness of the nearest neighbour method, the Mahalanobis distance, the Kullback-Leibler divergence, the Marczewski-Steinhaus distance measure.

1. THE BASIC NOTIONS

Let n multivariate statistical populations $\Pi_1, \Pi_2, ..., \Pi_n$ be given distributed according to density functions $f_1, f_2, ..., f_n$, respectively. The starting point of the hierarchical cluster analysis procedures is constructing a distance matrix **D**, elements of which express distances between each of the two populations Π_i and $\Pi_j (i, j = 1, 2, ..., n)$.

* University of Łódź, Chair of Statistical Methods.

$\mathbf{D} = [d_{ij}]$ $i = 1, 2, ..., n$; $j = 1, 2, ..., n$;

One of the most popular distance formulae applied to measuring distances between statistical populations is the Mahalanobis distance measure (M a h a l a n o b i s 1936), defined under assumption that the populations are normally distributed and have a common covariance matrix $\Pi_i \sim N(\mu_i, \sum)$ for $i = 1, 2, ..., n$. The Mahalanobis distance between two populations Π_i and Π_j takes the form

$$\Delta(i, j) = (\mu_i - \mu_j)^T \sum^{-1} (\mu_i - \mu_j) \tag{1.1}$$

K u l l b a c k and L e i b l e r (1951) introduced a distance measure between statistical populations called "divergence". The Kullback-Leibler divergence is a more general measure than the Mahalanobis distance. It can be used without limitation to the case of normal populations with equal covariance matrices.

Let two multivariate populations Π_i and Π_j have the respective probability densities $f_i(x_1, x_2, ..., x_k)$ and $f_j(x_1, x_2, ..., x_k)$ which are equivalent, i.e.

$$\int_A f_i(x_1, x_2, ..., x_k) = 0 \iff \int_A f_j(x_1, x_2, ..., x_k) = 0$$

for any $A \in B(R^k)$. Then the divergence between Π_i and Π_j was defined by Kullback and Leibler in the following form

$$J(i, j) = \int_{R^k} (f_i(\mathbf{x}) - f_j(\mathbf{x})) \cdot \log \frac{f_i(\mathbf{x})}{f_j(\mathbf{x})} d\mathbf{x} \tag{1.2}$$

where $\mathbf{x} = (x_1, x_2, ..., x_k)$.

In the case of normal populations (when $\Pi_i \sim N(\mu, \Sigma_i)$ and $\Pi_j \sim N(\mu, \Sigma_j)$), the divergence formula becomes

$$J(i, j) = \tfrac{1}{2} Tr[(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})] + \tfrac{1}{2}(\mu_i - \mu_j)^T (\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) \tag{1.3}$$

and in the particular case, when $\Sigma_i = \Sigma_j = \Sigma$ the divergence $J(i, j)$ has the form

$$J(i, j) = (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j) = \Delta(i, j) \tag{1.4}$$

One can easily see from the equality (1.4) that the Mahalanobis distance is a special case of the Kullback-Leibler divergence for normal populations with a common covariance matrix.

2. DEPARTURE FROM THE COVARIANCE EQUALITY ASSUMPTION – A SIMULATION STUDY

The Mahalanobis distance is a distance measure often applied, used in the cluster analysis for measuring distances between statistical populations. However, its use is limited to the case of normal populations with the equal covariance matrices. In practice, the Mahalanobis distance is employed even when the assumptions are not satisfied. The following question arises: how much disregarding the above mentioned assumptions affects the final results of clustering? Do they deviate much from the correct results or not? Let us consider the following example.

Example

Let 4 univariate normal populations Π_1, Π_2, Π_3, Π_4 be given

$\Pi_1 \sim N(5, 1)$ $\Pi_2 \sim N(1, 5)$ $\Pi_3 \sim N(6, 9)$ $\Pi_4 \sim N(0, 15)$

It can be easily seen, that the given standard deviations differ much from one another. Thus, the variance equality assumption is not satisfied. In that case the Mahalanobis distance formula provides us with the wrong distance matrix of the given set of objects Π_1, Π_2, Π_3, Π_4. In spite of this, let us not regard the above mentioned assumption and try to evaluate the distance matrix by means of the Mahalanobis distance formula. For this purpose it is necessary to adopt a variance value which would be common for all the populations Π_1, Π_2, Π_3, Π_4. In practice such a common variance is evaluated as a mean of all variances. The common variance becomes

$$\sigma = \tfrac{1}{2}(\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2) = 83$$

According to the Mahalanobis distance formula (1.1), adjusted to the univariate case, we obtain the following distances

	Π_1	Π_2	Π_3
Π_2	0.193		
Π_3	0.012	0.301	
Π_4	0.301	0.012	0.434

Now, using one of the well-known hierarchical cluster analysis methods (e.g. nearest neighbour method) the following family of clusters is obtained

$$A = \{(\Pi_1, \Pi_3), (\Pi_2, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4)\} \tag{2.1}$$

The results can be presented also graphically in the form of "a tree with a root".

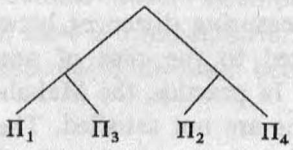

Figure 1. Graphical representation of the cluster family A

The family A in (2.1) and its graphical representation (see Fig. 1) are the final results of clustering.

We cannot forget however that the results may deviate much from the correct ones, because of the dissatisfied assumption concerning the equality of the population variances. It seems to be more reasonable to use in that example the Kullback-Leibler divergence formula (1.3) derived for normally distributed populations with unequal covariance matrices. The correct distances calculated for the same set of objects by means of the Kullback-Leibler formula (1.3), adjusted to the univariate case, are the following

	Π_1	Π_2	Π_3
Π_2	19.840		
Π_3	40.012	1.429	
Π_4	124.058	3.578	0.871

This leads to the following family of clusters

$$A' = \{(\Pi_3, \Pi_4), (\Pi_2, \Pi_3, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4)\} \tag{2.2}$$

represented in Fig. 2.

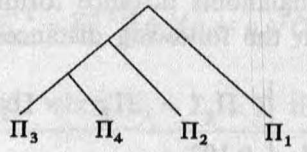

Figure 2. Graphical representation of the cluster family A'

We can see that the last results differ from the previous ones (see Fig. 1). The cluster family A differs from the cluster family A', although both of these families were obtained for the same set of objects

$\Pi_1, \Pi_2, \Pi_3, \Pi_4$. But the important question is, how much they differ and how to measure the similarity between the sets A and A'? In order to answer the question we need to find a measure which could express the degree of similarity between the families A and A'.

For this purpose we applied the so-called Marczewski-Steinhaus distance measure, defined for two families of subsets of the same set.

The Marczewski-Steinhaus distance measure

The Marczewski-Steinhaus distance measure is defined for two families of subset of the same set (Marczewski and Steinhaus 1958, Karoński and Palka 1977). Let us denote by F_i the i-th cluster of the family A and by E_i the i-th cluster of the family A'. The distance between the families A and A' takes the form

$$d(A, A') = \frac{1}{n-1} \min_{p \in P} \sum_{i=1}^{n-1} \frac{card(F_i - E_{p,i}) + card(E_{p,i} - F_i)}{card(F_i \cup E_{p,i})} \tag{2.3}$$

and $d(A, A') \in \langle 0, 1 \rangle$,

where p is the permutation of the first $n-1$ integers and P is the set of all such permutations.

Let us evaluate the Marczewski-Steinhaus distance $d(A, A')$ for two families of subsets (2.1) and (2.2) given in the example. The first family is the following

$$A = \{(\Pi_1, \Pi_3), (\Pi_2, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4)\}$$

and the second one is

$$A' = \{(\Pi_3, \Pi_4), (\Pi_2, \Pi_3, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4)\}$$

Now we consider 6 permutations of subsets of the family A'. Thus

$$A'_{p_1} = \{(\Pi_3, \Pi_4), (\Pi_2, \Pi_3, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4)\}$$
$$A'_{p_2} = \{(\Pi_3, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4), (\Pi_2, \Pi_3, \Pi_4)\}$$
$$A'_{p_3} = \{(\Pi_2, \Pi_3, \Pi_4), (\Pi_3, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4)\}$$
$$A'_{p_4} = \{(\Pi_2, \Pi_3, \Pi_4), (\Pi_1, \Pi_2, \Pi_3, \Pi_4), (\Pi_3, \Pi_4)\}$$
$$A'_{p_5} = \{(\Pi_1, \Pi_2, \Pi_3, \Pi_4), (\Pi_2, \Pi_3, \Pi_4), (\Pi_3, \Pi_4)\}$$
$$A'_{p_6} = \{(\Pi_1, \Pi_2, \Pi_3, \Pi_4), (\Pi_3, \Pi_4), (\Pi_2, \Pi_3, \Pi_4)\}$$

According to the formula (2.3) we obtain the following schemes of calculations for the first permutation p_1 (see Tab. 1).

Table 1

Schemes of middle calculations of the Marczewski-Steinhaus distance for two families of clusters: the family A given in (2.1) and the first permutation A'_{p_1} of the family A' given in (2.2).

F_i	F_1	F_2	F_3
	(Π_1, Π_3)	(Π_2, Π_4)	$(\Pi_1, \Pi_2, \Pi_3, \Pi_4)$
$E_{p_1,i}$	$E_{p_1,1}$	$E_{p_1,2}$	$E_{p_1,3}$
	(Π_3, Π_4)	(Π_2, Π_3, Π_4)	$(\Pi_1, \Pi_2, \Pi_3, \Pi_4)$
$F_i - E_{p_1,i}$	Π_1	$\varnothing$	$\varnothing$
$c1 = card(F_i - E_{p_1,i})$	1	0	0
$E_{p_1,i} - F_i$	Π_4	Π_3	$\varnothing$
$c2 = card(E_{p_1,i} - F_i)$	1	1	0
$F_i \cup E_{p_1,i}$	(Π_1, Π_3, Π_4)	(Π_2, Π_3, Π_4)	$(\Pi_1, \Pi_2, \Pi_3, \Pi_4)$
$c3 = card(F_i \cup E_{p_1,i})$	3	3	4
$\frac{c1+c2}{c3}$	$\frac{2}{3}$	$\frac{1}{3}$	0

We obtain that

$$S_1 = \sum_{i=1}^{3} \frac{card(F_i - E_{p_1,i}) + card(E_{p_1,i} - F_i)}{card(F_i \cup E_{p_1,i})} = \frac{2}{3} + \frac{1}{3} + 0 = 1$$

Continuing the calculations for all the permutations of the clusters of the set A' we obtain

$$S_2 = \frac{2}{3} + \frac{2}{4} + \frac{1}{4} = \frac{17}{12} \qquad S_5 = \frac{2}{4} + \frac{1}{3} + \frac{2}{4} = \frac{4}{3}$$

$$S_3 = \frac{3}{4} + \frac{2}{3} + 0 = \frac{17}{12} \qquad S_6 = \frac{2}{4} + \frac{2}{3} + \frac{1}{4} = \frac{17}{12}.$$

$$S_4 = \frac{3}{4} + \frac{2}{4} + \frac{2}{4} = \frac{7}{4}$$

Finnally, the Marczewski-Steinhaus distance has the value

$$d(A, A') = \frac{1}{3} \min\left\{1, \frac{17}{12}, \frac{17}{12}, \frac{7}{4}, \frac{4}{3}, \frac{17}{12}\right\} = \frac{1}{3} \cdot 1 = 0.33.$$

Thus, the distance between the family A and the family A' or between the trees G and G' is equal to 0.33. It follows from the analyzed example that the results of the cluster analysis, based on the Mahalanobis distance, can deviate even much from the correct results, if the assumption concerning the variance equality is not satisfied.

Simulation study

In this section we present the results of a computer simulation study performed similarly as described in the example but for a larger number of univariate populations. Let us assume that all populations are normally distributed with the expected values as follows

$m_1 = 4.86$ $m_6 = 4.91$ $m_{11} = 4.96$
$m_2 = 4.87$ $m_7 = 4.92$ $m_{12} = 4.97$
$m_3 = 4.88$ $m_8 = 4.93$ $m_{13} = 4.98$
$m_4 = 4.89$ $m_9 = 4.94$ $m_{14} = 4.99$
$m_5 = 4.90$ $m_{10} = 4.95$ $m_{15} = 5.00$

The aim is to study sensitivity of the cluster analysis methods (with the Mahalanobis distance matrix applied) to departure from the variance equality assumption. The results of such an investigation for the nearest neighbour method are presented in the Tab. 2.

Table 2

The Marczewski-Steinhaus distance values expressing robustness of the nearest neighbour method to departure from variance equality assumption

The variances of the populations								the Marczewski-Steinhaus distance
σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2	...	σ_{15}^2	
4	4	4	4	4	4	...	4	0.00
6	4	4	4	4	4	...	4	0.14
8	6	4	4	4	4	...	4	0.24
8	8	4	4	4	4	...	4	0.24
8	6	6	4	4	4	...	4	0.24
8	8	6	4	4	4	...	4	0.28
8	8	8	4	4	4	...	4	0.28
8	8	6	6	4	4	...	4	0.31
8	8	8	6	6	4	...	4	0.29
6	6	8	8	8	4	...	4	0.34

The numbers in the second column of the Tab. 2 represent the values of the Marczewski-Steinhaus distance between two families of clusters. Both families were obtained for the same set of populations by means of the nearest neighbour method but with the different distance matrices applied. In the first case the Mahalanobis distance formula (1.1) was applied under assumption that all population variances are equal. In the second case the Kullback-Leibler distance formula (1.3) was used, (the so-called Generalized

Mahalanobis Distance), the use of which is extended on normal populations with various covariance matrices.

3. FINAL REMARKS

The simulation results lead to the conclusion that the nearest neighbour method based on the Mahalanobis distance measure is not robust to departure from the variance equality assumption.

REFERENCES

Beran R. (1977): *Minimum Hellinger distance estimates for parametric models*, The Annals of Stat., Vol. 5, p. 445–463.

Bhattacharyya A. (1943): *On a measure of divergence between two statistical populations defined by their probability distributions*, Bull. Calcutta Math. Soc., Vol. 35, p. 99–109.

Everitt B. S. (1979): *A Monte Carlo investigation of the robustness of Hotteling's one- and two-sample T^2 tests*, Journal Amer. Stat. Assoc., Vol. 74, p. 48–51.

Holloway L. S., Dunn O. J. (1967): *The robustness of Hotteling's T^2*, Journal Amer. Stat. Assoc., Vol. 62, p. 124–136.

Hopkins J. W., Clay P. P. F. (1963): *Some empirical distributions of bivariate T^2 and homoscedasticity criterion M under unequal variance and leptocurtosis*, Journal Amer. Stat. Assoc., Vol. 58, p. 1048–1053.

Johnson M. E., Wang C., Ramberg J. S. (1979): *Robustness of Fisher's linear discriminant function to departures from normality*, Informal Report LA-8068-MS, Los Alamos Scientific Laboratory, University of California, October 1979.

Kobayashi H. M. (1970): *Distance measures and asymptotic relative efficiency*, IEEE Trans. Inform. Theory, Vol. IT-16, p. 288–291.

Kailath T. (1967): *The divergence and Bhattacharyya distance measures in signal selection*, IEEE Trans. Communication Technology, Vol. COM-15, p. 52–60.

Karoński M., Palka A. (1977): *On Marczewski-Steinhaus type distance between hypergraphs*, „Zastosowania Matematyki", XVI, 1, p. 47–57.

Koichi I. (1969): *On the effect of heteroscedasticity and non-normality upon some multivariate test procedures, Multivariate Analysis*, Vol. 2, Academic Press, New York, p. 87–120.

Kullback S., Leibler R. A. (1951): *On information and sufficiency*, Annals of Math. Stat. Vol. 22, p. 79–86.

Kullback S. (1952): *An application of information theory to multivariate analysis*, Annals of Math. Stat. Vol. 23, p. 88–102.

Mahalanobis P. C. (1936): *On the generalized distance in statistics*, Proc. Nat. Inst. Sci. India, Vol. 12, p. 49–55.

Marczewski E., Steinhaus H. (1958): *On a certain distance of sets and the corresponding distance of functions*, Coll. Math. Vol. 6, p. 319–327.

Agnieszka Rossa

MIARY ODLEGŁOŚCI POMIĘDZY POPULACJAMI STATYSTYCZNYMI I ICH ZASTOSOWANIE W ANALIZIE SKUPIEŃ – BADANIE MONTE CARLO

W pracy zawarte zostały wyniki symulacyjnego badania dotyczącego jednej z metod hierarchicznego grupowania populacji statystycznych, tj. metody najbliższego sąsiedztwa. Punktem wyjścia jest konstrukcja macierzy odległości pomiędzy obiektami (tu pomiędzy populacjami statystycznymi). Celem pracy było zbadanie odporności wspomnianej metody aglomeracyjnej na odejście od założeń warunkujących zastosowanie określonej miary odległości. W badaniu uwzględnione zostały dwie miary odległości: odległość Mahalanobisa, zdefiniowana dla populacji normalnych o jednakowych macierzach kowariancji oraz odległość Kullbacka-Leiblera, będąca uogólnieniem odległości Mahalanobisa na przypadek populacji o dowolnych rozkładach. W pracy główny nacisk położony został na badanie odporności wspomnianej metody aglomeracyjnej na odejście od założenia o równości macierzy kowariancji. Badanie symulacyjne przeprowadzone zostało w odniesieniu do ustalonego z góry zbioru 15 jednowymiarowych populacji normalnych, których wariancje zmieniane były w kolejnych krokach. Celem badania było ustalenie stopnia różnic pomiędzy rodzinami skupień otrzymanymi dla danego zbioru populacji lecz przy użyciu innej macierzy odległości. Jako miarę stopnia różnic pomiędzy otrzymanymi rodzinami skupień wykorzystano odległość Marczewskiego-Steinhausa.