Comparing populations based on squared Euclidean distance between sets of variables

Dominika Polko-Zając[1]

Abstract

In economics as well as in other researches there is often a need to detect the differences between multidimensional populations. This paper concerns the problem of comparing such populations using distance between two analysed sets of objects being characterized by many variables. The study use squared Euclidean distance measure but method can be used with any kind of distance measure between objects. In order to identify differences permutation tests were used. The method is illustrated by analysing economics data sets. Included empirical example contains data from Central Statistical Office of Poland. All calculations were done in R program.

Keywords: *multidimensional data, distance approach, permutation tests, Monte Carlo study*

JEL Classification: C12, C15, C30

DOI: 10.14659/SEMF.2018.01.39

1 Introduction

Multidimensional methods are important part of statistical methods. The purpose of the research is usually to detect the similarity or existing differences between the examined objects being characterized by many diagnostic variables. This stems mainly from the fact that in many areas of empirical research they relate to phenomena of complex, multidimensional structure.

Statistical methods of multidimensional analysis e.g. cluster analysis are used to divide one set of data into homogeneous groups of objects. The level of assessment of the similarity/distance of objects allows to include them in the same group of objects or to conclude that there are no similarities between the objects of the study. Cluster analysis allows to organize objects when their structure is unknown (classification within groups). This structure should only be discovered while having multidimensional data about objects. The division of the set of objects is based on a certain distance measure between the examined objects $d(x_i, x_j)$, where $i, j = 1,...,n.$. The calculated distance matrices containing distances for each of the object pairs allow to evaluate the differences between these objects. Distance

[1] Corresponding author: University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics, ul. 1 Maja 50, 40-287 Katowice, Poland, e-mail: dominika.polko@ue.katowice.pl

measurements are characterized by the fact that the increase in value means an increase in the diversity between the examined objects.

The article considers a different approach. Samples (groups) were randomly taken from *k* populations of any continuous distributions. The purpose of the study was to test the differences between the *k* populations using for that the distance between objects in the considered samples. Proposed method can be applied even when the sample size is small.

2 Dissimilarity (distance) measures

Comparing the examined objects in data sets, e.g. countries, regions, etc., measures which determine the similarity or differences between pairs of these objects are calculated. The measures used most often are divided into measures of proximity and measures of distance. Literature provides numerous suggestions of measures applicable in multidimensional methods. Applying different measures of similarity of objects is not a problem when all variables describing objects are measured on a scale of one type. Tables with numerous distance measures for variables measured on an interval, order and nominal scale are presented by Walesiak and Gatnar (2009). When the quantitative variables are analysed (interval), the following distance measures are distinguished to measure the distance between points $x=(x_1,x_2...x_n)$ and $y=(y_1,y_2...y_n)$ in the *n*–dimensional space, e.g.:

- Manhattan

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i|, \tag{1}$$

- Euclidean

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}, \tag{2}$$

- Squared Euclidean

$$d(x,y)=\sum_{i=1}^{n}(x_i-y_i)^2, \tag{3}$$

- Chebychev

$$d(x,y)=\max_{1\le i\le n}|x_i-y_i|, \tag{4}$$

- Minkowski

$$d(x,y)=\sqrt[p]{\sum_{i=1}^{n}|x_i-y_i|^p}. \tag{5}$$

Method described in the next paragraphs can be used with any kind of distance measure between objects but further considerations (example) use only squared Euclidean distance measure.

3 Distances between groups

Applying multidimensional analysis, various methods are used to calculate distances between groups. These methods are used to optimize the pre–made division into groups of objects. There are five commonly used hierarchical clustering methods (Everitt et al., 2011):

- average between groups linkage, calculated as the average distance between all pairs of objects belonging to groups A and B,

$$d(A,B)=\frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d\left(O_{A_i},O_{B_j}\right)}{n_A \cdot n_B}, \tag{6}$$

- average within groups linkage, determined as the arithmetic mean of the distance between all possible pairs of objects belonging to groups A and B,

$$d(A,B)=\frac{\sum_{i=2}^{n_A}\sum_{p=1}^{i} d\left(O_{A_i},O_{A_p}\right)+\sum_{j=2}^{n_B}\sum_{q=1}^{j} d\left(O_{B_j},O_{B_q}\right)+\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d\left(O_{A_i},O_{B_j}\right)}{\frac{n_A\left(n_A-1\right)}{2}+\frac{n_B\left(n_B-1\right)}{2}+n_A \cdot n_B}, \tag{7}$$

- centroid method, calculated as the distance between the centroids of objects of groups A and B,

$$d(A,B)=d\left(\overline{X}_A,\overline{X}_B\right), \tag{8}$$

  where: $\overline{X}_A,\overline{X}_B$ are midpoints of objects of groups A and B,

- median method, determined as the median distance between objects belonging to groups A and B

$$d(A,B)=\underset{i,j}{median}\left\{d\left(O_{A_i},O_{B_j}\right)\right\}, i=1,...,n_A,\ j=1,...,n_B, \tag{9}$$

- Ward's method, where distance between groups defined as sum of squares within groups, after fusion, summed over all variables,
- methods based on the vicinity of objects, e.g. the nearest neighbourhood (single linkage), i.e. the distance between the nearest objects belonging to groups A and B respectively, or the furthest neighbourhood (complete linkage), i.e. the distance between the most distant objects belonging to groups A and B respectively.

For all methods of calculating the distance between groups the distance matrix is calculated using the suitable measures presented in the second part of this article. The methods discussed above excluding Centroid method, Median method and Ward's method can be used with any kind of similarity or distance measure between objects. Above mentioned three methods use squared Euclidean distances.

The goal of those methods is to obtain the most concentrated values in groups, so often in multidimensional analyses, methods based on determining the average distance between objects within groups are used. In the method presented in the article, the goal is not to obtain homogeneous groups but testing existing differences between them. The analysis used selected measures of inter–group distances or their modifications to determine test statistics during the verification of the hypothesis about the similarity of the studied populations based on sets of variables.

4 Testing differences based on distances between groups

Let's assume that there are two sets of objects $A=\{O_1,O_2,...,O_{n_A}\}$ and $B=\{O_1,O_2,...,O_{n_B}\}$ of sizes n_A and n_B respectively. Each object is described using p diagnostic variables. Comparing populations based on two sets of variables the null hypothesis can be stated as follows: "samples were taken from multidimensional populations having the same distributions", which suggests the equality of two distributions. The alternative hypothesis is formulated: "samples were taken from multidimensional populations having different distributions". In a two–sample testing problem formally these hypotheses can be written as follows

$$H_0 : F = G, \tag{10}$$

and the alternative

$$H_1 : F \neq G \tag{11}$$

where:

F, G – continuous but unknown multidimensional distributions of populations.

To test the null hypothesis the permutation test can be used. A test statistic is computed using two sets of independent p–dimensional observations. Permutation tests in general take a test statistic T used for a parametric test, or one derived intuitively (Baker, 1995). Unfortunately

many classical statistical methods do not have their counterparts for multidimensional data. To test null hypothesis against alternative hypothesis the following test statistic can be used

$$T^{(1)} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d\left(O_{A_i}, O_{B_j}\right)}{n_A \cdot n_B}, \tag{12}$$

where,

$d\left(O_{A_i}, O_{B_j}\right)$ distances between *i*-th object belonging to group A and *j*-th object belonging to group B in multidimensional space R^p, i.e. distance between $O_{A_i} = \left(x_{i1}, x_{i2}, ..., x_{ip}\right)$ and $O_{B_j} = \left(y_{j1}, y_{2j}, ..., y_{jp}\right)$, where $i, j = 1, ..., n$.

It is also possible to test the differences between populations using as a test statistic the distance between the centroids of the population

$$T^{(2)} = d\left(\overline{X}_A, \overline{X}_B\right), \tag{13}$$

where: $\overline{X}_A, \overline{X}_B$ are midpoints of objects of groups A and B.

The distance between centroids, or midpoints in a multidimensional space, can also be expressed directly from the distances between items in each group. The formula to find the centroid distance were presented by Apostol and Mnatsakanian (2003)

$$T^{(3)} = \frac{1}{n_A \cdot n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d^2\left(O_{A_i}, O_{B_j}\right) - \frac{1}{n_A^2}\sum_{i=2}^{n_A}\sum_{p=1}^{i} d^2\left(O_{A_i}, O_{A_p}\right) - \frac{1}{n_B^2}\sum_{j=2}^{n_B}\sum_{q=1}^{j} d^2\left(O_{B_j}, O_{B_q}\right), \tag{14}$$

where,

$\sum_{i=2}^{n_A}\sum_{p=1}^{i} d^2\left(O_{A_i}, O_{A_p}\right)$ is the sum of squared distances between objects in group A,

$\sum_{j=2}^{n_B}\sum_{q=1}^{j} d^2\left(O_{B_j}, O_{B_q}\right)$ is the sum of squared distances between objects in group B,

$\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d^2\left(O_{A_i}, O_{B_j}\right)$ is the sum of squared distances between objects in group A and those in group B.

In order to identify differences between sets of variables permutation tests were used. Tests based on permutations of observations were introduced by R. A. Fisher in 1930's (Welch, 1990). Because of the need to perform complex calculations method was widely used only in recent decades, when computing capabilities of computers increased. Currently, the problem of using the permutation tests in statistical analysis is popular among researchers. The most

important references are Good (1994), Good (2005), Good (2006), Pesarin (2001), Basso et al. (2009), Pesarin and Salmaso (2010) and Kończak (2016).

Most of multidimensional two–sample tests perform poorly for high dimensional data and many of them are not applicable when the dimension of the data exceeds the sample size (Biswas and Ghosh, 2014). Permutation tests do not require additional assumptions about the form of the distribution in the population; are suitable for small sample sizes and are robust to outliers. The goal of the test is to verify hypothesis at certain level of significance to discover a differences between data sets. After the value of the statistic T_0 had been calculated, N permutations of variables were performed and values T_i ($i = 1, 2, \ldots, N$) were determined. The decision concerning a verified hypothesis is made on the basis of *ASL* (*achieved significance level*) value (Efron and Tibshirani, 1993):

$$ASL = P_{H_0}\{T \geq T_0\}. \tag{15}$$

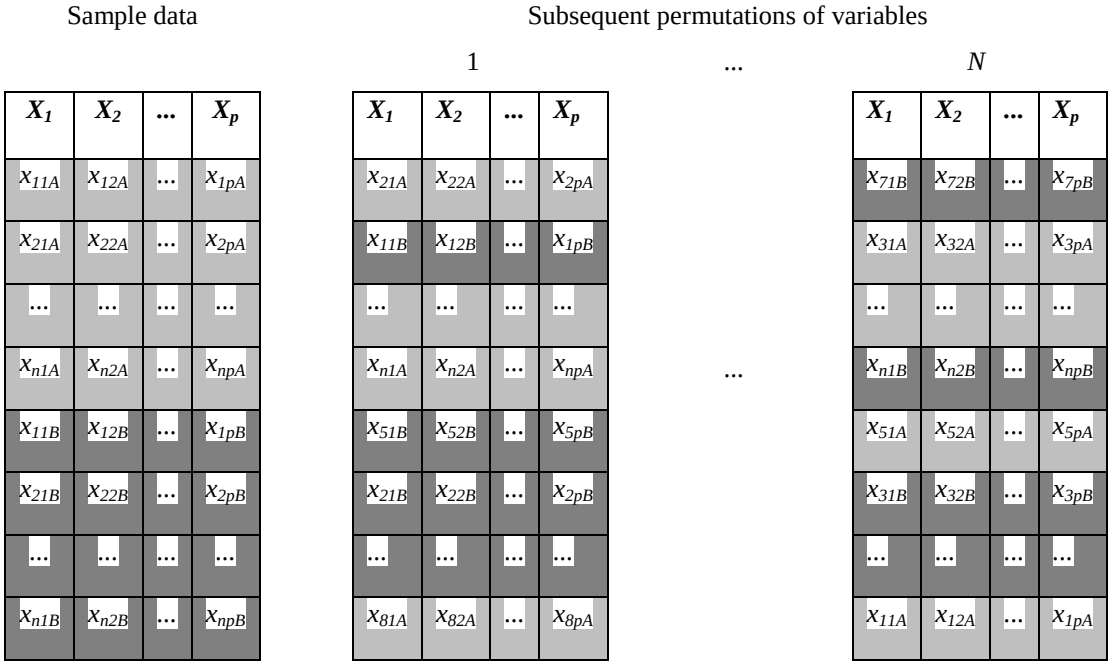

Fig. 1. The scheme of permutation variables.

On the basis of the random variable of large size (it is recommended in most cases the number of permutation to be greater than 1000) taken from the set of all possible permutations of the data set the *ASL* is determined using formula (Kończak, 2016)

$$\hat{ASL} = \frac{card\{i : T_i \geq T_0\}}{N}. \tag{16}$$

The smaller the value of *ASL*, the stronger evidence against H_0. Formally we choose a significance level α and reject H_0 if *ASL* is less than α.

The steps in the permutation test conducted to determine the significance of the difference between two sets of variables are as follows:

1. Assume the level of significance α;
2. Calculate the value of the statistics T_0 for the sample data;
3. Proceed permutations of data. A random permutation of data can be obtained by recreating the original data and destroying existing structure of variables (see. Fig. 1). Calculate test statistic values T_i for each permutation of data.
4. Create empirical distribution of T_i, where (i = 1, 2, … , N) and locate calculated value of T_0 on this distribution and estimate *ASL* value.

The simulation study was performed using R program (R Core Team, 2016). The author also presents another proposal for comparing multidimensional populations based on two sets of variables using permutation tests (Polko–Zając, 2017).

5 Empirical example

To illustrate the possibilities of application the method for the analysis of economic data, data from the Local Data Bank of the Central Statistical Office were used. The study concerned a comparison of the situation on the labour market. For this purpose, data for Polish voivodeships were used for two compared periods: 2004 and 2016 (Table 1). The data set contained five diagnostic variables:

X_1 – unemployment rate (in %),

X_2 – the percentage of the long–term unemployed, i.e. those unemployed for over a year in the total number of unemployed,

X_3 – the percentage of unemployed among people under 25 in the total number of unemployed,

X_4 – the percentage of unemployed people with no experience or length of employment up to 1 year in the total number of unemployed,

X_5 – the percentage of unemployed people with higher education in the total number of unemployed.

Table 1. Data of the situation on the labour market in 2004 and in 2016.

Voivodeship	Year 2004					Year 2016				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
Dolnośląskie	22.4	49.6	21.4	34.6	4.4	7.2	36.7	10.5	26.4	12.0
Kujawsko-Pomorskie	23.6	52.9	25.8	34.3	3.3	12.0	41.9	14.4	31.7	8.7
Lubelskie	17.8	54.7	28.0	49.9	7.4	10.3	45.4	16.0	46.1	15.4
Lubuskie	25.6	47.3	21.8	31.6	3.4	8.6	33.1	12.4	28.2	10.2
Łódzkie	19.5	54.8	22.4	36.3	5.4	8.5	43.1	11.2	29.3	11.8
Małopolskie	15.0	51.2	28.8	39.3	5.4	6.6	39.9	15.8	33.3	16.3
Mazowieckie	14.7	56.0	22.7	41.7	5.6	7.0	45.4	12.3	34.0	15.4
Opolskie	20.0	50.0	21.6	29.6	4.3	9.0	37.3	13.0	28.4	11.9
Podkarpackie	19.1	55.3	26.3	42.2	5.5	11.5	45.2	15.1	37.1	15.0
Podlaskie	16.1	50.2	25.6	43.2	6.5	10.3	46.0	14.6	38.5	14.5
Pomorskie	21.4	53.3	23.4	30.6	4.3	7.1	35.6	14.2	29.6	12.9
Śląskie	16.9	50.1	24.2	41.0	5.3	6.6	37.9	11.3	32.9	13.7
Świętokrzyskie	22.0	53.0	25.4	41.0	7.3	10.8	37.7	15.1	37.1	16.3
Warmińsko-Mazurskie	29.2	52.3	23.4	36.6	3.4	14.2	38.9	13.5	30.9	9.4
Wielkopolskie	15.9	48.4	27.4	34.1	4.3	4.9	35.0	14.6	27.5	13.0
Zachodniopomorskie	27.5	51.5	21.0	40.6	4.1	10.9	36.5	11.9	33.2	10.8

Source: Central Statistical Office of Poland (GUS).

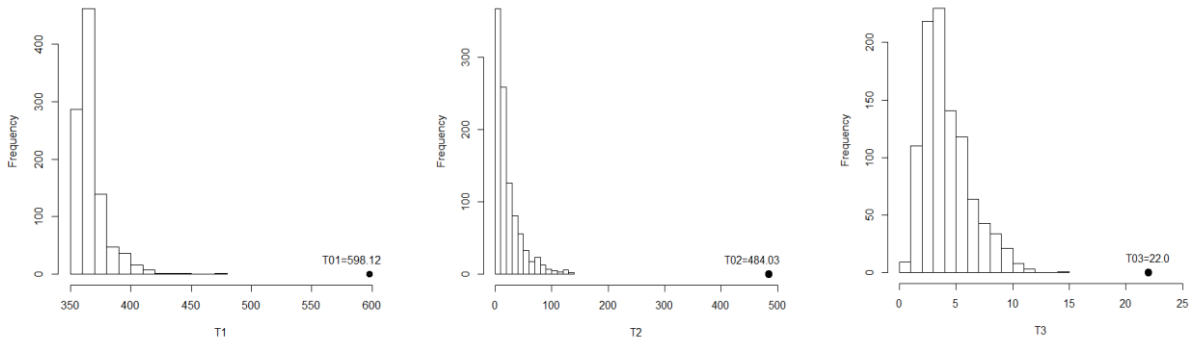

Fig. 2. Empirical distributions of statistics $T^{(1)}$, $T^{(2)}$ and $T^{(3)}$.

To test null hypothesis (10) the permutation test was used. Significance level $\alpha = 0.05$ was assumed and N=1000 permutations of variables were performed. As test statistics (12) – (14) were used. Statistics values were determined using the squared Euclidean distance measure (3). Values of test statistics calculated for the sample data are: T_{01}=598.12, T_{02}=484.03 and T_{03}=22.0. Empirical distributions of statistics were presented on Fig. 2. *ASL* values calculated with empirical distributions of statistics equal to zero so they are lower than assumed significance level. Verified hypothesis H_0 should be rejected in favour of alternative hypothesis, which means that there is the significant difference between situation on the labour market in 2004 and in 2016.

Conclusion

Many parametric and nonparametric methods are available for the multidimensional two–sample testing problem. Article deals with a permutation distance approach to multidimensional problem of hypothesis testing. The limitation of commonly used classical statistical methods makes the simulation methods being used in a variety of data analyses. The paper presents a method for testing distance between two analysed sets of objects characterized by many variables. In order to identify differences between populations permutation test was proposed. Introduced test based on comparing groups obtained by permuting the group's membership. The procedure using the permutation test is used to estimate the distribution of the test statistic. The proposed method is illustrated by an empirical example from real application for economics data sets.

References

Apostol, T. M. & Mnatsakanian, M. A. (2003). Sums of squares of distances in m-space. *Math. Assoc. Am. Monthly, 110*, 516.

Baker, R. D. (1995). 2 permutation tests of equality of variances. *Statistics and Computing*, *5*, 289–296.

Basso, D., Pesarin, F., Salmaso, L. & Solari, A. (2009). *Permutation Tests for Stochastic Ordering and ANOVA*. Heidelberg: Springer Science + Business Media.

Biswas, M. & Ghosh, A. K. (2014). A nonparametric two–sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, *123*, 160–171.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster Analysis*. *5th edition*. John Wiley & Sons.

Good, P. I. (1994). *Permutation Tests: A Practical Guide for Testing Hypotheses*. New York: Springer–Verlag.

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer Science Business Media.

Good, P. (2006). *Resampling Methods. A Practical Guide to Data Analysis*. Boston–Basel–Berlin: Birkhauser.

Kończak, G. (2016). *Testy permutacyjne. Teoria i zastosowania*. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Pesarin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics*. Chichester: John Wiley & Sons, Ltd.

Pesarin, F. & Salmaso, L. P. (2010). *Permutation Tests for Complex Data. Theory, Applications and Software*. Chichester: John Wiley & Sons, Ltd.

Polko–Zając, D. (2017). On comparing populations based on two sets of variables. In: Papież, M. & Śmiech S. (Eds.), *The 11th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio–Economic Phenomena*. Conference Proceedings. Cracow: Foundation of the Cracow University of Economics, 349-358. ISBN: 978–83–65173–85–0 (HTML).

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Walesiak, M. & Gatnar, E. (2009). *Statystyczna analiza danych z wykorzystaniem programu R*. Warszawa: Wydawnictwo Naukowe PWN.

Welch, W. J. (1990). Construction of Permutation Tests. *Journal of the American Statistical Association. Theory and Methods, 85*(411), 693–698.