ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 285, 2013

Jacek Stelmach[*]

TESTING DIFFERENCES BETWEEN POPULATIONS WITH EIGENVECTORS

Abstract. Testing differences between multivariate populations is one of a crucial problems in statistical investigations. The most known – MANOVA tests being parametric ones need to fulfill the assumptions about the conformity with multivariate normal distribution. Very often these assumptions are practically unrealistic or the verification, especially for small number of observations is hard.

This paper presents an approach, based on permutation tests (no needs of verification mentioned assumptions), where proposed test statistics base on the properties of eigenvectors. The investigations were carried out for simulated and real multivariate datasets, where the permutation tests were compared with variable-based and MANOVA test statistics.

Key words: permutation tests, multivariate analysis, eigenvectors.

I. INTRODUCTION

The testing of differences between populations is one of a crucial point in multivariate statistical inference. It enables exploration of how independent predictors influence a patterning of response on the dependent variable or to monitor the change of the investigated phenomenon. The most known method to test it is Multivariate Analysis of Variance (MANOVA) that gives the tool to test the equality of mean vectors for several groups. This method however – as a parametric one – requires fulfilling the assumptions, sometimes too stringent in examined cases. Ito (1980) says, that MANOVA is not sensitive to deviations from multivariate normality, however it doesn't concern the small sizes of observations.[1] Then non-parametric methods, based on permutation tests can be preferable. This paper presents a permutation test with the statistics based on the eigenvectors properties. Monte Carlo methods are used to estimate the power of the test, in cases where investigated populations differ not only with the mean vectors or variance.

[*] Ph.D. student, Department of Statistics, University of Economics, Katowice

[1] P. K. Ito, *Robustness of ANOVA and MANOVA test procedures,* Amsterdam: North Holland 1980, p. 220.

II. PROBLEM DESCRIPTION

Taking into considerations r multivariate populations (with p dimensions), that are observed as multivariate samples with n number of observations, the null hypothesis can be written as:

$$H_0 = F_1(x) = F_2(x) = ... = F_r(x) \tag{1}$$

where: $F_i(x)$ is a distribution of i^{th} population.

The most known and important group of tests is Multivariate Analysisi of Variance (MANOVA) where test statistics fall in two categories:

– distance based: a function of distances between the samples,

– variable based: a function of summary statistics created for each variable.

MANOVA enables verification of null hypotheses testing the equality of mean vectors for investigated populations. And as a parametric methods, needs to fulfill following assumptions:

1. normal distribution of the variable within groups,
2. homogeneity of variance across the range of variables,
3. homogeneity of variance/covariance matrix.

When these assumptions cannot be fulfilled, non-parametric tests are the alternative that can be carried out. Very often however, their critical values were calculated only for small number of variables[2]. Additionally, non-parametric tests – Wald-Wolfowitz, Mann-Whitney, Wilcoxon, Kruskal-Wallis, Mantel's – verify the equality of mean vectors, as well.

In batch processes, there is a need to test the difference between populations in several cases – taking into considerations the differences between the 'shape' of multivariate ellipsoids of received samples (i. e. between raw materials from different sources or different locations, indication of the failures in technological installations). No known literature has given a suggestion of proper test, able to distinguish such differences.

It was a main reason to look for such statistics, described in this article, that could be able to test the difference between the 'shapes' of given observations.

III. PROPOSED STATISTICS BASED ON EIGENVECTORS

Principal Component Analysis (PCA) allows reducing the dimensionality of original data set creating the smaller number of components. The first eigenvector and corresponding eigenvalue calculated during PCA determines a direction of the largest variance of data. The rest of eigenvectors determine the direction

[2] Cz. Domański, K. Pruska, *Nieklasyczne metody statystyczne,* Polskie Wydawnictwo Ekonomiczne, Warszawa 2000, p. 184.

of the rest of variance according to their eigenvalues. So if the data sets derive from the same population or from populations with very similar distributions, their eigenvectors should indicate the same directions – the same as for the eigenvectors created for the sum of all data sets – see figure 1.

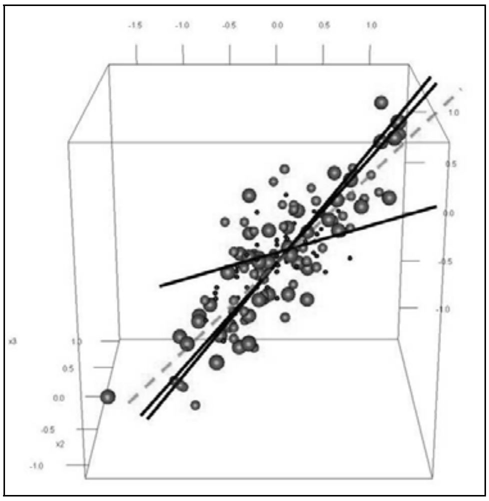

Figure 1. Three first eigenvectors calculated for three data sets indicate the directions of the largest variance. Two of them are very similar. However the third one indicates the different direction from the others and from the first eigenvector calculated for the sum of data set (dashed line)

Source: own calculations.

It should allow verifying the null hypothesis taking into consideration the 'shape' of multivariate ellipsoids of the populations. There were created test statistics based on the length of the vectors created as the difference between corresponding eigenvectors (taking into consideration only the eigenvectors with corresponding eigenvalues greater than 1):

$$EV = \sum_{i=1}^{r}\left(\sum_{j=1}^{m}\left(\frac{1}{m}\left\|V_{ij} - V_{sj}\right\|\sum_{k=1}^{m}\lambda_{ik}\right)\right) \qquad (2)$$

where:

m – number of eigenvalues greater than 1,

V_{ij} – j^{th} eigenvector calculated for i^{th} population,

V_{sj} – j^{th} eigenvector calculated for the sum of all data sets,

λ_{ik} – k^{th} eigenvalue calculated for i^{th} population.

After first calculation, test statistics presented above was supplemented to four statistics based on first eigenvectors only and calculated for Manhattan and Euclidean distances:

$$EV1\_e = \sum_{i=1}^{r} \sqrt{\sum_{j=1}^{p} (v_{ij} - v_{sj})^2} \tag{3}$$

$$EV1\_a = \sum_{i=1}^{r} \sum_{i=1}^{p} \left| v_{ij} - v_{sj} \right| \tag{4}$$

$$EV2\_e = \sum_{i=1}^{r} \sqrt{\sum_{j=1}^{p} (v_{ij} - \bar{v}_j)^2} \tag{5}$$

$$EV2\_a = \sum_{i=1}^{r} \sum_{i=1}^{p} \left| v_{ij} - \bar{v}_j \right| \tag{6}$$

Proposed test statistics have not calculated critical values, so permutations test will be carried out. It gives another opportunity – to avoid any assumptions typical for parametric tests.

IV. PERMUTATION TESTS

The idea of permutation test was worked out by R. A. Fisher. This test doesn't need any knowledge of the distribution of test statistics because instead of using any theoretical distribution, *ASL* (Achieved Significance Level) is estimated by Monte Carlo sampling from permutation distribution. And the power of permutation test is similar to parametric test, see Good P. I. (1994). The permutation tests sequence used in investigations is, as below:

1. Calculate the value of chosen statistics for tested sample – T^*.
2. Proceed a permutation (*N* times, in most cases it is recommended to be N>1000)[3] of data sets that destroys existing dependencies and parameters in data sets.
3. Calculate test statistics value for these permutations and create empirical distribution – T_i, where i=1, 2,…,N.
4. Locate calculated value of T^* on this distribution and estimate *p-value* as *ASL*:

[3] *The practice of business statistics, Companion chapter 18 – Bootstrap methods and permutation tests*, Hesterberg T., Monaghan S., Moore D. S., Clipson A., Epstein R.: W. H. Freeman and Company, New York 2003, s. 45.

$$ASL \approx \frac{card\{T_i : T^* \leq T_i\}}{N} \tag{7}$$

5. If received *ASL* value is more than assumed value of *α* level (for one-sided rejected region), null hypothesis cannot be rejected.

V. INVESTIGATIONS

The experiment was carried out for simulated and real data sets. Simulated data sets with 200 observations were prepared for three test cases:

1. Multivariate normal distribution with 5 dimensions and 5 groups; unit covariance/variance matrix, null mean vector, simulated change of mean vector for first group: *μ=[x,x,x,x,x]; x*=0, 0.1, 0.2, …0.6.
2. Multivariate normal distribution with 5 dimensions and 5 groups; unit covariance/variance matrix, null mean vector, simulated change of variance for the first group, the variance increasing, kaving been multiplied by: *x*=1.1, 1.2, …1.6.
3. Multivariate normal distribution with 5 dimensions and 5 groups; null mean vector, covariance/variance matrix as in table 1. Simulated change of parameter *x*=0.1, 0.2,…0.8.

Table 1. Variance/covariance matrix for third case of simulated data set

Matrix for first group					Matrix for remaining groups				
1	$-x$	0	0	0	1	x	0	0	0
$-x$	1	$-x$	0	0	x	1	x	0	0
0	$-x$	1	$-x$	0	0	x	1	x	0
0	0	$-x$	1	$-x$	0	0	x	1	x
0	0	0	$-x$	1	0	0	0	x	1

Source: own work.

Real data sets were taken from *dataset* package of R-CRAN software. Data sets: *iris, wines* and *vehicle* were used without any modification and after its centering (null mean vector).

The investigations were carried out for

- five statistics described with equations: 2-6 using permutation tests
- MANOVA test with statistics *Wilks lambda*:

$$\Lambda = \frac{|S_e|}{|S_h + S_e|} \tag{8}$$

where:

S_h – the hypothesis sum of squares and cross products matrix,
S_e –the error sums of squares and cross products matrix.

- two statistics using permutation tests: first according to Wilks lambda equation (8) and second - a sum of F statistics univariate ANOVA – for each variable, where:

$$F = \sum_{i=1}^{p} \frac{SSTR_i(n-r)}{SSE_i(r-1)} \tag{9}$$

where:

$SSTR_i$ – sum of squares for treatments for i^{th} variable,
SSE_i – sum of square errors for i^{th} variable.

VI. RESULTS

The results presented in table 2 were calculated with Monte Carlo method: from simulated data sets, the subsets with different size (20, 30, 50 and 100 observations) were randomly sampled (1000 times). A number of cases with the rejection of the null hypotheses (α=0.05), as a percentage of all sampled subsets indicates the performance of the tests. Table 3 presents *ASL* (permutations tests) or *p-value* (MANOVA) calculated for real data sets with α=0.10.

Table 2. A percentage of rejection of the null hypothesis (α=0.05), for simulated data sets (in %)

Data set 1																								
Parameter	$x = 0.1$				$x = 0.2$				$x = 0.3$				$x = 0.4$				$x = 0.5$				$x = 0.6$			
Subset sizes	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100
Test statistics																								
EV	4	8	4	1	4	3	4	4	3	3	4	3	4	6	5	5	5	4	5	8	9	7	6	8
EV1_e	6	3	4	1	4	3	4	4	3	3	5	3	4	6	6	5	5	5	5	10	7	7	6	10
EV1_a	7	3	3	3	4	4	4	7	4	4	3	6	4	6	6	5	6	4	5	8	7	5	5	7
EV2_e	3	6	1	4	3	6	2	5	3	4	1	3	3	1	4	2	4	5	2	2	3	5	1	1
EV2_a	4	8	4	6	1	3	5	4	5	5	5	5	3	1	5	3	5	2	3	1	3	4	1	3
Λ	4	6	4	2	12	10	11	16	11	19	27	66	20	38	75	99	45	76	91	100	61	94	100	100
Sum_F	4	8	4	1	12	10	9	15	8	21	27	70	25	37	78	100	53	81	95	100	75	97	100	100
Data set 2																								
Parameter	$x = 0.1$				$x = 0.2$				$x = 0.3$				$x = 0.4$				$x = 0.5$				$x = 0.6$			
Subset sizes	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100
Test statistics																								
EV	4	8	4	2	4	3	4	4	3	3	5	5	4	6	5	5	7	6	5	5	9	7	6	8
EV1_e	7	6	6	6	7	7	5	5	5	4	6	4	4	9	7	5	8	7	8	3	7	6	7	7
EV1_a	7	6	5	5	6	7	4	5	5	4	6	5	5	6	6	5	7	7	5	2	6	6	6	6
EV2_e	4	5	8	2	5	4	8	4	5	4	10	3	8	4	10	2	6	6	10	9	9	5	8	9
EV2_a	4	8	4	4	5	8	6	6	7	11	6	6	7	9	6	6	8	9	7	7	8	4	9	8
Λ	2	6	4	3	6	4	1	3	4	5	4	1	4	5	4	1	4	5	5	4	6	6	5	6
Sum_F	1	5	4	1	2	6	5	1	3	6	5	1	3	6	5	1	4	6	4	4	6	5	3	5
Data set 3																								
Parameter	$x = 0.1$				$x = 0.2$				$x = 0.5$				$x = 0.6$				$x = 0.7$				$x = 0.8$			
Subset sizes	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100	20	30	50	100
Test statistics																								
EV	0	5	6	4	6	3	5	6	3	5	1	3	0	3	2	12	6	11	18	28	8	15	27	55
EV1_e	0	5	6	4	6	3	5	6	3	5	1	3	0	3	2	13	6	11	19	28	8	15	28	55
EV1_a	1	3	4	3	8	5	6	6	3	5	2	3	0	3	2	13	5	11	20	28	8	14	26	51
EV2_e	8	7	11	5	3	5	2	3	1	3	5	1	2	2	2	14	7	10	18	28	7	15	24	52
EV2_a	11	4	14	8	6	7	6	6	1	4	2	4	6	3	2	14	4	9	17	28	6	14	22	24
Λ	2	3	0	0	5	2	0	0	2	0	2	0	6	5	1	0	7	6	2	0	6	5	3	0
Sum_F	1	3	0	0	5	1	0	0	1	0	2	0	4	5	1	0	2	2	1	0	3	5	2	0

Source: own work.

Table 3. *ASL/p-value* calculated for real data sets
(bold for cases where null hypotheses is not rejected)

Data sets	Test statistics						
	EV	EV1_e	EV1_a	EV2_e	EV1_a	Λ	Sum_F
iris	0.006	0.006	0.006	0.000	0.000	0.000	0.000
centered *iris*	0.060	0.082	0.083	0.001	**0.502**	**1.000**	**1.000**
vehicle	0.000	0.002	0.002	0.001	0.001	0.000	0.000
centered *vehicle*	0.006	0.006	0.006	0.007	0.007	**1.000**	**1.000**
wines	0.000	0.000	0.000	0.026	0.026	0.000	0.000
centered *wines*	0.002	0.000	0.000	**0.533**	**0.532**	**1.000**	**1.000**

Source: own work.

VII. CONCLUSIONS

Test statistics based on the eigenvectors are able to recognize the difference between multivariate populations if diversity concerns not only a mean vector or a variance of examined populations. According to results of simulations, better performance was indicated for the statistics that use first eigenvectors only and measure their difference with the first eigenvector of a sum of all observations. Permutation tests allow estimating *p-value* (as an *ASL* value) with no additional calculations concerning critical values and don't need any knowledge of the distribution of analyzed populations.

REFERENCES

Domański Cz., Pruska K., *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa 2000.

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

Good P. I. (1994) *Permutation Tests: A practical guide for testing Hypotheses*, Springer-Verlag, N. York.

Ito P. K. *Robustness of ANOVA and MANOVA test procedures* zamieszczony [w:] Krishnaiah P. R. (ed.) *Handbook of statistics 1. Analysis of variance* (p. 199–236). Amsterdam: North Holland, 1980.

The practice of busines statistics, Companion chapter 18 0 Bootstrap methods and permutation tests, Hesterberg T., Monaghan S., Moore D. S., Clipson A., Epstein R., W. H. Freeman and Company, New York 2003.

Jacek Stelmach

O TESTOWANIU RÓŻNIC POMIĘDZY POPULACJAMI ZA POMOCĄ WEKTORÓW WŁASNYCH

Testowanie różnic pomiędzy populacjami wielowymiarowymi jest jednym z kluczowych problemów w badaniach statystycznych. Najbardziej znane – testy MANOVA, jako parametryczne wymagają spełnienia założenia o zgodności z rozkładem normalnym wielowymiarowym. Bardzo często założenia te są praktycznie nierealne lub ich weryfikacja, szczególnie dla małej ilości obserwacji jest trudna.

Artykuł ten przedstawia podejście, oparte o testy permutacyjne (co zwalnia z weryfikacji powyższych założeń), gdzie proponowane statystyki testowe oparte są o własności wektorów własnych. Badania zostały przeprowadzone dla symulowanych i rzeczywistych zestawów danych, gdzie testy permutacyjne zostały porównane z testami opartymi na analizie zmiennych i statystykach testowych w MANOVA.