On a Non-Parametric Two-Sample Test of Equality of Location Parameters of Multivariate Populations

F. Rublík

Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovakia
Email: umerrubl@savba.sk

Abstract. *The paper investigates the behaviour of the power of the test, used in a monograph on modern nonparametric methods for testing the equality of location parameters of two multi-dimensional distributions. It is shown by means of simulations that the test has bad sensitivity to the violations of the null hypothesis.*

Keywords: Non-parametric Test, Location Parameter, Rank Statistic

1. Introduction

The two sample setting, used in this paper, is as follows. $\mathbf{X}$ is a random sample of size m from the distribution of the k-dimensional random vector $\boldsymbol{\xi}+\boldsymbol{\mu_Y}$, $\mathbf{Y}$ is a random sample of size n from the distribution of the k-dimensional random vector $\boldsymbol{\xi}+\boldsymbol{\mu_Y}$, where $\boldsymbol{\xi}$ are k-dimensional random fluctuations and $\boldsymbol{\mu_X}$, $\boldsymbol{\mu_Y}$ are location parameters. Thus

$$\mathbf{X} = \left(\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_m}\right) = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1m} \\ X_{21} & X_{22} & \ldots & X_{2m} \\ \vdots & & & \\ X_{k1} & X_{k2} & \ldots & X_{km} \end{pmatrix},$$

$$\mathbf{Y} = \left(\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}\right) = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1n} \\ Y_{21} & Y_{22} & \ldots & Y_{2n} \\ \vdots & & & \\ Y_{k1} & Y_{k2} & \ldots & Y_{kn} \end{pmatrix},$$

i.e., the observations are k-dimensional column vectors and variables are rows.

2. Subject and Methods

The subject of the paper is testing of the null hypothesis

$$H_0: \ \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y \tag{1}$$

of the equality of the location parameters of the two multivariate populations. First we explain the formula for the test statistic, proposed in [2].

Let $j \in \{1, \ldots, k\}$ be a fixed index. Combine the recorded values of the j-th variable from $\mathbf{X}$ and $\mathbf{Y}$ and compute their ranks W_{jt}. Thus

$$W_{j1}, W_{j2}, \ldots, W_{jN}, \quad N = m + n$$

denotes the vector of midranks of the j-th row of the matrix $\Big(\mathbf{X}, \mathbf{Y}\Big)$, i.e., the midranks of the j-th variable, and obviously $1 \leq W_{ji} \leq N$. Then

$$W_j = \sum_{t=1}^{m} W_{jt}, \quad j = 1, \ldots, k$$

denotes the sum of ranks of the j-th variable obtained from the first random sample $\mathbf{X}$. Put

$$W_{sum} = \sum_{j=1}^{k} W_j, \quad E = \frac{km(N+1)}{2}, \qquad S_i = \sum_{j=1}^{k} W_{ji}, \;\; i = 1, \ldots, N.$$

Thus S_i denotes the sum of the i-th column of the matrix (W_{ji}) with columns $i = 1, \ldots, N$ and rows $j = 1, \ldots, k$, i.e., the sum of ranks of the i-th observation. Further, let

$$\overline{S} = \frac{1}{N} \sum_{i=1}^{N} S_i, \qquad \sigma_S^2 = \frac{1}{N} \sum_{i=1}^{N} (S_i - \overline{S})^2, \quad var = \frac{mn}{N-1} \sigma_S^2. \tag{2}$$

The null hypothesis (1) is on pp. 203-205 of [2] tested by means of the test statistic

$$Z = \frac{W_{sum} - E}{\sqrt{var}} \tag{3}$$

and the test consists in referring the statistic (3) to the standard normal distribution. There are two possibilities of doing this, by the one-sided test (4) or by the two-sided test (5), i.e.,

$$reject\ H_0\ if\ Z > u_{1-\alpha}, \tag{4}$$
$$reject\ H_0\ if\ |Z| > u_{1-\alpha/2}, \tag{5}$$

where u_β denotes the β-th quantile of the standard normal $N(0,1)$ distribution.

The hypothesis (1) can be tested also by means of the Lawley-Hotelling statistic. This statistic is defined in the 2-sample setting by the formulas

$$T = m(\overline{X} - \overline{U})' S^{-1} (\overline{X} - \overline{U}) + n(\overline{Y} - \overline{U})' S^{-1} (\overline{Y} - \overline{U}), \quad \overline{X} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{X_i}, \quad \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y_i}, \tag{6}$$

$$S = \frac{1}{m+n-2} \left(\sum_{i=1}^{m} (\mathbf{X_i} - \overline{X})(\mathbf{X_i} - \overline{X})' + \sum_{i=1}^{n} (\mathbf{Y_i} - \overline{Y})(\mathbf{Y_i} - \overline{Y})' \right), \tag{7}$$

and has the asymptotic chi-square distribution with k degrees of freedom, provided that the null hypothesis (1) holds and the distribution of $\boldsymbol{\xi}$ has a regular covariance matrix. The null hypothesis (1) is rejected if $T > \chi_k^2(1-\alpha)$, where $\chi_k^2(1-\alpha)$ is the $(1-\alpha)$-th quantile of the chi-square distribution with k degrees of freedom.

In contemporary statistics there is often considered the use of tests not requiring the existence of the covariance matrix, because such testing rules usually do not fail to yield reliable results also in the case of sampling from heavy tailed distributions.

One of such test statistics was presented in [3], but since its computation in the case $m = n = 30$ requires much of the computer time, it is not included into the presented simulations. A simpler test statistic, based on the spatial median, was presented in [6] and studied also in [7].

The spatial median $\hat{\boldsymbol{\mu}} \in R^k$ of the k-dimensional vectors $\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_n}$ is defined by the equality

$$\sum_{i=1}^{n} \|\mathbf{Z_i} - \hat{\boldsymbol{\mu}}\| = \min\{\sum_{i=1}^{n} \|\mathbf{Z_i} - \mathbf{M}\|;\, \mathbf{M} \in R^k\}. \tag{8}$$

If these k-dimensional vectors do not lie on any line in R^k and are mutually distinct (which holds for sampling from continuous distribution), then according to [4] their spatial median is uniquely determined and can be computed by means of the results from [9].

Let $\hat{\boldsymbol{\mu}}_{\mathbf{X}}$ be the spatial median of the vectors $\mathbf{X_1}, \ldots, \mathbf{X_m}$, $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$ be the spatial median of $\mathbf{Y_1}, \ldots, \mathbf{Y_n}$ and $N = m + n$. Put

$$\overline{\boldsymbol{\mu}} = (m\hat{\boldsymbol{\mu}}_{\mathbf{X}} + n\hat{\boldsymbol{\mu}}_{\boldsymbol{Y}})/N.$$

Further, let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ denote the spatial median of the pooled random sample $(\mathbf{Z_1}, \ldots, \mathbf{Z_N}) = (\mathbf{X_1}, \ldots, \mathbf{X_m}, \mathbf{Y_1}, \ldots, \mathbf{Y_n})$ and

$$\hat{\mathbf{D}}_{\mathbf{1}} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\|\mathbf{Z_i} - \hat{\boldsymbol{\mu}}\|}\Big[\mathbf{I_d} - \frac{(\mathbf{Z_i} - \hat{\boldsymbol{\mu}})(\mathbf{Z_i} - \hat{\boldsymbol{\mu}})'}{\|\mathbf{Z_i} - \hat{\boldsymbol{\mu}}\|^2}\Big], \quad \hat{\mathbf{D}}_{\mathbf{2}} = \frac{1}{N}\sum_{i=1}^{N} \frac{(\mathbf{Z_i} - \hat{\boldsymbol{\mu}})(\mathbf{Z_i} - \hat{\boldsymbol{\mu}})'}{\|\mathbf{Z_i} - \hat{\boldsymbol{\mu}}\|^2},$$

where $\mathbf{I_d} = \mathrm{diag}\ (1, \ldots .1)$ is the $d \times d$ unit matrix. Let $\hat{\mathbf{V}} = \hat{\mathbf{D}}_{\mathbf{1}}^{-1}\hat{\mathbf{D}}_{\mathbf{2}}\hat{\mathbf{D}}_{\mathbf{1}}^{-1}$. The test statistic M_1 from [6] is

$$M_1 = m(\hat{\boldsymbol{\mu}}_{\mathbf{X}} - \overline{\boldsymbol{\mu}})'\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{X}} - \overline{\boldsymbol{\mu}}) + n(\hat{\boldsymbol{\mu}}_{\mathbf{Y}} - \overline{\boldsymbol{\mu}})'\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{Y}} - \overline{\boldsymbol{\mu}}).$$

If the distribution of $\boldsymbol{\xi}$ has a density f with respect to the Lebesgue measure on R^k and f is bounded on every bounded subset of R^k, then according to the results of [6] the rule rejecting H_0 if $M_1 > \chi^2_k(1-\alpha)$ is the test of (1) at the asymptotic significance level α.

Another test statistic for testing (1) was presented in [8]. As explained on p. 334 of [6], this statistic is in the two-sample setting given by the formula

$$W_1 = \frac{mnk}{N}\left\|\frac{1}{m}\sum_{i=1}^{m} U(\mathbf{X_i} - \hat{\boldsymbol{\theta}}) - \frac{1}{n}\sum_{j=1}^{n} U(\mathbf{Y_j} - \hat{\boldsymbol{\theta}}\ )\right\|^2, \tag{9}$$

where $\hat{\boldsymbol{\theta}}$ is the spatial median of the data $(\mathbf{X}, \mathbf{Y})$ and $U(\mathbf{Z}) = \mathbf{Z}/\|\mathbf{Z}$ if $\|\mathbf{Z}\| > 0$, and $U(\mathbf{Z}) = \mathbf{0_{k\times 1}}$ otherwise. If the random vector $\boldsymbol{\xi}$ has a density with respect to the Lebesgue measure, then according to [8] the rule rejecting H_0 if $W_1 > \chi^2_k(1-\alpha)$ is the test of (1) at the asymptotic significance level α.

3. Simulation Results

In the following table $\hat{P}_Q$ denotes the simulation results of the probability of rejection of (1) by the test based on the statistic Q, when the sample sizes are $m = 30$ and $n = 30$, respectively. The following estimates were obtained from trials consisting of 5000 simulations, the rejection was carried out at the asymptotic significance level $\alpha = 0.05$. The dimension of the random vector was $k = 3$, the random vector $\boldsymbol{\xi}$ was assumed to have independent components, all of them were either $N(0,1)$ distributed (Normal case) or all of them had the Cauchy $C(0,1)$ distribution (Cauchy case). The location parameters are $\boldsymbol{\mu}_X = (0,0,0)'$, and $\boldsymbol{\mu}_Y = c(0.1, -0.1, 0.1)'$. Thus when $c = 0$, then the null hypothesis (1) holds. The sampling from the distributions with the values $c = 0, 1, \ldots, 7$ was used to demonstrate the behaviour

of the tests when the distance from the null hypothesis is increasing. The notation $\hat{P}_{Z,1S}$ denotes the simulation estimate of the probability of (4), where $\alpha = 0.05$, and $\hat{P}_{Z,2S}$ denotes the simulation estimate of the probability of (5), where $\alpha = 0.05$, both for $m = n = 30$.

	Normal case							
c	0	1	2	3	4	5	6	7
$\hat{P}_{Z,1S}$	0.047	0.028	0.017	0.011	0.007	0.004	0.002	0.000
$\hat{P}_{Z,2S}$	0.047	0.056	0.073	0.097	0.135	0.186	0.249	0.319
$\hat{P}_T$	0.068	0.102	0.197	0.386	0.609	0.811	0.931	0.982
$\hat{P}_{M_1}$	0.043	0.067	0.128	0.263	0.454	0.664	0.839	0.938
$\hat{P}_{W_1}$	0.049	0.074	0.157	0.303	0.523	0.730	0.878	0.958

	Cauchy case							
c	0	1	2	3	4	5	6	7
$\hat{P}_T$	0.024	0.024	0.027	0.032	0.040	0.051	0.061	0.077
$\hat{P}_{M_1}$	0.062	0.064	0.085	0.113	0.155	0.205	0.266	0.341
$\hat{P}_{W_1}$	0.053	0.060	0.077	0.111	0.167	0.244	0.338	0.445

4. Discussion

As far as the tests based on T, M_1 or W_1 are concerned, the overall picture is analogous as in the 3-sample case, investigated in the simulation study of [6]. For distributions having covariance matrix the use of the Lawley-Hotelling test can be expected to yield good result, even though a better agreement of the size of this test with the nominal value will require sample sizes larger than $\min(m, n) = 30$. But if the possibility of observations coming from heavy tailed distributions has to be considered, then the Lawley-Hotelling test cannot be used, because as the results of the simulations show (the Cauchy case), it is insensitive to violations of the null hypothesis in such a case. For heavy tailed distributions from the tests considered in the previous simulations appears to yield mildly moderately best results the test based on W_1. A disadvantage of the tests based on M_1 or on W_1 is that they are not affine invariant. A test with this property is presented in the forthcomig paper [5]. The simulations show, that in general, both tests (4) and (5) are either insensitive or only weekly sensitive to violations of the null hypothesis (1).

5. Conclusions

Because of their bad performance, the tests (4) and (5), proposed in the monograph [2], should not be used for testing the null hypothesis (1) against the general alternative, that (1) does not hold. It is recommendable to carry out this testing by the tests based on W_1, M_1 or by the Lawley-Hotelling test.

Acknowledgements

The research was supported by the grant 1/0077/09 from VEGA and by the grant 2/0019/10 from VEGA.

References

[1] Chaudhuri, P. Multivariate location estimation using extension of R-estimates through U-statistics type approach. *Annals of Statistics*, 20 (1992), 897-519.

[2] Higgins, J, J.: An Introduction to Modern Statistics. Thomson Books/Cole, Pacific Grove, 2004.

[3] Hettmansperger, T. P., Möttönen J. and Oja, H. Affine invariant multivariate rank tests for several samples. *Statistica Sinica,* 8(1998), 785-800.

[4] Milasevic, P. and Ducharme, G. R. Uniqueness of the spatial median. *Annals of Statistics*, 15(1987), 1332-1333.

[5] Rublík, F. and Somorčík, J. Affine equivariant spatial median and its use in the multivariate multi-sample location problem. *Submitted for publication.*

[6] Somorčík, J. Tests Using Spatial Median. *Austrian Journal of Statistics,* 35(2006), 331-338.

[7] Somorčík, J. Performance of some spatial median tests under elliptical symmetry. *Measurement Science Review (online journal),* 7(2007), Section 1, No. 4, 43-50.

[8] Um, Y. and Randles, R. H. Nonparametric tests for the multivariate multi-sample location problem. *Statistica Sinica,* 8(1998), 801-812.

[9] Vardi, Y. and Zhang, C. H. The multivariate L_1-median and associated data depth. Proceedings of the National Academy of Sciences of the United States of America, 97(2000), 1423-1426.