A randomized nonparametric statistic for multivariate multisample testing hypothesis

Hidetoshi Murakami
Department of Mathematics, National Defense Academy, JAPAN
e-mail:murakami@nda.ac.jp

Abstract

We consider a multivariate multisample testing hypothesis, which is one of the most common types of statistical problems. Recent progress in computerized measurement technology has permitted the accumulation of multivariate data. The analysis of multivariate data is particularly important in studies of biological data, image data, and functional data. For the multivariate data, it is important to determine how to represent the rank based on the observation distances. One of the multivariate multisample testing problems based on the Jurečková - Kalina ranks of distance is discussed in this paper. A randomized multivariate multisample nonparametric statistic is proposed for the equality of two continuous distribution functions for the two-sided alternatives. Simulations are used to investigate the power of the suggested statistic for the two-sided alternative with various population distributions.

Keywords: Jurečková-Kalina's ranks, Multivariate multisample rank test, Power comparison

1. Introduction

We consider a multivariate multisample testing hypothesis, which is one of the most important types of statistical problems. In many applications, the underlying distribution is not known sufficiently to assume normality. Therefore, we require considering a nonparametric testing hypothesis. Nonparametric statistics have been used with great success by many authors. One of the best-known multisample nonparametric tests is the Kruskal-Wallis statistic.

Recent progress in computerized measurement technology has permitted the accumulation of multivariate data. In health studies, for example, each observation on a patient is actually a whole array of measurements that describe the health status of the person at a particular point in time. Thus, we are naturally led to consider vector-valued observations in dealing with data from these settings. If each component of the vectors is only studied marginally, then certain outliers, strongly influential points, and useful relationships among variables may not be detected. A multivariate examination of the data is very appropriate.

In this paper, the multisample nonparametric statistics are extended to the multivariate version of statistics. Let $\boldsymbol{X}_k = (\boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kn_k})$ be k independent samples from p-variate populations with continuous distribution functions F_{kp} with mean vectors $\boldsymbol{\mu}_k$. One of the problems for our hypothesis is given as follows:

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k \quad \text{against} \quad H_1 : \text{not } H_0.$$

To test this hypothesis, we developed the multivariate multisample nonparametric statistics. In Section 2, we introduce the multivariate multisample nonparametric statistics based on Jurečková-Kalina's ranks of distances (Jurečková and Kalina, 2012). In Section 3, we investigate the power of the proposed statistic. To compare the power of the multivariate multisample nonparametric statistics, we carry out simulation studies of various distributions. The simulations involve 100,000 repetitions. Finally, we conclude this paper in Section 4.

2. Multivariate multisample nonparametric statistics

In this section, we consider the multivariate multisample nonparametric statistics. For the multivariate data, it is important to determine how to represent the rank based on the observation distances. Recently, Jurečková and Kalina (2012) proposed a rank test based on the distances of observations. We extend their ranks of distance to the multisample version.

Let $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ be the pooled sample, where $\boldsymbol{Z}_{s_1} = \boldsymbol{x}_{s_1}^{(1)}$ for $s_1 = 1, \ldots, n_1$ and $\boldsymbol{Z}_{n_1+s_2} = \boldsymbol{x}_{s_2}^{(2)}$ for $s_2 = 1, \ldots, n_2$, $\cdots$, $\boldsymbol{Z}_{n_1+n_2+\cdots+n_{k-1}+s_k} = \boldsymbol{x}_{s_k}^{(k)}$ for $s_k = 1, \ldots, n_k$ and $N = n_1 + \cdots + n_k$. Jurečková and Kalina (2012) considered the distances $\{\ell_{ij} = L(\boldsymbol{Z}_i, \boldsymbol{Z}_j), i = 1, \ldots, N, j = 1, \ldots, N, j \neq i\}$ for every fixed i, where $L(\cdot, \cdot)$ denotes the Euclidean distance. Then, conditionally given $\boldsymbol{Z}_{s_1}$, the vector $\{\ell_{s_1 h_1}, h_1 = 1, \ldots, n_1, h_1 \neq s_1\}$ is a random sample from the distribution function $F_{1p}(z|\boldsymbol{Z}_{s_1}) = F_{1p}$. Similarly, the vector $\{\ell_{n_1+\cdots+n_{k-1}+s_k, h_k}, h_k = 1, \ldots, n_k, h_k \neq n_1 + \cdots + n_{k-1} + s_k\}$ is a random sample from the distribution function $F_{kp}(z|\boldsymbol{Z}_{n_1+\cdots+n_{k-1}+s_k}) = F_{kp}$. Assuming that the distribution functions F_{kp} are continuous, the rank of ℓ_{ij} is given by

$$R^i = (R_1^i, \ldots, R_{i-1}^i, R_{i+1}^i, \ldots, R_N^i),$$

where $j = 1, \ldots, N$ and $j \neq i$. Then we consider the rank statistics as follows:

- Kruskal-Wallis statistic (Gibbons and Chakraborti, 2010):

$$T_{i1}^{(p)} = \frac{12}{N(N+1)} \sum_{\beta=1}^{k} n_\beta \left(W_\beta - \frac{N+1}{2} \right)^2,$$

  where

$$W_\beta = \frac{1}{n_\beta} \sum_{\gamma=1}^{n_\beta} R_{Y+\gamma}^i \quad \text{and} \quad Y = \sum_{\alpha=1}^{\beta-1} n_\alpha.$$

- Multisample median statistic (Hájek *et al.*, 1999):

$$T_{i2}^{(p)} = 4 \sum_{\beta=1}^{k} \frac{1}{n_\beta} \left(A_\beta - \frac{n_\beta}{2} \right)^2,$$

  where

$$A_\beta = \sum_{\gamma=1}^{n_\beta} \frac{1}{2} \left\{ \text{sign} \left(\sum_{\gamma=1}^{n_\beta} R_{Y+\gamma}^i - \frac{N+1}{2} \right) + 1 \right\}.$$

- Multisample Mood statistic (Rublík, 2007):

$$T_{i3}^{(p)} = \frac{180}{N(N+1)(N^2-4)} \sum_{\beta=1}^{k} n_\beta \left(M_\beta - \frac{N^2-1}{12} \right)^2,$$

  where

$$M_\beta = \frac{1}{n_\beta} \sum_{\gamma=1}^{n_\beta} \left(R_{Y+\gamma}^i - \frac{N+1}{2} \right)^2.$$

- Multisample Lepage-type statistic (Rublík, 2007):

$$T_{i4}^{(p)} = T_{i1}^{(p)} + T_{i3}^{(p)}.$$

Then, the statistic $T_{iq}^{(p)}$, $q = 1, \ldots, 4$, is equally distributed for $i = 1, \ldots, N$ under the null hypothesis. A randomization of $T_{1q}^{(p)}, \ldots, T_{Nq}^{(p)}$ maintains the simple structure of the test. Thus, we obtain

$$\mathbb{P}(T_q^{(p)} = T_{iq}^{(p)}) = \frac{1}{N}, \quad i = 1, \ldots, N, \tag{1}$$

where the randomization in (1) is independent of the observations. For any C_q,

$$\mathbb{P}(T_q^{(p)} > C_q) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}(T_{iq}^{(p)} > C_q),$$

and the statistic rejets H_0 if $T_q^{(p)} > C_q$.

3. Simulation study

In this section, we investigate the behavior of the $T_q^{(p)}$ statistics in simulation studies with R 2.14.1. The simulations involve 100,000 repetitions, and the significance level is 5%. To compare the power of the multivariate nonparametric statistics, we carry out a simulation study of different populations with various distributions. In this paper, we focus on the case of $n_1 = n_2 = n_3 = 5$ and $n_1 = 15$, $n_2 = 10$, $n_3 = 5$ for $k = 3$ and $p = 3$. Additionally, we describe the following distributions:

1. $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: the multivariate normal distribution.
2. $t(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \delta_k)$: the multivariate t distribution with δ degrees of freedom
3. $SN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\nu}_k)$: the multivariate skew normal distribution with shape parameter $\boldsymbol{\nu}$ (Azzalini and Dalla-Valle, 1996).

We examined the power at which the location and correlation parameters differed. Herein, we define the p dimensional matrix as follows:

$$I^{(3)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \rho_1^{(3)} = \begin{pmatrix} 1 & 0.4 & 0.2 \\ 0.4 & 1 & 0.3 \\ 0.4 & 0.3 & 1 \end{pmatrix}, \qquad \rho_1^{(3)} = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.3 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{pmatrix}$$

Case 1		
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\boldsymbol{\mu}_2 = \mathbf{0}$	$\boldsymbol{\mu}_3 = \mathbf{0}$
$\boldsymbol{\Sigma}_1 = I^{(3)}$	$\boldsymbol{\Sigma}_2 = I^{(3)}$	$\boldsymbol{\Sigma}_3 = I^{(3)}$

Case 2		
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\boldsymbol{\mu}_2 = \mathbf{0}$	$\boldsymbol{\mu}_3 = \mathbf{0}$
$\boldsymbol{\Sigma}_1 = I^{(3)}$	$\boldsymbol{\Sigma}_2 = \rho_1^{(3)}$	$\boldsymbol{\Sigma}_3 = \rho_2^{(3)}$

Case 3		
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\boldsymbol{\mu}_2 = \mathbf{1.0}$	$\boldsymbol{\mu}_3 = \mathbf{2.0}$
$\boldsymbol{\Sigma}_1 = I^{(3)}$	$\boldsymbol{\Sigma}_2 = I^{(3)}$	$\boldsymbol{\Sigma}_3 = I^{(3)}$

Case 4		
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\boldsymbol{\mu}_2 = \mathbf{1.0}$	$\boldsymbol{\mu}_3 = \mathbf{2.0}$
$\boldsymbol{\Sigma}_1 = I^{(3)}$	$\boldsymbol{\Sigma}_2 = \rho_1^{(3)}$	$\boldsymbol{\Sigma}_3 = \rho_2^{(3)}$

Case 6		
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\boldsymbol{\mu}_2 = \mathbf{2.0}$	$\boldsymbol{\mu}_3 = \mathbf{4.0}$
$\boldsymbol{\Sigma}_1 = I^{(3)}$	$\boldsymbol{\Sigma}_2 = I^{(3)}$	$\boldsymbol{\Sigma}_3 = I^{(3)}$

Case 6		
$\boldsymbol{\mu}_1 = \mathbf{0}$	$\boldsymbol{\mu}_2 = \mathbf{2.0}$	$\boldsymbol{\mu}_3 = \mathbf{4.0}$
$\boldsymbol{\Sigma}_1 = I^{(3)}$	$\boldsymbol{\Sigma}_2 = \rho_1^{(3)}$	$\boldsymbol{\Sigma}_3 = \rho_2^{(3)}$

Table 1 lists the results of the simulation of the multivariate normal distribution.

Table 1: The Multivariate Normal Distribution
Case of $n_1 = n_2 = n_3 = 5$

	Case 1	Case 2	case 3	case 4	case 5	case 6
$T_1^{(3)}$	**0.049**	**0.049**	0.536	0.505	0.949	0.936
$T_2^{(3)}$	**0.039**	**0.039**	0.399	0.375	0.827	0.815
$T_3^{(3)}$	**0.050**	**0.051**	0.048	0.041	0.383	0.297
$T_4^{(3)}$	**0.050**	**0.051**	0.428	0.392	0.907	0.883

Case of $n_1 = 15$, $n_2 = 10$ and $n_3 = 5$

	Case 1	Case 2	case 3	case 4	case 5	case 6
$T_1^{(3)}$	**0.050**	**0.052**	0.732	0.705	0.996	0.994
$T_2^{(3)}$	**0.040**	**0.042**	0.592	0.554	0.959	0.946
$T_3^{(3)}$	**0.050**	**0.057**	0.196	0.156	0.678	0.629
$T_4^{(3)}$	**0.050**	**0.057**	0.673	0.660	0.992	0.991

For a case of non-normal distribution, we treat the multivariate t distribution with 2 degrees of freedom in Table 2.

Table 2: The Multivariate t Distribution
Case of $n_1 = n_2 = n_3 = 5$

	Case 1	Case 2	case 3	case 4	case 5	case 6
$T_1^{(3)}$	**0.048**	**0.049**	0.257	0.258	0.637	0.627
$T_2^{(3)}$	**0.038**	**0.039**	0.225	0.223	0.617	0.589
$T_3^{(3)}$	**0.050**	**0.051**	0.072	0.066	0.165	0.140
$T_4^{(3)}$	**0.050**	**0.051**	0.226	0.218	0.615	0.588

Case of $n_1 = 15$, $n_2 = 10$ and $n_3 = 5$

	Case 1	Case 2	case 3	case 4	case 5	case 6
$T_1^{(3)}$	**0.050**	**0.052**	0.421	0.413	0.859	0.853
$T_2^{(3)}$	**0.041**	**0.041**	0.373	0.353	0.790	0.774
$T_3^{(3)}$	**0.050**	**0.054**	0.123	0.115	0.385	0.334
$T_4^{(3)}$	**0.050**	**0.054**	0.396	0.386	0.859	0.849

We use the multivariate skew normal distribution to simulate an asymmetrical distribution in Table 3. We assume $\boldsymbol{\nu} = \mathbf{4}$ in this paper.

The results shown in Tables reveal that the $T_4^{(3)}$ statistic was more efficient than the $T_2^{(3)}$ statistic. Moreover, the $T_1^{(3)}$ statistic was the more powerful than the $T_4^{(3)}$ statistic for a shifted location parameter. Therefore, the $T_1^{(3)}$ statistic was more suitable than the $T_2^{(3)}$, $T_3^{(3)}$ and $T_4^{(3)}$ statistics for the parameters associated with the multivariate normal, t and skew normal distributions.

Table 3: The Multivariate Skew Normal Distribution
Case of $n_1 = n_2 = n_3 = 5$

	Case 1	Case 2	case 3	case 4	case 5	case 6
$T_1^{(3)}$	**0.049**	**0.055**	0.683	0.770	0.978	0.992
$T_2^{(3)}$	**0.039**	**0.043**	0.588	0.669	0.840	0.859
$T_3^{(3)}$	**0.050**	**0.051**	0.098	0.122	0.572	0.590
$T_4^{(3)}$	**0.050**	**0.055**	0.587	0.676	0.963	0.983

Case of $n_1 = 15$, $n_2 = 10$ and $n_3 = 5$

	Case 1	Case 2	case 3	case 4	case 5	case 6
$T_1^{(3)}$	**0.050**	**0.062**	0.838	0.903	0.999	1.000
$T_2^{(3)}$	**0.040**	**0.049**	0.736	0.809	0.988	0.994
$T_3^{(3)}$	**0.050**	**0.049**	0.346	0.354	0.779	0.791
$T_4^{(3)}$	**0.050**	**0.056**	0.796	0.866	0.998	1.000

4. Concluding remarks

In this paper, we considered the multivariate multisample nonparametric statistics by applying Jurečková and Kalina's ranks of distances. By simulation studies, the multivariate Kruskal-Wallis statistic was shown to be more powerful than various multivariate multisample nonparametric statistics for the shifted location parameter under the multivariate normal, t and skew normal distributions. Therefore, the multivariate Kruskal-Wallis statistic is more competent than the $T_2^{(p)}$, $T_3^{(p)}$ and $T_4^{(p)}$ statistics based on Jurečková and Kalina's ranks of distances.

References

[1] Azzalini, A. and Dalla-Valle, A. (1996) "The multivariate skew-normal distribution", *Biometrika*, 83, 715-726.

[2] Gibbons, J. D. and Chakraborti, S. (2010) *Nonparametric statistical Inference*, 5th Edition. CRC Press, New York.

[3] Hájek, J., Sidák, Z. and Sen, P. K. (1999) *Theory of rank tests*, 2nd Edition. Academic Press, San Diego.

[4] Jurečková, J. and Kalina, J. (2012) "Nonparametric multivariate rank tests and their unbiasedness", *Bernoulli*, 18, 229-251.

[5] Rublík, F. (2007) "On the asymptotic efficiency of the multisample location-scale rank tests and their adjustment for ties", *Kybernetika*, 43, 279-306.