

# A Triangle Test for Equality of Distribution Functions in High Dimension

Zhenyu Liu

Reza Modarres

Department of Statistics

The George Washington University

Washington, D.C. 20052

## Abstract

A new nonparametric test statistic is proposed for testing the equality of two multivariate distributions in high-dimensional settings by comparing their interpoint distances. Given two  $p$ -dimensional random samples  $\mathbf{X}$  and  $\mathbf{Y}$ , a triangle can be formed by randomly selecting one data point from the  $\mathbf{X}$  sample and two data points from the  $\mathbf{Y}$  sample or one data point from the  $\mathbf{Y}$  sample and two data points from the  $\mathbf{X}$  sample. Our test statistic estimates the probability that the distance of two data points from the same distribution is the smallest, the middle or the largest in the triangle. We show that the test statistic is asymptotically distribution-free under the null hypothesis of equal, but unknown distribution functions. The triangle test is compared to other nonparametric tests through a simulation study. The statistic is well defined when  $p > n$ , and its computational complexity is independent of  $p$ , making it suitable for high-dimensional settings.

## Keywords

U-statistic; Hypothesis Test; Multivariate Analysis; Distribution Function.

## 1 Introduction

Suppose we have two independent random samples  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  and  $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$  from two  $p$ -dimension distributions with cdf  $F$  and  $G$ ,  $p \geq 1$ , respectively. A classical nonparametric inference problem is to test  $H_0: F = G$  versus the general alternative  $H_a: F \neq G$ .

For the univariate problem ( $p = 1$ ), several distribution-free tests are available, such as the Kolmogorov-Smirnov test, Wald-Wolfowitz runs test, and Wilcoxon rank sum test. Such test statistics are based on the ranks of the observations in the pooled sample. The concept of rank and ordering does not easily generalize in a multivariate setting ( $p > 1$ ) and extension of univariate statistics often yield tests that are not distribution free. Applying the Fisher's permutation principle, Bickel (1969) showed it is possible to construct a consistent distribution free multivariate Simirnov test by conditioning on the empirical cdf of the pooled sample.

Interpoint distances provide an alternative approach. They can be used to obtain consistent distribution-free statistics for multivariate two-sample problem of

testing  $H_0: F = G$ . Friedman and Rafsky (1979) constructed multivariate generalizations of the Wald-Wolfowitz runs test and Kolmogorov-Smirnov test based on the minimal spanning tree of sample interpoint distances. Henze and Penrose (1999) showed that the Friedman-Rafsky test is consistent against all alternatives. Bickel and Brieman (1983), Henze (1988), and Schilling (1986) developed consistent and asymptotically distribution free tests based on nearest-neighbor distances. Székely and Rizzo (2004) proposed a consistent permutation test based on the e-distance or energy-distance (Aslan and Zeoh, 2005) between two samples. These methods reduced the multidimensional problem to a one-dimensional problem by comparing sample interpoint distances.

In this paper, we propose a triangle test for testing the equality of two multivariate distribution functions, which is also based on sample interpoint distances. The test statistics are constructed on appealing geometric considerations and are shown to be asymptotically distribution free.

## 2 The Triangle Test

### 2.1 Motivation

Let  $\delta(\cdot, \cdot)$  be any appropriately chosen distance function such that

$$\delta(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}.$$

We form a triangle with randomly selected data points  $\mathbf{x}_1 \sim F$  and  $\mathbf{y}_1, \mathbf{y}_2 \sim G$ . Let  $\theta_1, \theta_2$  and  $\theta_3$  denote the probabilities of  $\delta(\mathbf{y}_1, \mathbf{y}_2)$  being the smallest, the middle and the largest in the triangle, respectively. Thus, we have

$$\theta_1 = \theta_2 = \theta_3 = \frac{1}{3} \quad \Leftrightarrow \quad \delta(\mathbf{x}_1, \mathbf{y}_1) \stackrel{D}{=} \delta(\mathbf{x}_1, \mathbf{y}_2) \stackrel{D}{=} \delta(\mathbf{y}_1, \mathbf{y}_2). \quad (1)$$

Similarly, for a triangle with randomly selected data points  $\mathbf{y}_1 \sim G$  and  $\mathbf{x}_1, \mathbf{x}_2 \sim F$  as vertices, we denote the probabilities of  $\delta(\mathbf{x}_1, \mathbf{x}_2)$  being the smallest, the middle and the largest in the triangle as  $\lambda_1, \lambda_2$  and  $\lambda_3$ , respectively. We have

$$\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3} \quad \Leftrightarrow \quad \delta(\mathbf{y}_1, \mathbf{x}_1) \stackrel{D}{=} \delta(\mathbf{y}_1, \mathbf{x}_2) \stackrel{D}{=} \delta(\mathbf{x}_1, \mathbf{x}_2). \quad (2)$$

Maa, Pearl, and Bartoszyński (1996) proved that, under mild conditions,  $F = G$  if and only if the interpoint distance within samples and between samples have the same univariate distribution. That is, if  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are iid with cdf  $F$  and  $\mathbf{y}_1$  and  $\mathbf{y}_2$  are iid with cdf  $G$ ,

$$\delta(\mathbf{x}_1, \mathbf{y}_1) \stackrel{D}{=} \delta(\mathbf{x}_1, \mathbf{y}_2) \stackrel{D}{=} \delta(\mathbf{y}_1, \mathbf{y}_2) \quad \Leftrightarrow \quad F = G.$$

Thus, if (1) or (2) is satisfied, we have

$$\delta(\mathbf{x}_1, \mathbf{y}_1) \stackrel{D}{=} \delta(\mathbf{x}_1, \mathbf{y}_2) \stackrel{D}{=} \delta(\mathbf{y}_1, \mathbf{y}_2)$$

and therefore, we have  $F = G$ .

## 2.2 Triangle Statistics

For two independent random samples  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  and  $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ , we define

$$\begin{aligned} U_k &= \frac{1}{n \binom{m}{2}} \sum_{r=1}^n \sum_{i < j}^m I_{1k}[x_r; y_i, y_j] \\ V_k &= \frac{1}{m \binom{n}{2}} \sum_{r'=1}^m \sum_{i' < j'}^n I_{2k}[y_{r'}; x_{i'}, x_{j'}] \\ \text{for } k &= 1, 2, 3, \quad r, i', j' = 1, \dots, n, \quad \text{and } r', i, j = 1, \dots, m, \end{aligned}$$

where  $I_{1k}[x_r; y_i, y_j] = 1$  if  $\delta(\mathbf{y}_i, \mathbf{y}_j)$  is the smallest ( $k = 1$ ), the middle ( $k = 2$ ), and the largest ( $k = 3$ ) in the triangle with vertices  $x_r, y_i$  and  $y_j$ ; and  $I_{2k}[y_{r'}; x_{i'}, x_{j'}] = 1$  if  $\delta(\mathbf{x}_{i'}, \mathbf{x}_{j'})$  is the smallest ( $k = 1$ ), the middle ( $k = 2$ ), and the largest ( $k = 3$ ) in the triangle with vertices  $y_{r'}, x_{i'}$  and  $x_{j'}$ .

$U_k$  and  $V_k$  are *two-sample U-statistics* with kernels  $h_k(\mathbf{x}_r; \mathbf{y}_i, \mathbf{y}_j) = I_{1k}[x_r; y_i, y_j]$  and  $g_k(\mathbf{y}_{r'}; \mathbf{x}_{i'}, \mathbf{x}_{j'}) = I_{2k}[y_{r'}; x_{i'}, x_{j'}]$ . They are unbiased estimators for  $\theta_k$  and  $\lambda_k$ , respectively. Under  $H_0: F = G$ , for  $k = 1, 2, 3$ ,

$$\begin{aligned} E(U_k) &= \theta_k = \frac{1}{3} \\ E(V_k) &= \lambda_k = \frac{1}{3}. \end{aligned}$$

Thus, any large deviation of  $U_k$  or  $V_k$  from  $\frac{1}{3}$  provides evidence that  $F \neq G$ ; hence we can use  $U_k$  or  $V_k$  individually or we can use functions of these statistics to test  $H_0: F = G$ . Note that, in general, the distributions of  $U_k$  and  $V_k$  depend on the cdfs  $F$  and  $G$ . Therefore, test statistics based on  $U_k$  and  $V_k$  are not distribution free.

Let  $N = n + m$  be the total number observations, and assume that as  $n, m \rightarrow \infty$ ,

$$\frac{n}{N} \rightarrow \rho, \quad \frac{m}{N} \rightarrow 1 - \rho, \quad 0 < \rho < 1. \quad (3)$$

We can show that  $U_k$  and  $V_k$  are asymptotically normal using the theory of two-sample  $U$ -statistic.

Recall that  $\theta_k$  is defined for a triangle with three random data points,  $\mathbf{x}_1, \mathbf{y}_1$  and  $\mathbf{y}_2$ , as vertices and it is symmetric in  $\mathbf{y}_1$  and  $\mathbf{y}_2$ . We define the following conditional probabilities for  $\theta_k$ :

$p_{1k}(\mathbf{x}_1) = E_{G,G}(I_{1k}[\mathbf{x}_1; y_1, y_2]) =$  Given  $\mathbf{x}_1$ , the conditional probability that  $\delta(\mathbf{y}_1, \mathbf{y}_2)$  being the smallest ( $k = 1$ ), the middle ( $k = 2$ ) and the largest ( $k = 3$ ) in the triangle, and

$q_{1k}(\mathbf{y}_1) = E_{F,G}(I_{1k}[x_1; \mathbf{y}_1, y_2]) =$  Given  $\mathbf{y}_1$ , the conditional probability that  $\delta(\mathbf{y}_1, \mathbf{y}_2)$  being the smallest ( $k = 1$ ), the middle ( $k = 2$ ) and the largest ( $k = 3$ ) in the triangle.

Similarly, let

$p_{2k}(\mathbf{y}_1) = E_{F,F}(I_{2k}[\mathbf{y}_1; x_1, x_2]) =$  Given  $\mathbf{y}_1$ , the conditional probability that  $\delta(\mathbf{x}_1, \mathbf{x}_2)$  being the smallest ( $k = 1$ ), the middle ( $k = 2$ ) and the largest ( $k = 3$ ) in the triangle with vertices  $\mathbf{y}_1, \mathbf{x}_1$  and  $\mathbf{x}_2$ , and

$q_{2k}(\mathbf{x}_1) = E_{F,G}(I_{2k}[y_1; \mathbf{x}_1, x_2]) =$  Given  $\mathbf{x}_1$ , the conditional probability that  $\delta(\mathbf{x}_1, \mathbf{x}_2)$  being the smallest ( $k = 1$ ), the middle ( $k = 2$ ) and the largest ( $k = 3$ ) in the triangle with vertices  $\mathbf{y}_1, \mathbf{x}_1$  and  $\mathbf{x}_2$ .

We define the following parameters for these conditional probabilities

$$\begin{aligned}\alpha_{1st} &= Cov(p_{1s}(\mathbf{x}_1), p_{1t}(\mathbf{x}_1)); & \alpha_{2st} &= Cov(p_{2s}(\mathbf{y}_1), p_{2t}(\mathbf{y}_1)); \\ \beta_{1st} &= Cov(q_{1s}(\mathbf{y}_1), q_{1t}(\mathbf{y}_1)); & \beta_{2st} &= Cov(q_{2s}(\mathbf{x}_1), q_{2t}(\mathbf{x}_1)); \\ \gamma_{1st} &= Cov(p_{1s}(\mathbf{x}_1), q_{2t}(\mathbf{x}_1)); & \gamma_{2st} &= Cov(p_{2s}(\mathbf{y}_1), q_{1t}(\mathbf{y}_1)).\end{aligned}$$

For  $k = 1, 2, 3$ , the projections of  $U_k - \theta_k$  and  $V_k - \lambda_k$  onto the sets of all functions of the form  $\sum_{i=1}^n p_i(\mathbf{x}_i) + \sum_{j=1}^m q_j(\mathbf{y}_j)$  are given by

$$\begin{aligned}\hat{U}_k &= \frac{1}{n} \sum_{i=1}^n (p_{1k}(\mathbf{x}_i) - \theta_k) + \frac{2}{m} \sum_{j=1}^m (q_{1k}(\mathbf{y}_j) - \theta_k); \\ \hat{V}_k &= \frac{1}{m} \sum_{j=1}^m (p_{2k}(\mathbf{y}_j) - \lambda_k) + \frac{2}{n} \sum_{i=1}^n (q_{2k}(\mathbf{x}_i) - \lambda_k).\end{aligned}$$

**Lemma 1.** *Under condition (3), for  $k = 1, 2, 3$ ,  $\hat{U}_k$  and  $\hat{V}_k$  have (jointly) asymptotically a multivariate normal distribution with null mean vector and dispersion matrix with elements*

$$\begin{aligned}u_{st} &= cov(\hat{U}_s, \hat{U}_t) = \frac{1}{n} \alpha_{1st} + \frac{4}{m} \beta_{1st}; \\ v_{st} &= cov(\hat{V}_s, \hat{V}_t) = \frac{1}{n} \alpha_{2st} + \frac{4}{m} \beta_{2st}; \\ w_{st} &= cov(\hat{U}_s, \hat{V}_t) = \frac{2}{n} \gamma_{1st} + \frac{2}{m} \gamma_{2ts}.\end{aligned}$$

**Lemma 2.** *Under condition (3), for  $k = 1, 2, 3$ ,*

$$\begin{aligned}U_k - \theta_k &\xrightarrow{P} \hat{U}_k; \\ V_k - \lambda_k &\xrightarrow{P} \hat{V}_k.\end{aligned}$$

**Theorem 1.** *Under condition (3), for  $k = 1, 2, 3$ ,  $(U_k - \theta_k)$  and  $(V_k - \lambda_k)$  have (jointly) asymptotically a multivariate normal distribution with null mean vector and dispersion matrix with elements*

$$\begin{aligned}u_{st} &= \frac{1}{n} \alpha_{1st} + \frac{4}{m} \beta_{1st}; \\ v_{st} &= \frac{1}{n} \alpha_{2st} + \frac{4}{m} \beta_{2st}; \\ w_{st} &= \frac{2}{n} \gamma_{1st} + \frac{2}{m} \gamma_{2ts}.\end{aligned}$$

Since the dispersion matrices depend on the unknown distributions  $F$  and  $G$ , we will find there consistent estimators and construct quadratic forms to obtain asymptotically distribution-free test statistics.

Let

$$\begin{aligned}\hat{p}_{1k}(x_r) &= \frac{1}{\binom{m}{2}} \sum_{i < j} I_{1k}[x_r; y_i, y_j]; & \hat{q}_{1k}(y_i) &= \frac{1}{n} \frac{1}{m-1} \sum_{r=1}^n \sum_{\substack{j=1 \\ j \neq i}}^m I_{1k}[x_r; y_i, y_j]; \\ \hat{p}_{2k}(y_r) &= \frac{1}{\binom{m}{2}} \sum_{i < j} I_{2k}[y_r; x_i, x_j]; & \hat{q}_{2k}(x_i) &= \frac{1}{m} \frac{1}{n-1} \sum_{r=1}^m \sum_{\substack{j=1 \\ j \neq i}}^n I_{2k}[y_r; x_i, x_j];\end{aligned}$$

$$\begin{aligned}
\hat{\alpha}_{1st} &= \frac{1}{n-1} \sum_{r=1}^n [\hat{p}_{1s}(x_r) - U_s] \cdot [\hat{p}_{1t}(x_r) - U_t]; \\
\hat{\beta}_{1st} &= \frac{1}{m-1} \sum_{i=1}^m [\hat{q}_{1s}(y_i) - U_s] \cdot [\hat{q}_{1t}(y_i) - U_t]; \\
\hat{\alpha}_{2st} &= \frac{1}{m-1} \sum_{r=1}^m [\hat{p}_{2s}(y_r) - V_s] \cdot [\hat{p}_{2t}(y_r) - V_t]; \\
\hat{\beta}_{2st} &= \frac{1}{n-1} \sum_{i=1}^n [\hat{q}_{2s}(x_i) - V_s] \cdot [\hat{q}_{2t}(x_i) - V_t]; \\
\hat{\gamma}_{1st} &= \frac{1}{n-1} \sum_{r=1}^n [\hat{p}_{1s}(x_r) - U_s] \cdot [\hat{q}_{2t}(x_r) - V_t]; \\
\hat{\gamma}_{2st} &= \frac{1}{m-1} \sum_{r=1}^m [\hat{p}_{2s}(y_r) - V_s] \cdot [\hat{p}_{1t}(y_r) - U_t].
\end{aligned}$$

**Lemma 3.** Under condition (3), for  $s, t = 1, 2, 3$ ,

$$\begin{aligned}
\hat{\alpha}_{1st} &\xrightarrow{P} \alpha_{1st}; & \hat{\beta}_{1st} &\xrightarrow{P} \beta_{1st}; \\
\hat{\alpha}_{2st} &\xrightarrow{P} \alpha_{2st}; & \hat{\beta}_{2st} &\xrightarrow{P} \beta_{2st}; \\
\hat{\gamma}_{1st} &\xrightarrow{P} \gamma_{1st}; & \hat{\gamma}_{2st} &\xrightarrow{P} \gamma_{2st}.
\end{aligned}$$

Under the null  $H_0: F = G$ , for  $k, s, t = 1, 2, 3$ ,

$$p_{1k}(\mathbf{x}_1) \stackrel{D}{=} p_{2k}(\mathbf{y}_1); \quad q_{1k}(\mathbf{y}_1) \stackrel{D}{=} q_{2k}(\mathbf{x}_1);$$

and

$$\alpha_{1st} = \alpha_{2st}; \quad \beta_{1st} = \beta_{2st}; \quad \gamma_{1st} = \gamma_{2st}.$$

Let

$$\mathbf{U} = \begin{pmatrix} U_1 - \frac{1}{3} \\ U_2 - \frac{1}{3} \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} V_1 - \frac{1}{3} \\ V_2 - \frac{1}{3} \end{pmatrix}$$

and denote

$$\hat{\alpha}_1 = \begin{pmatrix} \hat{\alpha}_{111} & \hat{\alpha}_{112} \\ \hat{\alpha}_{121} & \hat{\alpha}_{122} \end{pmatrix}, \hat{\alpha}_2 = \begin{pmatrix} \hat{\alpha}_{211} & \hat{\alpha}_{212} \\ \hat{\alpha}_{221} & \hat{\alpha}_{222} \end{pmatrix}, \hat{\alpha} = (\hat{\alpha}_1 + \hat{\alpha}_2)/2.$$

$$\hat{\beta}_1 = \begin{pmatrix} \hat{\beta}_{111} & \hat{\beta}_{112} \\ \hat{\beta}_{121} & \hat{\beta}_{122} \end{pmatrix}, \hat{\beta}_2 = \begin{pmatrix} \hat{\beta}_{211} & \hat{\beta}_{212} \\ \hat{\beta}_{221} & \hat{\beta}_{222} \end{pmatrix}, \hat{\beta} = (\hat{\beta}_1 + \hat{\beta}_2)/2;$$

$$\hat{\gamma}_1 = \begin{pmatrix} \hat{\gamma}_{111} & \hat{\gamma}_{112} \\ \hat{\gamma}_{121} & \hat{\gamma}_{122} \end{pmatrix}, \hat{\gamma}_2 = \begin{pmatrix} \hat{\gamma}_{211} & \hat{\gamma}_{212} \\ \hat{\gamma}_{221} & \hat{\gamma}_{222} \end{pmatrix}, \hat{\gamma} = (\hat{\gamma}_1 + \hat{\gamma}_2)/2.$$

$$\hat{\Sigma}_1 = \frac{1}{n} \hat{\alpha}_1 + \frac{4}{m} \hat{\beta}_1, \quad \hat{\Sigma}_2 = \frac{1}{m} \hat{\alpha}_2 + \frac{4}{n} \hat{\beta}_2, \quad \hat{\Gamma} = \frac{2}{n} \hat{\gamma} + \frac{2}{m} \hat{\gamma}',$$

$$\hat{\Sigma} = \frac{1}{n} \hat{\alpha} + \frac{4}{m} \hat{\beta}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}, \quad \hat{\Omega} = \begin{pmatrix} \hat{\Sigma} & \hat{\Gamma} \\ \hat{\Gamma}' & \hat{\Sigma} \end{pmatrix}.$$

**Theorem 2.** Assuming condition (3) and under the null  $H_0: F = G$ ,

$$\begin{aligned} \mathbf{T}_1 &= \mathbf{U}' \hat{\Sigma}_1^{-1} \mathbf{U} \xrightarrow{D} \chi_2^2, \\ \mathbf{T}_2 &= \mathbf{V}' \hat{\Sigma}_2^{-1} \mathbf{V} \xrightarrow{D} \chi_2^2, \\ \mathbf{Q} &= \mathbf{W}' \hat{\Omega}^{-1} \mathbf{W} \xrightarrow{D} \chi_4^2. \end{aligned}$$

### 3 Simulation Study

We conducted a small simulation study to examine the empirical power of the test statistics  $T_1$ ,  $T_2$  and  $Q$  for testing the equality of two distribution functions. The Euclidean Distance was chosen for  $\delta(\cdot, \cdot)$ ; hence, the statistics are invariant under location and/or scale change, as well as under rotation. The performance of our tests was compared to the performances of Puri-Sen rank test (1971) and Székely and Rizzo's e-distance test (2004) against location shift alternatives and scale alternatives for bivariate normal distributions.

For small samples, the statistics are implemented by conditioning on the pooled observed samples to obtain a distribution free test procedure. In each case, the empirical power was estimated from simulation of 1000 pairs of samples and the test decision was based on 500 random permutations of the vector of sample membership. For large samples, our test statistics and Puri-Sen rank test were evaluated using their asymptotic distributions: chi-square distribution with 2 or 4 degree of freedom for triangle statistics and chi-square distribution with  $p$  degree of freedom for Puri-Sen rank test. The empirical power for these tests are calculated from simulation of 1000 pairs of bivariate normal samples. The asymptotic distribution of e-distance test is not available.

Table 1: Observed significance levels (percent rejections) at  $\alpha = 0.05$  of bivariate normal  $F = G = N_2((0, 0), I)$ .

n	m	PS	E	$T_1$	$T_2$	Q
10	10	3	4	5	6	6
15	15	5	5	5	6	4
20	20	7	7	5	4	5
30	30	5	NA	7	6	6
50	50	7	NA	5	4	5
100	100	4	NA	5	5	7

Table 1 lists the observed significances levels (percentage of rejection under  $H_0$ ). All tests in this simulation study achieved approximately correct empirical significance. Table 2 lists the empirical power of all tests against location shift alternatives. In each case, the triangle tests are least powerful and the e-distance is most powerful among these tests. Table 3 lists the empirical power of all tests against scale alternatives. In each case, the triangle tests are most powerful where the other two test statistics perform equally poorly. When there is a scale change in  $G$ , the distances from points in  $\mathbf{Y}$  sample to points in  $\mathbf{X}$  sample decrease or increase together, consequently, the test statistics are very sensitive to the change. This explains why the triangle test is so powerful in detecting small scale change. On the other hand, when there is a small shift in location, the distances of some points

Table 2: Empirical powers at  $\alpha = 0.05$  of bivariate normal location alternatives  $F = N_2((0, 0), I)$ ,  $G = N_2((0, \delta), I)$ .

		$\delta = 0.5$					$\delta = 1.0$					$\delta = 2.0$				
n	m	PS	E	$T_1$	$T_2$	Q	PS	E	$T_1$	$T_2$	Q	PS	E	$T_1$	$T_2$	Q
10	10	12	14	6	6	6	38	46	9	9	10	90	94	42	37	56
15	15	18	22	6	5	4	63	68	10	12	12	100	100	71	72	81
20	20	25	27	6	4	3	77	83	13	14	16	100	100	84	78	95
30	30	39	NA	6	6	4	92	NA	19	20	25	100	NA	95	93	100
50	50	55	NA	6	5	4	100	NA	29	27	48	100	NA	100	99	100
100	100	89	NA	8	10	12	100	NA	56	58	94	100	NA	100	100	100

Table 3: Empirical powers at  $\alpha = 0.05$  of bivariate normal scale alternatives  $F = N_2((0, 0), I)$ ,  $G = N_2((0, 0), \sigma^2 I)$ .

		$\sigma = 1.3$					$\sigma = 1.5$					$\sigma = 2$				
n	m	PS	E	$T_1$	$T_2$	Q	PS	E	$T_1$	$T_2$	Q	PS	E	$T_1$	$T_2$	Q
10	10	3	4	19	13	13	4	6	19	18	20	5	6	54	51	60
15	15	5	6	21	21	26	5	5	41	27	30	6	6	71	69	76
20	20	6	7	33	34	36	5	6	47	35	38	6	6	82	80	86
30	30	5	NA	42	47	45	5	NA	72	56	58	5	NA	97	97	98
50	50	6	NA	33	34	36	6	NA	92	82	83	7	NA	100	100	100
100	100	5	NA	92	98	77	5	NA	100	100	100	6	NA	100	100	100

in  $\mathbf{Y}$  sample to points in  $\mathbf{X}$  sample decrease, whereas those of the others increase and, consequently, the test statistics will not sensitive to a small shift in location.

Since Puri-Sen rank test is only applicable for cases with  $p < n$ , the simulation study result suggest that for high dimensional data ( $p > n$ ), we can apply e-distance test or triangle tests. The e-distance test is more powerful against location shift alternatives while our triangle test are more powerful against scale alternatives.

## Summary

For testing the equality of two multivariate distributions, the tests based on sample interpoint distances are very attractive since their computational complexity is independent of the dimension and are suitable for high dimensional data, especially for cases when  $p > n$ . Also, as pointed out by Bartoszyński, Pearl, and Lawrence (1997), the tests based on sample interpoint distances are broadly applicable to discrete, continuous, multivariate or even nonnumerical data with a proper method of computing distances between the objects in the sample. The simulation showed that our triangle tests are very powerful in detecting the scale difference in multivariate distributions.

As defined in section 2.2, the conditional probability that given  $\mathbf{y}_1$ ,  $\delta(\mathbf{x}_1, \mathbf{x}_2)$  being the largest ( $k = 3$ ) in the triangle provides a method to order data points from  $G$  distribution w.r.t.  $F$ . The larger this probability is, the closer the point is w.r.t.  $F$ . This probability may be used as a useful alternative to the data depth discussed by Liu, Parelius, and Singh (1999), which orders multivariate data from the center outward. Therefore, the triangle data depth can be used to develop a multivariate Wilcoxon rank sum test and to define a nonparametric classification rule.

## References

- [1] Bartoskzyński, R., Pearl, D. K. and Lawrence, J. (1997). A multidimensional goodness-of-fit test based on interpoint distances. *J. Ameri. Statist. Assoc.*, 97, 577-586.
- [2] Bickel, P. J. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Statist.*, 40, 1-23.
- [3] Bickel, P. J. and Breiman, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Probab.*, 11, 185-214.
- [4] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalization of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.*, 7, 697-717.
- [5] Gibbons, J. D. (1971). *Nonparametric Statistical Inference*. McGraw-Hill.
- [6] Henze, N. (1988). A multivariate two-sample test based on the the number of nearest neighbor type coincidences. *Ann. Statist.*, 16, 772-783.
- [7] Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. *Ann. Statist.*, 27, 290-298.
- [8] Liu, R., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference (with Discussion). *Ann. Statist.*, 27, 783-858.
- [9] Maa, J. F., Pearl, D. K. and Bartoskzyński, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann. Statist.*, 24, 1069-1074.
- [10] Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley and Sons.
- [11] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Ameri. Statist. Assoc.*, 81, 799-806.
- [12] Szekely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension, *InterStat*, Nov. (5).
- [13] Aslan, B. and Zeoh, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy , *J. Statist. Comp. and Simul.*, vol. 75, No. 2, 109-119.