

# On Distribution-Free Tests for the Multivariate Two-Sample Location-Scale Model

Valentin Rousson

University of Zürich, Zürich, Switzerland

E-mail: rousson@ifspm.unizh.ch

Received July 8, 1999; published online July 19, 2001

In this paper, we propose simple exact procedures for testing both a location shift and/or a scale change between two multivariate distributions. Our tests are strictly distribution-free and can be made either scale invariant or rotation invariant. Our approach combines a generalization of the Wilcoxon test based on projections of the data onto the first principal component, a generalization of the Siegel–Tukey test based on the concept of data depth, and a bivariate test for the location problem proposed by K. V. Mardia (1967, *J. Roy. Statist. Soc. Ser. B* **29**, 320–342). In addition, we show that the limiting null distribution of a test statistic proposed by R. Y. Liu and K. Singh (1993, *J. Amer. Statist. Assoc.* **88**, 252–260) does not depend on the depth considered. © 2001 Elsevier Science

AMS 1991 subject classifications: 62H15; 62G10.

*Key words and phrases:* data depth; multivariate orderings; nonparametric methods; principal component analysis; rank tests.

## 1. INTRODUCTION

We consider two independent random samples  $X = \{X_1, \dots, X_m\}$  and  $Y = \{Y_1, \dots, Y_n\}$  drawn from unknown continuous  $p$ -variate distributions  $F$  and  $G$ , respectively. Throughout the paper, we wish to nonparametrically test the null hypothesis of the equality of these two distributions

$$H_0: F(x) = G(x) \quad \text{for all } x = (x_1, \dots, x_p).$$

The choice of the testing procedure will of course depend on the alternative of interest. In this paper, we consider a location model

$$H_1: \text{there exists } \theta = (\theta_1, \dots, \theta_p) \neq 0,$$

$$G(x) = F(x - \theta) \quad \text{for all } x = (x_1, \dots, x_p),$$

where  $0$  is the  $p$ -variate null vector, a scale model

$$H_2: \text{there exists } \sigma = (\sigma_1, \dots, \sigma_p) \neq 1,$$

$$G(x) = F(x/\sigma) \quad \text{for all } x = (x_1, \dots, x_p),$$

where  $x/\sigma = (x_1/\sigma_1, \dots, x_p/\sigma_p)$  and  $\mathbf{1}$  denotes a  $p$ -dimensional vector with all entries equal to 1, and a location-scale model

$H_{ls}$ : there exists  $\theta = (\theta_1, \dots, \theta_p) \neq 0$  or there exists  $\sigma = (\sigma_1, \dots, \sigma_p) \neq 1$ ,

$$G(x) = F((x - \theta)/\sigma) \quad \text{for all } x = (x_1, \dots, x_p),$$

where  $(x - \theta)/\sigma = ((x_1 - \theta_1)/\sigma_1, \dots, (x_p - \theta_p)/\sigma_p)$ . If one wishes to non-parametrically test  $H_0$  against  $H_l$ ,  $H_s$ , or  $H_{ls}$  in the univariate case ( $p = 1$ ), he may use the Wilcoxon rank-sum test, the Siegel–Tukey test or the Kolmogorov–Smirnov test, respectively, which are strictly distribution-free procedures. The aim of this paper is to propose a simple way of generalizing and combining these well-known univariate tests in order to provide another strictly distribution-free procedure which is valid under a location-scale model in the multivariate case ( $p \geq 1$ ).

Several nonparametric methods have been proposed to deal with the multivariate two-sample location problem. See Wald and Wolfowitz (1944), Chatterjee and Sen (1964), Puri and Sen (1966), Tamura (1966), Brown and Hettmansperger (1987), or Randles (1992) among others. Most of these methods are based on permutation tests so that the test statistics are only conditionally distribution-free, and the null distribution is either complicated to tabulate, or depends on the data. As a consequence, these procedures cannot be put into practice as exact tests. Instead they may use an approximative chi-square null distribution with  $p$  degrees of freedom. See also Puri and Sen (1971, Chap. 5) and Hettmansperger (1984, Chap. 6).

One unconditional nonparametric test for the location problem was given by Mardia (1967) in the bivariate case. His method centers the data with respect to the sample mean of the combined sample and orders the  $m + n$  angles between the plotted data with the  $x$ -axis. Then, the test statistic is defined as

$$U = \frac{2(m+n-1)}{mn} \left[ \left\{ \sum_{i=1}^m \cos(2\pi r_i/(m+n)) \right\}^2 + \left\{ \sum_{i=1}^m \sin(2\pi r_i/(m+n)) \right\}^2 \right],$$

where  $r_i$  denotes the rank of  $X_i$  in the combined sample with respect to this angle-ordering. The null hypothesis is rejected when  $U$  is too large. This statistic is affine invariant and critical values for small samples were provided when  $m + n \leq 18$ . For larger samples, an approximative chi-square distribution with 2 degrees of freedom was demonstrated. Randles and Peters (1990) and Peters and Randles (1991) proposed generalizations of this test without tabulating the null distribution.

If we are interested in a multivariate scale model we may use the concept of *data depth* introduced by Tukey (1975). The depth of a point  $x$  into a distribution  $F$ , noted here  $D(F, x)$ , is an indication of how “central”  $x$  is in  $F$ .

Larger depths are associated with more central points. This concept provides a center-outward ordering of the observations  $X_i$  of a data set  $X$ , by considering the ordering of their depths  $D(F_m, X_i)$  into their empirical distribution  $F_m$ . Different kinds of depths have been proposed in the literature, including those of Tukey (1975), Oja (1983), and Liu (1990). The most natural choice of depth is arguably the Euclidean one,

$$D_1(F, x) = [(x - \mu_F)' (x - \mu_F)]^{-1/2},$$

which is rotation invariant. A scale invariant version of the Euclidean depth is given by

$$D_2(F, x) = [(x - \mu_F)' \Sigma_F^{-1} (x - \mu_F)]^{-1/2},$$

where  $\mu_F$  is the expectation of  $F$  and  $\Sigma_F$  is the diagonal matrix which contains the marginal variances of  $F$ . If  $\Sigma_F$  denotes the variance-covariance matrix of  $F$ , we obtain the affine invariant Mahalanobis depth.

Let  $Z = \{Z_1, \dots, Z_{m+n}\}$  be the union of  $X$  and  $Y$  and let  $H_{m+n}$  be its empirical distribution. In order to test for a scale change between two multivariate distributions, Liu and Singh (1993) proposed to use the test statistic

$$W_1 = \sum_{i=1}^n \text{Rank}(Y_i) \text{ in } Z \text{ w.r.t. } D(H_{m+n}, Y_i),$$

which follows the Wilcoxon distribution under the null hypothesis (i.e., the distribution of the sum of the observations of a random sample of size  $n$  drawn without replacement from  $\{1, 2, \dots, m+n\}$ ). This test may be seen as a generalization of the Siegel–Tukey test which also defines a center-outward ordering.

This procedure of Liu and Singh is exact but not powerful against a location shift between  $F$  and  $G$ . As a consequence, it is not adequate if we consider a multivariate location-scale model. To overcome this drawback, Liu and Singh (1993) considered the statistic

$$W_2 = \sum_{i=1}^n \text{Rank}(Y_i) \quad \text{in } \{Y_i\} \cup X \text{ w.r.t. } D(F_m, Y_i),$$

where each  $Y_i$  is ranked individually in the data set  $\{Y_i\} \cup X$  rather than in  $Z$ . If the dispersion of the  $X_i$  is larger than the dispersion of the  $Y_i$ , the statistic

$$W'_2 = \sum_{i=1}^m \text{Rank}(X_i) \quad \text{in } \{X_i\} \cup Y \text{ w.r.t. } D(G_n, X_i)$$

should be used instead, where  $G_n$  is the empirical distribution of the  $Y_i$ . This procedure is able to detect a location shift and/or a scale change between  $F$  and  $G$  but is no longer an exact test. Under the null hypothesis, it was shown that the limiting distribution of  $W_2/(nm) - 1/m$  when  $\min(m, n) \rightarrow \infty$  is normal with mean  $1/2$  and with variance  $(m+n)/(12mn)$  if  $D$  is the Mahalanobis depth. Liu and Singh (1993) wrote that finding the limiting distributions for other kinds of depth remained an open problem.

We take here the opportunity to provide a simple argument which proves that this limiting null distribution is actually the same whatever the depth chosen. Consider the statistic

$$W_3 = \sum_{i=1}^n \text{Rank}(Y_i) \quad \text{in } Z \text{ w.r.t. } D(F_m, Y_i),$$

which is not distribution-free under the null hypothesis, but whose limiting null distribution is clearly the same as that of  $W_1$  (since both  $F_m$  and  $H_{m+n}$  converge to  $F$ ). As the null distribution of  $W_1$  is a Wilcoxon one, it converges towards a normal distribution with mean  $n(m+n+1)/2$  and variance  $mn(m+n+1)/12$  (see, e.g., Hettmansperger, 1984, p. 134). Observe that  $W_2 = W_3 - n(n-1)/2$ . As a consequence, the limiting null distribution of  $W_2/(nm) - 1/m$  is normal with mean  $1/2$  and with variance  $(m+n+1)/(12mn)$  regardless of the depth  $D$  considered.

Liu and Singh (1993) introduced still another procedure for the multivariate two-sample location-scale problem, which is defined only when  $m$  is much larger than  $n$  (e.g. in a quality control context). See also Liu (1995). There seems however to exist no practical and strictly distribution-free procedure to deal with this problem for arbitrary small sample sizes (if we except Vincze (1961) who suggested to apply the Kolmogorov–Smirnov test along a line selected at random). In this paper, we propose a simple procedure which attempts to fill this gap. Our test is introduced in Section 2 and numerical results are provided in Section 3. Section 4 illustrates the application of our test on a real multivariate data set while some final comments take place in Section 5.

## 2. THE PROPOSED TEST

Let  $Z = \{Z_1, \dots, Z_{m+n}\}$  be the union of  $X$  and  $Y$  as in Section 1. Under the null hypothesis, the joint density of  $Z$  is symmetric in  $Z_1, \dots, Z_{m+n}$ . The idea of our method is to replace the  $p$ -variate data set  $Z$  by a  $q$ -variate data set  $V = \{V_1, \dots, V_{m+n}\}$  (with  $q=1$  or  $2$ ) defined as  $V_i = t(Z_i)$  where the function  $t$  depends itself on the data, which we may write  $t(x) = t(x; Z_1, \dots, Z_{m+n})$ . If this function  $t$  is symmetric in  $Z_1, \dots, Z_{m+n}$ , the joint

density of  $V$  will also be symmetric in  $V_1, \dots, V_{m+n}$  under the null hypothesis (see Proposition 1 below). For example, if  $q=1$ , each of the  $(m+n)!$  possible orderings of the  $V_i$  will be equally probable under the null hypothesis. Hence we obtain a strictly distribution-free procedure for testing  $H_0$  by applying a strictly distribution-free two-sample test to the  $V_i$ .

**PROPOSITION 1.** *Consider two (multivariate) data sets  $Z = \{Z_1, \dots, Z_N\}$  and  $V = \{V_1, \dots, V_N\} = T(\{Z_1, \dots, Z_N\})$  where the transformation  $T$  is defined such that  $V_i = t(Z_i)$  for  $i=1, \dots, N$  with  $t(x) = t(x; Z_1, \dots, Z_N)$ . Denote the joint densities of  $Z$  and  $V$  by  $f_z$  and  $f_v$ , respectively. If  $f_z$  and  $t$  are symmetric in  $Z_1, \dots, Z_N$ , then  $f_v$  is symmetric in  $V_1, \dots, V_N$ .*

*Proof.* Let  $\sigma(1), \dots, \sigma(N)$  be any permutation of  $1, \dots, N$ . Consider the data set  $W = \{W_1, \dots, W_N\}$  such that  $W_i = t(Z_{\sigma(i)})$  for  $i=1, \dots, N$ . Note that  $W = T(\{Z_{\sigma(1)}, \dots, Z_{\sigma(N)}\})$  has the same joint distribution as  $V = T(\{Z_1, \dots, Z_N\})$  since  $f_z$  is symmetric in  $Z_1, \dots, Z_N$ . But as  $t$  is also symmetric in  $Z_1, \dots, Z_N$ , we have  $W_i = V_{\sigma(i)}$  for  $i=1, \dots, N$ . Thus  $f_v(V_{\sigma(1)}, \dots, V_{\sigma(N)}) = f_v(V_1, \dots, V_N)$  and so  $f_v$  is symmetric in  $V_1, \dots, V_N$ .

In order to be powerful under a location-scale model, we shall define the  $V_i$  such that they contain information about both the location and the scale of the  $Z_i$ . For summarizing the location of a multivariate data set into a lower dimension we may consider projection techniques such as principal component analysis. For characterizing the scale of the data we shall use the concept of data depth. More precisely, we shall use the inverse of the depths  $D_1$  or  $D_2$  defined in Section 1. Our general method may be described as follows:

- (1) Center the values  $Z_i$  ( $i=1, \dots, m+n$ ) such that they have a mean vector zero.
- (2) Calculate the projections  $U_i$  of the  $Z_i$  ( $i=1, \dots, m+n$ ) onto the first principal component of the combined sample.
- (3) Calculate the  $R_i$  as inverse depths of the  $Z_i$  ( $i=1, \dots, m+n$ ) into the empirical distribution of the combined sample (if the depth  $D_1$  is chosen, the  $R_i$  are the Euclidean distances between the  $Z_i$  and the origin).
- (4) Standardize in turn the  $U_i$  and the  $R_i$  such that they have zero mean and unit standard deviation.
- (5) Define  $V_i = \psi(U_i, R_i)$  ( $i=1, \dots, m+n$ ) for a well chosen  $q$ -variate function  $\psi$  (with  $q=1$  or  $q=2$ ).
- (6) Apply a distribution-free procedure to the  $V_i$  valid in dimension  $q$  (e.g., the Wilcoxon rank-sum test or the Kolmogorov–Smirnov test if  $q=1$ , or Mardia's test if  $q=2$ ).

One may verify that the definition of the  $V_i$  does not involve the particular ordering of the  $Z_i$ . In particular, we do not use any information that an observation  $Z_i$  is in  $X$  or in  $Y$ . This ensures the symmetry in its arguments of the function  $t$  defined earlier in this section and Proposition 1 applies under the null hypothesis. This proves the strictly nonparametric character of our test.

If we wish to use Mardia's test at Step 6 of our procedure, we simply take  $\psi$  as the bivariate identity function and apply Mardia's test to the bivariate data set  $(U_i, R_i)$ . We shall see in the next section that this approach is particularly promising.

If we wish to use an univariate test, such as the Wilcoxon rank-sum test or the Kolmogorov–Smirnov test, we may define  $\psi$  as a linear combination of  $U_i$  and  $R_i$  and calculate

$$V_i(\gamma) = \gamma U_i + \text{sign}(U_i, R_i)(1 - \gamma) R_i,$$

for a certain value of  $\gamma$  in  $[0, 1]$  and where  $\text{sign}(U_i, R_i)$  denotes the sign of the correlation of the bivariate data set  $(U_i, R_i)$ . The choice  $\gamma = 1$  leads to a powerful test under a location model, while  $\gamma = 0$  defines a powerful test under a scale model. The intermediate choice  $\gamma = 0.5$  corresponds to defining  $V_i$  as the projection of the bivariate elements  $(U_i, R_i)$  onto their first principal component. This choice is found adequate under a location-scale model since both  $U_i$  (location information) and  $R_i$  (scale information) are used for defining  $V_i$ .

Another possible choice of  $\gamma$  may be motivated as follows. In the univariate case, it is equivalent to apply the Wilcoxon rank-sum test to the  $V_i$  or to the  $Z_i$  when  $\gamma = 1$  (since  $V_i = Z_i$ ). On the other hand, applying the Wilcoxon rank-sum test to the  $V_i$  is equivalent to applying the Siegel–Tukey test to the  $Z_i$  when  $\gamma = 0$ . It is known that the univariate Wilcoxon test is more efficient than the univariate Siegel–Tukey test when the null distribution is normal. By comparing these nonparametric tests with their parametric analogues, namely the classical  $t$ - and  $F$ -tests, respectively, we have an asymptotic relative efficiency of 0.955 for the former and 0.608 for the latter (see, e.g., Gibbons, 1976, p. 30). Thus, the Wilcoxon test is in a sense 57% more powerful than the Siegel–Tukey test. In order to have a test which is well-balanced with respect to location and scale powerfulness, one could define  $V_i$  with  $\gamma = 1/(1 + 1.57) = 0.39$ .

A desirable property of a two-sample test would be to be affine invariant (the outcome of the test should not depend on the  $p$ -variate system of coordinates which expresses the  $Z_i$ ). Unfortunately, our method does not share this property. However, our test can be made either rotation invariant (or more generally invariant with respect to an orthogonal transformation) or scale invariant, depending on how Steps 2 and 3 of our procedure are

implemented. Note that in both cases, our test will be invariant under an homogeneous scale change (when the same scale change is applied to each variable). In fact, the rotation and scale invariance properties of our test depend directly on the rotation and scale invariance properties of the  $V_i$ , and hence of the  $U_i$  and  $R_i$ , as follows:

- Step 2. Recall that the definition of principal components uses eigenvectors of the variance-covariance or correlation matrix of the data. Use of the former option defines rotation invariant  $U_i$ , whereas use of the latter option defines scale invariant  $U_i$ . Note that the  $U_i$  cannot be made affine invariant.

- Step 3. Choice of the Euclidean depth  $D_1$  defines rotation invariant  $R_i$ , while choice of depth  $D_2$  defines scale invariant  $R_i$ . Note that the Mahalanobis depth or any other affine invariant depth would define affine invariant  $R_i$ .

This property of our test may be sufficient in praxis. In most applications the concern is either rotation invariance or scale invariance, but rarely both. Rotation invariance is desirable when each variable shares the same unit of measurement, for example when we are interested in the representation of items on a two-dimensional map. In such cases, we would like our test to be independent of the coordinates chosen (and thus to be rotation invariant), but invariance under a (nonhomogeneous) scale change is often not relevant. On the other hand, scale invariance is a necessary property when the units of the variables involved may not be compared with each other (such as a distance with a temperature), in which cases a rotation does not really make sense.

### 3. SIMULATION STUDY

In order to check the performance of our methods, we performed a small simulation study. We considered bivariate samples  $X$  and  $Y$  of sizes  $m = n = 9$ . The first sample was drawn from a product of two independent standardized normal or Laplace distributions. The second sample was generated such that we had a location shift  $\theta$  and a scale change  $\sigma$ . We considered the following four possibilities:

- (1)  $\theta = (0, 0)$  and  $\sigma = (1, 1)$
- (2)  $\theta = (1, 1)$  and  $\sigma = (1, 1)$
- (3)  $\theta = (0, 0)$  and  $\sigma = (2, 2)$
- (4)  $\theta = (1, 1)$  and  $\sigma = (2, 2)$ .

Case (1) corresponds to the null hypothesis. Cases (2), (3), and (4) correspond to a location model, a scale model and a location-scale model, respectively.

Among exact procedures, we tried the Wilcoxon and Kolmogorov–Smirnov tests applied to the rotation invariant  $V_i(\gamma)$  defined in Section 2 with  $\gamma = 0, 0.39, 0.5, 1$ , as well as Mardia's procedure applied to the original data set and to the rotation invariant data set  $(U_i, R_i)$ . The level 0.9 was used throughout. Note that due to their discreteness, the true levels of the Wilcoxon, the Kolmogorov–Smirnov and Mardia's tests were 0.906, 0.966 and 0.900, respectively. For obtaining a fairer comparison with the other methods, the Kolmogorov–Smirnov tests were randomized such that their true levels were 0.900.

In addition, we tried the nonexact procedure based on  $W_2$  or  $W'_2$  (see Section 1). The Euclidean depth  $D_1$  was used. The statistic  $W_2$  was selected when the variance of the  $R_i$  corresponding to the  $X_i$  was smaller than the variance of the  $R_i$  corresponding to the  $Y_i$ . Otherwise the statistic  $W'_2$  was selected. Both a normal and a Wilcoxon approximation were calculated. The Wilcoxon version of this test was then calibrated using a nominal level of 0.93 in order to be compared with the exact tests.

Tables I and II give the percentage of acceptance of the null hypothesis achieved by each of these methods based on 1000 samples drawn according to cases (1), (2), (3), and (4) with bivariate normal and bivariate Laplace distributions, respectively. From the first columns of these tables, we see that the Wilcoxon approximation to  $W_2$  was more accurate than the normal approximation under the null hypothesis, but the proportion of acceptance remained well below the nominal level.

TABLE I

Percentage of Acceptance of the Null Hypothesis for 10 Methods and 4 Cases When the Population Is Bivariate Normal with  $m = n = 9$

Test	Case (1)	Case (2)	Case (3)	Case (4)
Wilcox. $V_i(1)$	90.9	22.5	90.2	59.5
Wilcox. $V_i(0.5)$	89.5	56.5	49.6	32.3
Wilcox. $V_i(0.39)$	90.5	74.8	39.0	29.1
Wilcox. $V_i(0)$	91.4	95.6	32.0	43.2
Kolm–Smirn. $V_i(1)$	90.3	31.0	82.1	54.2
Kolm–Smirn. $V_i(0.5)$	89.8	53.9	57.5	36.7
Kolm–Smirn. $V_i(0)$	90.7	93.5	38.4	49.1
Mardia orig. data	89.9	30.3	84.5	49.1
Mardia $(U_i, R_i)$	90.8	46.3	44.7	34.4
Approx. normal $W_2$	83.3	38.7	18.2	14.6
Approx. Wilcox. $W_2$	86.3	42.6	21.0	15.7
Calibrated $W_2$	90.3	50.1	28.7	20.3

TABLE II

Percentage of Acceptance of the Null Hypothesis for 10 Methods and 4 Cases When the Population Is Bivariate Laplace with  $m = n = 9$

Test	Case (1)	Case (2)	Case (3)	Case (4)
Wilcox. $V_i(1)$	91.2	25.8	92.1	55.4
Wilcox. $V_i(0.5)$	90.2	52.7	61.3	46.5
Wilcox. $V_i(0.39)$	90.9	73.0	52.0	47.5
Wilcox. $V_i(0)$	91.6	91.7	49.8	56.1
Kolm-Smirn. $V_i(1)$	89.6	27.2	85.3	47.7
Kolm-Smirn. $V_i(0.5)$	89.4	53.3	67.2	49.2
Kolm-Smirn. $V_i(0)$	92.8	90.2	54.7	60.0
Mardia orig. data	89.7	16.8	83.3	36.7
Mardia ( $U_i, R_i$ )	90.9	33.9	59.5	36.0
Approx. normal $W_2$	83.3	28.1	35.1	16.7
Approx. Wilcox. $W_2$	86.0	31.2	37.7	18.7
Calibrated $W_2$	89.6	36.6	44.4	22.7

For the bivariate normal location model, the Wilcoxon test applied to the  $V_i(1)$  was the most powerful among all the tests considered, including the nonexact one. When the distribution was bivariate Laplace, however, it did not perform as well as Mardia's procedure, which is known to have high power for heavy-tailed distributions (see Peters and Randles, 1991). Under a scale model, the Wilcoxon test applied to the  $V_i(0)$  was the best

TABLE III

Percentage of Acceptance of the Null Hypothesis for 10 Methods and Cases (A), (B), and (C) with  $m = n = 30$

Test	Case (A)	Case (B)	Case (C)
Wilcox. $V_i(1)$	89.9	88.5	83.1
Wilcox. $V_i(0.5)$	81.2	45.0	70.6
Wilcox. $V_i(0.39)$	77.2	38.8	73.9
Wilcox. $V_i(0)$	77.0	46.3	79.5
Kolm-Smirn. $V_i(1)$	87.4	61.4	61.1
Kolm-Smirn. $V_i(0.5)$	82.6	32.8	55.2
Kolm-Smirn. $V_i(0)$	75.3	26.8	54.1
Mardia orig. data	89.5	53.0	50.2
Mardia ( $U_i, R_i$ )	79.3	22.0	42.8
Approx. normal $W_2$	76.6	47.0	73.1
Approx. Wilcox. $W_2$	77.3	47.2	73.9
Calibrated $W_2$	81.3	50.9	76.7

among exact procedures. For the bivariate normal location-scale model, the Wilcoxon test applied to the  $V_i(0.39)$  was especially powerful. Nevertheless, in the bivariate Laplace case, Mardia's procedures again performed better.

In order to be powerful under all these models, Mardia's procedure applied to the data set  $(U_i, R_i)$ , and the Wilcoxon test applied to the  $V_i(\gamma)$  with  $\gamma$  around 0.4 or 0.5, were found to be good compromises between tests specifically developed for a location model and tests specifically developed for a scale model. The nonexact test based on  $W_2$  performed well too, but its calibration may be problematic in practice (when the distribution is unknown).

The randomized Kolmogorov–Smirnov tests were found close to the Wilcoxon ones. Moreover, these tests are interesting if one is interested in detecting other kinds of difference than a location shift or a scale change. Table III provides results based on 1000 samples drawn from the following three situations:

(A) The first sample follows a product of two independent normal and the second a product of two independent Laplace distributions.

(B) The first sample follows a product of two independent normal and the second a product of two independent chi-square (with 1 d.f.) distributions.

(C) The first sample follows a product of two independent Laplace and the second a product of two independent chi-square (with 1 d.f.) distributions.

TABLE IV

Percentage of Acceptance of the Null Hypothesis for 9 Methods and 4 Cases When the Population Is Univariate Normal with  $m = n = 30$

Test	Case (1)	Case (2)	Case (3)	Case (4)
Wilcox. $V_i(1)$	90.7	2.1	89.9	25.0
Wilcox. $V_i(0.5)$	90.5	36.4	30.8	11.8
Wilcox. $V_i(0.39)$	91.7	58.6	18.5	11.2
Wilcox. $V_i(0)$	89.7	94.3	12.6	21.4
Kolm–Smirn. $V_i(1)$	90.6	5.4	60.3	11.8
Kolm–Smirn. $V_i(0.5)$	90.3	22.2	39.5	12.2
Kolm–Smirn. $V_i(0)$	90.4	92.7	13.9	23.7
Mardia $(U_i, R_i)$	91.9	24.9	22.6	11.1
Approx. normal $W_2$	88.4	33.8	11.0	5.2
Approx. Wilcox. $W_2$	88.7	34.5	11.6	5.3
Calibrated $W_2$	89.5	36.5	12.8	6.0

TABLE V

Percentage of Acceptance of the Null Hypothesis for 9 Methods and 4 Cases When the Population Is Univariate Laplace with  $m = n = 30$

Test	Case (1)	Case (2)	Case (3)	Case (4)
Wilcox. $V_i(1)$	89.8	0.9	90.3	10.5
Wilcox. $V_i(0.5)$	88.7	18.2	43.6	21.2
Wilcox. $V_i(0.39)$	88.7	47.5	34.0	25.5
Wilcox. $V_i(0)$	89.7	89.0	29.1	43.4
Kolm-Smirn. $V_i(1)$	89.2	0.6	72.3	4.7
Kolm-Smirn. $V_i(0.5)$	89.7	11.5	51.6	19.7
Kolm-Smirn. $V_i(0)$	88.9	2.1	37.2	46.1
Mardia ( $U_i, R_i$ )	89.5	33.9	40.1	3.1
Approx. normal $W_2$	86.8	10.9	25.1	3.1
Approx. Wilcox. $W_2$	87.2	11.1	25.7	3.5
Calibrated $W_2$	90.0	13.6	31.7	4.8

For all these cases, the distributions were standardized such that the vector means were equal to (1,1) and the marginal variances to (2,2). Sample sizes were taken to be  $m = n = 30$ . The true levels of the Wilcoxon and the Kolmogorov-Smirnov tests were here 0.901 and 0.929, and the latter were randomized. One can see their advantage over the Wilcoxon tests for cases (B) and (C), while case (A) corresponds to a case for which all tests failed to be powerful. Mardia's procedure was applied using its chi-square null approximation. It was found once again very powerful when applied to the data set ( $U_i, R_i$ ). The statistic  $W_2$  was calibrated using a nominal level 0.92 but was not as performant as under cases (1)–(4).

Finally we tried all these methods under the univariate analogues of situations (1)–(4) with sample sizes  $m = n = 30$  (except the original Mardia's procedure which is strictly bivariate). Tables IV and V refer to normal and Laplace cases, respectively. It is interesting to note that Mardia's procedure applied to the data set ( $U_i, R_i$ ) was found competitive with the randomized Kolmogorov-Smirnov test under a location-scale model. Moreover, it had not to be randomized for achieving exactness since its true level was very close to the nominal one.

#### 4. REAL DATA EXAMPLE

We illustrate our methodology on the Tibetan skulls data set studied in Morant (1923). This data set consists of  $p = 5$  measurements (all in millimetres)

$Z_1$ : Greatest length of skull

$Z_2$ : Greatest horizontal breadth of skull

$Z_3$ : Height of skull

$Z_4$ : Upper face height

$Z_5$ : Face breadth, between outermost points of cheek bones

TABLE VI

Tibetan Skulls Data Set and Standardized Values of  $U_i$  and  $R_i$  for Both the Rotation and the Scale Invariant Versions of Our Tests

$Z_{i1}$	$Z_{i2}$	$Z_{i3}$	$Z_{i4}$	$Z_{i5}$	$U_i$ rot. inv.	$R_i$	$U_i$ scale inv.	$R_i$
190.5	152.5	145.0	73.5	136.5	1.22	0.98	1.21	1.09
172.5	132.0	125.5	63.0	121.0	-1.55	0.94	-1.83	1.21
167.0	130.0	125.5	69.5	119.5	-1.78	1.27	-1.74	1.16
169.5	150.5	133.5	64.5	128.0	-0.91	0.63	-0.87	0.72
175.0	138.5	126.0	77.5	135.5	-0.23	-0.62	-0.06	-0.48
177.5	142.5	142.5	71.5	131.0	-0.06	-0.55	0.08	-0.37
179.5	142.5	127.5	70.5	134.5	-0.12	-1.04	-0.23	-0.93
179.5	138.0	133.5	73.5	132.5	-0.07	-1.85	-0.05	-1.95
173.5	135.5	130.5	70.0	133.5	-0.60	-0.87	-0.59	-0.99
162.5	139.0	131.0	62.0	126.0	-1.74	1.16	-1.70	1.06
178.5	135.0	136.0	71.0	124.0	-0.57	-0.47	-0.60	-0.52
171.5	148.5	132.5	65.0	146.5	-0.06	0.80	-0.04	0.88
180.5	139.0	132.0	74.5	134.5	0.09	-1.78	0.11	-1.81
183.0	149.0	121.5	76.5	142.0	0.58	0.57	0.54	0.84
169.5	130.0	131.0	68.0	119.0	-1.59	0.98	-1.59	0.88
172.0	140.0	136.0	70.5	133.5	-0.48	-0.81	-0.33	-1.05
170.0	126.5	134.5	66.0	118.5	-1.63	1.32	-1.69	1.34
182.5	136.0	138.5	76.0	134.0	0.31	-1.04	0.39	-0.96
179.5	135.0	128.5	74.0	132.0	-0.23	-1.13	-0.27	-1.09
191.0	140.5	140.5	72.5	131.5	0.72	-0.12	0.53	-0.33
184.5	141.5	134.5	76.5	141.5	0.77	-0.62	0.81	-0.69
181.0	142.0	132.5	79.0	136.5	0.40	-1.01	0.57	-0.82
173.5	136.5	126.0	71.5	136.5	-0.50	-0.55	-0.48	-0.57
188.5	130.0	143.0	79.5	136.0	0.84	0.44	0.91	0.64
175.0	153.0	130.0	76.5	142.0	0.34	0.50	0.61	0.57
196.0	142.5	123.5	76.0	134.0	0.93	0.75	0.54	0.51
200.0	139.5	143.5	82.5	146.0	2.21	1.94	2.17	1.81
185.0	134.5	140.0	81.5	137.0	0.75	-0.15	0.96	0.16
174.5	143.5	132.5	74.0	136.5	-0.11	-0.98	0.07	-1.15
195.5	144.0	138.5	78.5	144.0	1.71	0.96	1.62	0.70
197.0	131.5	135.0	80.5	139.0	1.37	0.98	1.18	0.72
182.5	131.0	135.0	68.5	136.0	0.01	-0.64	-0.21	-0.55

Note. Rows 1-17 and 18-32 refer to type X and type Y, respectively.

made on 32 skulls divided into two groups. The  $m = 17$  skulls of type  $X$  came from graves in Sikkim and neighbouring area of Tibet, while the  $n = 15$  skulls of type  $Y$  were picked up on a battlefield in the Lhasa district and were believed to be those of native soldiers from the eastern province of Khams. Hand *et al.* (1994, p. 111) noted that these skulls were of particular interest because it was thought at the time that Tibetans from Khams might be survivors of a particular fundamental human type, unrelated to the Mongolian and Indian types which surrounded them.

When there is some doubt about the origin of a group of skulls, as those of type  $Y$ , it may be useful to test whether they belong to the same population as another group of skulls whose origin is more certain (as those of type  $X$ ). If skulls of type  $Y$  came from another population as skulls of type  $X$ , they would typically differ in location with respect to at least some of the measurements. On the other hand, if skulls of type  $Y$  came from a mixture of several distinct populations (as it may arise on a battlefield), their measurements would exhibit an higher dispersion compared to type  $X$  (they would differ in scale). Our tests which can detect both a location shift and scale change between two samples may be of particular interest to answer such a question.

The data are given in Table VI, along with the  $U_i$  and the  $R_i$  for both the location and scale invariant version of our tests. We may indeed hesitate here between these two options. Since all measurements were made in millimetres, one could argue that scale invariance is not really necessary and advocate rotation invariance in order to be somewhat robust with respect to how the five variables have been defined to characterize a skull. On the other hand, one may prefer scale invariance since dispersion may differ among variables. Table VII provides the  $p$ -values resulting from our different tests. We observe that location differed significantly between the

TABLE VII

$p$ -Values for Both Versions of Our Tests Applied to  
the Tibetan Skulls Data Set of Table VI

Test	Rot. inv.	Scale inv.
Wilcox. $V_i(1)$	0.0003	0.0005
Wilcox. $V_i(0.5)$	0.08	0.02
Wilcox. $V_i(0.39)$	0.18	0.08
Wilcox. $V_i(0)$	0.53	0.35
Kolm-Smirn. $V_i(1)$	0.002	0.003
Kolm-Smirn. $V_i(0.5)$	0.08	0.01
Kolm-Smirn. $V_i(0)$	0.55	0.11
Mardia ( $U_i, R_i$ )	0.003	0.006

two types of skulls while scale did not (see the tests which use  $V_i(1)$  and  $V_i(0)$ , respectively). Conclusion of the tests which use  $V_i(0.5)$  varied (at the 5% significance level) whether implemented with their rotation or scale invariance versions. By way of contrast, Mardia's test applied to the data set  $(U_i, R_i)$  clearly rejected the null hypothesis in both cases and proved to be robust with respect to this invariance issue, which is certainly another advantage of this test.

## 5. CONCLUSION

In this paper, we have proposed some strictly nonparametric methods which combine the ideas of linear ordering and center-outward ordering. Contrary to the already existing methods, they can be powerful under a location-scale model and may be used in any dimension, including dimension one. Moreover, their null distributions are well known and already tabulated in the statistical literature. A simulation study has shown their good performance in the univariate and bivariate cases. The version of our test which uses Mardia's procedure was found particularly attractive.

## ACKNOWLEDGMENTS

Most of this research was carried out when the author was visiting the Centre for Mathematics and its Applications at the Australian National University, Canberra. The author is grateful to the Swiss National Science Foundation (Grant 81NE-54413) for its financial support, as well as to Peter Hall, John Braun and two referees for their careful reading of the manuscript and their valuable comments which led to an improvement of the paper.

## REFERENCES

1. B. M. Brown and T. P. Hettmansperger, Affine invariant rank methods in the bivariate location model, *J. Roy. Statist. Soc. Ser. B* **49** (1987), 301–310.
2. S. K. Chatterjee and P. K. Sen, Nonparametric tests for the bivariate two sample location problem, *Calcutta Statist. Assoc. Bull.* **13** (1964), 18–58.
3. J. D. Gibbons, "Nonparametric Methods for Quantitative Analysis," Holt, Rinehart and Winston, New York, 1976.
4. D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski, (Eds.), "A Handbook of Small Data Sets," Chapman and Hall, London, 1994.
5. T. P. Hettmansperger, "Statistical Inference Based on Ranks," Wiley, New York, 1984.
6. R. Y. Liu, On a notion of data depth based on random simplices, *Ann. Statist.* **18** (1990), 405–414.
7. R. Y. Liu and K. Singh, A quality index based on data depth and multivariate rank tests, *J. Amer. Statist. Assoc.* **88** (1993), 252–260.

8. R. Y. Liu, Control charts for multivariate processes, *J. Amer. Statist. Assoc.* **90** (1995), 1380–1387.
9. K. V. Mardia, A non-parametric test for the bivariate two-sample location problem, *J. Roy. Statist. Soc. Ser. B* **29** (1967), 320–342.
10. G. M. Morant, A first study of the Tibetan skull, *Biometrika* **14** (1923), 193–260.
11. H. Oja, Descriptive statistics for multivariate distribution, *Statist. Probab. Lett.* **1** (1983), 327–333.
12. D. Peters and R. H. Randles, A bivariate signed rank test for the two-sample location problem, *J. Roy. Statist. Soc. Ser. B* **53** (1991), 493–504.
13. M. L. Puri and P. K. Sen, On a class of multivariate multisample rank-order tests, *Sankhyā Ser. A* **28** (1966), 353–375.
14. M. L. Puri and P. K. Sen, “Nonparametric Methods in Multivariate Analysis,” Wiley, New York, 1971.
15. R. H. Randles, A two sample extension of the multivariate interdirection sign test, in “ $L_1$ -Statistical Analysis and Related Methods” (Y. Dodge, Ed.), pp. 295–302, North-Holland, Amsterdam, 1992.
16. R. H. Randles and D. Peters, Multivariate rank tests for the two-sample location problem, *Commun. Statist. Theory Methods* **19** (1992), 4225–4238.
17. R. Tamura, Multivariate nonparametric several-sample tests, *Ann. Math. Statist.* **37** (1966), 611–618.
18. J. W. Tukey, Mathematics and picturing data, in “Proc. of the 1974 International Congress of Mathematicians, Vancouver,” Vol. 2, pp. 523–531, 1975.
19. I. Vincze, On two-sample tests based on order statistics, in “Proc. 4th Berkeley Symp. Math. Statist. and Probab.,” Vol. 1, pp. 695–705, 1961.
20. A. Wald and J. Wolfowitz, Statistical tests based on permutation of observations, *Ann. Math. Statist.* **15** (1944), 358–372.