Statistical Science
2004, Vol. 19, No. 4, 686–696
DOI 10.1214/088342304000000594
© Institute of Mathematical Statistics, 2004

New Nonparametric Tests of Multivariate Locations and Scales Using Data Depth

Jun Li and Regina Y. Liu

Abstract. Multivariate statistics plays a role of ever increasing importance in the modern era of information technology. Using the center-outward ranking induced by the notion of data depth, we describe several nonparametric tests of location and scale differences for multivariate distributions. The tests for location differences are derived from graphs in the so-called DD plots (depth vs. depth plots) and are implemented through the idea of permutation tests. The proposed test statistics are scale-standardized measures for the location difference and they can be carried out without estimating the scale or variance of the underlying distributions. The test for scale differences introduced in Liu and Singh (2003) is a natural multivariate rank test derived from the center-outward depth ranking and it extends the Wilcoxon rank-sum test to the testing of multivariate scale. We discuss the properties of these tests, and provide simulation results as well as a comparison study under normality. Finally, we apply the tests to compare airlines' performances in the context of aviation safety evaluations.

Key words and phrases: Data depth, DD plot, multivariate location difference, multivariate scale difference, permutation test, multivariate rank test, Wilcoxon rank-sum test.

1. INTRODUCTION

Recent advances in computer technology have facilitated the collection of massive multivariate data in many industries. The demand for effective multivariate analyses has never been greater. Most existing multivariate analysis still relies on the assumption of normality, which is often difficult to justify in practice. Using the center-outward ranking induced by the notion of data depth, we describe several nonparametric tests for location and scale differences in multivariate distributions. These tests are completely nonparametric and have broader applicability than the existing tests. They can even be moment-free and thus valid for testing parameters which are not defined using moments, such as the locations of Cauchy distributions.

For testing location differences, our test statistics are constructed from the graphs observed in the DD plots (depth vs. depth plots). The DD plot, introduced by Liu, Parelius and Singh (1999), is a two-dimensional graph which can serve as a simple diagnostic tool for visual comparisons of two samples of any dimension. Different distributional differences, such as location, scale, skewness or kurtosis differences, are associated with different graphical patterns in DD plots. In this paper, we focus on the pattern associated with the location difference in the DD plots and we propose two tests for testing possible location differences between two samples. Since the data depth is affine-invariant, it provides a scale-standardized measure of the position of any data point relative to the center of the distribution. This property allows us to view our depth-based test statistics as scale-standardized measures for the location difference. Consequently, the tests can be carried out without the difficulty of estimating the variance of the null sampling distributions. Instead, we derive the decision rules by obtaining p-values using the idea of permutation.

For testing multivariate scales, we review a new rank test proposed by Liu and Singh (2003). This rank

Jun Li is a Ph.D. candidate and Regina Y. Liu is Professor, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA (e-mail: junli@stat.rutgers.edu, rliu@stat.rutgers.edu).

test is derived from the center-outward ranking induced by data depth assigned to the combined sample. It is constructed in a way similar to the Wilcoxon rank-sum test and can be carried out using either the Wilcoxon rank-sum table or simulations. It is a completely nonparametric test for testing the scale expansion or contraction. It includes the Ansari–Bradley and the Siegel–Tukey tests as special cases for testing the equality of variances in the univariate setting.

The tests discussed in this paper are appealing since they are guided visually by the DD plot and admit a full theoretical justification. Most importantly, they are easy to implement regardless of the dimensionality of the data. For all the proposed tests, we present several simulation studies, including power comparisons between our proposed nonparametric tests and some existing parametric tests. The performance of our tests is generally comparable to the parametric tests under the multivariate normal setting, with only minor loss of efficiency. However, our tests dramatically outperform the parametric tests under the multivariate Cauchy setting. This is in part because our tests are moment-free and thus are more suitable for dealing with parameters not derived from moments, such as those in the case of Cauchy distributions.

The rest of the paper is organized as follows. In Section 2 we give a brief review of the notion of data depth, depth-induced multivariate rankings and DD plots. In Section 3 we describe two tests for location differences. These tests are referred to as the T-based test and the M-based test, and they are derived from the graphs in DD plots which reflect location changes between two distributions. We then carry out the tests by using Fisher's permutation test to determine p-values. We justify the validity of the tests by showing that the distribution of the obtained p-values follows approximately the uniform distribution U[0, 1] under the null hypothesis and also that it decreases to 0 under the alternative hypothesis. Results from several simulation studies are presented. Under the normality assumption, our tests are comparable to the Hotelling T^2 test. We devote Section 4 to the scale comparison of two multivariate distributions, which includes a depth-induced multivariate rank test introduced in Liu and Singh (2003) and a graphical display of scale curve (Liu, Parelius and Singh, 1999). In Section 5 we apply our tests to compare the performance of 10 airlines using the airline performance data collected by the Federal Aviation Administration (FAA). Finally, we provide some concluding remarks in Section 6.

2. NOTATION AND BACKGROUND MATERIAL

We begin with a brief description of the notion of data depth and its properties.

2.1 Data Depth and Center-Outward Ranking of Multivariate Data

A *data depth* is a way to measure the "depth" or "outlyingness" of a given point with respect to a multivariate data cloud or its underlying distribution. It gives rise to a natural *center-outward ordering* of the sample points in a multivariate sample. This ordering gives rise to new and easy ways to quantify the many complex multivariate features of the underlying distribution, including *location*, *quantiles*, *scale*, *skewness* and *kurtosis*. This ordering in effect turn provides a new nonparametric multivariate inference scheme (cf. Liu, Parelius and Singh, 1999), which includes several graphical methods for comparing multivariate distributions or samples. Some of the methods in Liu, Parelius and Singh (1999) motivated the comparison methods presented in this paper. Before we show how the depth and its ordering can be used to construct multivariate nonparametric tests, we first use the simplicial depth proposed by Liu (1990) as an example of depth measure (1) to describe the general concept of data depth and its corresponding center-outward ordering, and (2) to introduce necessary notation.

The word "depth" was first used by Tukey (1975) to picture data, and the far reaching ramifications of depth in ordering and analyzing multivariate data was observed and elaborated by Liu (1990), Donoho and Gasko (1992), Liu, Parelius and Singh (1999) and others. Many existing notions of data depth were listed in Liu, Parelius and Singh (1999).

Let $\{Y_1, \ldots, Y_m\}$ be a random sample from the distribution $G(\cdot)$ in $\mathbb{R}^k$, $k \geq 1$. We begin with the bivariate setting $k = 2$. Let $\triangle(a, b, c)$ denote a triangle with vertices a, b and c. Let $I(\cdot)$ be the indicator function, that is, $I(A) = 1$ if A occurs and $I(A) = 0$ otherwise. Given the sample $\{Y_1, \ldots, Y_m\}$, the sample simplicial depth of $y \in R^2$ is defined as

$$D_{G_m}(y) = \binom{m}{3}^{-1} \sum_{(*)} I\big(y \in \triangle(Y_{i_1}, Y_{i_2}, Y_{i_3})\big), \tag{2.1}$$

which is the fraction of the triangles generated from the sample that contain the point y. Here $(*)$ runs over all possible triplets of $\{Y_1, \ldots, Y_m\}$. A large value of $D_{G_m}(y)$ indicates that y is contained in many triangles generated from the sample, and thus it lies deep within the data cloud. On the other hand, a small

$D_{G_m}(y)$ indicates an outlying position of y. Thus D_{G_m} is a measure of the depth (or outlyingness) of y w.r.t. the data cloud $\{Y_1, \ldots, Y_m\}$.

The above simplicial depth can be generalized to any dimension k as

$$D_{G_m}(y) = \binom{m}{k+1}^{-1} \sum_{(*)} I\big(y \in s[Y_{i_1}, \ldots, Y_{i_{k+1}}]\big), \tag{2.2}$$

where $(*)$ runs over all possible subsets of $\{Y_1, \ldots, Y_m\}$ of size $k+1$. Here $s[Y_{i_1}, \ldots, Y_{i_{k+1}}]$ is the closed simplex whose vertices are $\{Y_{i_1}, \ldots, Y_{i_{k+1}}\}$, that is, the smallest convex set determined by $\{Y_{i_1}, \ldots, Y_{i_{k+1}}\}$. When the distribution G is known, then the simplicial depth of y w.r.t. to G is defined as

$$D_G(y) = P_G\{y \in s[Y_1, \ldots, Y_{k+1}]\}, \tag{2.3}$$

where $Y_1, \ldots, Y_{k+1}$ are $k+1$ random observations from G. Depth $D_G(y)$ measures how deep y is w.r.t. G. A fuller motivation together with the basic properties of $D_G(\cdot)$ can be found in Liu (1990). In particular, it is shown that $D_G(\cdot)$ is affine-invariant and that $D_{G_m}(\cdot)$ converges uniformly and strongly to $D_G(\cdot)$. The affine invariance ensures that our proposed inference methods are coordinate-free, and the convergence of D_{G_m} to D_G allows us to approximate $D_G(\cdot)$ by $D_{G_m}(\cdot)$ when G is unknown.

For the given sample $\{Y_1, Y_2, \ldots, Y_m\}$, we calculate all the depth values $D_{G_m}(Y_i)$ and then order the Y_i's according to their ascending depth values. Denoting by $Y_{[j]}$ the sample point associated with the jth smallest depth value, we obtain the sequence $\{Y_{[1]}, Y_{[2]}, \ldots, Y_{[m]}\}$ which is the depth order statistics of the Y_i's, where $Y_{[m]}$ is the *deepest* point and $Y_{[1]}$ is the most outlying point. Here, a smaller rank is associated with a more outlying position w.r.t. the underlying distribution G. Note that the order statistics derived from depth are different from the usual order statistics in the univariate case, since the latter are ordered from the smallest sample point to the largest, while the former starts from the *middle* sample point and moves outward in all directions. This property is illustrated in Figure 1, which shows the depth ordering of a random sample of 500 points drawn from a bivariate normal distribution. The plus (+) marks the deepest point, and the most inner convex hull encloses the deepest 20% of the sample points. The convex hull expands further to enclose the next deepest 20% by each expansion. Such nested convex hulls, determined by the decreasing depth value, also indicate that the depth ordering is from the center outward.

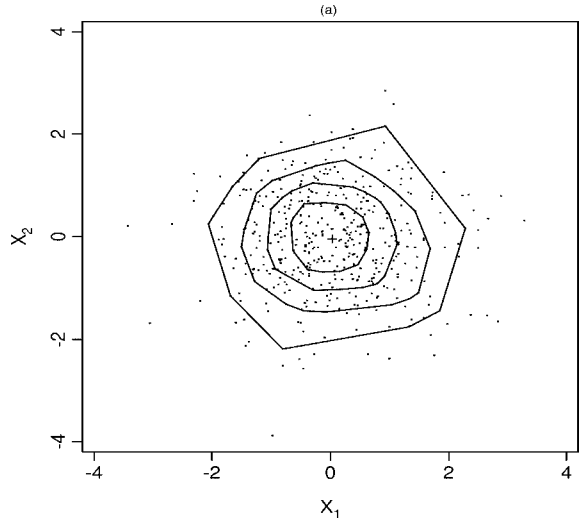

FIG. 1. *Depth contours for a bivariate normal sample.*

When the distribution G is known, $D_G(y)$ leads to an ordering of all points in $\mathbb{R}^k$ from the deepest point outward. The deepest point here is the maximizer of $D_G(\cdot)$ (or the average of the maximizers if there is more than one), which is denoted by $\boldsymbol{\mu}^*$. Clearly, $\boldsymbol{\mu}^*$ can be viewed as a location parameter of the distribution G, and it coincides with the mean (and the center of symmetry) if G is symmetric.

2.2 DD Plots for Graphical Comparisons of Multivariate Samples

Let $\{X_1, \ldots, X_n\}$ $(= \mathbf{X})$ and $\{Y_1, \ldots, Y_m\}$ $(= \mathbf{Y})$ be two random samples drawn, respectively, from F and G, where F and G are two continuous distributions in $\mathbb{R}^k$. Comparisons of the two samples can be conveniently studied in the framework of testing the null hypothesis

$$H_0 : F = G.$$

Depending on the specific difference we seek between F and G, we can choose a proper alternative hypothesis to carry out the test. The so-called *depth vs. depth plots* were proposed by Liu, Parelius and Singh (1999) for graphical comparisons of two multivariate samples. Specifically, the DD plot is the plot of $DD(F_n, G_m)$, where

$$DD(F_n, G_m) = \big\{\big(D_{F_n}(x), D_{G_m}(x)\big),\ x \in \{\mathbf{X} \cup \mathbf{Y}\}\big\}. \tag{2.4}$$

This is the empirical version of

$$DD(F, G) = \big\{\big(D_F(x), D_G(x)\big),\ \text{for all } x \in \mathbb{R}^k\big\}. \tag{2.5}$$

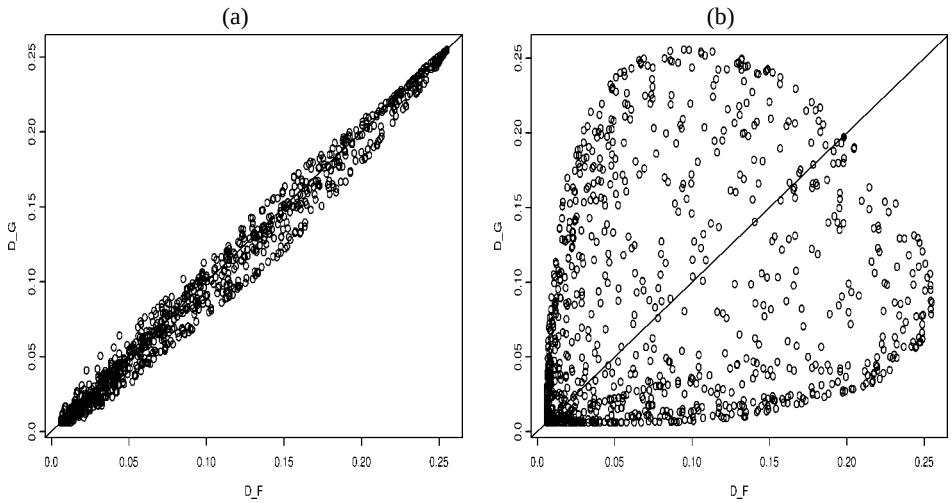

FIG. 2. *DD plots of* (a) *identical distributions and* (b) *location shift.*

Note that $DD(F, G)$ as well as $DD(F_n, G_m)$ are always subsets of $\mathbb{R}^2$ no matter how large is the dimension k of the data. The two-dimensional graphs of DD plots are easy to visualize and they turn out to be convenient tools for graphical comparison of multivariate samples. If $F = G$, then $D_F(x) = D_G(x)$ for all $x \in \mathbb{R}^k$, and thus the resulting $DD(F, G)$ is simply a line segment on the 45° line in the DD plot, from $(0, 0)$ to $(\max_t D_F(t), \max_t D_G(t))$. This is illustrated by the simulation result in Figure 2(a), which is the DD plot of two samples drawn from the bivariate normal distribution with mean $(0, 0)$. Deviations from the 45° line segment in DD plots would suggest that there are differences between the distributions F and G. As it turns out, each particular pattern of deviation from the diagonal line can be attributed to a specific type of difference between the two distributions. For example, as shown in Figure 2(b), in presence of a location shift in the two samples, the DD plot generally has a leaf-shaped figure, with the leaf stem anchoring at the lower left corner point $(0, 0)$ and the cusp lying on the diagonal line pointing toward the upper right corner. (The variations of the leaf shape reflect the magnitude of the location shift as well as the symmetry of the underlying distributions, as we discuss further in Section 3.) Figures 3(a) and (b) shows yet different patterns of DD plots (the half-moon pattern and the wedge-like pattern) that are indicative, respectively, of scale and skewness differences.

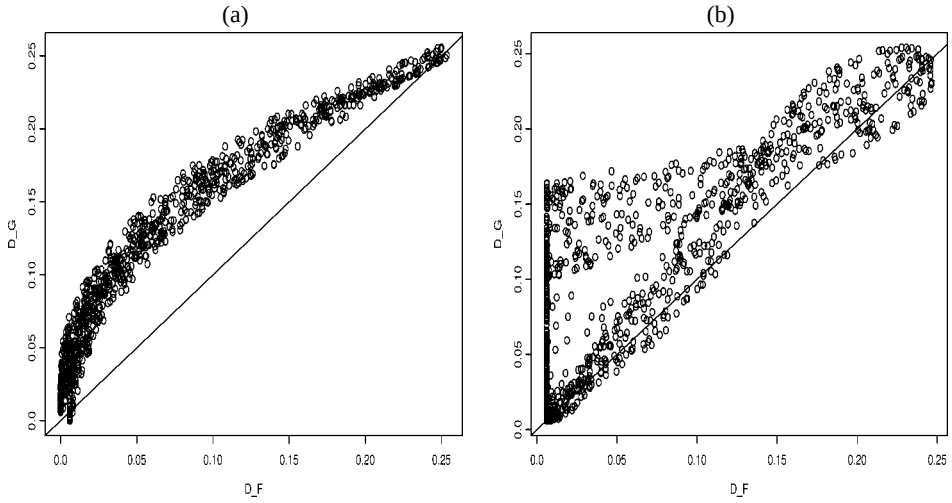

FIG. 3. *DD plots of* (a) *scale increase and* (b) *skewness difference.*

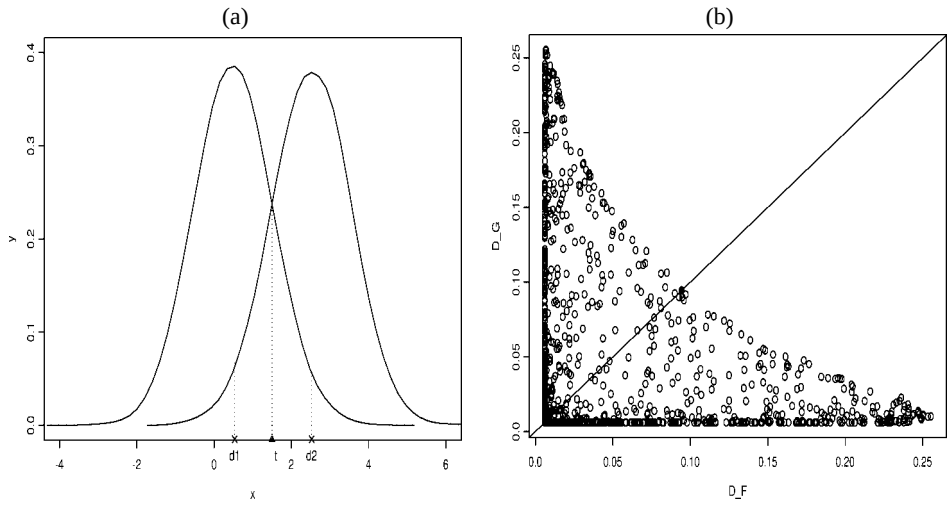

FIG. 4. (a) *Two distributions with a location shift.* (b) *DD plot of large location shift.*

3. TESTS OF LOCATION DIFFERENCES USING DD PLOTS

As described above, DD plots can serve as diagnostic tools for detecting visually the difference between two samples of any dimension. To make DD plots rigorous testing tools, we need to construct test statistics which can capture deviation patterns in DD plots and establish the null distributions for those statistics. In this section, we focus specifically on deriving test statistics from DD plots for testing location differences. Some brief comments on testing other distributional differences are made in the concluding remarks.

Recall that $\mathbf{X} \equiv \{X_1, \ldots, X_n\} \sim F$ and $\mathbf{Y} \equiv \{Y_1, \ldots, Y_m\} \sim G$ are two given samples in $\mathbb{R}^k$. For convenience, we assume that $n = m$, although the inference methods described in this paper remain valid otherwise. Assume that F and G are identical except for a possible location shift θ [i.e., $G(\cdot) = F(\cdot - \theta)$]. The hypotheses of interest are then

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_a : \theta \neq 0.$$

Note that F and G are not required to be symmetric. If they are, their deepest points (i.e., the location parameters) coincide with the centers of symmetry (as well as the means).

Under H_0, the DD plot $DD(F_n, G_m)$ should be clustered along the diagonal line, as seen in Figure 2(a). In case there is a location shift from F to G, the DD plot exhibits a leaf shape with its tip pulling away from the upper right corner point along the diagonal line toward the lower left corner point $(0, 0)$. Note that, using the simplicial depth, $(\max_t D_F(t), \max_t D_G(t)) = (2^{-k}, 2^{-k})$, the maximum DD value achievable by the deepest point under the null hypothesis. For example, when $k = 2$, $1/4$ is the achievable maximum for $D_F(\cdot)$ and $D_G(\cdot)$. The larger the location shift is, the closer the tip of the leaf is pulled diagonally downward to $(0, 0)$. Figure 4(a) illustrates this phenomenon in a simple univariate setting, using two symmetric density functions with a location shift. The crossing of the two densities occurs at the point t. The cusp point of the leaf-shaped DD plot in Figure 2(b), marked by a solid dot, corresponds to $(D_F(t), D_G(t))$. If there are no location shifts, then the two densities coincide and t is the deepest point for both F and G. Hence, $D_F(t) = D_G(t) = 1/2$ and the cusp point hits exactly the right upper corner point. If the location shift widens, then the cusp point pulls further toward $(0, 0)$, as seen in the DD plots in Figures 2(b) and 4(b).

3.1 T-Based Test: Monitor Shrinking Cusp Point

The above observation suggests that the closer the cusp point is to $(0, 0)$, the more likely there is a location shift between the two underlying distributions. This suggests considering the distance between the cusp point and $(0, 0)$ as our testing statistic. Before we define this distance more precisely, we note that the sample versions of depth D_{F_n} and D_{G_n} are both discrete, and thus the cusp point of the DD plot may not fall exactly on the diagonal line, especially if $m \neq n$. Therefore, we first introduce the following notation to define the relative positions of any two points in $\mathbb{R}^2$ and to derive a convenient approximation of the cusp point.

For (a_1, b_1) and (a_2, b_2) in $\mathbb{R}^2$, we define

$$(a_1, b_1) \succeq (a_2, b_2) \quad \text{if } a_1 \geq a_2 \text{ and } b_1 \geq b_2,$$
$$(a_1, b_1) \prec (a_2, b_2) \quad \text{otherwise.}$$

Define the set

$$\mathbf{Q} \equiv \{Z \in \mathbf{X} \cup \mathbf{Y} : \text{there does not exist } W \in \mathbf{X} \cup \mathbf{Y} \text{ s.t. } (D_{F_n}(W), D_{G_n}(W)) \succeq (D_{F_n}(Z), D_{G_n}(Z))\}.$$

Then the cusp point is identified or approximated by the point $(D_{F_n}(Z_c), D_{G_n}(Z_c))$ that satisfies $Z_c \in \mathbf{Q}$ and $|D_{F_n}(Z_c) - D_{G_n}(Z_c)| \leq |D_{F_n}(Z) - D_{G_n}(Z)|$ for all $Z \in \mathbf{Q}$. Let

$$T_n = (D_{F_n}(Z_c) + D_{G_n}(Z_c))/2. \tag{3.1}$$

Then the distance from the cusp point to $(0, 0)$ is approximately equal to $\sqrt{2}T_n$. This is equivalent to working with T_n. Intuitively, the larger the location shift between the two distributions, the smaller the T_n value and thus the stronger the evidence against H_0. To determine when T_n is small enough to reject H_0 decisively, we need to derive the null distribution of T_n. The derivation of this null distribution turns out to be quite demanding and we plan to carry it out as a separate project in the future. Alternatively, we propose here to use Fisher's permutation test to determine the following p-value and complete our test procedure. (The idea of using Fisher's permutation test to obtain the p-value of a test is well known; for reference see, e.g., Chapter 15 of Efron and Tibshirani, 1993.)

Let

$$p_n^T = P_{H_0}(T_n < T_{\text{obs}}), \tag{3.2}$$

where T_{obs} is the observed value of T_n based on the given sample $\mathbf{X} \cup \mathbf{Y}$. The p_n^T is also referred to as the achieved significance level.

REMARK 3.1. If the sample size is sufficiently large, the definition of T_n in (3.1) can be approximated by

$$T_n = \max_{Z \in \mathbf{X} \cup \mathbf{Y}} \{D_{F_n}(Z) : D_{F_n}(Z) = D_{G_n}(Z)\}. \tag{3.3}$$

REMARK 3.2. If the underlying distributions F and G are symmetric, the population version of T_n, denoted by T, is the depth (under either F or G) of the midpoint of the line segment that connects the two centers of symmetry. In the setting of Figure 4(a), $T = D_F(t)$.

Without the null distribution of T_n, we can proceed and use the permutation method to approximate the T-based p-value defined in (3.2). The procedure is as follows.

1. Permute the combined sample $\mathbf{X} \cup \mathbf{Y}$ B times. Here B is sufficiently large. For each permutation, we treat the first n elements as the X sample and the remaining elements as the Y sample. Denote the outcome of the ith permutation by $\mathbf{X}_i^* = \{X_{i1}^*, \ldots, X_{in}^*\}$ and $\mathbf{Y}_i^* = \{Y_{i1}^*, \ldots, Y_{in}^*\}$ for $i = 1, \ldots, B$.
2. Obtain the DD plot for each $\mathbf{X}_i^* \cup \mathbf{Y}_i^*$ and evaluate the corresponding T_n value [following (3.1) or (3.3)], which is then denoted by T_i^*, $i = 1, \ldots, B$.

The empirical distribution of $T_i^*, i = 1, \ldots, B$, can be used as an approximate of the null distribution of T_n. Consequently, under H_0, the p_n^T defined in (3.2) can be approximated by

$$p_{n,B}^T = \sum_{i=1}^{B} I_{\{T_i^* \leq T_{\text{obs}}\}}/B. \tag{3.4}$$

The above permutation procedure contains, in principle, all $n!$ permutations. If n is not too large or computational speed is not a concern, then we let $B = n!$.

In theory, for any testing procedure, a valid p-value follows a uniform distribution on $[0, 1]$, U$[0, 1]$, under H_0. It is clear that our p-value, $p_{n,B}^T$, is valid, since it is derived from the permutation test. Our simulation results also show that, under H_0, the histograms of $p_{n,B}^T$ are reasonably close to U$[0, 1]$.

We have shown that the T-based test described above works well under the null hypothesis and allows us to control the type I error using U$[0, 1]$. We now proceed to evaluate the power of this test under the alternative hypothesis. Ideally, the power of the T-based test grows (or, equivalently, $p_{n,B}^T$ decreases) as the location shift increases. To this end, we conduct some simulation experiments under both bivariate normal and exponential distributions. Figure 5 shows the histograms of $p_{n,B}^T$ for the bivariate normal case, where $F = \text{N}((0, 0), \mathbf{I})$ and $G = \text{N}((\mu, \mu), \mathbf{I})$ with (μ, μ) equal to $(0.1, 0.1)$, $(0.2, 0.2)$ and $(0.3, 0.3)$ for G, respectively, from top down. Clearly, as the location shift grows larger, the histograms from top down become more skewed to the right and the $p_{n,B}^T$ value leans more toward 0. This shows that the power of the T-based test grows as the location shift increases, which is a desirable property. Similar patterns of histograms of $p_{n,B}^T$ are observed for the bivariate exponential case, where the mean for F is $(1, 1)$ and is $(0.9, 0.9)$, $(0.75, 0.75)$ and $(0.5, 0.5)$ for G. In both settings, we let $n = 100$, $B = 500$ and use the simplicial depth to obtain DD plots. For each case, the experiment is repeated 100 times to obtain 100 corresponding $p_{n,B}^T$'s.

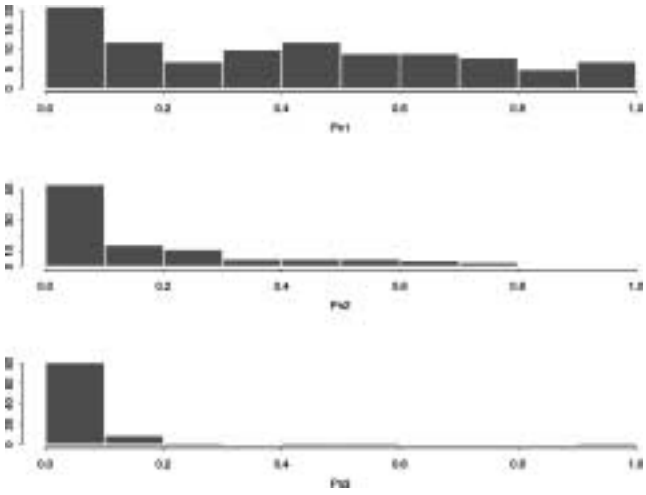

FIG. 5. *Histograms of $p^T_{n,B}$ under H_a, where $H_0 : (\mu, \mu) = (0, 0)$ and $H_a : (\mu, \mu) = (0.1, 0.1), (0.2, 0.2)$ and $(0.3, 0.3)$ (bivariate normal case).*

3.2 M-Based Test: Monitor the Maximum Depth Points

We next propose another test based on the DD plot for detecting a location change in two multivariate distributions. This test statistic is more in the form of a location estimator. In the context of data depth, the location parameter of a distribution is defined as the deepest point (see, e.g., Liu, Parelius and Singh, 1999). If the two distributions F and G are identical, they should have the same deepest point. On the other hand, if there is a location change, the deepest point of the distribution F would no longer be the deepest point of the distribution G and thus it attains a smaller depth value w.r.t. G. The larger the location change is, the smaller this depth becomes. This trend can also be observed from the DD plots in Figures 2(b) and 4(b). This observation motivates our second test below, in which the test statistic monitors directly the depth values of the deepest points of the underlying distributions. Let

$$M_n = \min\{D_{F_n}(Z_{G_n}), D_{G_n}(Z_{F_n})\}, \tag{3.5}$$

where Z_{G_n} and Z_{F_n} are the deepest points among $\mathbf{X} \cup \mathbf{Y}$ with respect to G_n and F_n, respectively.

REMARK 3.3. In theory, we may also consider other functions of the two depths such as the maximum in (3.5). However, we choose to work with the minimum of the two depths, because the minimum is more sensitive to the location change and it can achieve more power.

We now proceed and carry out the test by determining its achieved significance level, defined as

$$p^M_n = P_{H_0}(M_n < M_{\text{obs}}). \tag{3.6}$$

Again, we turn to Fisher's permutation test to estimate the p-value p^M_n. The procedure consists of the two steps outlined for the T-based test, except that each permutation replication is now used to evaluate the M_n as defined in (3.5). Denote by M^*_i the M_n value obtained in the ith permutation. The p-value p^M_n is then approximated by

$$p^M_{n,B} = \sum_{i=1}^{B} I_{\{M^*_i \le M_{\text{obs}}\}}/B. \tag{3.7}$$

Discussions of the validity and power of the T-based test in terms of the proposed p-value apply similarly to the proposed $p^M_{n,B}$ based on the M-based test described above. Again, the histograms of the 100 simulated $p^M_{n,B}$'s under the standard bivariate normal and exponential distributions appear close to U[0, 1] under H_0, and they skew more to the right as the location shift widens, as observed in Figure 5.

3.3 Power Comparisons: T and M Tests versus Hotelling T^2

Since both T- and M-based tests are completely nonparametric, it should be interesting to compare them to known parametric tests to see their loss of efficiency, if any. The first comparison is with the Hotelling (1947) T^2 test under the normality assumption where $F = \mathrm{N}((0, 0), \mathbf{I})$ and $G = \mathrm{N}((\mu, \mu), \mathbf{I})$. Each test is repeated 1000 times and the simplicial depth is used to compute needed depth values. The power of the T-based (or M-based) tests is estimated by the proportion of the simulated $p^T_{n,B}$'s (or $p^M_{n,B}$'s) which are less than the nominal type I error $\alpha = 0.05$. Table 1 lists the estimated power for $\mu = 0, 0.1, 0.2, 0.3, 0.4, 0.5$. The results clearly show that both T- and M-based tests perform comparably to the Hotelling T^2 test, even though the former are completely nonparametric and do not utilize the normality assumption.

TABLE 1
Power comparison under bivariate normal distributions

μ	0	0.1	0.2	0.3	0.4	0.5
T-based	0.054	0.109	0.373	0.714	0.933	0.993
M-based	0.060	0.113	0.386	0.710	0.921	0.988
Hotelling T^2	0.059	0.124	0.410	0.765	0.953	0.995

TABLE 2
Power comparison under bivariate Cauchy distributions

μ	0	0.1	0.2	0.3	0.4	0.5
T-based	0.052	0.060	0.114	0.154	0.214	0.350
M-based	0.046	0.072	0.118	0.214	0.324	0.522
Hotelling T^2	0.020	0.010	0.020	0.034	0.022	0.052

We also conducted the same comparison study for the bivariate Cauchy distributions with the location parameter (μ, μ). Clearly, both T- and M-based tests outperform the Hotelling T^2. This can be attributed to the fact that the first two tests using the simplicial depth are moment-free approaches and thus more suitable for testing location parameters not derived from moments, such as in the case of Cauchy distributions. The results in Table 2 seem to suggest also that the M-based test is more powerful than the T-based test in the Cauchy case. We plan to investigate further the difference between the T- and M-based tests, including their robustness properties as well as their capability to cope with asymmetric underlying distributions.

4. RANK TESTS FOR SCALE EXPANSION OR CONTRACTION

Let, again, $\mathbf{X} \equiv \{X_1, \ldots, X_n\} \sim F$ and $\mathbf{Y} \equiv \{Y_1, \ldots, Y_m\} \sim G$ be two given samples in $\mathbb{R}^k$. Assume that F and G are identical except for a possible scale difference. For simplicity, assume that we are interested in testing if G has a larger scale in the sense that the scale of G is an expansion of that of F. In other words, the hypotheses of interest are

$$\begin{aligned} &H_0 : F \text{ and } G \text{ have the same scale} \\ &H_a : G \text{ has a larger scale.} \end{aligned} \tag{4.1}$$

Combine the two samples, that is, let $\mathbf{W} \equiv \{W_1, W_2, \ldots, W_{n+m}\} \equiv \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$. If G has a larger scale, then the X_i's are more likely to cluster tightly around the center of the combined sample, while the Y_i's are more likely to scatter at outlying positions. This outlyingness can be easily captured by data depth. Following this observation, Liu and Singh (2003) developed depth-induced rank tests to compare scales among two or multiple multivariate samples. In this paper we provide a brief review of their rank test for two samples. Detailed discussions and justifications can be found in Liu and Singh (2003).

4.1 Larger Scale—More Outlying Data—Smaller Ranks

Using any measure of depth, we can compute the depth values of the points in the combined sample $\mathbf{W}$. We then assign ranks to the combined sample $\mathbf{W}$ according to the ascending depth values, namely, lower ranks to the points with lower depth values. Specifically, we let $r(Y_i)$ be the center-outward rank of Y_i within the combined sample, that is,

$$\begin{aligned} r(Y_i) = \#\{W_j \in \mathbf{W} : D_{n+m}(W_j) \le D_{n+m}(Y_i), \\ j = 1, 2, \ldots, n+m\}, \end{aligned} \tag{4.2}$$

and we let the sum of the ranks for the sample $\mathbf{Y}$ be

$$R(\mathbf{Y}) = \sum_{i=1}^{m} r(Y_i). \tag{4.3}$$

Here, $D_{n+m}(\bullet)$ is the sample depth value of $\bullet$ measured w.r.t. $\{W_1, W_2, \ldots, W_{n+m}\}$. Under H_0, if there are no ties, $\{r(Y_1), \ldots, r(Y_m)\}$ can be viewed as a random sample of size m drawn without replacement from the set $\{1, \ldots, n+m\}$. If H_a is true, then the Y_i's tend to be more outlying, and thus assume smaller depth values and thus smaller ranks. In other words, we should reject H_0 if the rank sum $R(\mathbf{Y})$ is too small. The critical values for carrying out this test can be implemented using the Wilcoxon rank-sum procedure as if one is testing a negative location shift in the univariate setting. For a review of the Wilcoxon rank-sum test and its tabulated distributions for different sample size combinations, see, for example, Hettmansperger (1984). When n and m are sufficiently large, following the large-sample approximation, we can reject H_0 if $R^* \le z_\alpha$ for an α-level test. Here

$$R^* = \frac{R(\mathbf{Y}) - \{m(n+m+1)/2\}}{\{nm(n+m+1)/12\}^{1/2}}. \tag{4.4}$$

The depth ranking of sample points, due to its center-outward nature, often leads to ties, especially in high dimension cases. To use the tables provided for the Wilcoxon rank-sum test, we may consider the random tie-breaking scheme. However, we can actually carry out the test and obtain its *exact* p-value without breaking ties by the following approach. Since powerful computing facilities are easily available nowadays, we can use computers to obtain the exact distributions for the observed ranks, with or without ties. Specifically, we permute all the observed ranks (possibly including ties), calculate the sum of the first m ranks in each permutation, and finally tabulate such rank sums and their

corresponding frequencies in the total number of permutations. This distribution allows us to determine the exact p-value of our test, which is simply the proportion of the rank sums which are less than or equal to the observed rank sum in (4.3). As an illustrative example, we assume that $n = m = 2$ and that the ranks for the combined sample turn out to be $\{1, 2, 2, 4\}$ with a tie. The sampling distribution of the rank sum $R \equiv R(\mathbf{Y})$ is

$$\begin{aligned} &P(R=3) = 8/24, \quad P(R=4) = 4/24, \\ &P(R=5) = 4/24, \quad P(R=6) = 8/24. \end{aligned} \tag{4.5}$$

Therefore, if the observed rank sum is 4, then the p-value is $P(R \le 4) = 0.5$. For large samples, the distribution of the rank sum can be approximated by considering large enough numbers of permutations. In Table 3, we present some simulation results to examine the power of the rank test. Here the samples are from three bivariate distributions: Cauchy, normal and exponential, each with the component variance σ^2. We assume $n = m$, and consider $n = 20$ and $n = 30$. In each case 5000 random permutations of the observed ranks were used to approximate the sampling null distribution, and the rank test at significance level 0.05 was repeated 1000 times.

The results in Table 3 show that the power achieved by the rank test for scale expansions is quite respectable, especially in the nonnormal cases. A power comparison between the above rank test and a χ^2 test under the normality assumption can be found in Liu and Singh (2003). The results there show some minor loss of efficiency of the rank test in the normal case. Liu and Singh (2003) also discussed in detail the properties of this rank test as well as several approaches for dealing with large numbers of ties in the depth ranking. Moreover, they also generalized the rank test to the case of multiple samples.

Note that the rank test described above can be viewed as the multivariate generalization of Ansari–Bradley and Siegel–Tukey tests for testing the equality of variance in the univariate setting. Both tests try to assign smaller ranks to the data points which are more outlying toward two tails, although the Siegel–Tukey test avoids ties by alternating ranks.

TABLE 3
Simulated power of the rank test for scale expansions ($\alpha = 0.05$)

	$n = 30$			$n = 20$		
σ	**Cauchy**	**Normal**	**Exp**	**Cauchy**	**Normal**	**Exp**
1–1	0.056	0.044	0.049	0.054	0.051	0.043
1–1.2	0.345	0.325	0.218	0.261	0.242	0.188
1–2	0.996	0.994	0.940	0.966	0.940	0.813

If we are interested in testing whether or not G has a smaller (contracted) scale, then we should reject the null when the rank sum is too large.

The rank test above is easily implementable and is completely nonparametric. Its p-value yields a decisive decision rule. The test result can be independently verified visually by two graphical tools: One is the DD plot [see Figure 3(a) and the discussion in Section 2.2]; the other is the scale curve introduced by Liu, Parelius and Singh (1999). The sample scale curve derived from a sample of size n is defined as

$$S_n(p) = \text{volume}\,\{C_{n,p}\} \quad \text{for } 0 \le p \le 1. \tag{4.6}$$

Here $C_{n,p}$ is the convex hull that contains the $\lceil np \rceil$ deepest points. Roughly speaking, the scale curve measures the volume expansion of the nested depth contours, as seen in Figure 1, as the contours grow to enclose more probability mass. This plot of $S_n(p)$ versus p shows the scale of the distribution as a simple curve in the plane, which is easily visualized and interpreted. When comparing the scales of two samples, if one scale curve is consistently above the other, then the sample with the higher scale curve is more spread out and thus has a larger or expanded scale.

5. APPLICATION TO AIRLINE PERFORMANCE DATA

We apply all tests described so far to an analysis of an airline performance data set collected by the FAA. It consists of several monthly performance measures of the top 10 air carriers from July 1993 to May 1998. The performance measures include the fractions of nonconformity in airworthiness and operation surveillance. A small nonconformity fraction is a desirable feature. Several depth-induced multivariate control charts (Liu, 1995; Cheng, Liu and Luxhøj, 2000) have been used to monitor and compare the performances of all 10 airlines. For illustration, the T- and M-based tests are used to determine whether there is a significant difference in location (referred to as expected target performance in the aviation safety domain) in the distributions that underlie two air carriers. In comparing air carriers 1 and 4, their scatter plots in Figure 6(a) show a clear location shift to the upper right in carrier 4. The deepest point of carrier 4, marked by a solid triangle, is more to the upper right than that of carrier 1,

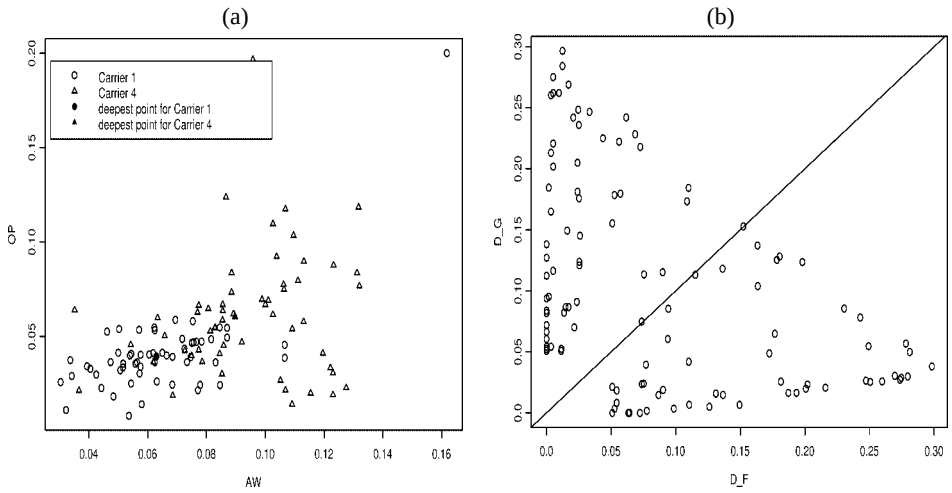

FIG. 6. (a) *Scatter plot and* (b) *DD plot for carriers* 1 *and* 4.

marked by •. The DD plot for the two carriers in Figure 6(b) has the cusp point pulled down toward (0, 0) to the midrange of the plot and clearly indicates a location difference in the two distributions. Using both T- and M-based tests, we found the approximated p-values to be nearly 0, which confirms a significant location shift in the two distributions.

In judging airline performance, in addition to examining the expected target performance (i.e., the location of the distribution) of the airlines, the stability of the performance within the airlines is also a major concern. This measure of stability is simply the measure of scale or variation of the performance distribution. Thus, comparing performance stability amounts to comparing the scales of distributions. Larger or more expanded scales mean less stable performance. We proceed and compare the scales of carriers 1 and 4. The p-value is 0.00038 using the test statistic in (4.3), which clearly supports the conclusion that carrier 4 has a larger scale than carrier 1. In other words, the performances of carrier 4 are more scattered and hence less stable. This same conclusion can also be reached by examining the two graphs in Figure 7. Figure 7(a) is the DD plot of carriers 1 and 4 after centering the data respectively at their deepest points, removing the effect of location difference. It shows a pattern which combines Figure 3(a) and (b). This suggests that there are both scale and skewness differences between the two carriers. Figure 7(b) displays the scale curves, as defined in (4.6), of four carriers. Obviously, the scale curve of carrier 4 lies consistently above all others, including that of carrier 1. The findings are also sup-

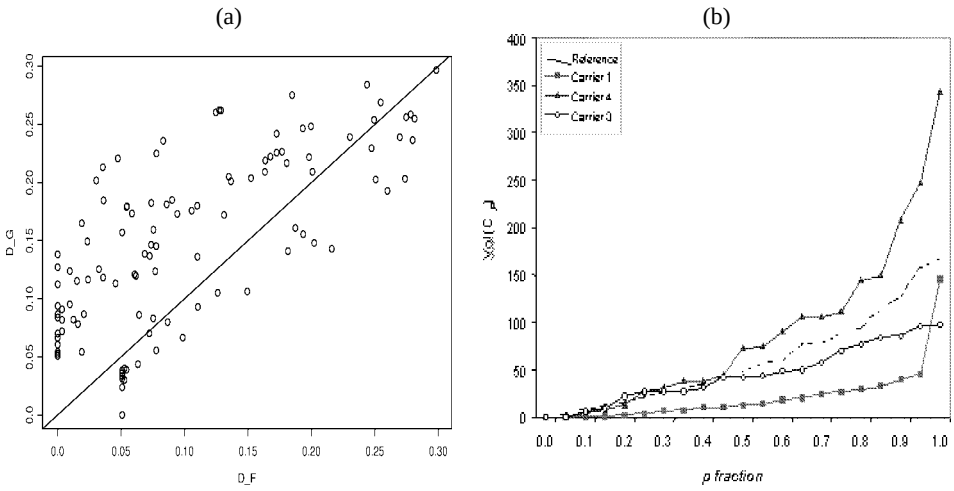

FIG. 7. (a) *DD plot for carriers* 1 *and* 4 *after centering.* (b) *Scale curves for air carriers.*

ported by the scatter plots in Figure 6(a), which show more scattered data for carrier 4. In summary, the performance of carrier 4 is inferior to that of carrier 1, in that carrier 4 has significantly higher target nonconformity ratios and it is also much less stable overall. Possible causes should be identified and corrective measures should be taken.

6. CONCLUDING REMARKS

Although our illustrative examples are in $\mathbb{R}^2$, all tests discussed in this paper apply to any dimension.

The DD plot of the rank test in Section 3 can be constructed using any notion of data depth which is affine-invariant. Some notions of depth may be more suitable than others in capturing a certain feature of a distribution. For example, if the underlying distribution is close to elliptical, then it is more efficient to use the Mahalanobis depth. Otherwise, more geometric depths such as the simplicial depth or the half-space depth (Tukey, 1975) may be more desirable since they do not require specific distributional structures or moment conditions. Details on some of these conditions for different depths can be found in Liu and Singh (1993) and Zuo and Serfling (2000). Note that it can be shown that the M-based test using Mahalanobis (1936) depth is asymptotically equivalent to the Hotelling T^2 test when comparing elliptical distributions. In other cases, the M-based test is more robust.

Concerning the issue of computational feasibility in computing depth, although the exact sample simplicial depth value in any dimension can be computed by solving a system of linear equations, more efficient algorithms are desirable. Rousseeuw and Ruts (1996) provided an efficient algorithm for computing both the simplicial and the half-space depths in $\mathbb{R}^2$. Developing efficient algorithms in the case of higher dimensions has recently generated much interest in computational geometry. It is reasonable to expect rapid progress in this direction.

Some depth rank tests have been proposed by Liu and Singh (1993) for testing simultaneously location and scale changes. It may be worthwhile to compare these rank tests separately with the T-based and M-based tests for testing location changes, and with the rank test described in (4.3) for testing scale changes.

Several graphical diagnostic tools stemming from DD plots for the two-sample problem have been proposed by Hettmansperger, Oja and Visuri (1999) and Liu, Parelius and Singh (1999). Their associated inferences need to be developed to make the graphical tools rigorous tests. Combining proper statistics derived from graphical tools with the permutation test idea may prove to be a helpful step in developing these tests.

ACKNOWLEDGMENTS

This research is supported in part by grants from the National Science Foundation, the National Security Agency and the Federal Aviation Administration.

REFERENCES

CHENG, A., LIU, R. and LUXHØJ, J. (2000). Monitoring multivariate aviation safety data by data depth: Control charts and threshold systems. *IIE Trans. Operations Engineering* **32** 861–872.

DONOHO, D. and GASKO, M. (1992). Breakdown properties of location estimates based on half-space depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

HETTMANSPERGER, T. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

HETTMANSPERGER, T., OJA, H. and VISURI, S. (1999). Discussion of "Multivariate analysis by data depth: Descriptive statistics, graphics and inference," by R. Liu, J. Parelius and K. Singh. *Ann. Statist.* **27** 845–854.

HOTELLING, H. (1947). Multivariate quality control: Illustrated by the air testing of sample bomb sight. In *Selected Techniques of Statistical Analysis for Scientific and Industrial Research, and Production and Management Engineering* (C. Eisenhart, M. Hastay and W. Wallis, eds.) 111–184. McGraw–Hill, New York.

LIU, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414.

LIU, R. (1995). Control charts for multivariate processes. *J. Amer. Statist. Assoc.* **90** 1380–1387.

LIU, R., PARELIUS, J. and SINGH, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Ann. Statist.* **27** 783–858.

LIU, R. and SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* **88** 252–260.

LIU, R. and SINGH, K. (2003). Rank tests for comparing multivariate scale using data depth: Testing for expansion or contraction. Unpublished manuscript.

LIU, R., SINGH, K. and TENG, J. (2004). DDMA-charts: Nonparametric multivariate moving average control charts based on data depth. *Allg. Stat. Arch.* **88** 235–258.

MAHALANOBIS, P. (1936). On the generalized distance in statistics. *Proc. Nat. Acad. Sci. India* **12** 49–55.

MOSLER, K. (2002). *Multivariate Dispersion, Central Regions and Depth*: *The Lift Zonoid Approach*. *Lecture Notes in Statist.* **165**. Springer, New York.

ROUSSEEUW, P. and RUTS, I. (1996). Algorithm AS 307: Bivariate location depth. *Appl. Statist.* **45** 516–526.

TUKEY, J. (1975). Mathematics and the picturing of data. *Proc. International Congress of Mathematicians* **2** 523–531. Canadian Math. Congress, Montreal.

ZUO, Y. and SERFLING, R. (2000). Structural properties and convergence results for contours of sample statistical depth functions. *Ann. Statist.* **28** 483–499.