

Multivariate quantile-quantile plots and related tests using spatial quantiles

Subhra Sankar Dhar¹, Probal Chaudhuri¹ and Biman Chakraborty²

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Calcutta^{1,1} and School of Mathematics, University of Birmingham²

e mails: dsubhra@gmail.com¹, probal@isical.ac.in¹, B.Chakraborty@bham.ac.uk²

Abstract

The univariate quantile-quantile (Q-Q) plot is a well-known graphical tool for examining whether two data sets are generated from the same distribution or not. It is also used to determine how well a specified probability distribution fits a given sample. In this article, we consider an extension of Q-Q plot for multivariate data based on spatial quantiles introduced and studied by Chaudhuri (1996) and Koltchinskii (1997). The usefulness of the proposed graphical tool is illustrated on different real and simulated data some of which are fairly high dimensional. We also propose some statistical tests for distributions of multivariate samples based on spatial quantiles and study their performance compared to some other tests available in the existing literature.

Keywords and phrases: Characterization of distributions, consistency of tests, levels and powers of tests, one-sample and two-sample problems, quantile difference plots, tests for distributions.

1 Introduction

Quantile-quantile (Q-Q) plot is a diagnostic tool, which is widely used to assess the distributional similarities and differences between two independent univariate samples. It is also a popular device for checking the appropriateness of a specified probability distribution for a given univariate data. There are a few generalizations of Q-Q plot for bivariate and multivariate samples that have been proposed in the literature. For bivariate samples, Marden (1998)

proposed a version of Q-Q plot, which is based on drawing arrows from the spatial quantiles (see, e.g., Breckling and Chambers (1988), Chaudhuri (1996) and Koltchinskii (1997)) in one sample to the corresponding spatial quantiles in another sample in a two-sample problem (or to the corresponding spatial quantiles of a specified probability distribution in a one-sample problem). If the arrows turn out to be small in all directions, it is an indication that the two independent samples have similar distributions (or, in the case of a one-sample problem, it indicates that the given sample does not deviate much from the specified probability distribution). Marden (1998) also mentioned a test based on arrow lengths.

Friedman and Rafsky (1979, 1981) proposed a different procedure for distributional comparison of two multivariate samples. Their methodology is based on the idea of minimal spanning tree. They used the minimal spanning tree fitted to the pooled sample and ranked the data points according to their positions on the tree. Then the Wald-Wolfowitz-Smirnov run test was used on the runs of the occurrences of data points, which corresponds to one sample, in that ranking of the data points in the pooled sample.

Easton and McCulloch (1990) proposed a multivariate Q-Q plot based on the idea of matching a multivariate data set with a multivariate reference sample using an appropriate permutation of the data. Their procedure is based on the permutation of the data that leads to minimum sum of Euclidean distances between the matching data points from the two given samples. They assumed equality of the sizes for the two samples, and in order to assess how well a specified probability distribution fits a given multivariate sample, they used samples simulated from the specified distribution. In order to solve the optimization problem involved in matching the two samples, they used an iterative algorithm.

Liu, Parelius and Singh (1998) proposed an alternative visualization methodology, namely, DD-plot, for comparing two multivariate data sets based on the concept of data depth. They computed the depth values of each data point in the combined sample once with respect to the first sample and then again with respect to the second sample. Then they plotted those pairs of depth values for each data point in a two-dimensional plot.

It will be appropriate to note here none of the graphical tools developed by Marden (1998), Friedman and Rafsky (1981) and Liu et al. (1998) coincide with the usual univariate q-q plot when they are applied to univariate data even though the univariate versions of those graphical tools are closely related to the usual q-q plot. Hence, strictly speaking, none of them can be taken as a natural multivariate extension of q-q plot. Further, unlike Friedman and Rafsky (1979, 1981) and Marden (1998), neither Liu et al. (1998) nor Easton and McCulloch (1990) mentioned any statistical test for comparing distributions of two multivariate samples. On the other hand, several distributional tests for univariate data, namely, Kolmogorov-Smirnov test, Cramer-Smirnov-von-Mises test (see, e.g., Serfling (1980)), Shapiro-Wilk test (see Shapiro and Wilk (1965) and Leslie, Stephens and Fotopoulos (1986)) and Anderson-Darling test (Anderson and Darling (1952, 1954)), etc. have been discussed in the literature. Multivariate extensions of a few of these tests have also been studied in the literature (see, e.g., Bickel (1969), Deheuvels (1981) and Justel, Pena and Zamar (1997)).

In this article, motivated by the characterization of distributions based on spatial quantiles (see Koltchinskii (1997)), we propose an extension of Q-Q plot using spatial quantiles for multivariate data. We also propose and study some statistical tests for distributions based on spatial quantiles.

2 Q-Q plot

Let $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ be two d -dimensional data sets, where $\mathbf{x}_i = (x_i^1, \dots, x_i^d)$ and $\mathbf{y}_j = (y_j^1, \dots, y_j^d)$. Suppose that $\mathbf{u}_1, \dots, \mathbf{u}_n$ and $\mathbf{u}_{n+1}, \dots, \mathbf{u}_{n+m}$ are spatial ranks of data sets \mathcal{X} and \mathcal{Y} , respectively, where spatial rank of $\mathbf{z} \in \mathbb{R}^d$ with respect to the data cloud formed by the observations $\mathbf{x}_1, \dots, \mathbf{x}_n$ (or $\mathbf{y}_1, \dots, \mathbf{y}_m$) is defined as $n^{-1} \sum_{i:\mathbf{x}_i \neq \mathbf{z}} \|\mathbf{z} - \mathbf{x}_i\|^{-1} (\mathbf{z} - \mathbf{x}_i)$ (or $m^{-1} \sum_{i:\mathbf{y}_i \neq \mathbf{z}} \|\mathbf{z} - \mathbf{y}_i\|^{-1} (\mathbf{z} - \mathbf{y}_i)$) (see, e.g., Chaudhuri (1996), Mottonen and Oja (1995) and Serfling (2002, 2004)). Further, for $k = 1, \dots, n + m$, let $Q_{\mathcal{X}}(\mathbf{u}_k) = (Q_{\mathcal{X},1}(\mathbf{u}_k), \dots, Q_{\mathcal{X},d}(\mathbf{u}_k))$ and $Q_{\mathcal{Y}}(\mathbf{u}_k) = (Q_{\mathcal{Y},1}(\mathbf{u}_k), \dots, Q_{\mathcal{Y},d}(\mathbf{u}_k))$ be \mathbf{u}_k -th spa-

tial quantiles of data sets \mathcal{X} and \mathcal{Y} , respectively, where $Q_{\mathcal{X}}(\mathbf{u}_k)$ and $Q_{\mathcal{Y}}(\mathbf{u}_k)$ are defined as $\arg \min_{Q \in \mathbb{R}^d} n^{-1} \sum_{i=1}^n \{ \|\mathbf{x}_i - Q\| + \langle \mathbf{u}_k, (\mathbf{x}_i - Q) \rangle \}$ and $\arg \min_{Q \in \mathbb{R}^d} m^{-1} \sum_{i=1}^m \{ \|\mathbf{y}_i - Q\| + \langle \mathbf{u}_k, (\mathbf{y}_i - Q) \rangle \}$ (see Chaudhuri (1996, p. 863)), respectively. Here $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product. One may use the algorithm given in Chaudhuri (1996, p. 864) to compute spatial quantiles. In the case of d -dimensional data, for any fixed $l \in \{1, \dots, d\}$, we can plot $(Q_{\mathcal{X},l}(\mathbf{u}_k), Q_{\mathcal{Y},l}(\mathbf{u}_k))$ for all $k = 1, \dots, n + m$. If the original data are d -dimensional, we will have d scatter plots, and when $d = 1$, we get the usual univariate q-q plot. Note that unlike the procedure of matching of the two samples used by Easton and McCulloch (1990), our approach of matching the two samples using the spatial ranks does not require equality of the sizes of the two samples nor does it require one to solve any complex optimization problem. It follows from a well-known characterization result in Koltchinskii (1997, p.446) that $F = G \Leftrightarrow Q_F(\mathbf{u}) = Q_G(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^d$ such that $\|\mathbf{u}\| < 1 \Leftrightarrow Q_{F,l}(\mathbf{u}) = Q_{G,l}(\mathbf{u})$ for all $l = 1, \dots, d$, where F and G are two multivariate distributions, and $Q_F(\mathbf{u})$ and $Q_G(\mathbf{u})$ are the spatial quantiles associated with F and G , respectively. This fact implies that for two random samples, if the points in each of the above-mentioned d scatter plots remain tightly clustered around the 45° line passing through the origin, it is an indication that the two samples are generated from the same distribution. This is one of the main motivation behind considering spatial quantiles as for many other multivariate quantiles like marginal quantiles (or more generally, l_p quantiles considered by Chakraborty (2001)), such a characterization result is not available.

Now, we demonstrate our methodology using some examples. We generated 100 i.i.d. observations from $F = N_2(\boldsymbol{\mu}_1, \Sigma_1)$ and $G = N_2(\boldsymbol{\mu}_2, \Sigma_2)$ distributions, where $N_2(\boldsymbol{\mu}, \Sigma)$ denotes the bivariate normal distribution with location parameter $\boldsymbol{\mu}$ and scatter matrix Σ . First, we considered $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (0, 0)$, and $\Sigma_1 = \Sigma_2 = I_2 = 2 \times 2$ identity matrix, i.e., when F and G are standard bivariate normal distributions. Next, we considered $\boldsymbol{\mu}_1 = (0, 0)$, $\boldsymbol{\mu}_2 = (2, 2)$, and $\Sigma_1 = \Sigma_2 = I_2$, i.e., G is a location shift of F . We also considered $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (0, 0)$, $\Sigma_1 = I_2$, and $\Sigma_2 = 2I_2$, i.e., F and G have the same location but different scales. The Q-Q

plots for the above cases are displayed in Figure 1.

In each scatter plot in the first row of Figure 1, the points are clustered tightly around the 45° line passing through the origin. In each of the diagrams in the second row, the points are tightly clustered around the 45° straight line passing through the point $(0, 2)$, which indicates that one distribution is a location shift of another distribution, and the shift is 2. Finally, in each of the plots in the third row, the points are clustered around a line through the origin having slope different from 45° , which indicates that one distribution can be obtained from the other by scale transformation.

We next illustrate how the proposed graphical tool can be used for checking whether a specified distribution F_0 fits a given data set or not. Let $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a d -dimensional data set, where $\mathbf{x}_i = (x_i^1, \dots, x_i^d)$ and $\mathbf{u}_1, \dots, \mathbf{u}_n$ are spatial ranks of the data points $\mathbf{x}_i, i = 1, \dots, n$. Suppose that $Q_{F_0}(\mathbf{u}_k) = (Q_{F_0,1}(\mathbf{u}_k), \dots, Q_{F_0,d}(\mathbf{u}_k))$ is the \mathbf{u}_k -th spatial quantile of the specified distribution F_0 , where $k = 1, \dots, n$. As in the preceding case, for a fixed $l \in \{1, \dots, d\}$, since $Q_{\mathcal{X}}(\mathbf{u}_k) = \mathbf{x}_k$, here we plot $(x_k^l, Q_{F_0,l}(\mathbf{u}_k))$ for all $k = 1, \dots, n$. Here also, we will have d -scatter plots when original data points are d -dimensional, and if F_0 fits the data well, the points in each scatter plot will be clustered tightly around the 45° line passing through the origin. Clearly, when $d = 1$, we will again get the standard univariate q-q plot for the one-sample problem.

We now consider the reference distribution $F_0 = N_2(\mathbf{0}, I_2)$. Marden (1998) provided a result for computing spatial quantiles of the multivariate normal distribution, and we have used that for our calculation. We generated 100 i.i.d. observations from each of $N_2(\boldsymbol{\mu}, \Sigma)$, $C_2(\boldsymbol{\mu}, \Sigma)$ and $L_2(\boldsymbol{\mu}, \Sigma)$, where $N_2(\boldsymbol{\mu}, \Sigma)$, $C_2(\boldsymbol{\mu}, \Sigma)$ and $L_2(\boldsymbol{\mu}, \Sigma)$ denote the bivariate normal, Cauchy (i.e., $f(\mathbf{x}) = c\{1 + (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\}^{-1}$) and Laplace (i.e., $f(\mathbf{x}) = c \exp[-\{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\}^{1/2}]$) distributions with location parameter $\boldsymbol{\mu}$ and scatter matrix Σ , respectively. In the following Q-Q plots, we consider $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I_2$.

It is clearly evident from the diagrams in the first row of Figure 2 that the reference distribution (i.e., standard bivariate normal) fits the data well as the points in those diagrams

Figure 1: Q-Q plots when samples are generated from the same distribution (first row), from distributions having different locations but the same scale (second row), and from distributions having same location but different scales (third row).

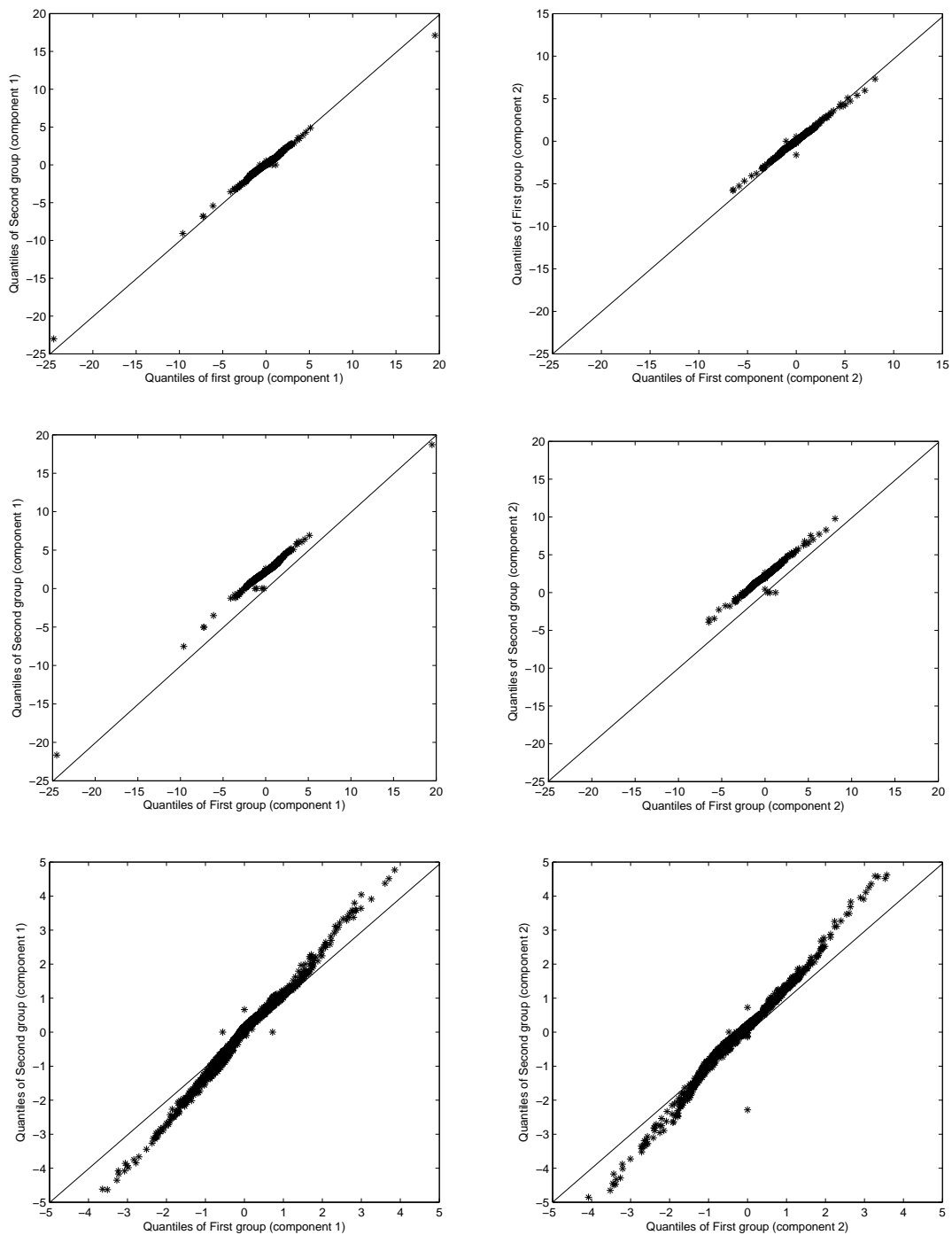
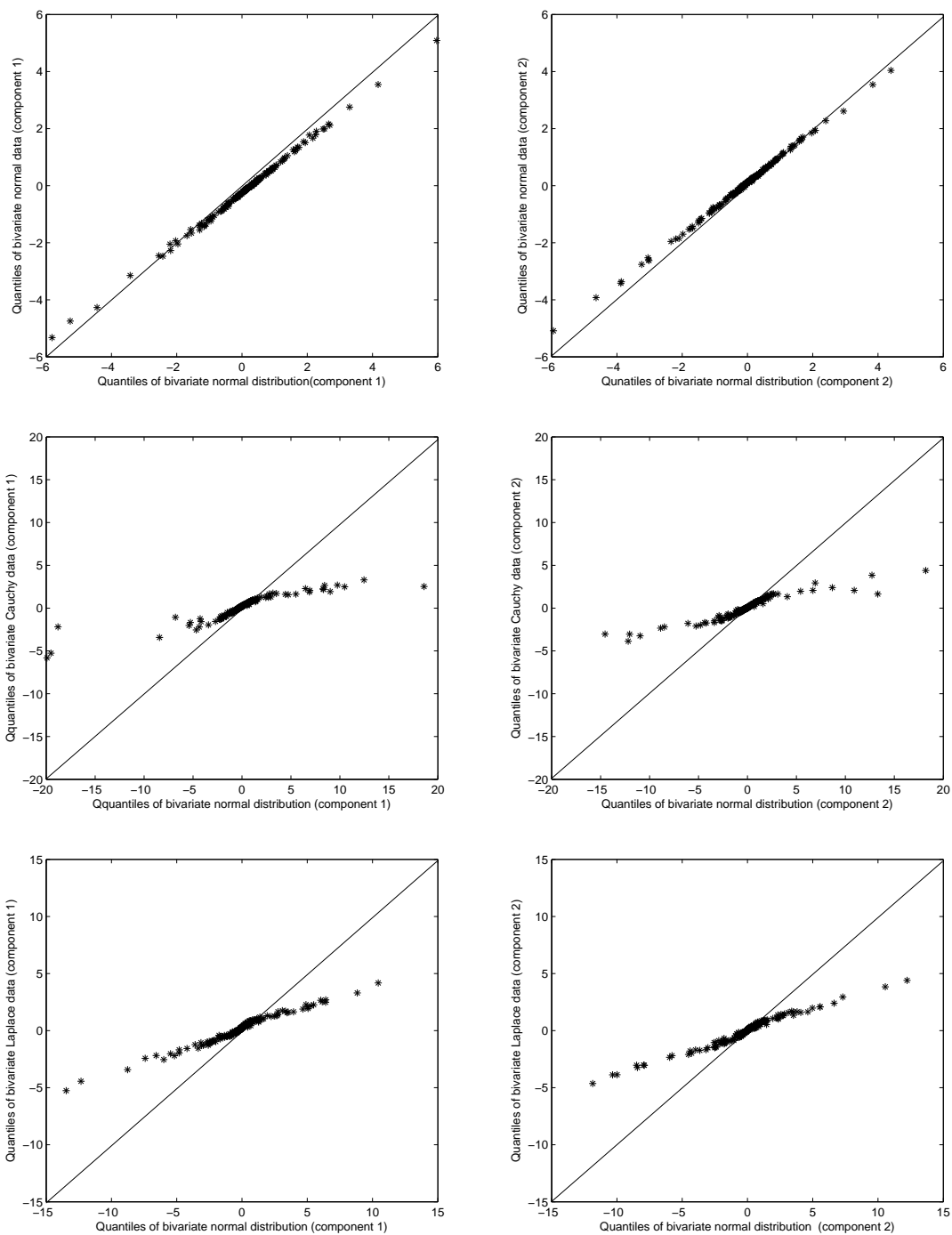


Figure 2: Q-Q plots when data are generated from bivariate normal (first row), bivariate Cauchy (second row) and bivariate Laplace (third row) distributions, respectively.



are tightly clustered around the 45° line passing through the origin. On the other hand, in each scatter plot in the second and the third rows, the points are significantly deviating from the 45° line passing through the origin, which is an indication that the reference distribution does not fit the data well.

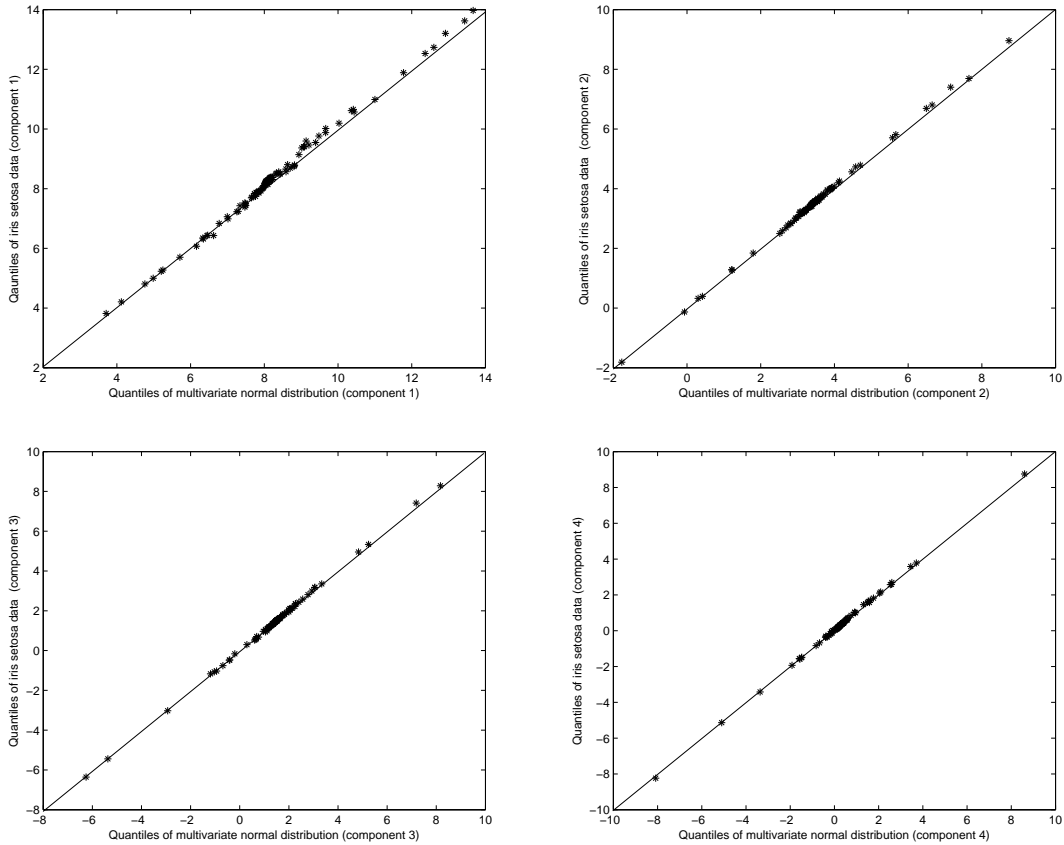
3 Analysis of real and simulated data

In this section, we use four real data sets, namely, iris data, hemophilia data, sunspot number data and sea level pressures data and some simulated data generated from certain Gaussian processes for illustrating our methodology. The iris data is available in <http://archive.ics.uci.edu/ml>, hemophilia data can be obtained from the “rrcov” package in the software *R*, sunspot number data is available in <http://www.ngdc.noaa.gov/stp/SOLAR/ftpsunspot> and sea level pressures data can be obtained from <http://www.cpc.noaa.gov/data/indices/darwin> and <http://www.cpc.noaa.gov/data/indices/tahiti>.

Iris data: In this case, there are three multivariate data sets corresponding to three different varieties of iris, namely, *Iris setosa*, *Iris virginica* and *Iris versicolor*, each of size 50. In each data point, there are four measurements, namely, sepal length, sepal width, petal length and petal width. We would like to determine how close each data set is to a 4-dimensional Gaussian distribution. In this case, we take $F = N_4(\mathbf{0}, I_4)$ as the reference distribution, and keeping the issue of robustness in mind, we standardize the data using the minimum covariance determinant (MCD) estimates of location and scatter matrix (see Rousseeuw and Leroy (1987)) instead of the usual mean and the variance-covariance matrix.

It is visible in Figures 3 through 5 that almost all scatter plots are clustered tightly around the 45° straight line passing through $(0, 0)$ except the scatter plot for the sepal length of *Iris virginica*, which deviates to some extent from the 45° line passing through the origin (see first diagram in Figure 4). Overall, it is indicated from the diagrams that the measurements on all three types of iris behave reasonably like multivariate normal distributions.

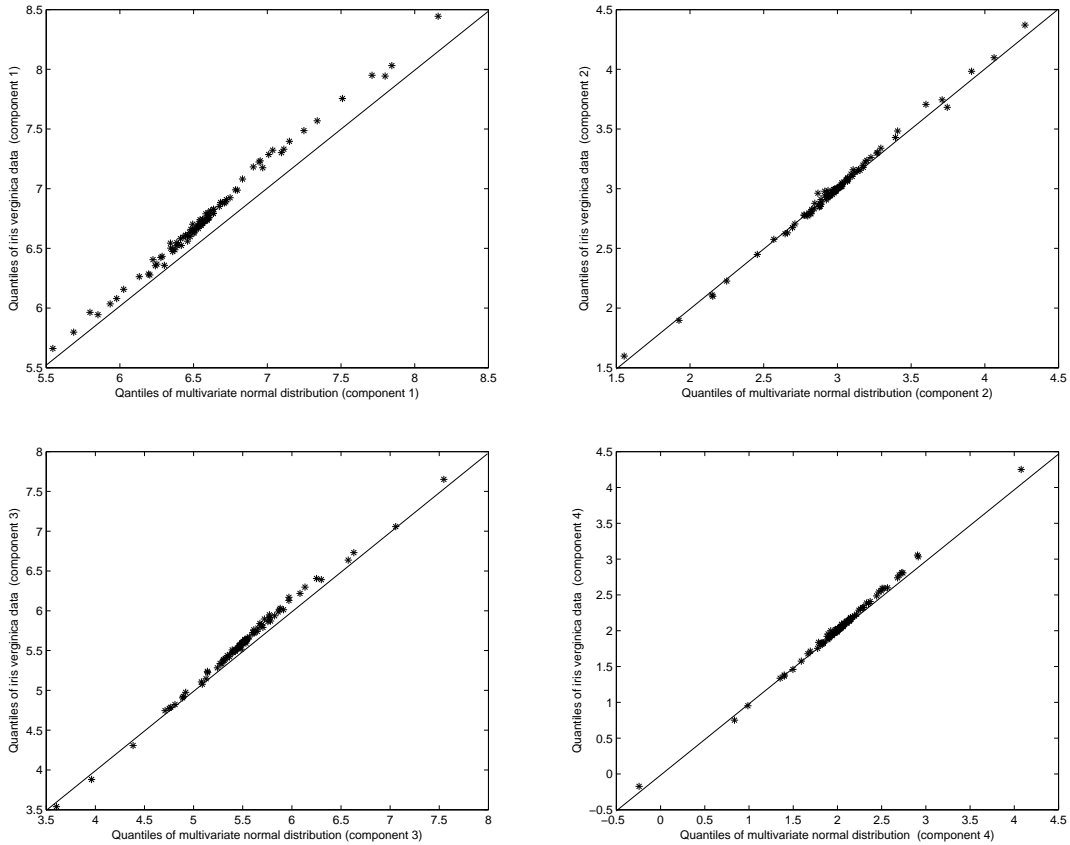
Figure 3: Q-Q plots of *Iris setosa* with multivariate normal as the reference distribution.



Hemophilia data: The hemophilia data set contains two variables, namely, AHF activity and AHF antigen on 75 women, who belong to two groups. Among the 75 women considered, 30 are non-carriers (first group), and the remaining 45 of them are known as Hemophilia A carriers (second group). In Figure 6, we display Q-Q plots for these two groups of data. The points in the diagrams in Figure 6 are not clustered around the 45° straight line passing through the origin, which is a consequence of the difference between the distributions associated with the two groups.

Monthly sunspot number data: The sunspot number data contains monthly average number of sunspots during the period 1749 to 2009. Out of those of 261 years, data for 1749 and 2009 are incomplete. We have removed the observations corresponding to 1749 and 2009

Figure 4: Q-Q plots of *Iris virginica* with multivariate normal as the reference distribution.



and carried out our analysis on the observations for the remaining 259 years. We divided the data into two samples. One sample contains six dimensional data corresponding to the six months January, February, March, October, November and December, and the other one consists of six dimensional data corresponding to the months April, May, June, July, August and September. Here each coordinate of the six dimensional data corresponds to a specific month, and the motivation behind splitting the data into two parts corresponding to the periods October-March and April-September comes from the fact that one equinox in a year occurs on March 20/21 and another on September 22/23. In Figure 7, we have six scatter plots. The points in each scatter plot are deviating significantly from the 45° straight line passing through the origin. Further, in each scatter plot, the points appear to lie on a line

Figure 5: Q-Q plots of *Iris versicolor* with multivariate normal as the reference distribution.

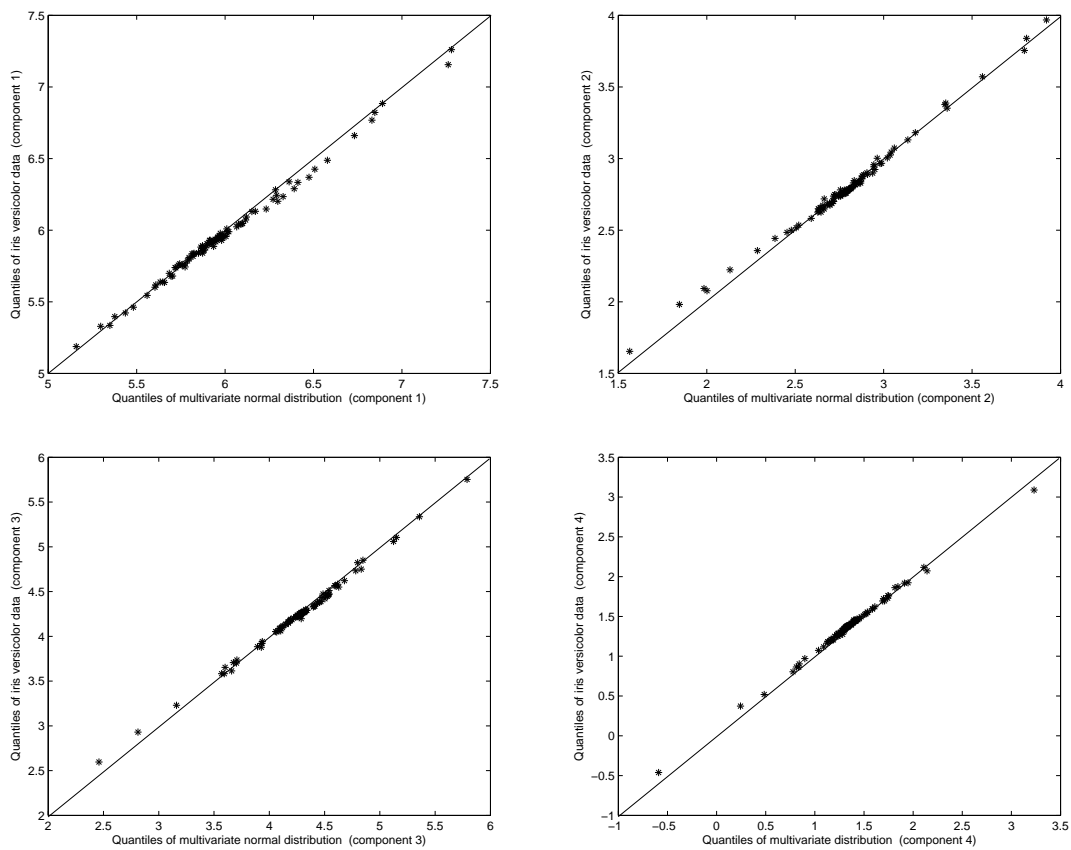


Figure 6: Q-Q plots for hemophilia data.

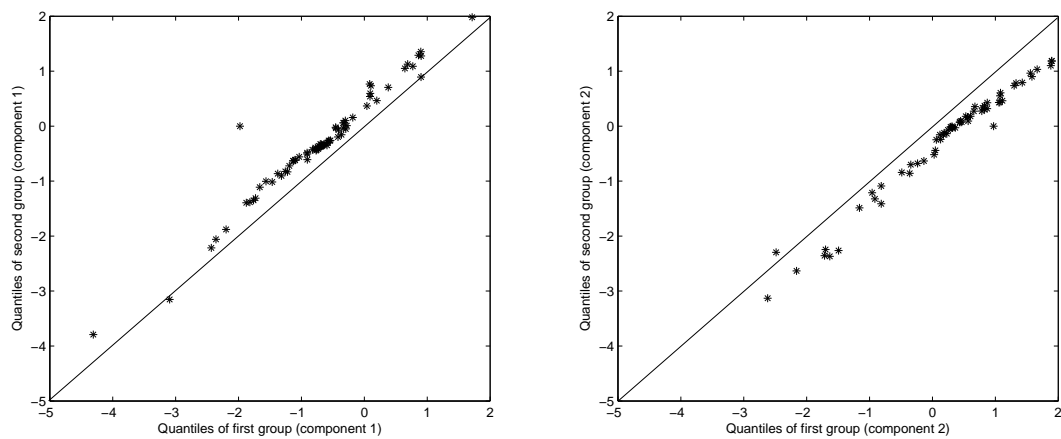
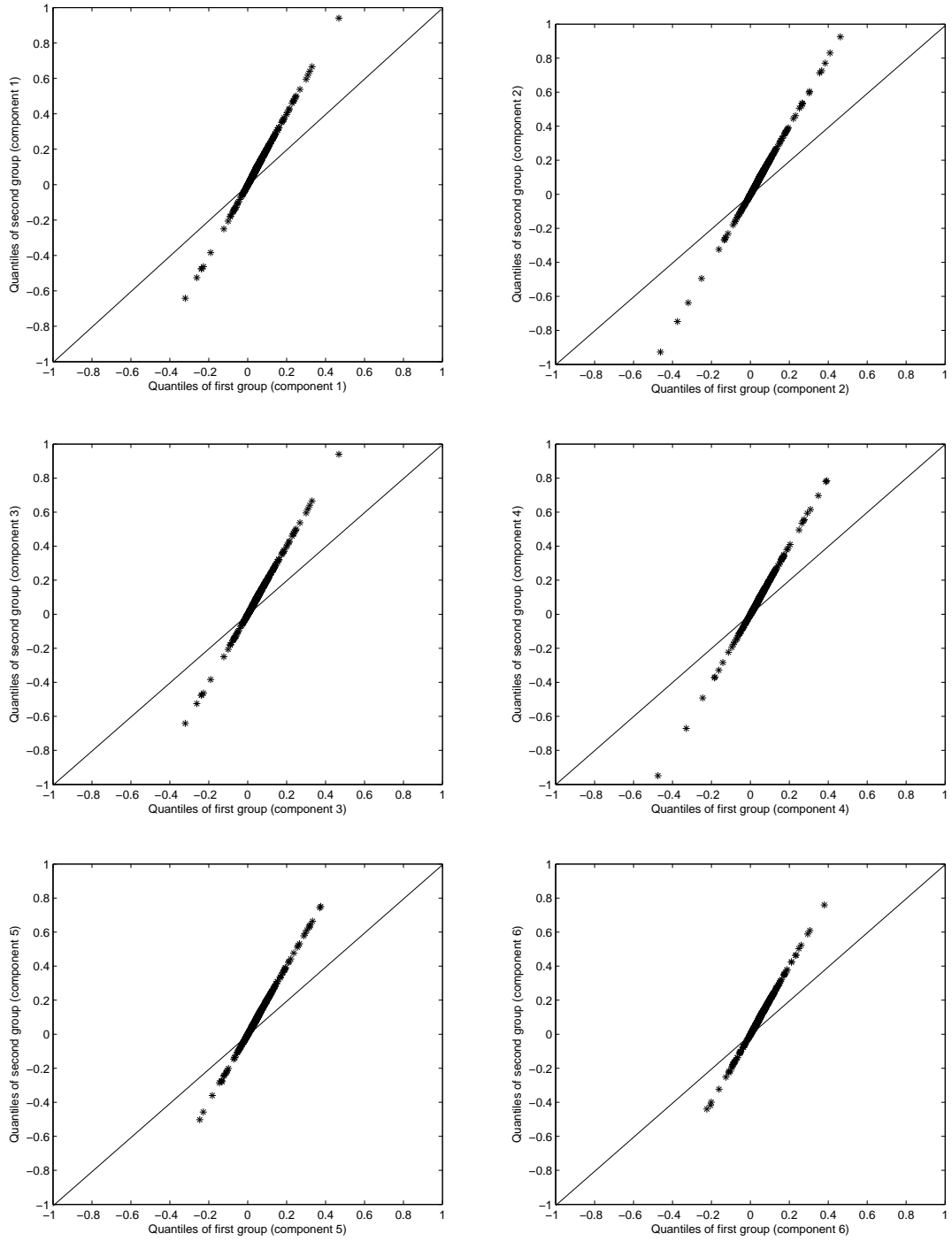


Figure 7: Q-Q plots for monthly sunspot number data.



with a slope different from 45° and an intercept different from zero. This is an indication that two samples differ both in their locations as well as their scales.

3.1 Plots for high dimensional real and simulated data: an alternative approach

Data on sea level pressures: This data set consists of monthly sea level pressures from two different islands in southern Pacific ocean, namely, Darwin ($13^\circ S$, $131^\circ E$) and Tahiti ($17^\circ S$, $149^\circ W$) during the period of 1850 to 2008. Here we have a two-sample problem with each sample corresponding to an island and containing 159 twelve dimensional observations. For this data, each data point corresponds to a year, and each coordinate of a data point corresponds to an observation of a particular month. In this case, it is not convenient to display and visually examine twelve different scatter plots, and we can plot (see Figure 8) $(l, Q_{X,l}(\mathbf{u}_k) - Q_{Y,l}(\mathbf{u}_k))$ for all $k = 1, \dots, 159 + 159 = 318$, $l = 1, \dots, 12$, where $Q_{X,l}(\mathbf{u}_k)$ and $Q_{Y,l}(\mathbf{u}_k)$ are defined as in Section 2. Note that, in this way, we get a single two dimensional plot, where there are twelve vertical lines parallel to one another, and points are plotted along those lines. It is amply indicated in Figure 8 that the distributions corresponding to two samples are significantly different, and there are differences in their locations as well as scales. On each vertical line, the points are distributed in such a way that the center of each distribution is significantly different from zero. Further, the spreads of the distribution of the points on different vertical lines appear to be quite different.

Observations simulated from Gaussian process: We generated 10 i.i.d. observations from each of three pairs of Gaussian processes with different choices of mean functions: $m_1(t)$, $m_2(t)$, and covariance kernels: $k_1(s, t)$, $k_2(s, t)$, where $s, t \in [0, 1]$. In our study, we consider an equally spaced partition $\{t_1, \dots, t_{20}\}$ of $[0, 1]$ and sample the observations at those time points. First, we consider two standard Brownian motions, i.e., $m_1(t) = m_2(t) = 0$ and $k_1(s, t) = k_2(s, t) = \min(s, t)$. Next, we consider $m_1(t) = 0$, $m_2(t) = 2$, and $k_1(s, t) = k_2(s, t) = \min(s, t)$, and in the third case, our choices of param-

Figure 8: Quantile difference plot for data on sea level pressures.

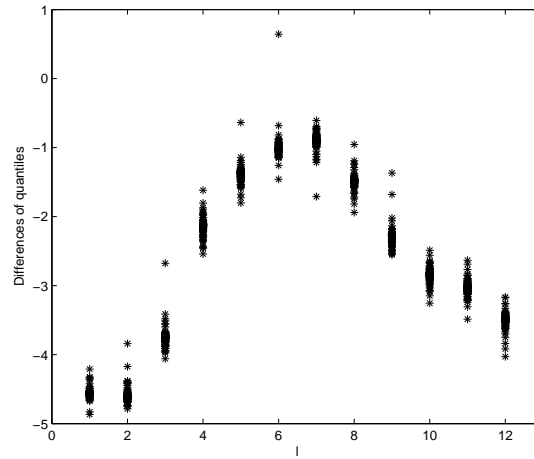
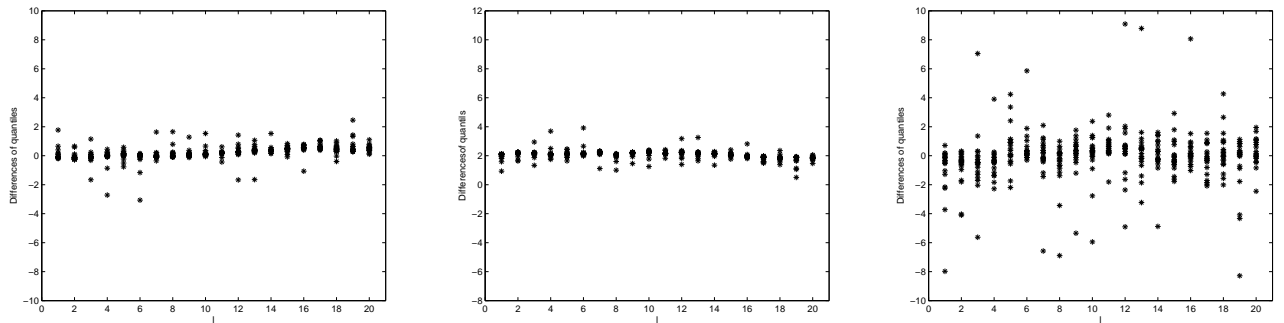


Figure 9: Quantile difference plots when two Brownian motions are identical (first diagram from the left), two processes having different mean functions but same covariance kernels (second diagram) and two processes having the same mean function but different covariance kernels (third diagram).



eters are $m_1(t) = m_2(t) = 0$, $k_1(s, t) = \min(s, t)$, and $k_2(s, t) = 2 \min(s, t)$. The plots of quantile differences for the above cases are displayed in Figure 9. In the first diagram (from the left) in Figure 9, the points in each vertical line are tightly clustered around the horizontal axis passing through the origin, which indicates that the samples are obtained from similar distributions. On the other hand, the difference between the distributions in their locations and scales, are reflected in the other two diagrams.

Figure 10: Quantile difference plots for *Iris setosa* (left most), *Iris virginica* (middle) and *Iris versicolor* (right).

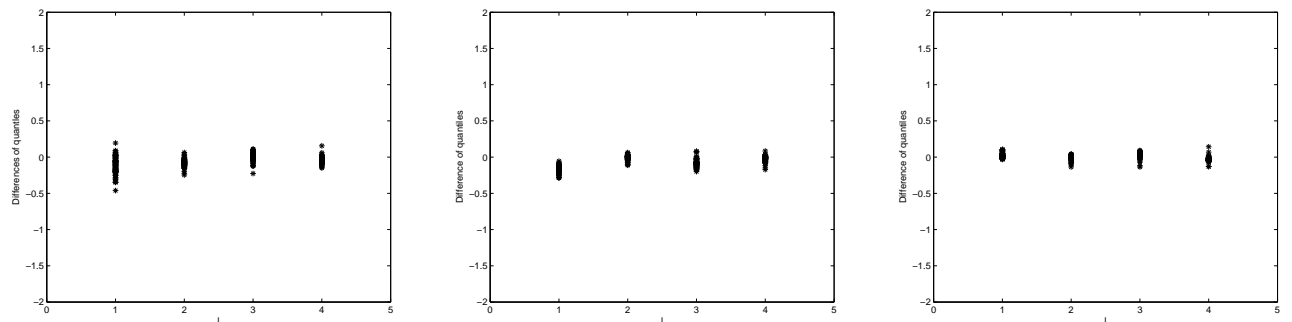


Figure 11: Quantile difference plots for hemophilia data.

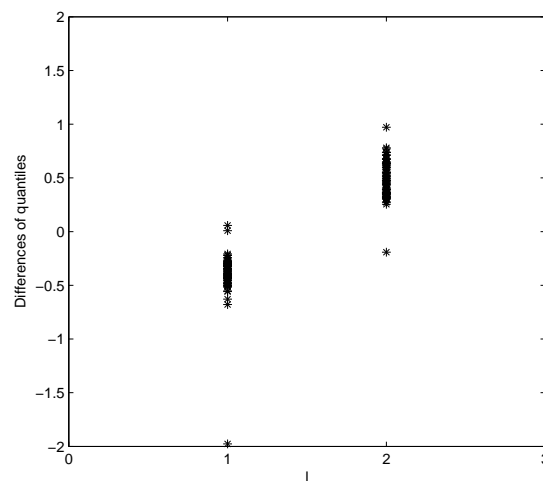
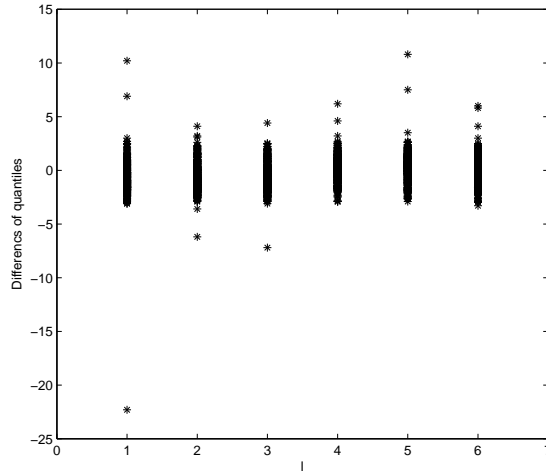


Figure 12: Quantile difference plots for monthly sunspot number data.



Note that this alternative approach based on quantile differences is related to the arrow plots considered by Marden (1998) for bivariate data. The arrows in the arrow plot represent the vector differences of two sets of bivariate quantiles. However, Marden's plots are limited to bivariate data while our quantile difference plots can be conveniently used for multivariate data with dimensions two or larger. In Figures 10, 11 and 12, we display the quantile difference plots for iris data, hemophilia data and monthly sunspot number data that we have already mentioned and discussed. It is evident from Figure 10 that the measurements on the three iris species behave reasonably like observations from multivariate normal distributions. On the other hand, differences between the underlying distributions of the two groups in hemophilia data and monthly sunspot number data corresponding to the two periods (October-March and April-September) are clearly reflected in Figures 11 and 12, respectively.

4 Distributional tests based on spatial quantiles

The plots based on differences of quantiles considered in Section 3 along with the characterization of distributions by spatial quantiles (Koltchinskii (1997)) mentioned in Section 2 motivated us to consider some tests for distributions based on the differences of spatial quantiles in one-sample and two-sample problems. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be i.i.d. observations from an unknown distribution F with a continuous probability density function, and suppose that we want to test $H_0 : F = F_0 (\Leftrightarrow Q_F(\mathbf{u}) = Q_{F_0}(\mathbf{u}) \text{ for all } \mathbf{u})$ against the alternatives $H_1 : F \neq F_0 (\Leftrightarrow Q_F(\mathbf{u}) \neq Q_{F_0}(\mathbf{u}) \text{ for some } \mathbf{u})$, where F_0 is a specified distribution with a continuous probability density function. It follows from the asymptotic results in Chaudhuri (1996, Theorem 3.1.2.) and Koltchinskii (1997, Theorem 5.7) that, for any $0 < b < 1$ and $\mathbf{u} \in S_1(b) = \{\mathbf{u} : \|\mathbf{u}\| < b\}$, the process $\sqrt{n}\{\hat{Q}_F(\mathbf{u}) - Q_F(\mathbf{u})\}$ converges in distribution to the Gaussian process $Z(\mathbf{u})$ with zero mean and covariance kernel

$$k(\mathbf{u}_1, \mathbf{u}_2) = [D_1^F\{Q(\mathbf{u}_1)\}]^{-1}[D_2^F\{Q(\mathbf{u}_1), Q(\mathbf{u}_2), \mathbf{u}_1, \mathbf{u}_2\}][D_1^F\{Q(\mathbf{u}_2)\}]^{-1}.$$

Here $D_1^F(Q) = E_F[\|\mathbf{X} - Q\|^{-1}\{I_d - \|\mathbf{X} - Q\|^{-2}(\mathbf{X} - Q)(\mathbf{X} - Q)^T\}]$, and $D_2^F(Q_1, Q_2, \mathbf{u}, \mathbf{v}) = E_F[\{\|\mathbf{X} - Q_1\|^{-1}(\mathbf{X} - Q_1) + \mathbf{u}\}\{\|\mathbf{X} - Q_2\|^{-1}(\mathbf{X} - Q_2) + \mathbf{v}\}^T]$. Then, the continuity of the integral functional implies that the asymptotic distribution of $n \int_{\mathbf{u} \in S_1(b)} \|\hat{Q}_F(\mathbf{u}) - Q_F(\mathbf{u})\|^2 d\mathbf{u}$ is the same as the distribution of $\int_{\mathbf{u} \in S_1(b)} \|Z(\mathbf{u})\|^2 d\mathbf{u}$. Hence, a test that rejects H_0 whenever $n \int_{\mathbf{u} \in S_1(b)} \|\hat{Q}_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\|^2 d\mathbf{u} > c_\alpha$, where c_α is the $(1 - \alpha)$ -th quantile ($0 < \alpha < 1$) of the distribution of $\int_{\mathbf{u} \in S_1(b)} \|Z(\mathbf{u})\|^2 d\mathbf{u}$, will have asymptotic level α . Further, if $Q_F(\mathbf{u}) \neq Q_{F_0}(\mathbf{u})$ for some $\mathbf{u} \in S_1(b)$, we have $\int_{\mathbf{u} \in S_1(b)} \|Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\|^2 d\mathbf{u} > 0$ in view of the continuity of $Q_F(\mathbf{u})$ and $Q_{F_0}(\mathbf{u})$. Consequently, the power of the above mentioned test will tend to 1 as $n \rightarrow \infty$ in that case. The integral $n \int_{\mathbf{u} \in S_1(b)} \|\hat{Q}_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\|^2 d\mathbf{u}$ can be approximated by the average $V_n = (1/N) \sum_{i=1}^N n \|\hat{Q}_F(\mathbf{u}_i) - Q_{F_0}(\mathbf{u}_i)\|^2$, where the \mathbf{u}_i 's are i.i.d. random vectors with the uniform distribution on the sphere $S_1(b)$, and N is appropriately large. V_n can be used as a test statistic for testing H_0 against H_1 . Now, in view of the well-known orthogonal

decomposition of a finite dimensional multivariate normal distribution, the distribution of V_n can be approximated by the distribution of $\sum_{k=1}^{Nd} M_k^2$, where the M_k 's are independent random variables, and M_k follows $N(0, \lambda_k)$ distribution for $k = 1, \dots, Nd$. Here λ_k 's are the eigen values of the $Nd \times Nd$ dimensional covariance matrix (we denote it as K) having $d \times d$ dimensional (i, j) -th block $k(\mathbf{u}_i, \mathbf{u}_j)$, where $i, j = 1, \dots, N$. In order to find out the cut off point c_α of the tests, one needs to estimate the eigen values λ_k 's of the $Nd \times Nd$ dimensional matrix K . Note that the expression for K (see the expression of $k(\mathbf{u}_1, \mathbf{u}_2)$) involves spatial quantiles and expectation with respect to the underlying distribution. As mentioned in Section 2, we can estimate population spatial quantiles by their sample versions and use sample averages for estimating expectations. Using the estimated eigen values $\hat{\lambda}_k$'s, one can simulate Gaussian random variables M_k 's with zero means and $\hat{\lambda}_k$'s as the variances for $k = 1, \dots, Nd$. One can generate several Monte-carlo replications of $\sum_{k=1}^{Nd} M_k^2$ and depending on the specified level of the test, choose a cut off point of the test as the appropriate quantile of the empirical distribution of that sum. In our numerical work, we chose $b = 0.999$, $N = 1000$ and used 1000 Monte-carlo replications to obtain the empirical distribution of $\sum_{i=1}^{Nd} M_k^2$. We have made those choices based on our empirical experience keeping in mind the computational cost involved and the performance of the test.

Similarly, for a two-sample problem with unknown absolutely continuous distributions F and G having continuous probability density functions, suppose that we want to test $H_0^* : F = G (\Leftrightarrow Q_F(\mathbf{u}) = Q_G(\mathbf{u}) \text{ for all } \mathbf{u})$ against the alternatives $H_1^* : F \neq G (\Leftrightarrow Q_F(\mathbf{u}) \neq Q_G(\mathbf{u}) \text{ for some } \mathbf{u})$. If the sample sizes n and m are such that as $n, m \rightarrow \infty$, $n/(n+m) \rightarrow \gamma \in (0, 1)$, then, arguing in a similar way as in the one-sample problem, one can show that the integral $(n+m) \int_{\mathbf{u} \in S_1(b)} \|(\hat{Q}_F(\mathbf{u}) - \hat{Q}_G(\mathbf{u})) - (Q_F(\mathbf{u}) - Q_G(\mathbf{u}))\|^2 d\mathbf{u}$ converges weakly to $\int_{\mathbf{u} \in S_1(b)} \|Z_1(\mathbf{u})\|^2 d\mathbf{u}$, where $Z_1(\mathbf{u})$ is a Gaussian process with zero mean and covariance kernel

$$k_1(\mathbf{u}_1, \mathbf{u}_2) = \frac{[D_1^F \{Q(\mathbf{u}_1)\}]^{-1} [D_2^F \{Q(\mathbf{u}_1), Q(\mathbf{u}_2), \mathbf{u}_1, \mathbf{u}_2\}] [D_1^F \{Q(\mathbf{u}_2)\}]^{-1}}{\gamma}$$

$$+ \frac{[D_1^G\{Q(\mathbf{u}_1)\}]^{-1}[D_2^G\{Q(\mathbf{u}_1), Q(\mathbf{u}_2), \mathbf{u}_1, \mathbf{u}_2\}][D_1^G\{Q(\mathbf{u}_2)\}]^{-1}}{(1 - \gamma)}.$$

Here, a test, which rejects H_0^* for an appropriately large value of $(n + m) \int_{\mathbf{u} \in S_1(b)} \|\hat{Q}_F(\mathbf{u}) - \hat{Q}_G(\mathbf{u})\|^2 d\mathbf{u}$, will have similar asymptotic level and power properties like the test proposed in the one-sample problem. As in the one sample problem, the average $T_{n,m} = (1/N) \sum_{i=1}^N (n + m) \|\hat{Q}_F(\mathbf{u}_i) - \hat{Q}_G(\mathbf{u}_i)\|^2$, which is a convenient approximation for $(n + m) \int_{\mathbf{u} \in S_1(b)} \|\hat{Q}_F(\mathbf{u}) - \hat{Q}_G(\mathbf{u})\|^2 d\mathbf{u}$, can be used as a test statistic for testing H_0^* against H_1^* . Here, the \mathbf{u}_i 's are again i.i.d. random vectors with the uniform distribution on $S_1(b)$, and we have chosen $N = 1000$ and $b = 0.999$ in our numerical works. Also, the implementation of this test will be similar to that of the test mentioned in the one-sample problem in the sense that the distribution of the test statistic $T_{n,m}$ can be approximated by a sum of squares of independent normal random variables all with zero mean but different variances, which can be estimated using the covariance kernel $k_1(\mathbf{u}_1, \mathbf{u}_2)$.

4.1 Demonstration of proposed tests in real and simulated data

We have implemented our proposed tests on some simulated and real data that we have already discussed in Sections 2 and 3. For those simulated data sets, we have used 1000 Monte-carlo replications to compute the powers of the tests. We have considered the same sample sizes and the dimensions as in Sections 2 and 3. The results are summarized in Tables 1 through 4.

It is indicated by the figures in Tables 1 and 2 that the observed level of the test is close to its nominal level of the test, and the test is quite powerful when the alternative hypothesis is true. The p-values reported in Table 3 indicate that, according to our test, the measurements on all three varieties of iris follow normal distributions reasonably well. For hemophilia data, the p-value reported in Table 4 is small but not significantly small. This indicates some difference in the distributions corresponding to the non-carriers and Hemophilia-A carriers – however, the difference is statistically not very significant. On the

Table 1: Estimated powers of our test for different simulated data sets in one-sample problems for 5% nominal level.

Sample distribution (F)	Specified distribution (F_0)	Estimated power
$N_2(\mathbf{0}, I_2)$	$N_2(\mathbf{0}, I_2)$	4.6%
$C_2(\mathbf{0}, I_2)$	$N_2(\mathbf{0}, I_2)$	87.3%
$L_2(\mathbf{0}, I_2)$	$N_2(\mathbf{0}, I_2)$	55.5%

Table 2: Estimated powers of our test for different simulated data sets in two-sample problems for 5% nominal level.

First sample distribution (F)	Second sample distribution (G)	Estimated power
$N_2(\mathbf{0}, I_2)$	$N_2(\mathbf{0}, I_2)$	4.5%
$N_2(\mathbf{0}, I_2)$	$N_2(\mathbf{2}, I_2)$	76.7%
$N_2(\mathbf{0}, I_2)$	$N_2(\mathbf{0}, 2 * I_2)$	54.9%
Standard Brownian motion	Standard Brownian motion	4.4%
Standard Brownian motion	Gaussian process with mean function 2 and covariance function $k(s, t) = \min(s, t)$	69.1%
Standard Brownian motion	Gaussian process with mean function 0 and covariance function $k(s, t) = 2 \min(s, t)$	43.2%

Table 3: p -values of the proposed tests for different real data sets for one-sample problem.

Data set	Sample	Specified distribution (F_0)	p -value
Iris Data	Iris Setosa	$N_4(\mathbf{0}, I_4)$	0.72
	Iris Verginica	$N_4(\mathbf{0}, I_4)$	0.54
	Iris Versicolor	$N_4(\mathbf{0}, I_4)$	0.55

Table 4: p -values of the proposed tests for different real data sets for two-sample problem.

Data set	First sample	Second sample	p -value
Hemophilia data	non-carriers	Hemophilia A carriers	0.16
Sunspot number data	First period of six months	Second period of six months	0.07
Sea level pressures data	Darwin	Tahiti	0.00

other hand, for the other real data sets mentioned in Table 4, we get significantly small p -values implying statistically significant differences between the multivariate distributions under comparison, and the results are consistent with the scatter plots in Figures 7 and 8.

In the next subsection, we will make a thorough comparative study of the levels and the powers of our tests and some other tests that exist in the literature for multivariate as well as univariate data.

4.2 Comparison with some other tests

A few multivariate tests for distributions have been proposed for one-sample and two-sample problems, and among those tests, Kolmogorov-Smirnov test and Cramer-Smirnov-von-Mises test are possibly most well-known. Bickel (1969) discussed some asymptotic properties of two-sample multivariate Kolmogorov-Smirnov test (we denote it as $KS(\text{two})$), and Justel, Pena and Zamar (1997) investigated the performance of bivariate one-sample Kolmogorov-Smirnov test (we denote it as $KS(\text{one})$) in several examples. The asymptotic distributions of the multivariate Cramer-Smirnov-von-Mises test statistics were derived by Deheuvels (1981) in one- and two-sample problems (we denote them as $CSVM(\text{one})$ and $CSVM(\text{two})$ for one- and two-sample problems, respectively). Besides, Baringhaus and Franz (2004) developed a multivariate two sample test (we denote it as BF) based on certain averages of interpoint Euclidean distances, and they showed that the BF test reduces to the $CSVM(\text{two})$ test for univariate data (see Baringhaus and Franz (2004, p. 198)). However, they assumed finite first moment, and consequently, their proposed test is not expected to perform well for data arising from heavy-tailed distributions with infinite first moment (e.g., multivariate Cauchy distribution). Recently, Baringhaus and Franz (2010) proposed a modified and improved version of their test based on suitable transformations of interpoint Euclidean distances. In our numerical study, following their recommendation, we have considered the test (we denote it as BF^*) based on the logarithmic transformation of interpoint distances.

We carried out a detailed simulation study of the levels and the powers of the tests mentioned above and our tests for multivariate as well as univariate data. The results are summarized in Tables 5 and 6. It is evident from the figures in Tables 5 and 6 that our tests, the KS tests, the $CSVM$ tests and the BF^* test are all comparable in terms of their powers and levels in all the cases considered in our simulation study. However, our two-sample test is more powerful than the BF test for multivariate data. The computations for BF and BF^* tests have been done using the “Cramer” package in the Statistical software *R*, and the KS and the $CSVM$ tests for multivariate data have been implemented using the

Table 5: Estimated powers of one-sample tests for 5% nominal level.

Tests	$F = N_d(\mathbf{0}, I_d)$ and $F_0 = N_d(\mathbf{0}, I_d)$								
	$d = 1$			$d = 2$			$d = 3$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
KS(one)	0.049	0.053	0.049	0.037	0.041	0.046	0.036	0.042	0.043
CSVM (one)	0.052	0.053	0.052	0.035	0.043	0.048	0.035	0.042	0.045
Our test	0.048	0.053	0.055	0.036	0.042	0.045	0.035	0.041	0.045
Tests	$F = C_d(\mathbf{0}, I_d)$ and $F_0 = N_d(\mathbf{0}, I_d)$								
	$d = 1$			$d = 2$			$d = 3$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
KS(one)	0.932	1	1	0.367	0.561	0.613	0.324	0.503	0.557
CSVM (one)	0.916	1	1	0.372	0.588	0.633	0.342	0.549	0.586
Our test	0.899	0.998	1	0.363	0.588	0.622	0.321	0.550	0.600
Tests	$F = L_d(\mathbf{0}, I_d)$ and $F_0 = N_d(\mathbf{0}, I_d)$								
	$d = 1$			$d = 2$			$d = 3$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
KS(one)	0.665	0.771	0.800	0.273	0.503	0.542	0.255	0.465	0.531
CSVM (one)	0.671	0.799	0.832	0.267	0.515	0.543	0.245	0.462	0.532
Our test	0.699	0.811	0.855	0.284	0.531	0.565	0.263	0.481	0.552

Table 6: Estimated powers of two-sample tests for 5% nominal level.

Tests	$F = N_d(\mathbf{0}, I_d)$ and $G = N_d(\mathbf{0}, I_d)$								
	$d = 1$			$d = 2$			$d = 3$		
	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$
KS(two)	0.035	0.048	0.053	0.033	0.043	0.045	0.032	0.039	0.043
CSVM(two)	0.039	0.045	0.047	0.035	0.042	0.047	0.034	0.041	0.044
BF	0.039	0.045	0.047	0.034	0.041	0.043	0.033	0.038	0.042
BF*	0.039	0.044	0.048	0.036	0.043	0.043	0.033	0.037	0.040
Our test	0.032	0.046	0.049	0.033	0.040	0.043	0.032	0.038	0.041
Tests	$F = C_d(\mathbf{0}, I_d)$ and $G = N_d(\mathbf{0}, I_d)$								
	$d = 1$			$d = 2$			$d = 3$		
	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$
KS(two)	0.374	0.519	0.618	0.393	0.452	0.595	0.383	0.465	0.575
CSVM (two)	0.375	0.450	0.607	0.393	0.456	0.592	0.369	0.477	0.582
BF	0.375	0.450	0.607	0.335	0.401	0.499	0.287	0.353	0.465
BF*	0.383	0.516	0.611	0.379	0.485	0.587	0.381	0.455	0.571
Our test	0.382	0.522	0.618	0.382	0.479	0.601	0.377	0.468	0.588
Tests	$F = L_d(\mathbf{0}, I_d)$ and $G = N_d(\mathbf{0}, I_d)$								
	$d = 1$			$d = 2$			$d = 3$		
	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$	$n = 10$ $m = 10$	$n = 50$ $m = 50$	$n = 100$ $m = 100$
KS(two)	0.253	0.393	0.477	0.275	0.375	0.487	0.271	0.367	0.521
CSVM(two)	0.244	0.401	0.457	0.281	0.388	0.507	0.276	0.381	0.525
BF	0.244	0.401	0.457	0.242	0.332	0.399	0.241	0.330	0.488
BF*	0.275	0.427	0.522	0.285	0.411	0.509	0.276	0.379	0.543
Our test	0.271	0.432	0.533	0.281	0.401	0.519	0.285	0.388	0.551

Table 7: Estimated powers of some univariate one-sample tests for 5% nominal level.

$F = N(0, 1)$ and $F_0 = N(0, 1)$								
SW test			AD test			Our test		
$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
0.050	0.049	0.049	0.042	0.047	0.047	0.048	0.053	0.055
$F = C(0, 1)$ and $F_0 = N(0, 1)$								
SW test			AD test			Our test		
$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
0.011	0.377	0.879	0.017	0.290	0.556	0.899	0.998	1
$F = L(0, 1)$ and $F_0 = N(0, 1)$								
SW test			AD test			Our test		
$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
0.009	0.117	0.243	0.015	0.178	0.301	0.699	0.811	0.855

Table 8: Comparison of powers of our two-sample test and MST-run test when $F = N_d(\mathbf{0}, I_d)$ and $G = N_d(\frac{\Delta}{\sqrt{d}}\mathbf{1}_d, \sigma I_d)$ for 5% nominal level. Here d , Δ and σ are as in Friedman and Rafsky (1981, p.706), and $\mathbf{1}_d = (1, \dots, 1)_{1 \times d}$.

	$d = 1$	$d = 2$	$d = 5$	$d = 10$	$d = 20$
	$\Delta = 0.3, \sigma = 1$	$\Delta = 0.5, \sigma = 1$	$\Delta = 0.75, \sigma = 1$	$\Delta = 1.0, \sigma = 1$	$\Delta = 1.2, \sigma = 1$
Our test	0.332	0.554	0.701	0.832	0.993
MST-run test	0.180	0.350	0.640	0.780	0.860
	$\Delta = 0, \sigma = 1.3$	$\Delta = 0, \sigma = 1.2$	$\Delta = 0, \sigma = 1.2$	$\Delta = 0, \sigma = 1.1$	$\Delta = 0, \sigma = 1.075$
Our test	0.251	0.171	0.263	0.075	0.144
MST-run test	0.240	0.140	0.210	0.090	0.130

asymptotic distributions (see, e.g., Bickel (1969) and Deheuvels (1981)) of the corresponding test statistics. For univariate data, the computations of the KS and the CSVM tests are done using the “stats” and the “CvM2SL2Test” packages in the Statistical software *R*, respectively.

Shapiro-Wilks test (we denote it as SW, see Shapiro and Wilk (1965) and Leslie et al. (1986)) and Anderson-Darling test (we denote it as AD test, see Anderson and Darling (1952, 1954)) are two other well-known tests for checking the normality of univariate data, and it will be appropriate to compare the performance of our one-sample test with that of the SW and the AD tests in the univariate case. The levels and the powers of those tests and our test are reported in Table 7. The computations of the SW and the AD tests are done using the “stats” and “nortest” packages in the Statistical software *R*. It is evident from the figures in Table 7 that both of the SW and the AD tests are outperformed by our test when data are obtained from non-Gaussian distributions.

Recall from the Introduction that Friedman and Rafsky (1979) proposed a multivariate generalizations of the Wald-Wolfowitz run test using the idea of minimum spanning tree (we denote it as MST-run), and they developed a graphical tool for comparing two multivariate samples (see Friedman and Rafsky (1981)). We have compared the powers of our two-sample test with those of the MST-run test in different distributions (namely, multivariate normal with different locations and scales) considered by Friedman and Rafsky (1981, p.706). The results are reported in Table 8, and it is clear that the MST-run test does not perform well

compared to our test (see figures in Table 8).

Acknowledgment: The research of the first author is partially supported by a grant from Council of Scientific and Industrial Research (CSIR), Government of India.

References

- [1] Anderson, T. W. and Darling, D. A. (1952) Asymptotic theory of certain ‘goodness of fit’ criteria based on stochastic processes. *The Annals of Mathematical Statistics*, **23**, 193–212.
- [2] Anderson, T. W. and Darling, D. A. (1954) A test of goodness of fit. *Journal of the American Statistical Association*, **49**, 765–769.
- [3] Baringhaus, L. and Franz, C. (2004) On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**, 190–206.
- [4] Baringhaus, L. and Franz, C. (2010) Rigid motion invariant two-sample tests. *Statistica Sinica*, **20**, 1333–1361.
- [5] Bickel, P. (1969) A distribution free version of Smirnov two sample test in the p-variate case. *The Annals of Statistics*, **40**, 1–23.
- [6] Breckling, J. and Chambers, R. (1988) M-Quantiles. *Biometrika*, **75**, 761–777.
- [7] Chakraborty, B. (2001) On Affine Equivariant Multivariate Quantiles, *The Annals of the Institute of Statistical Mathematics*, **53**, 380–403.
- [8] Chaudhuri, P. (1996) On a Geometric Notion of Quantiles for Multivariate Data. *Journal of the American Statistical Association*, **91**, 862–872.
- [9] Deheuvels, P. (1981) An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, **11**, 102–113.
- [10] Easton, G. S. and McCulloch, R. E. (1990) A Multivariate Generalization of Quantile-Quantile Plots. *Journal of the American Statistical Association*, **85**, 376–386.
- [11] Friedman, J. H. and Rafsky, L. C. (1979) Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, **7**, 697–717.
- [12] Friedman, J. H. and Rafsky, L. C. (1981) Graphics for the Multivariate Two-Sample problem. *Journal of the American Statistical Association*, **76**, 277–287.
- [13] Justel, A., Pena, D. and Zamar, R. (1997) A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics and Probability Letters*, **35**, 251–259.
- [14] Koltchinskii, V. (1997) M-Estimation, Convexity and Quantiles. *The Annals of Statistics*, **25**, 435–477.
- [15] Leslie, J. R., Stephens, M. A. and Fotopoulos, S. (1986) Asymptotic Distribution of the Shapiro-Wilk W for Testing for Normality. *The Annals of Statistics*, **14**, 1497–1506.
- [16] Liu, R., Parelius, J. M. and Singh, K. (1999) Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference. *The Annals of Statistics*, **27**, 783–840.
- [17] Marden, J. (1998) Bivariate Q-Q plots and spider web plots. *Statistica Sinica*, **8**, 813–826.
- [18] Mottonen, J. and Oja, H. (1995) Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, **5**, 201–213.
- [19] Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.
- [20] Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.
- [21] Serfling, R. (2002) Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, **56**, 214–232.
- [22] Serfling, R. (2004) Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference*, **123**, 259–278.
- [23] Shapiro, S. S. and Wilk, M. B. (1965) Analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.