Bernoulli **20**(3), 2014, 1484–1506
DOI: 10.3150/13-BEJ530

Comparison of multivariate distributions using quantile–quantile plots and related tests

SUBHRA SANKAR DHAR[1], BIMAN CHAKRABORTY[2] and PROBAL CHAUDHURI[3]

[1] *Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur 208016, India. E-mail: dsubhra@gmail.com*
[2] *School of Mathematics, University of Birmingham, United Kingdom. E-mail: B.Chakraborty@bham.ac.uk*
[3] *Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata-700108, India. E-mail: probal@isical.ac.in*

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 develop and study a multivariate version of the Q–Q plot based on the spatial quantile. The usefulness of the proposed graphical device is illustrated on different real and simulated data, some of which have fairly large dimensions. We also develop certain statistical tests that are related to the proposed multivariate Q–Q plot and study their asymptotic properties. The performance of those tests are compared with that of some other well-known tests for multivariate distributions available in the literature.

Keywords: characterization of distributions; contiguous alternatives; Gaussian process; Pitman efficacy; spatial quantiles; tests for distributions; the level and the power of test

1. Introduction

The univariate quantile–quantile (Q–Q) plot is a diagnostic tool, which is widely used to assess the distributional similarities and differences between two independent samples (see, e.g., Gnanadesikan and Wilk [18], Gnanadesikan [17] and Chambers *et al.* [9]). As discussed in Doksum [12], Doksum and Sievers [13] and Koenker ([23], pages 31 and 32), there are some fundamental connections between the Q–Q plot and the two-sample problem involving a semiparametric treatment effect model. The Q–Q plot is also a popular device for checking the appropriateness of a specified probability distribution for a given univariate data. While the univariate Q–Q plot has a long history as a graphical tool for data analysis, there are only limited attempts in the literature to generalize the Q–Q plot for multivariate samples. One can construct the Q–Q plot for multivariate data using the marginal quantiles. However, a Q–Q plot based on the marginal quantiles fails to capture the nature of dependence among the marginals of a multivariate distribution. Such a Q–Q plot can only compare the marginal distributions, but it is inadequate for a proper comparison of two multivariate distributions because the marginal quantiles do not characterize a multivariate distribution (see the supplemental article (Dhar, Chakraborty and Chaudhuri [11]) for an illustrative example).

1350-7265 © 2014 ISI/BS

Breckling and Chambers [7], Chaudhuri [10] and Koltchinskii [24] extensively studied a multivariate quantile, which is popularly known as the spatial quantile. Koltchinskii ([24], Corollary 2.9, page 446) established that these spatial quantiles characterize multivariate distributions. In this article, we propose an extension of the Q–Q plot using the spatial quantiles for multivariate data. As we will see in subsequent sections, these Q–Q plots are in many ways natural generalizations of the univariate Q–Q plot. In particular, for a d-dimensional multivariate data, there will be d two-dimensional plots, where the points in each plot cluster around a straight line with slope $= 1$ and intercept $= 0$ *if and only if* the two multivariate distributions under comparison are identical.

Motivated by the one-sample Q–Q plot, Shapiro and Wilk [33] proposed a test for normality of univariate data. We also propose and study some statistical tests for multivariate distributions, which are related to our multivariate Q–Q plots. In our numerical and asymptotic studies, those tests turn out to have either comparable or superior performance when compared with the Kolmogorov–Smirnov and the Cramer–von Mises tests for multivariate distributions.

2. Multivariate Q–Q plots

Recall that a univariate Q–Q plot based on two samples with sizes n and m consists of r ($r = n+m$ if $n \neq m$ and $r = n$ if $n = m$) points in the two-dimensional plane, where for $i = 1, 2, \ldots, r$, the two coordinates of the ith point are the (i/r)th quantiles of the two samples. Here, in order to compare the quantiles, one has to match the quantiles of one data set with the corresponding quantiles of another data set. Easton and McCulloch [14] made an attempt to solve a similar matching problem for multivariate data. Their procedure was based on the permutation of the data that produced the minimum sum of the Euclidean distances between the matching data points in the two given samples. Consequently, in order to assess how well a specified probability distribution fits a given multivariate sample, they used a sample simulated from the specified distribution. The Q–Q plots proposed by them can be used in two-sample problems only if the two samples have the same size. In this paper, we use a matching procedure based on the spatial rank and the spatial quantile. The procedure is computationally simple and can be used in a two-sample problem even if the two samples do not have the same size. Further, in the case of a one-sample problem, where one tries to test whether a specified distribution fits the data well or not, the construction of our Q–Q plot does not require generation of a sample from the specified distribution.

The spatial rank of $\mathbf{z} \in \mathbb{R}^d$ with respect to the data cloud formed by the observations $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is defined as $n^{-1} \sum_{i:\mathbf{x}_i \neq \mathbf{z}} \|\mathbf{z} - \mathbf{x}_i\|^{-1}(\mathbf{z} - \mathbf{x}_i)$ (see, e.g., Möttönen and Oja [29], Chaudhuri [10] and Serfling [32]). For a random vector $\mathbf{x}$ with a probability distribution F on $\mathbb{R}^d$, the d-dimensional spatial quantile $Q_F(\mathbf{u}) = (Q_{F,1}(\mathbf{u}), \ldots, Q_{F,d}(\mathbf{u}))$ is defined as $Q_F(\mathbf{u}) = \arg\min_{Q \in \mathbb{R}^d} E\{\Phi(\mathbf{u}, \mathbf{x} - Q) - \Phi(\mathbf{u}, \mathbf{x})\}$ (see Chaudhuri [10] and Koltchinskii [24]). Here $\Phi(\mathbf{u}, \mathbf{s}) = \|\mathbf{s}\| + \langle \mathbf{u}, \mathbf{s} \rangle$, $\mathbf{u} \in B^d = \{\mathbf{v}\colon\ \mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\| < 1\}$, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, and $\| \cdot \|$ is the Euclidean norm induced by the inner product. For a random sample $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the empirical spatial quantile $Q_{\mathcal{X}}(\mathbf{u}) = (Q_{\mathcal{X},1}(\mathbf{u}), \ldots, Q_{\mathcal{X},d}(\mathbf{u}))$ is obtained by replacing F with its empirical version F_n. When different coordinate variables in a multivariate data are measured in different units, the spatial quantiles and the spatial ranks are usually computed after standardizing each coordinate variable appropriately. Note that when our objective is to compare the distributions of two random vectors $\mathbf{x}$ and $\mathbf{y}$, the problem is equivalent to comparing the distributions of $\mathbf{A}^{-1}\mathbf{x}$ and $\mathbf{A}^{-1}\mathbf{y}$, where $\mathbf{A}$ is any appropriate positive definite matrix used to standardize the variables.

We now consider a one-sample multivariate problem involving a d-dimensional data set $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, where $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d})$ has distribution F, and let F_0 be a specified probability distribution on $\mathbb{R}^d$. Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the 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 $k = 1, \dots, n$. Note that since $Q_{\mathcal{X}}(\mathbf{u}_k) = \mathbf{x}_k$, where $Q_{\mathcal{X}}(\mathbf{u}_k)$ is the $\mathbf{u}_k$th empirical spatial quantile of the data set $\mathcal{X}$, a natural way of matching the quantiles of the data set with those of the specified probability distribution will be by setting the correspondence between $\mathbf{x}_k$ and $Q_{F_0}(\mathbf{u}_k)$ (see also Marden [27,28]). Consider the set of points in $\mathbb{R}^2$ defined as $S_{n,i}(\mathcal{X}, F_0) = \{(x_{k,i}, Q_{F_0,i}(\mathbf{u}_k)){:}\ k = 1, \dots, n\}$, where $Q_{F_0,i}(\mathbf{u}_k)$ and $x_{k,i}$ are the ith components of $Q_{F_0}(\mathbf{u}_k)$ and $\mathbf{x}_k$, respectively, and $i = 1, \dots, d$. In particular, when $d = 1$, $S_{n,1}(\mathcal{X}, F_0)$ coincides with the set of points that form the univariate Q–Q plot for the one-sample problem. Theorem 2.1, stated below, ensures that for all $i = 1, \dots, d$, the points in the ith two-dimensional plot will lie close to a straight line with slope $= 1$ and intercept $= 0$ *if and only if* $F = F_0$.

Theorem 2.1. *Suppose that F_0 is a specified distribution having a positive density function, which is bounded on every bounded subset of $\mathbb{R}^d$ $(d \geq 2)$, and the same is true for F, the true distribution of the data. Assume that $S_{n,i}(\mathcal{X}, F_0)$ is constructed using the $\mathbf{u}_k$'s lying in any given closed ball in $\mathbb{R}^d$ with the center at the origin and the radius strictly smaller than one. Let $L(\varepsilon)$ be the collection of points that lie in an ε-neighborhood of a straight line with slope $= 1$ and intercept $= 0$. Then, for every $\varepsilon > 0$, we have*

$$\lim_{n\to\infty} P\left(\bigcap_{i=1}^{d}\left[S_{n,i}(\mathcal{X}, F_0) \subseteq L(\varepsilon)\right]\right) = 1,$$

if and only if $F = F_0$.

An implication of Theorem 2.1 is that the plots constructed using $S_{n,i}(\mathcal{X}, F_0)$ for $i = 1, \dots, d$ can be used, just like the univariate Q–Q plot, to determine whether the specified distribution F_0 fits the data well or not. In practice, F_0 may involve some unspecified parameters that need to be estimated from the data. For instance, there may be some unknown location and scatter parameters associated with F_0, and we can estimate them using standard techniques like the maximum likelihood method. In such a case, we can make an affine transformation of the data using the maximum likelihood estimates of the location and the scatter parameters. In view of the asymptotic consistency of the maximum likelihood estimate under appropriate conditions, the assertion in Theorem 2.1 about the linearity of the Q–Q plots remains valid if we construct the Q–Q plots using such transformed data, and the data are actually generated from F_0. One may also use other consistent estimates of the location and the scale parameters having high breakdown points (e.g., the minimum covariance determinant estimates; see Rousseeuw and Leroy [30]), which are robust against outliers. It will be appropriate to point out that Easton and McCulloch [14] also proposed an affine transformation of the data before constructing their Q–Q plots in the one-sample problem. Their proposal is not related in any way to the maximum likelihood estimation based on the specified distribution F_0, and it involves an iterative algorithm for computing the affine transformation. Easton and McCulloch [14] did not consider the case when the specified distribution involves unknown parameters other than the location and the scatter parameters. Any such parameter can be estimated by the maximum likelihood method using F_0 and the data.

We next consider the two-sample multivariate problem involving two independent d-dimensional data sets, namely, $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$ has distribution F, and $\mathbf{y}_j = (y_{j,1}, \ldots, y_{j,d})$ has distribution G. Suppose that $\mathbf{u}_1, \ldots, \mathbf{u}_n$ and $\mathbf{u}_{n+1}, \ldots, \mathbf{u}_{n+m}$ are the spatial ranks of these observations within their respective data sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. As in the case of the one-sample problem, $Q_{\mathcal{X}}(\mathbf{u}_k) = \mathbf{x}_k$ for $k = 1, \ldots, n$, and $Q_{\mathcal{Y}}(\mathbf{u}_k) = \mathbf{y}_k$ for $k = n+1, \ldots, n+m$. We compute $Q_{\mathcal{X}}(\mathbf{u}_k)$ for $k = n+1, \ldots, n+m$ and $Q_{\mathcal{Y}}(\mathbf{u}_k)$ for $k = 1, \ldots, n$ using the algorithm given in Chaudhuri ([10], pages 864 and 865). Then, we can match the two sets of quantiles by setting the correspondence between $Q_{\mathcal{X}}(\mathbf{u}_k)$ and $Q_{\mathcal{Y}}(\mathbf{u}_k)$ for $k = 1, \ldots, n+m$. As in the case of the one-sample problem, one may construct the Q–Q plots for the two-sample problem as a collection of d two-dimensional plots, where each plot corresponds to a component of the spatial quantile. Let $S_{n,m,i}(\mathcal{X}, \mathcal{Y}) = \{(Q_{\mathcal{X},i}(\mathbf{u}_k), Q_{\mathcal{Y},i}(\mathbf{u}_k))\colon\ k = 1, \ldots, (n+m)\}$, where $Q_{\mathcal{X},i}(\mathbf{u}_k)$ and $Q_{\mathcal{Y},i}(\mathbf{u}_k)$ are the ith components of $Q_{\mathcal{X}}(\mathbf{u}_k)$ and $Q_{\mathcal{Y}}(\mathbf{u}_k)$, respectively, and $i = 1, \ldots, d$. Note that when $d = 1$, our proposed multivariate matching coincides with the usual way of matching the univariate quantiles in a two-sample problem, and the points in $S_{n,m,1}(\mathcal{X}, \mathcal{Y})$ are same as those used in constructing the univariate two-sample Q–Q plot. Theorem 2.2, stated below, ensures that for all $i = 1, \ldots, d$, the points in the ith two-dimensional plot will lie close to a straight line with slope $= 1$ and intercept $= 0$ *if and only if* $F = G$.

Theorem 2.2. *Suppose that F and G have positive density functions, which are bounded on every bounded subset of $\mathbb{R}^d$ $(d \geq 2)$, and $S_{n,m,i}(\mathcal{X}, \mathcal{Y})$ is constructed using the $\mathbf{u}_k$'s lying in any given closed ball in $\mathbb{R}^d$ with the center at the origin and the radius strictly smaller than one. Further, let $L(\varepsilon)$ be the collection of points that lie in an ε-neighborhood of a straight line with slope $= 1$ and intercept $= 0$, and assume that $n, m \to \infty$ in such a way that $\lim_{n,m\to\infty} \frac{n}{(n+m)} = \lambda \in (0,1)$. Then, for every $\varepsilon > 0$, we have*

$$\lim_{n,m\to\infty} P\left(\bigcap_{i=1}^{d} \bigl[S_{n,m,i}(\mathcal{X}, \mathcal{Y}) \subseteq L(\varepsilon)\bigr]\right) = 1,$$

if and only if $F = G$.

In view of the equivariance of the spatial quantiles under location and homogeneous scale transformations, the assertions in Theorems 2.1 and 2.2 will also hold for the straight line with slope $= \sigma$ and intercept $= \mu_i$ $(i = 1, \ldots, d)$ *if and only if* $F(\mathbf{x}) = F_0((\mathbf{x} - \boldsymbol{\mu})/\sigma)$ and $F(\mathbf{x}) = G((\mathbf{x} - \boldsymbol{\mu})/\sigma)$, respectively, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$ and $\sigma > 0$.

We now briefly discuss some earlier attempts to develop graphical tools for comparing multivariate distributions. For bivariate data, Marden [27,28] proposed a version of the Q–Q plot, which is based on drawing arrows from the spatial quantiles 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). However, such an arrow plot can be drawn only for a bivariate data. Also, when the two samples are related to each other by a location and a homogeneous scale transformation, such arrow plots cannot detect that unlike our Q–Q plots. Friedman and Rafsky [16] proposed a different visualization procedure for comparing the distributions of two multivariate samples. Their methodology is based on the idea of a minimal spanning tree. Liu, Parelius and Singh [26] proposed an alternative visualization device called the DD-plot for comparing two multivariate data sets based on the concept of data depth. However, none of these graphical tools developed by Marden [27,28], Friedman and Rafsky [16] and Liu, Parelius and Singh [26] will coincide with the usual univariate Q–Q plot when they are applied to the univariate data, and none of them can be taken as a natural multivariate extension of the univariate Q–Q plot.

3. Tests for comparing multivariate distributions

For each two-dimensional plot in our Q–Q plots, the overall deviation of the points from the straight line with slope $= 1$ and intercept $= 0$ can be measured by $\int\{Q_{\mathcal{X},i}(\mathbf{u}) - Q_{F_0,i}(\mathbf{u})\}^2\,d\mathbf{u}$ and $\int\{Q_{\mathcal{X},i}(\mathbf{u}) - Q_{\mathcal{Y},i}(\mathbf{u})\}^2\,d\mathbf{u}$ for the one-sample and the two-sample problems, respectively, where $i = 1, \ldots, d$. These deviations in d different plots can be aggregated as $\sum_{i=1}^{d}\int\{Q_{\mathcal{X},i}(\mathbf{u}) - Q_{F_0,i}(\mathbf{u})\}^2\,d\mathbf{u} = \int\|Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\|^2\,d\mathbf{u}$ and $\sum_{i=1}^{d}\int\{Q_{\mathcal{X},i}(\mathbf{u}) - Q_{\mathcal{Y},i}(\mathbf{u})\}^2\,d\mathbf{u} = \int\|Q_{\mathcal{X}}(\mathbf{u}) - Q_{\mathcal{Y}}(\mathbf{u})\|^2\,d\mathbf{u}$ for the one-sample and the two-sample problems, respectively. These aggregated quantities can be taken as the total deviations in our Q–Q plots. These measures of total deviations can be used to construct tests for comparing multivariate distributions. Such tests will be rotationally invariant in view of the rotational equivariance of the spatial quantiles.

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ consist of i.i.d. observations from an unknown distribution F having a density function, which is assumed to be bounded on every bounded subset of $\mathbb{R}^d$ $(d \geq 2)$. Suppose that we want to test H_0: $F = F_0(\Leftrightarrow Q_F(\mathbf{u}) = Q_{F_0}(\mathbf{u})$ for all $\mathbf{u} \in B^d)$ against the alternative H_1: $F \neq F_0(\Leftrightarrow Q_F(\mathbf{u}) \neq Q_{F_0}(\mathbf{u})$ for some $\mathbf{u} \in B^d)$, where F_0 is a specified distribution having a density function, which is bounded on every bounded subset of $\mathbb{R}^d$ $(d \geq 2)$. In order to test H_0 against H_1, we can use the test statistic $V_n = n\int\|Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\|^2\,d\mathbf{u}$, where the integral is over a closed ball with the center at the origin and the radius strictly smaller than one. Note that the test statistic V_n (as well as the test statistic $T_{n,m}$ considered later in this section) can be viewed as the sum of the arrow lengths in the arrow plot considered by Marden [27] for a bivariate data.

Consider now a multivariate Gaussian process $Z_1(\mathbf{u})$ having zero mean and the covariance kernel

$$k_1(\mathbf{u}_1, \mathbf{u}_2) = \left[D_1\left\{Q_{F_0}(\mathbf{u}_1)\right\}\right]^{-1}\left[D_2\left\{Q_{F_0}(\mathbf{u}_1), Q_{F_0}(\mathbf{u}_2), \mathbf{u}_1, \mathbf{u}_2\right\}\right]\left[D_1\left\{Q_{F_0}(\mathbf{u}_2)\right\}\right]^{-1}.$$

Here $D_1\{Q_{F_0}(\mathbf{u})\} = E_{F_0}[\|\mathbf{x} - Q_{F_0}(\mathbf{u})\|^{-1}\{I_d - \|\mathbf{x} - Q_{F_0}(\mathbf{u})\|^{-2}(\mathbf{x} - Q_{F_0}(\mathbf{u}))(\mathbf{x} - Q_{F_0}(\mathbf{u}))^T\}]$, $D_2\{Q_{F_0}(\mathbf{u}), Q_{F_0}(\mathbf{v}), \mathbf{u}, \mathbf{v}\} = E_{F_0}[\{\|\mathbf{x} - Q_{F_0}(\mathbf{u})\|^{-1}(\mathbf{x} - Q_{F_0}(\mathbf{u})) + \mathbf{u}\}\{\|\mathbf{x} - Q_{F_0}(\mathbf{v})\|^{-1}(\mathbf{x} - Q_{F_0}(\mathbf{v})) + \mathbf{v}\}^T]$. Henceforth, I_d denotes the $d \times d$ identity matrix, all vectors are assumed to be column vectors, and the superscript T denotes the transpose of a vector. Let $\mathcal{V} = \int\|Z_1(\mathbf{u})\|^2\,d\mathbf{u}$, where the integral is over the same closed ball as in the definition of V_n. We now state a theorem describing the asymptotic behaviour of the test based on V_n.

Theorem 3.1. *Let* $c_1(\alpha)$ *be the* $(1-\alpha)$*th quantile* $(0 < \alpha < 1)$ *of the distribution of* $\mathcal{V}$. *A test, which rejects* H_0 *for* $V_n > c_1(\alpha)$, *will have asymptotic size* α. *Further, when* H_1 *is true, the asymptotic power of the test will be one if the integral defining* V_n *is taken over an appropriately large closed ball in* $\mathbb{R}^d$.

In order to implement our test, we need to compute V_n, and we have approximated the integral that appears in this test statistic by an average of the integrand over 1000 i.i.d. Monte Carlo replications obtained from the random generations of $\mathbf{u}$ from the uniform distribution on a closed ball with the center at the origin and the radius $= 0.99$. In view of the asymptotic Gaussian distribution of the process $\sqrt{n}\{Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}$ under H_0 and the well-known orthogonal decomposition of a finite-dimensional multivariate normal distribution, the distribution of the test statistic V_n under H_0: $F = F_0$ can be approximated by a weighted sum of chi-square random variables each with one degree of freedom. In our numerical work, we have computed $c_1(\alpha)$ by generating 1000 Monte Carlo replications from a weighted sum of chi-square variables, where the weights are the eigenvalues of the covariance matrices of appropriate normal random vectors. Note that the covariance matrices involve the spatial quantiles and certain expectations under the specified distribution F_0, and those can be computed numerically.

Let us next consider a two-sample problem with two independent sets of i.i.d. observations $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ from the distributions F and G, respectively. We assume the same conditions on the density functions of F and G as for the density functions of F and F_0 in the one-sample problem discussed above. In this two-sample problem, our hypotheses are H_0^*: $F = G (\Leftrightarrow Q_F(\mathbf{u}) = Q_G(\mathbf{u})$ for all $\mathbf{u} \in B^d)$ and H_1^*: $F \neq G (\Leftrightarrow Q_F(\mathbf{u}) \neq Q_G(\mathbf{u})$ for some $\mathbf{u} \in B^d)$. In order to test H_0^* against H_1^*, one can use the test statistic $T_{n,m} = (n+m) \int \|Q_{\mathcal{X}}(\mathbf{u}) - Q_{\mathcal{Y}}(\mathbf{u})\|^2 \, d\mathbf{u}$, where the integral is over a closed ball with the center at the origin and the radius strictly smaller than one.

Let $Z_2(\mathbf{u})$ be a multivariate Gaussian process having zero mean and the covariance kernel

$$k_2(\mathbf{u}_1, \mathbf{u}_2) = \frac{[D_1\{Q_F(\mathbf{u}_1)\}]^{-1}[D_2\{Q_F(\mathbf{u}_1), Q_F(\mathbf{u}_2), \mathbf{u}_1, \mathbf{u}_2\}][D_1\{Q_F(\mathbf{u}_2)\}]^{-1}}{\lambda(1-\lambda)},$$

where λ is as defined in the statement of Theorem 2.2, and D_1, D_2 are as defined before the statement of Theorem 3.1. Define $\mathcal{T} = \int \|Z_2(\mathbf{u})\|^2 \, d\mathbf{u}$, where the integral is over the same closed ball as in the definition of $T_{n,m}$. We now state a theorem describing the asymptotic behaviour of the test based on $T_{n,m}$.

Theorem 3.2. *Let* $c_2(\alpha)$ *be the* $(1-\alpha)$*th quantile* $(0 < \alpha < 1)$ *of the distribution of* $\mathcal{T}$. *A test, which rejects* H_0^* *for* $T_{n,m} > c_2(\alpha)$, *will have asymptotic size* α. *Further, when* H_1^* *is true, the test will have asymptotic power one if the integral defining* $T_{n,m}$ *is taken over an appropriately large closed ball in* $\mathbb{R}^d$.

For numerical implementation, one can compute $T_{n,m}$ and $c_2(\alpha)$ for the two-sample problem in a similar way as we have computed V_n and $c_1(\alpha)$, respectively, in the one-sample problem. However, here we have estimated the unknown quantities (i.e., the spatial quantiles and certain expectations under H_0^*) appearing in the covariance kernel based on the combined sample of the $\mathbf{x}$'s and the $\mathbf{y}$'s. In Sections 5 and 6, we have compared the performance of our tests with that of the Kolmogorov–Smirnov and the Cramer–von Mises tests for multivariate distributions. For numerical implementation, we have used R codes that are available from the first author of the paper.

4. Demonstration of multivariate Q–Q plots using simulated and real data

We begin with the one-sample problem and consider two simulated data sets each consisting of 100 i.i.d. observations. The observations in the first set were generated from the trivariate normal distribution having zero mean and scatter matrix $\Sigma = ((\sigma_{ij}))_{1\le i,j\le 3}$ with $\sigma_{i,i} = 1$ for $i = 1, 2, 3$, $\sigma_{1,2} = 0.5$, $\sigma_{1,3} = 0.2$ and $\sigma_{2,3} = 0.3$. For the second set, the observations were generated from the trivariate Laplace distribution with p.d.f. $f(\mathbf{x}) = (1/8\pi)\exp^{-\|\mathbf{x}\|}$. For both of them, we considered the trivariate normal distribution as the specified distribution F_0 with unknown parameters $\boldsymbol{\mu}$ and Σ. Following the remarks after Theorem 2.1, $\boldsymbol{\mu}$ and Σ were estimated from each data set using the sample mean vector and the sample dispersion matrix, respectively, which are the maximum likelihood estimates in this case. We standardized the data sets using these estimates and compared the spatial quantiles of the standardized data with those of the standard trivariate normal distribution. We computed the spatial quantiles for standard trivariate normal distributions using the results in Marden ([27], pages 824 and 825). The Q–Q plots for the two simulated data sets are displayed in Figure 1.

It is clearly evident from the plots in the first row of Figure 1 that the specified distribution fits the data well as the points in those plots are tightly clustered around the straight line with

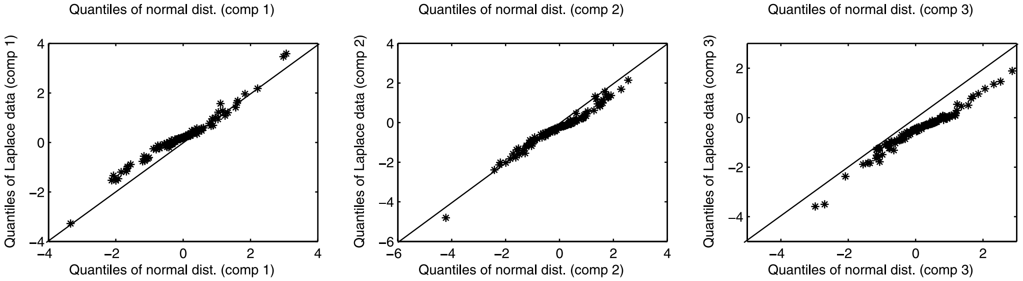

Figure 1. The Q–Q plots for the one-sample examples, where the specified distribution is trivariate normal. The plots in the first and the second rows are for the examples, where the distributions of the data are trivariate normal and trivariate Laplace, respectively.

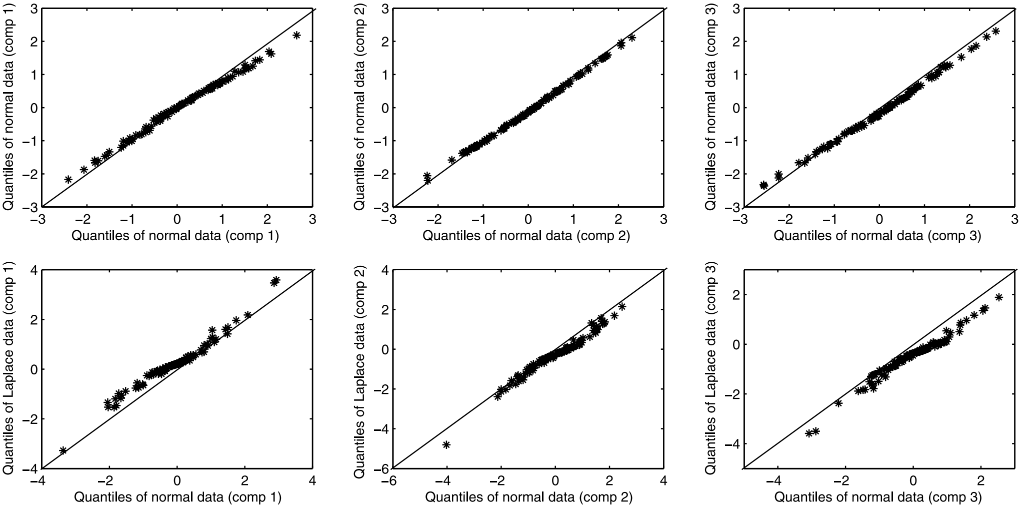

Figure 2. The Q–Q plots for the two-sample problem. The plots in the first row for an example, where the samples are generated from the same distribution, and those in the second row for an example, where the samples are generated from different distributions.

slope $= 1$ and intercept $= 0$. On the other hand, in each Q–Q plot in the second row, the points are significantly deviating from the straight line with slope $= 1$ and intercept $= 0$, and the points are actually clustered around a nonlinear curve. We have also computed the p-values for the one-sample test discussed in Section 3 for testing H_0: $F = F_0$ against H_1: $F \neq F_0$ for these two simulated data sets. We have obtained a high p-value $= 0.784$ for the first sample whereas the p-value for the second example is 0.049, which is quite small.

We next consider two simulated data sets to demonstrate our Q–Q plots for the two-sample problem. In both the data sets, the distribution of the first sample F was chosen to be the standard trivariate normal distribution while G, the distribution of the second sample, was taken to be the standard trivariate normal in one set and the trivariate Laplace distribution in the other set. The size of each sample was 100. The Q–Q plots for the two data sets are displayed in Figure 2. In each plot in the first row of Figure 2, the points are tightly clustered around the straight line with slope $= 1$ and intercept $= 0$. On the other hand, the points are significantly deviating from the straight line with slope $= 1$ and intercept $= 0$ in each plot in the second row of Figure 2. We also carried out the two-sample test described in Section 3 for testing H_0^*: $F = G$ against H_1^*: $F \neq G$, and we obtained a high p-value $= 0.731$ for the first data set whereas a small p-value $= 0.048$ was obtained for the second data set.

4.1. Detection of special features using multivariate Q–Q plots

We now consider a two-sample problem, where the first sample consists of 100 i.i.d. observations from the standard trivariate normal distribution (F), and the second sample consists of

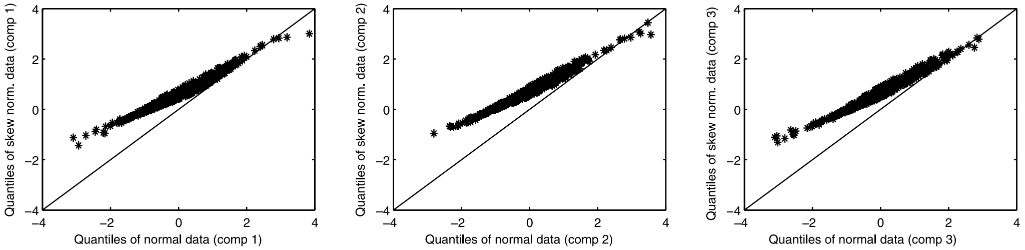

Figure 3. The Q–Q plots for the two-sample problem, where the first sample is generated from the standard trivariate normal distribution, and the second sample is generated from a trivariate skew-normal distribution.

100 i.i.d. observations from a trivariate skew-normal distribution (G) (see Azzalini and Dalla Valle ([4], page 717)). The p.d.f. of the trivariate skew-normal distribution is given by $f(\mathbf{z}) = 2\phi_3(\mathbf{z};\Omega)\Phi(\alpha^T\mathbf{z})$, where $\mathbf{z} \in \mathbb{R}^3$, $\alpha^T = \frac{\lambda^T\Psi^{-1}\Delta^{-1}}{\sqrt{1+\lambda^T\Psi^{-1}\lambda}}$, $\Delta = \mathrm{diag}(\sqrt{1-\delta_1^2}, \sqrt{1-\delta_2^2}, \sqrt{1-\delta_3^2})$, $\lambda = (\frac{\delta_1}{\sqrt{1-\delta_1^2}}, \frac{\delta_2}{\sqrt{1-\delta_2^2}}, \frac{\delta_3}{\sqrt{1-\delta_3^2}})^T$, and $\Omega = \Delta(\Psi + \lambda\lambda^T)\Delta$. Here $\phi_3(\mathbf{z};\Omega)$ denotes the p.d.f. of a trivariate normal distribution with standardized marginals and correlation matrix Ω, and Φ is the distribution function of the standard univariate normal distribution. In this study, we have considered $\delta_1 = \delta_2 = \delta_3 = 0.9$ and $\Psi = I_d$. The Q–Q plots for this two-sample problem are displayed in Figure 3, and we see a heavier tail in one direction in each plot in this figure. This is an indication that one sample is generated from a more skewed distribution than the other. Also, the small p-value $= 0.048$ obtained using our two-sample test for testing H_0^*: $F = G$ against H_1^*: $F \neq G$ implies that the two distributions are significantly different in this data set.

We next consider an example to demonstrate how our Q–Q plots can be used to detect outliers present in the data. We again consider a two-sample problem, where the first sample consists of 100 i.i.d. observations from the standard trivariate normal distribution. The second sample consists of 97 i.i.d. observations from the standard trivariate normal distribution and the remaining three data points in the sample are $(10, 10, 10)$, $(9, 9, 9)$ and $(8, 8, 8)$. The Q–Q plots for this data set are displayed in Figure 4. The presence of three outliers in the second sample is clearly indicated by the plots in Figure 4.

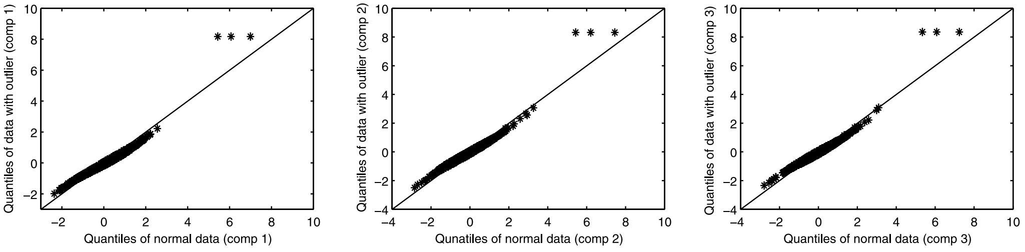

Figure 4. The Q–Q plots for the two-sample problem, where the first sample is generated from the standard trivariate normal distribution, and the second sample contains some outliers.

4.2. Analysis of real data

We first consider Fisher's *Iris data*, which is available in http://archive.ics.uci.edu/ml. In this data, there are three multivariate samples corresponding to three different varieties of Iris, namely, *Iris setosa*, *Iris virginica* and *Iris versicolor*. Each sample has size 50. In each sample, there are four measurements, namely, the sepal length, the sepal width, the petal length and the petal width. We would like to determine how close is the distribution of each sample to a four-dimensional normal distribution. This can be formulated as a one-sample problem, where F is the distribution of a sample, and the four-dimensional normal distribution is our specified distribution F_0. Note that F_0 involves an unknown mean $\boldsymbol{\mu}$ and an unknown dispersion Σ. For each species, following the remarks after Theorem 2.1, we estimated $\boldsymbol{\mu}$ and Σ by the sample mean vector and the sample dispersion matrix, which are maximum likelihood estimates. Then we standardized the data in each sample using the corresponding sample mean vector and the corresponding sample dispersion matrix. The Q–Q plots in Figure 5 were constructed using the spatial quantiles of a standardized sample and the spatial quantiles of the standard four-dimensional normal distribution.

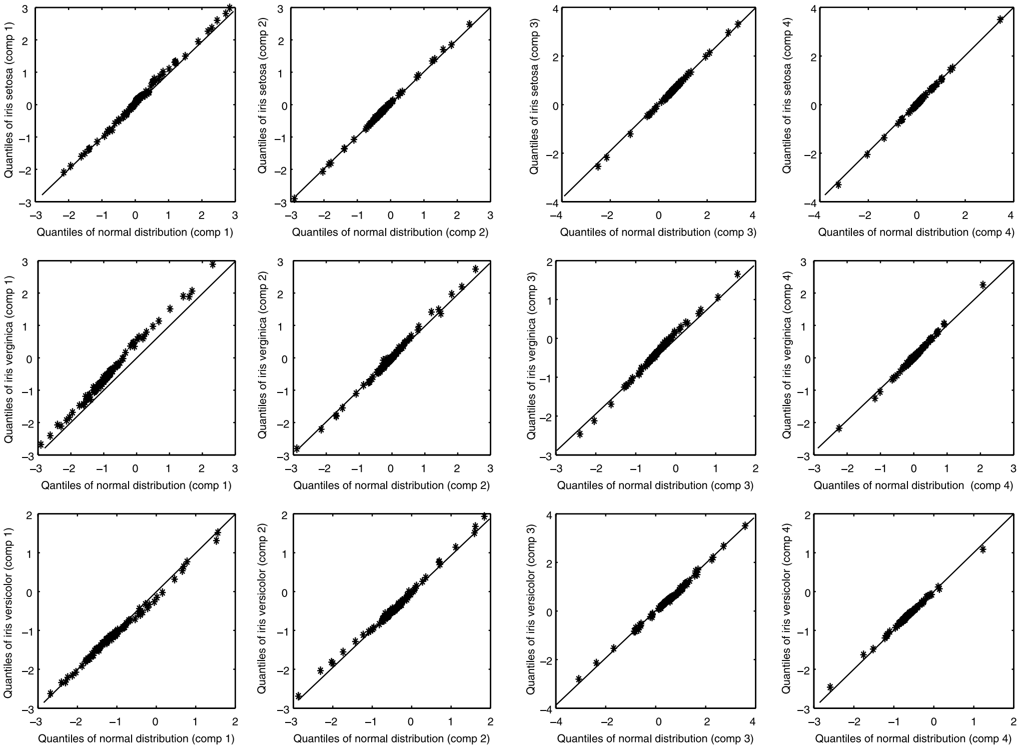

Figure 5. The Q–Q plots for *Iris setosa* (first row), *Iris virginica* (second row) and *Iris versicolor* (third row).

It is visible in the plots in Figure 5 that in almost all cases, the points are tightly clustered around the straight line with slope = 1 and intercept = 0 except in the first plot for *Iris virginica*, where the points deviate to some extent from that straight line. Our one-sample test for testing H_0: $F = F_0$ against H_1: $F \neq F_0$ led to very high p-values, namely, 0.841, 0.413 and 0.582 for *Iris setosa*, *Iris virginica* and *Iris versicolor*, respectively. These p-values imply that H_0 is to be accepted, and multivariate normal distributions seem to fit the data well for all three Iris species.

Our next real data set is the *Vertebral Column data*, which is available in http://archive.ics.uci.edu/ml/datasets/Vertebral+Column. This data set contains six variables on 310 patients, who belong to two groups. Among the 310 patients, 100 are normal, and the remaining 210 of them are abnormal. We view it as a two-sample problem with F as the distribution of the measurements corresponding to the normal patients, and G as the distribution of the measurements corresponding to the abnormal patients. In this study, we considered only two variables, namely, the pelvic incidence and the pelvic tilt as these two pelvic parameters are strongly associated with the severity and the stiffness of lumbosacral spondylolisthesis. Both the pelvic incidence and the pelvic tilt are angles and measured in the same unit, and no standardization of the data is necessary in order to compute and compare the spatial quantiles of these two samples. In Figure 6, we display the Q–Q plots for this data. The points in the Q–Q plots are clearly not clustered around any straight line. In fact, most of the points in each plot lie on a stretched S-shaped curve, which indicates that the distribution G associated with the abnormal patients has heavier tails than the distribution F associated with the normal patients. The p-value obtained using the two-sample test for testing H_0^*: $F = G$ against H_1^*: $F \neq G$ is 0.038, which also indicates that the two distributions are significantly different.

The third real data set that we consider is the *Monthly Sunspot number data*, which is available in http://www.ngdc.noaa.gov/stp/solar/ssndata.html. This data set contains monthly average number of sunspots during the period of 1749 to 2009. As data for 1749 and 2009 are incomplete, we have carried out our analysis on the observations for the remaining 259 (1750 to 2008) 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. 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. We treat this as a two-sample problem, where

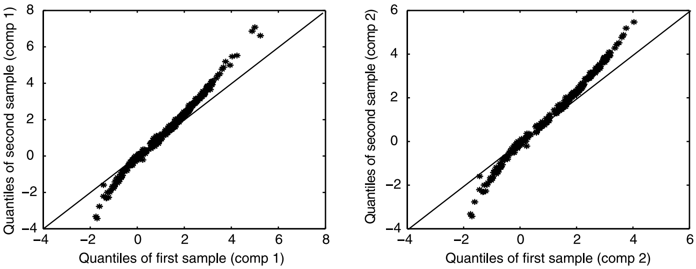

Figure 6. The Q–Q plots for the vertebral column data.

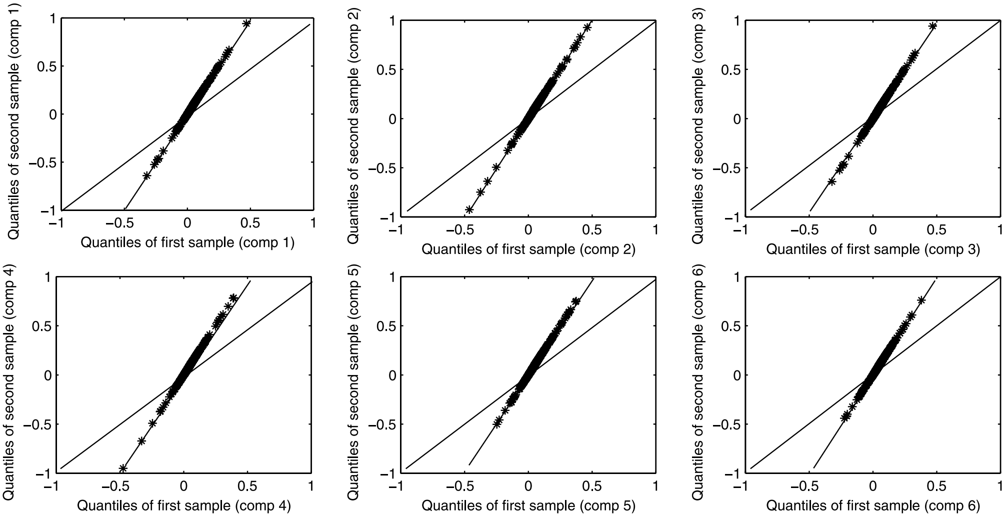

Figure 7. The Q–Q plots for the monthly sunspot number data.

F and G are the distributions corresponding to the sunspot numbers during the periods October–March and April–September, respectively. The Q–Q plots for the data are presented in Figure 7. In each of the plots, the points lie very close to a straight line with slope $= 2$ and intercept $= 0$. In view of the remark after Theorem 2.2, these plots indicate that the distributions F and G are related by the equation $F(\mathbf{x}) = G(\mathbf{x}/2)$. Hence, the two multivariate samples corresponding to the two periods October–March and April–September have distributions that differ only in the scales of the variables. The two distributions have the same location, and one distribution can be obtained from the other by a scale transformation using the scale factor 2. This fact was further confirmed when we carried out some alternative statistical analysis of the data such as the comparison of the marginal quantiles and the direct comparison of the means and the variances of the variables.

4.3. Multivariate Q–Q plots for data with large dimensions

When the dimension of the data is large, there will be too many two-dimensional plots, and it will be inconvenient to display and visually examine all of them. In that case, one can plot $(l, Q_{\mathcal{X},l}(\mathbf{u}_k) - Q_{\mathcal{Y},l}(\mathbf{u}_k))$ for $k = 1, \ldots, (n + m)$ and $l = 1, \ldots, d$ in a single two-dimensional plot with d vertical lines parallel to one another. We next demonstrate this procedure on some simulated and real data sets.

First, we consider a two-sample problem, where the data in each sample consists of 10 i.i.d. observations from a standard Brownian motion with its mean function $m(t) = 0$ and covariance kernel $k(s, t) = \min(s, t)$, where $s, t \in [0, 1]$ (note that $F = G$ here). For our second data set, one sample consists of 10 i.i.d. observations from a standard Brownian motion with its mean function

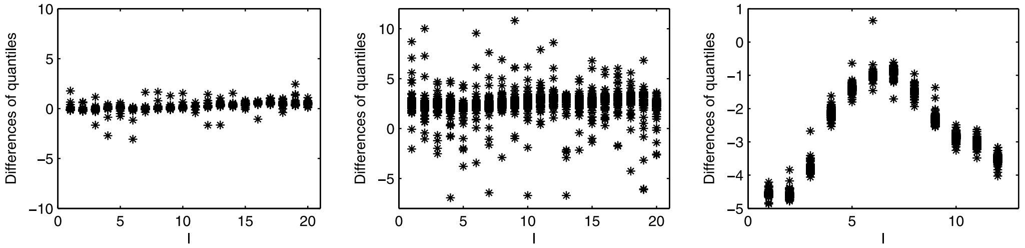

Figure 8. The quantile difference plots for the data on Brownian motions and sea level pressures.

$m_1(t) = 0$ and covariance kernel $k_1(s,t) = \min(s,t)$ as before (i.e., we have the same F as before). However, the second sample in the second data consists of 10 i.i.d. observations from a Brownian motion with its mean function $m_2(t) = 2$ and covariance kernel $k_2(s,t) = 2\min(s,t)$ (which corresponds to the distribution G). In our study, we considered equally spaced points $t_1, \ldots, t_{20}$ in [0, 1] and sampled the observations at those time points.

The fourth real data set that we consider is the *Sea Level Pressures data*, which is available in http://www.cpc.noaa.gov/data/indices/darwin and http://www.cpc.noaa.gov/data/indices/tahiti. This data set consists of monthly sea level pressures from two different islands in the southern Pacific ocean, namely, Darwin (13°S, 131°E) and Tahiti (17°S, 149°W) during the period 1850–2008. Thus, we have a two-sample problem with each sample corresponding to an island and containing 159 twelve-dimensional observations. Here F and G are the distributions of the multivariate observations corresponding to the two islands. For this data, each data point corresponds to a year, and each coordinate of a data point corresponds to an observation in a particular month.

The plots of the quantile differences for the above three data sets are displayed in Figure 8. In the first plot in Figure 8, the points in each vertical line are tightly clustered around a horizontal straight line passing through the origin, which indicates that the samples are obtained from similar distributions. It is further confirmed by the large p-value $= 0.623$ obtained using our two-sample test for testing H_0^*: $F = G$ against H_1^*: $F \neq G$. On the other hand, the difference in the locations and the scales of the two distributions F and G are clearly visible in the second plot in Figure 8. The p-value obtained using our two-sample test in this case is 0.042, which indicates significant difference between the two distributions and strong support in favour of H_1^*: $F \neq G$. It is also amply indicated by the third plot in Figure 8 as well as the small p-value $= 0.045$ obtained using our two-sample test that the distributions F and G for the two samples corresponding to the two islands Darwin and Tahiti are significantly different.

5. Finite sample level and power study for different tests

Here we carry out some simulation studies to compare our tests with the well-known multivariate extensions of the Kolmogorov–Smirnov (KS) and the Cramer–von Mises (CVM) tests (see, e.g., Burke [8] and Justel, Peña and Zamar [20]) in the one-sample and the two-sample problems. For testing H_0: $F = F_0$ against H_1: $F \neq F_0$, the KS and the CVM test statistics are $T_n^{(1)} = \sup_{\mathbf{x}\in\mathbb{R}^d}\sqrt{n}|F_n(\mathbf{x}) - F_0(\mathbf{x})|$ and $T_n^{(2)} = n\int_{\mathbf{x}\in\mathbb{R}^d}[F_n(\mathbf{x}) - F_0(\mathbf{x})]^2\,\mathrm{d}F_0(\mathbf{x})$, respectively, where $F_n(\mathbf{x})$ is the empirical version of $F(\mathbf{x})$. To test H_0^*: $F = G$ against H_1^*: $F \neq G$, the KS and the CVM test statistics are $T_{n,m}^{(1)} = \sup_{\mathbf{x}\in\mathbb{R}^d}\sqrt{n+m}|F_n(\mathbf{x}) - G_m(\mathbf{x})|$ and $T_{n,m}^{(2)} = (n+m)\int_{\mathbf{x}\in\mathbb{R}^d}[F_n(\mathbf{x}) - G_m(\mathbf{x})]^2\,\mathrm{d}M_{(n,m)}(\mathbf{x})$, respectively, where $(n+m)M_{(n,m)}(\mathbf{x}) = nF_n(\mathbf{x}) + mG_m(\mathbf{x})$, and F_n and G_m are the empirical versions of F and G, respectively. The KS and the CVM tests for multivariate data can be implemented using the asymptotic distributions of the corresponding test statistics.

For the one-sample problem, we have considered $F_0 = N_d$ and $F = (1-\beta)N_d + \beta C_d$ and $(1-\beta)N_d + \beta L_d$. Here $\beta \in [0, 1]$, N_d, L_d and C_d are the d-dimensional standard normal distribution, the d-dimensional Laplace distribution with p.d.f. $f(\mathbf{x}) = (\Gamma(d/2)/2\Gamma(d)\pi^{d/2})\exp^{-\|\mathbf{x}\|}$ and the d-dimensional Cauchy distribution with p.d.f. $f(\mathbf{x}) = (\Gamma((d+1)/2)/\sqrt{\pi}\Gamma(d/2))(1 + \|\mathbf{x}\|^2)^{-(d+1)/2}$, respectively. In the case of the two-sample problem, we have considered $F = N_d$ and $G = (1-\beta)N_d + \beta C_d$ and $(1-\beta)N_d + \beta L_d$.

In Figure 9, we have plotted the ratio between the empirical power of our test (numerator) and that of another test (denominator) for different values of the parameter β. It is evident from Figure 9 that our test is significantly more powerful than the KS test in all the cases considered in our simulation study. However, the CVM test performs better than our test in some cases, and our test outperforms the CVM test in some other cases.

Friedman and Rafsky [15] proposed a multivariate generalization of the Wald–Wolfowitz run test using the idea of minimum spanning tree (the MST-run test). We have compared the empirical powers of our two-sample test with those of the MST-run test for $F = N_d(\mathbf{0}, I_d)$ and

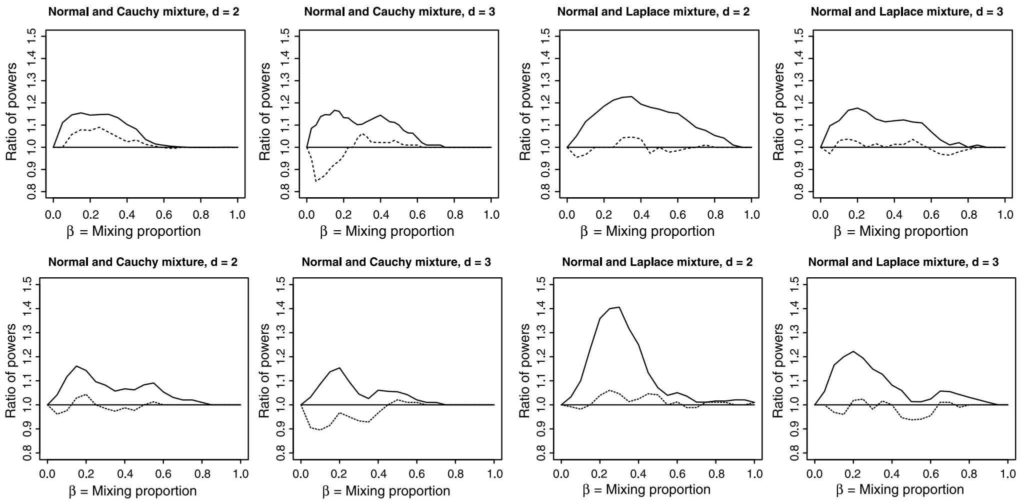

Figure 9. The graphs of the ratios of empirical powers based on 1000 Monte Carlo replications at 5% nominal level. The numerator in each ratio is the power of our test while the denominators of the ratios corresponding to the *solid* and the *dotted* curves are the powers of the KS and the CVM tests, respectively. The first row corresponds to the one-sample problem with $n = 10$, and the second row corresponds to the two-sample problem with $n = m = 10$.

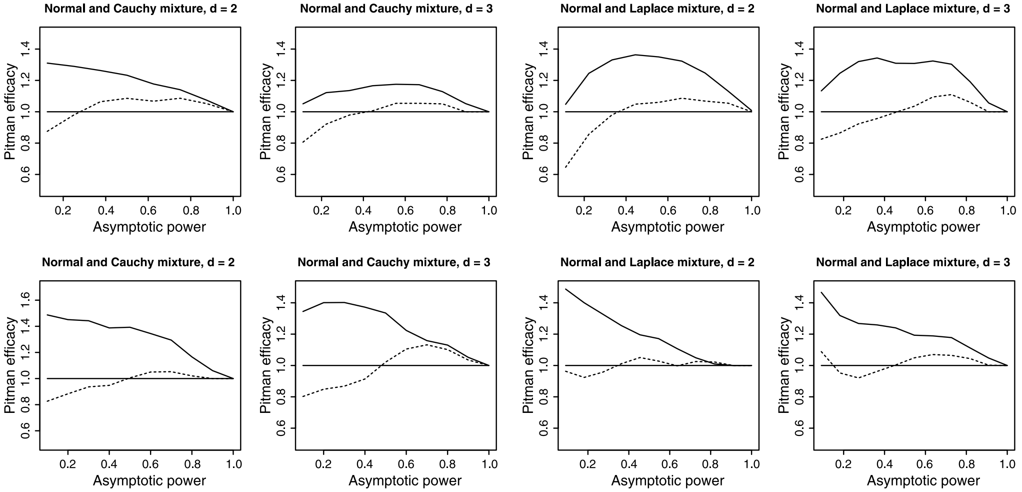

Figure 10. The Pitman efficacy of our test relative to the KS test (*solid* curve) and the CVM test (*dotted* curve) at 5% nominal level. The first row corresponds to the one-sample problem, and the second row corresponds to the two-sample problem.

$G = N_d(d^{-1/2}\Delta\mathbf{1}_d, \sigma I_d)$, where $N_d(\boldsymbol{\mu}, \Sigma)$ is the d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and dispersion Σ, $\mathbf{1}_d$ is the d-dimensional vector of 1's, and the values of Δ and σ are chosen as in Friedman and Rafsky ([16], page 706). For sample sizes $n = m = 100$ and 5% nominal level, the results are reported in Table 1, and it is clear that the MST-run test has inferior performance compared to our test.

For univariate data, our proposed tests in the one-sample and the two-sample problems lead to new tests that have previously not been considered in the literature. In addition to the KS and the CVM tests, there are several other tests that are available in the literature (see, e.g., Shapiro and Wilk [33], Anderson and Darling [3] and Ahmad [1,2]) for comparing the distributions of univariate data in the one-sample and the two-sample problems. We have discussed and compared

Table 1. Comparison of the empirical powers based on 100 Monte Carlo replications of our two-sample test and the MST-run test in different dimensions

	$d = 2$	$d = 5$	$d = 10$	$d = 20$
	$\Delta = 0.5, \sigma = 1$	$\Delta = 0.75, \sigma = 1$	$\Delta = 1.0, \sigma = 1$	$\Delta = 1.2, \sigma = 1$
Our test	0.55	0.70	0.83	0.99
MST-run test	0.35	0.64	0.78	0.86
	$\Delta = 0, \sigma = 1.2$	$\Delta = 0, \sigma = 1.2$	$\Delta = 0, \sigma = 1.1$	$\Delta = 0, \sigma = 1.075$
Our test	0.17	0.26	0.07	0.14
MST-run test	0.14	0.21	0.09	0.13

the performance of these tests for univariate data in detail in the supplemental article (see Dhar, Chakraborty and Chaudhuri [11]).

6. Asymptotic power study under contiguous alternatives

Since our tests, the KS and the CVM tests are all asymptotically consistent, a natural question is how the asymptotic powers of our tests and the KS and the CVM tests compare with one another under contiguous alternatives (see Hájek and Šidák [19]). In the case of the one-sample problem, the null hypothesis is given by H_0: $F(\mathbf{x}) = F_0(\mathbf{x})$, and we consider a sequence of contiguous alternatives H_n: $F(\mathbf{x}) = (1 - \gamma/\sqrt{n})F_0(\mathbf{x}) + (\gamma/\sqrt{n})H(\mathbf{x})$ for a fixed $\gamma > 0$ and $n = 1, 2, \ldots$. Consider a multivariate Gaussian process $Z_1'(\mathbf{u})$ with the mean function

$$m_1(\mathbf{u}) = \gamma\big[D_1\big\{Q_{F_0}(\mathbf{u})\big\}\big]^{-1} E_H\left\{\frac{\mathbf{x} - Q_{F_0}(\mathbf{u})}{\|\mathbf{x} - Q_{F_0}(\mathbf{u})\|} + \mathbf{u}\right\}$$

and the covariance kernel $k_1(\mathbf{u}_1, \mathbf{u}_2)$, where $k_1(\mathbf{u}_1, \mathbf{u}_2)$ is as defined before Theorem 3.1. Let $\mathcal{V}' = \int \|Z_1'(\mathbf{u})\|^2\, \mathrm{d}\mathbf{u}$, where the integral is over the same closed ball as in the definition of V_n in Section 3. We now state a theorem describing the asymptotic powers of the test based on V_n as well as the KS and the CVM tests under contiguous alternatives.

Theorem 6.1. *Assume that F_0 and H have continuous and positive densities f_0 and h, respectively, on $\mathbb{R}^d (d \geq 2)$, and $E_{F_0}\{\frac{h(\mathbf{x})}{f_0(\mathbf{x})} - 1\}^2 < \infty$. Then, the sequence of alternatives H_n form a contiguous sequence. Under such alternatives, the asymptotic power of the test based on V_n is given by $P_\gamma[\mathcal{V}' > c_1(\alpha)]$, where $c_1(\alpha)$ is as defined in Theorem 3.1 such that $P_{\gamma=0}[\mathcal{V}' > c_1(\alpha)] = \alpha$. Further, under those alternatives, the asymptotic powers of the tests based on $T_n^{(1)}$ and $T_n^{(2)}$ are given by $P_\gamma[\sup_{\mathbf{t}\in\mathbb{R}^d} |Z_1''(\mathbf{t})| > c_1^*(\alpha)]$ and $P_\gamma[\int_{\mathbf{t}\in\mathbb{R}^d} \{Z_1''(\mathbf{t})\}^2\, \mathrm{d}F_0(\mathbf{t}) > c_1^{**}(\alpha)]$, respectively, where $Z_1''(\mathbf{t})$ $(\mathbf{t} \in \mathbb{R}^d)$ is a Gaussian process with its mean function $m_1'(\mathbf{t}) = \gamma\{H(\mathbf{t}) - F_0(\mathbf{t})\}$ and covariance kernel $k_3(\mathbf{t}_1, \mathbf{t}_2) = F_0(\min(\mathbf{t}_1, \mathbf{t}_2)) - F_0(\mathbf{t}_1)F_0(\mathbf{t}_2)$. Here "min" denotes the coordinatewise minimum of the two vectors in $\mathbb{R}^d$, and $c_1^*(\alpha)$ and $c_1^{**}(\alpha)$ satisfy $P_{\gamma=0}[\sup_{\mathbf{t}\in\mathbb{R}^d} |Z_1''(\mathbf{t})| > c_1^*(\alpha)] = \alpha$ and $P_{\gamma=0}[\int_{\mathbf{t}\in\mathbb{R}^d} \{Z_1''(\mathbf{t})\}^2\, \mathrm{d}F_0(\mathbf{t}) > c_1^{**}(\alpha)] = \alpha$.*

Next, for the two-sample problem, the null hypothesis is given by H_0^*: $F(\mathbf{x}) = G(\mathbf{x})$, and we consider a sequence of alternatives $H_{n,m}^*$: $G(\mathbf{x}) = (1 - \gamma/\sqrt{n+m})F(\mathbf{x}) + (\gamma/\sqrt{n+m})H(\mathbf{x})$ for a fixed $\gamma > 0$ and $n, m = 1, 2, \ldots$. Consider a multivariate Gaussian process $Z_2'(\mathbf{u})$ with the mean function

$$m_2(\mathbf{u}) = -\gamma\big[D_1\big(Q_F(\mathbf{u})\big)\big]^{-1} E_H\left\{\frac{\mathbf{y} - Q_F(\mathbf{u})}{\|\mathbf{y} - Q_F(\mathbf{u})\|} + \mathbf{u}\right\}$$

and the covariance kernel $k_2(\mathbf{u}_1, \mathbf{u}_2)$. Here $k_2(\mathbf{u}_1, \mathbf{u}_2)$ is as defined before Theorem 3.2. Let $\mathcal{T}' = \int \|Z_2'(\mathbf{u})\|^2\, \mathrm{d}\mathbf{u}$, where the integral is over the same closed ball as in the definition of $T_{n,m}$ in Section 3. We now state a theorem describing the asymptotic powers of the test based on $T_{n,m}$ as well as the KS and the CVM tests under contiguous alternatives.

Theorem 6.2. *Assume that F and H have continuous and positive densities f and h, respectively, on $\mathbb{R}^d$ $(d \geq 2)$, $E_F\{\frac{h(\mathbf{y})}{f(\mathbf{y})} - 1\}^2 < \infty$, and $n, m \to \infty$ in such a way that $\lim_{n,m\to\infty} \frac{n}{(n+m)} = \lambda \in (0, 1)$. Then, the sequence of densities associated with alternatives $H^*_{n,m}$ form a contiguous sequence. Under such alternatives, the asymptotic power of the test based on $T_{n,m}$ is given by $P_\gamma[\mathcal{T}' > c_2(\alpha)]$, where $c_2(\alpha)$ is as defined in Theorem 3.2 such that $P_{\gamma=0}[\mathcal{T}' > c_2(\alpha)] = \alpha$. Further, under those alternatives, the asymptotic powers of the tests based on $T^{(1)}_{n,m}$ and $T^{(2)}_{n,m}$ are given by $P_\gamma[\sup_{\mathbf{t}\in\mathbb{R}^d} |Z''_2(\mathbf{t})| > c^*_2(\alpha)]$ and $P_\gamma[\int_{\mathbf{t}\in\mathbb{R}^d}\{Z''_2(\mathbf{t})\}^2\,\mathrm{d}F(\mathbf{t}) > c^{**}_2(\alpha)]$, respectively, where $Z''_2(\mathbf{t})$ $(\mathbf{t} \in \mathbb{R}^d)$ is a Gaussian process with its mean function $m'_2(\mathbf{t}) = -\gamma\{H(\mathbf{t}) - F(\mathbf{t})\}$ and covariance kernel $k_4(\mathbf{t}_1, \mathbf{t}_2) = \frac{F(\min(\mathbf{t}_1,\mathbf{t}_2))-F(\mathbf{t}_1)F(\mathbf{t}_2)}{\lambda(1-\lambda)}$. Here also "min" denotes the coordinatewise minimum of the two vectors in $\mathbb{R}^d$, and $c^*_2(\alpha)$ and $c^{**}_2(\alpha)$ are such that $P_{\gamma=0}[\sup_{\mathbf{t}\in\mathbb{R}^d} |Z''_2(\mathbf{t})| > c^*_2(\alpha)] = \alpha$ and $P_{\gamma=0}[\int_{\mathbf{t}\in\mathbb{R}^d}\{Z''_2(\mathbf{t})\}^2\,\mathrm{d}F(\mathbf{t}) > c^{**}_2(\alpha)] = \alpha$.*

Theorems 6.1 and 6.2 enable us to derive the Pitman efficacies of our tests relative to the KS and the CVM tests. The Pitman efficacy (see, e.g., Serfling [31] and Lehmann and Romano [25]) of our test relative to another test for varying choices of the asymptotic power (determined by γ) is given by $(\gamma'/\gamma)^2$, where γ and γ' are such that the asymptotic power of our test under contiguous alternatives $(1 - \gamma/\sqrt{n})F_0(\mathbf{x}) + (\gamma/\sqrt{n})H(\mathbf{x})$ (or $(1 - \gamma/\sqrt{n+m})F(\mathbf{x}) + (\gamma/\sqrt{n+m})H(\mathbf{x})$) is the same as the asymptotic power of the other test under contiguous alternatives $(1 - \gamma'/\sqrt{n})F_0(\mathbf{x}) + (\gamma'/\sqrt{n})H(\mathbf{x})$ (or $(1 - \gamma'/\sqrt{n+m})F(\mathbf{x}) + (\gamma'/\sqrt{n+m})H(\mathbf{x})$).

In order to compute the critical values and the powers of our one-sample and two-sample tests, we have used 1000 simulations of each Gaussian process and approximated the integral of the squared norm of a multivariate Gaussian process by the average of the squared norms of some appropriate multivariate normal random vectors. In this numerical study, we could compute the true covariance matrices as the underlying distributions were known. We have computed the critical value and the asymptotic power of the CVM test in a similar way. However, in the case of the KS test, we have approximated the supremum of a Gaussian process by a maximum over 1000 simulations of the process.

In Figure 10, we have plotted the Pitman efficacy of our test for different values of the asymptotic power. It is clearly indicated by Figure 10 that our test and the CVM test outperform the KS test in terms of the Pitman efficacy in all the cases considered here. However, between our test and the CVM test, one has superior performance in some cases while the other has superior performance in some other cases, and there is only a small difference in their performance.

Appendix: Proofs

Proof of Theorem 2.1. In view of the results in Chaudhuri [10] and Koltchinskii [24], we have

$$\sup_{\mathbf{u}} \| Q_{\mathcal{X}}(\mathbf{u}) - Q_F(\mathbf{u}) \| = o_P(1), \tag{A.1}$$

where the supremum is taken over any given closed ball with the center at the origin and the radius strictly smaller than one. When $F = F_0$, we have $Q_F(\mathbf{u}) = Q_{F_0}(\mathbf{u})$ for all $\|\mathbf{u}\| < 1$. This along with the uniform convergence result in (A.1) leads to the proof of the "*if* part" of the theorem.

Next, consider some $\mathbf{u}$ with $\|\mathbf{u}\| < 1$. It follows from the conditions in the theorem that with probability tending to one, the spatial rank vectors $\mathbf{u}_k$'s form a dense subset of the unit ball around the origin as $n \to \infty$. Since

$$\lim_{n\to\infty} P\left(\bigcap_{i=1}^{d}\left[S_{n,i}(\mathcal{X}, F_0) \subseteq L(\varepsilon)\right]\right) = 1$$

for every $\varepsilon > 0$, we must have $Q_F(\mathbf{u}) = Q_{F_0}(\mathbf{u})$ in view of (A.1). It now follows from the characterization of multivariate distributions by the spatial quantiles (see Corollary 2.9 in Koltchinskii ([24], page 446)) that $F = F_0$. This completes the proof of the "*only if* part" of the theorem. □

Proof of Theorem 2.2. It follows from the results in Chaudhuri [10] and Koltchinskii [24] that for the two independent samples $\mathcal{X}$ and $\mathcal{Y}$, we have $\sup_{\mathbf{u}} \|(Q_{\mathcal{X}}(\mathbf{u}), Q_{\mathcal{Y}}(\mathbf{u})) - (Q_F(\mathbf{u}), Q_G(\mathbf{u}))\| = o_P(1)$ when $n, m \to \infty$ in such a way that $\lim_{n,m\to\infty} \frac{n}{(n+m)} = \lambda \in (0, 1)$. Here the supremum is taken over any given closed ball with the center at the origin and the radius strictly smaller than one. Then the proof of the theorem follows by similar arguments as in the proof of Theorem 2.1. □

Proof of Theorem 3.1. As proved in Koltchinskii [24], the centered and normalized stochastic process $\sqrt{n}\{Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}$ converges weakly to the Gaussian process $Z_1(\mathbf{u})$ (defined in Section 3) under H_0. Here $\mathbf{u}$ lies in any given closed ball with the center at the origin and the radius strictly smaller than one. It follows from the continuity of the integral functional that V_n converges in distribution to $\mathcal{V}$. Consequently, the asymptotic level of the test will be α.

The asymptotic power of the test is given by $\lim_{n\to\infty} P_{H_1}[V_n > c_1(\alpha)]$. Now, note that $V_n > c_1(\alpha)$ if and only if $n\int \|\{Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\} - \{Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}\|^2\,d\mathbf{u} > c_1(\alpha) + n[\int \langle \{Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}, \{Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}\rangle\,d\mathbf{u} - 2\int \langle \{Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}, \{Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}\rangle\,d\mathbf{u}]$. Here the integrals are over a closed ball with the center at the origin and the radius strictly smaller than one as before.

When $F \neq F_0$, in view of the characterization property of the spatial quantiles (see Corollary 2.9 in Koltchinskii [24]), we have $Q_F(\mathbf{u}) \neq Q_{F_0}(\mathbf{u})$ for some $\mathbf{u}$ with $\|\mathbf{u}\| < 1$. The uniform convergence of $Q_{\mathcal{X}}(\mathbf{u})$ to $Q_F(\mathbf{u})$ and the continuity of the spatial quantiles $Q_F(\mathbf{u})$ and $Q_{F_0}(\mathbf{u})$ as functions of $\mathbf{u}$ imply that $c_1(\alpha) + n[\int \langle \{Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}, \{Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}\rangle\,d\mathbf{u} - 2\int \langle \{Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}, \{Q_F(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}\rangle\,d\mathbf{u}]$ tends to $-\infty$ in probability as $n \to \infty$. Hence, $P_{H_1}[V_n > c_1(\alpha)] \to 1$ as $n \to \infty$. This completes the proof. □

Proof of Theorem 3.2. Arguing in a similar way as in the proof of Theorem 3.1 and using the weak convergence results in Koltchinskii [24], and the independence of the two samples, if $n, m \to \infty$ in such a way that $\lambda = \lim_{n,m\to\infty} \frac{n}{(n+m)} \in (0, 1)$, one can show that $T_{n,m}$ converges in distribution to $\mathcal{T}$ under H_0^*, and consequently, the asymptotic level of the test that rejects H_0^* when $T_{n,m} > c_2(\alpha)$ will be α. Next, the asymptotic power of the test is given by $P_{H_1^*}[T_{n,m} > c_2(\alpha)]$. Using similar arguments as in the second part of the proof of Theorem 3.1, one can establish that $P_{H_1^*}[T_{n,m} > c_2(\alpha)] \to 1$ as $n, m \to \infty$. □

Proof of Theorem 6.1. The logarithm of the likelihood ratio for testing H_0 against H_n is

$$
\begin{aligned}
L_n &= \sum_{i=1}^{n} \log \frac{(1-\gamma/\sqrt{n}) f_0(\mathbf{x}_i) + (\gamma/\sqrt{n}) h(\mathbf{x}_i)}{f_0(\mathbf{x}_i)} = \sum_{i=1}^{n} \log\left[1 + (\gamma/\sqrt{n})\left\{\frac{h(\mathbf{x}_i)}{f_0(\mathbf{x}_i)} - 1\right\}\right] \\
&= \frac{\gamma}{\sqrt{n}} \sum_{i=1}^{n}\left\{\frac{h(\mathbf{x}_i)}{f_0(\mathbf{x}_i)} - 1\right\} - \frac{\gamma^2}{2n}\sum_{i=1}^{n}\left\{\frac{h(\mathbf{x}_i)}{f_0(\mathbf{x}_i)} - 1\right\}^2 + R_n \qquad \text{(A.2)} \\
&= \frac{\gamma}{\sqrt{n}} \sum_{i=1}^{n} k_i - \frac{\gamma^2}{2} \times \frac{1}{n}\sum_{i=1}^{n} k_i^2 + R_n,
\end{aligned}
$$

where $k_i = \frac{h(\mathbf{x}_i)}{f_0(\mathbf{x}_i)} - 1$. Note that $R_n \xrightarrow{P} 0$ as $n \to \infty$ since $\sigma^2 := E_{F_0}[\frac{h(\mathbf{x})}{f_0(\mathbf{x})} - 1]^2 < \infty$. Further, by a straightforward application of the central limit theorem, the first term in (A.2) is asymptotically normal with its mean $= 0$ and variance $= \gamma^2\sigma^2$, and the second term in (A.2) converges in probability to $\frac{\gamma^2}{2}\sigma^2$ by the weak law of large numbers. So, using Slutsky's theorem, L_n is asymptotically normal with mean $= -\frac{\gamma^2}{2}\sigma^2$ and variance $= \gamma^2\sigma^2$. This ensures the contiguity of the sequence H_n using the corollary to Lecam's first lemma in Hájek and Šidák ([19], pages 204).

Now, we consider $\mathbf{u}_1, \ldots, \mathbf{u}_k$ in a given closed ball with the center at the origin and the radius strictly smaller than one, and $\mathbf{t}_1, \ldots, \mathbf{t}_l \in \mathbb{R}^d$. Then, under H_0, one can establish that the joint distribution of $\sqrt{n}\{Q_{\mathcal{X}}(\mathbf{u}_1) - Q_{F_0}(\mathbf{u}_1), \ldots, Q_{\mathcal{X}}(\mathbf{u}_k) - Q_{F_0}(\mathbf{u}_k), F_n(\mathbf{t}_1) - F_0(\mathbf{t}_1), \ldots, F_n(\mathbf{t}_l) - F_0(\mathbf{t}_l), L_n/\sqrt{n}\}$ is asymptotically multivariate normal. This follows using the Bahadur type linear expansion of $\{Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}$ (see Chaudhuri [10]), the expansion of L_n (see (A.2) above) and the fact that $F_n(\mathbf{t}) - F_0(\mathbf{t})$ is a simple average of i.i.d. random variables. Note that for any $p = 1, \ldots, k$, the covariance between $\sqrt{n}\{Q_{\mathcal{X}}(\mathbf{u}_p) - Q_{F_0}(\mathbf{u}_p)\}$ and L_n is

$$
\begin{aligned}
&\frac{\gamma}{n} E_{F_0}\left[\sum_{i=1}^{n}\left\{D_1\bigl[Q_{F_0}(\mathbf{u}_p)\bigr]^{-1}\left\{\frac{\mathbf{x}_i - Q_{F_0}(\mathbf{u}_p)}{\|\mathbf{x}_i - Q_{F_0}(\mathbf{u}_p)\|} + \mathbf{u}_p\right\}\right\} \times \left\{\frac{h(\mathbf{x}_i)}{f_0(\mathbf{x}_i)} - 1\right\}\right] \\
&\quad = \gamma\bigl[D_1\bigl\{Q_{F_0}(\mathbf{u}_p)\bigr\}\bigr]^{-1} E_H\left\{\frac{\mathbf{x} - Q_{F_0}(\mathbf{u}_p)}{\|\mathbf{x} - Q_{F_0}(\mathbf{u}_p)\|} + \mathbf{u}_p\right\} = m_1(\mathbf{u}_p),
\end{aligned}
$$

because $E_{F_0}\{\frac{\mathbf{x} - Q_{F_0}(\mathbf{u}_p)}{\|\mathbf{x} - Q_{F_0}(\mathbf{u}_p)\|} + \mathbf{u}_p\} = \mathbf{0}$. Also, one can show that for any $j = 1, \ldots, l$, the covariance between $\sqrt{n}\{F_n(\mathbf{t}_j) - F_0(\mathbf{t}_j)\}$ and L_n is $m_1'(\mathbf{t}_j) = \gamma\{H(\mathbf{t}_j) - F_0(\mathbf{t}_j)\}$.

Now, by a straightforward application of Lecam's third lemma (see Hájek and Šidák [19], page 208), one can establish that under contiguous alternatives, $\sqrt{n}\{Q_{\mathcal{X}}(\mathbf{u}_1) - Q_{F_0}(\mathbf{u}_1), \ldots, Q_{\mathcal{X}}(\mathbf{u}_k) - Q_{F_0}(\mathbf{u}_k)\}$ is asymptotically kd-dimensional multivariate normal with the mean vector having the d-dimensional pth block $m_1(\mathbf{u}_p)$ $(p = 1, 2, \ldots, k)$, and its $kd \times kd$-dimensional covariance matrix is obtained from the covariance kernel k_1, which is given before Theorem 3.1. Further, the spatial quantile process satisfies the tightness condition under contiguous alternatives in view of the fact that it is tight under H_0. The tightness under H_0 follows from the weak convergence of the spatial quantile process (see Koltchinskii [24]). So, the spatial quantile process

$\sqrt{n}\{Q_{\mathcal{X}}(\mathbf{u}) - Q_{F_0}(\mathbf{u})\}$ converges to $Z_1'(\mathbf{u})$ under H_n, where $Z_1'(\mathbf{u})$ is a Gaussian process with its mean function $m_1(\mathbf{u})$ and covariance kernel $k_1(\mathbf{u}_1, \mathbf{u}_2)$. Hence, under H_n, the asymptotic power of the test based on V_n is $P_\gamma[\mathcal{V}' > c_1(\alpha)]$.

Similarly, using the weak convergence of the stochastic process $\sqrt{n}\{F_n(\mathbf{t}) - F(\mathbf{t})\}$ under H_0 to a Gaussian process (see, e.g., Bickel and Wichura [6]) together with Lecam's third lemma, one can show that under contiguous alternatives, $\sqrt{n}\{F_n(\mathbf{t}_1) - F_0(\mathbf{t}_1), \dots, F_n(\mathbf{t}_l) - F_0(\mathbf{t}_l)\}$ is asymptotically l-dimensional multivariate normal with the mean vector having the jth component $m_1'(\mathbf{t}_j)$ ($j = 1, \dots, l$), and its $l \times l$-dimensional covariance matrix is obtained from the covariance kernel k_3, which is given in the statement of the theorem. Now, it follows from the finite-dimensional asymptotic distribution and the tightness of the process $\sqrt{n}\{F_n(\mathbf{t}) - F_0(\mathbf{t})\}$ under contiguous alternatives that the stochastic process $\sqrt{n}\{F_n(\mathbf{t}) - F_0(\mathbf{t})\}$ converges to $Z_1''(\mathbf{t})$ under H_n, where $Z_1''(\mathbf{t})$ is a Gaussian process with its mean function $m_1'(\mathbf{t})$ and covariance kernel $k_3(\mathbf{t}_1, \mathbf{t}_2)$. Consequently, under H_n, the asymptotic powers of the tests based on $T_n^{(1)}$ and $T_n^{(2)}$ are $P_\gamma[\sup_{\mathbf{t}\in\mathbb{R}^d} |Z_1''(\mathbf{t})| > c_1^*(\alpha)]$ and $P_\gamma[\int_{\mathbf{t}\in\mathbb{R}^d} \{Z_1''(\mathbf{t})\}^2 \,\mathrm{d}F_0(\mathbf{t}) > c_1^{**}(\alpha)]$, respectively. □

Proof of Theorem 6.2. The logarithm of the likelihood ratio for testing H_0^* against $H_{n,m}^*$ is

$$\begin{aligned} L_{n,m} &= \log \frac{\prod_{i=1}^n f(\mathbf{x}_i) \prod_{j=1}^m \{(1 - \gamma/\sqrt{n+m}) f(\mathbf{y}_j) + \gamma/\sqrt{n+m} h(\mathbf{y}_j)\}}{\prod_{i=1}^n f(\mathbf{x}_i) \prod_{j=1}^m f(\mathbf{y}_j)} \\ &= \sum_{j=1}^m \log\left\{1 + \frac{\gamma}{\sqrt{n+m}}\left(\frac{h(\mathbf{y}_j)}{f(\mathbf{y}_j)} - 1\right)\right\} \qquad \text{(A.3)} \\ &= \frac{\gamma}{\sqrt{n+m}} \sum_{j=1}^m k_j' - \frac{\gamma^2}{2(n+m)} \times \sum_{j=1}^m k_j'^2 + R_{n,m}, \end{aligned}$$

where $k_j' = \frac{h(\mathbf{y}_j)}{f(\mathbf{y}_j)} - 1$. Note that $R_{n,m} \xrightarrow{P} 0$ as $n, m \to \infty$ since $\sigma_*^2 := E_F\{\frac{h(\mathbf{y})}{f(\mathbf{y})} - 1\}^2 < \infty$. Using similar arguments as in the proof of Theorem 6.1, $L_{n,m}$ is asymptotically normal with mean $= -\frac{\gamma^2}{2}(1 - \lambda)\sigma_*^2$ and variance $= \gamma^2(1 - \lambda)\sigma_*^2$. This fact ensures the contiguity of the sequence of densities under $H_{n,m}^*$ using the corollary to Lecam's first lemma in Hájek and Šidák ([19], page 204).

Now, here also, we consider $\mathbf{u}_1, \dots, \mathbf{u}_k$ in a given closed ball with the center at the origin and the radius strictly smaller than one, and $\mathbf{t}_1, \dots, \mathbf{t}_l \in \mathbb{R}^d$. Then, under H_0, one can establish that the joint distribution of $\sqrt{n+m}\{Q_{\mathcal{X}}(\mathbf{u}_1) - Q_{\mathcal{Y}}(\mathbf{u}_1), \dots, Q_{\mathcal{X}}(\mathbf{u}_k) - Q_{\mathcal{Y}}(\mathbf{u}_k), F_n(\mathbf{t}_1) - G_m(\mathbf{t}_1), \dots, F_n(\mathbf{t}_l) - G_m(\mathbf{t}_l), L_{n,m}/\sqrt{n+m}\}$ is asymptotically multivariate normal. This asymptotic normality is a consequence of the independence of the two samples, the Bahadur type linear expansion of the difference of the spatial quantiles $Q_{\mathcal{X}}(\mathbf{u}) - Q_{\mathcal{Y}}(\mathbf{u})$ (see Chaudhuri [10]), the expansion of $L_{n,m}$ given in (A.3) and the fact that $F_n(\mathbf{t})$ and $G_m(\mathbf{t})$ are simple averages of i.i.d. random variables. Note that for any $p = 1, \dots, k$, the covariance between

$\sqrt{n+m}\{Q_{\mathcal{X}}(\mathbf{u}_p) - Q_{\mathcal{Y}}(\mathbf{u}_p)\}$ and $L_{n,m}$ is

$$\begin{aligned}
&E_F\Bigg[\sqrt{n+m}\Bigg[\frac{1}{n}\sum_{i=1}^{n}\bigg\{D_1\big[Q_F(\mathbf{u}_p)\big]^{-1}\bigg\{\frac{\mathbf{x}_i - Q_F(\mathbf{u}_p)}{\|\mathbf{x}_i - Q_F(\mathbf{u}_p)\|} + \mathbf{u}_p\bigg\}\bigg\} \\
&\qquad - \frac{1}{m}\sum_{j=1}^{m}\bigg\{D_1\big[Q_F(\mathbf{u}_p)\big]^{-1}\bigg\{\frac{\mathbf{y}_j - Q_F(\mathbf{u}_p)}{\|\mathbf{y}_j - Q_F(\mathbf{u}_p)\|} + \mathbf{u}_p\bigg\}\bigg\}\Bigg] \times \frac{\gamma}{\sqrt{n+m}}\sum_{j=1}^{m}\bigg\{\frac{h(\mathbf{y}_j)}{f(\mathbf{y}_j)} - 1\bigg\}\Bigg] \\
&= -\sqrt{n+m}E_F\Bigg[\frac{1}{m}\sum_{j=1}^{m}\bigg\{D_1\big[Q_F(\mathbf{u}_p)\big]^{-1}\bigg\{\frac{\mathbf{y}_j - Q_F(\mathbf{u}_p)}{\|\mathbf{y}_j - Q_F(\mathbf{u}_p)\|} + \mathbf{u}_p\bigg\}\bigg\} \\
&\qquad\qquad \times \frac{\gamma}{\sqrt{n+m}}\sum_{j=1}^{m}\bigg\{\frac{h(\mathbf{y}_j)}{f(\mathbf{y}_j)} - 1\bigg\}\Bigg] \qquad \text{(since } \mathbf{x} \text{ and } \mathbf{y} \text{ are independent)} \\
&= -\gamma\big[D_1^F\big(Q(\mathbf{u})\big)\big]^{-1}E_H\bigg\{\frac{\mathbf{y} - Q_F(\mathbf{u}_p)}{\|\mathbf{y} - Q_F(\mathbf{u}_p)\|} + \mathbf{u}_p\bigg\} = m_2(\mathbf{u}_p),
\end{aligned}$$

because $E_F\{\frac{\mathbf{y}-Q_F(\mathbf{u}_p)}{\|\mathbf{y}-Q_F(\mathbf{u}_p)\|} + \mathbf{u}_p\} = \mathbf{0}$. Arguing in a similar way as in the proof of Theorem 6.1, one can establish that under $H^*_{n,m}$, the process $\sqrt{n+m}\{Q_{\mathcal{X}}(\mathbf{u}) - Q_{\mathcal{Y}}(\mathbf{u})\}$ converges to $Z_2'(\mathbf{u})$, where $Z_2'(\mathbf{u})$ is a Gaussian process with its mean function $m_2(\mathbf{u})$ and covariance kernel $k_2(\mathbf{u}_1, \mathbf{u}_2)$, which is defined before Theorem 3.2. Hence, the asymptotic power of the test based on $T_{n,m}$ is $P_\gamma[\mathcal{T}' > c_2(\alpha)]$.

Also, under H_0^*, one can show that for any $j = 1, \dots, l$, the covariance between $\sqrt{n+m} \times \{F_n(\mathbf{t}_j) - G_m(\mathbf{t}_j)\}$ and $L_{n,m}$ is $m_2'(\mathbf{t}_j) = -\gamma\{H(\mathbf{t}_j) - F(\mathbf{t}_j)\}$. Further, under H_0^*, the stochastic process $\sqrt{n+m}\{F_n(\mathbf{t}) - G_m(\mathbf{t})\}$ converges to a Gaussian process with zero mean and the covariance kernel k_4, which is given in the statement of the theorem (see, e.g., Bickel and Wichura [6]). Now, it follows from the finite-dimensional asymptotic distributions and the tightness of the process $\sqrt{n+m}\{F_n(\mathbf{t}) - G_m(\mathbf{t})\}$ under contiguous alternatives that the stochastic process $\sqrt{n+m}\{F_n(\mathbf{t}) - G_m(\mathbf{t})\}$ converges to $Z_2''(\mathbf{t})$ under $H^*_{n,m}$, where $Z_2''(\mathbf{t})$ is a Gaussian process with its mean function $m_2'(\mathbf{t})$ and covariance kernel $k_4(\mathbf{t}_1, \mathbf{t}_2)$. Consequently, under $H^*_{n,m}$, the asymptotic power of the test based on $T^{(1)}_{n,m}$ is $P_\gamma[\sup_{\mathbf{t}\in\mathbb{R}^d} |Z_2''(\mathbf{t})| > c_2^*(\alpha)]$.

In the case of $T^{(2)}_{n,m}$, we first show that $(n+m)\int_{\mathbf{x}\in\mathbb{R}^d}[F_n(\mathbf{x}) - G_m(\mathbf{x})]^2\,\mathrm{d}(M_{n,m} - F) \xrightarrow{P} 0$ as $n, m \to \infty$ under H_0^*. For that, it is enough to prove that $T^{(2,1)}_{n,m} = (n+m)\int_{\mathbf{x}\in\mathbb{R}^d}[F_n(\mathbf{x}) - G_m(\mathbf{x})]^2\,\mathrm{d}(F_n - F) \xrightarrow{P} 0$ and $T^{(2,2)}_{n,m} = (n+m)\int_{\mathbf{x}\in\mathbb{R}^d}[F_n(\mathbf{x}) - G_m(\mathbf{x})]^2\,\mathrm{d}(G_m - G) \xrightarrow{P} 0$ as $n, m \to \infty$ under H_0^*. Now, it follows from the arguments in the proofs of the lemma on page 424 in Kiefer [21] and Theorem 2 in Kiefer and Wolfowitz [22] that $T^{(2,1)}_{n,m} \xrightarrow{P} 0$ and $T^{(2,2)}_{n,m} \xrightarrow{P} 0$ as $n, m \to \infty$ under H_0^*, and hence, $(n+m)\int_{\mathbf{x}\in\mathbb{R}^d}[F_n(\mathbf{x}) - G_m(\mathbf{x})]^2\,\mathrm{d}(M_{n,m} - F) \xrightarrow{P} 0$ as $n, m \to \infty$ under H_0^*. Therefore, $(n+m)\int_{\mathbf{x}\in\mathbb{R}^d}[F_n(\mathbf{x}) - G_m(\mathbf{x})]^2\,\mathrm{d}(M_{n,m} - F) \xrightarrow{P} 0$ as $n, m \to \infty$ under contiguous alternatives $H^*_{n,m}$. Hence, the asymptotic power of the test based on $T^{(2)}_{n,m}$ under $H^*_{n,m}$ is $P_\gamma[\int_{\mathbf{t}\in\mathbb{R}^d}\{Z_2''(\mathbf{t})\}^2\,\mathrm{d}F(\mathbf{t}) > c_2^{**}(\alpha)]$. □

Acknowledgments

The research of the first author is partially supported by a grant from the Council of Scientific and Industrial Research (CSIR), Government of India. The authors are thankful to two anonymous referees and an anonymous Associate Editor for several useful comments.

Supplementary Material

Supplement to "Comparison of multivariate distributions using quantile–quantile plots and related tests" (DOI: 10.3150/13-BEJ530SUPP; .pdf). In the supplement, we provide additional multivariate Q–Q plots and discuss the performance of various tests for univariate data.

References

[1] Ahmad, I.A. (1993). Modification of some goodness-of-fit statistics to yield asymptotically normal null distributions. *Biometrika* **80** 466–472. MR1243521

[2] Ahmad, I.A. (1996). Modification of some goodness of fit statistics. II. Two-sample and symmetry testing. *Sankhyā Ser. A* **58** 464–472. MR1659118

[3] Anderson, T.W. and Darling, D.A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.* **49** 765–769. MR0069459

[4] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726. MR1440039

[5] Bickel, P.J. (1967). Some contributions to the theory of order statistics. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley*, *Calif.*, 1965/66), *Vol. I*: *Statistics* 575–591. Berkeley, CA: Univ. California Press. MR0216701

[6] Bickel, P.J. and Wichura, M.J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656–1670. MR0383482

[7] Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika* **75** 761–771. MR0995118

[8] Burke, M.D. (1977). On the multivariate two-sample problem using strong approximations of the EDF. *J. Multivariate Anal.* **7** 491–511. MR0458704

[9] Chambers, J., Cleveland, W., Kleiner, B. and Tukey, P. (1983). *Graphical Methods for Data Analysis*. Belmont: Wadsworth.

[10] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.* **91** 862–872. MR1395753

[11] Dhar, S.S., Chakraborty, B. and Chaudhuri, P. (2013). Supplement to "Comparison of multivariate distributions using quantile–quantile plots and related tests." DOI:10.3150/13-BEJ530SUPP.

[12] Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist.* **2** 267–277. MR0356350

[13] Doksum, K.A. and Sievers, G.L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63** 421–434. MR0443210

[14] Easton, G.S. and McCulloch, R.E. (1990). A multivariate generalization of quantile-quantile plots. *J. Amer. Statist. Assoc.* **85** 376–386.

[15] Friedman, J.H. and Rafsky, L.C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. MR0532236

[16] Friedman, J.H. and Rafsky, L.C. (1981). Graphics for the multivariate two-sample problem. *J. Amer. Statist. Assoc.* **76** 277–287.

[17] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley. MR0440802

[18] Gnanadesikan, R. and Wilk, M.B. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55** 1–17.

[19] Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. New York: Academic Press. MR0229351

[20] Justel, A., Peña, D. and Zamar, R. (1997). A multivariate Kolmogorov–Smirnov test of goodness of fit. *Statist. Probab. Lett.* **35** 251–259. MR1484961

[21] Kiefer, J. (1959). K-sample analogues of the Kolmogorov–Smirnov and Cramér–V. Mises tests. *Ann. Math. Statist.* **30** 420–447. MR0102882

[22] Kiefer, J. and Wolfowitz, J. (1958). On the deviations of the empiric distribution function of vector chance variables. *Trans. Amer. Math. Soc.* **87** 173–186. MR0099075

[23] Koenker, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge: Cambridge Univ. Press. MR2268657

[24] Koltchinskii, V.I. (1997). M-estimation, convexity and quantiles. *Ann. Statist.* **25** 435–477. MR1439309

[25] Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*. New Delhi: Springer.

[26] Liu, R.Y., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27** 783–840.

[27] Marden, J.I. (1998). Bivariate qq-plots and spider web plots. *Statist. Sinica* **8** 813–826. MR1651510

[28] Marden, J.I. (2004). Positions and QQ plots. *Statist. Sci.* **19** 606–614. MR2185582

[29] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametr. Stat.* **5** 201–213. MR1346895

[30] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. New York: Wiley. MR0914792

[31] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley. MR0595165

[32] Serfling, R.J. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *J. Statist. Plann. Inference* **123** 259–278. MR2062982

[33] Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality: Complete samples. *Biometrika* **52** 591–611. MR0205384

[34] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. New York: Wiley. MR0838963

[35] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. London: Chapman & Hall. MR0848134

Received September 2011 and revised March 2013