# A new test for multivariate normality by combining extreme and non-extreme BHEP tests[*]

C. Tenreiro [†]

18th November 2014

**Abstract**

In this paper we propose a new multiple test procedure for assessing multivariate normality which combines BHEP (Baringhaus-Henze-Epps-Pulley) tests by considering extreme and non-extreme choices of the tuning parameter in the definition of the BHEP test statistic. Monte Carlo power comparisons indicate that the new test presents a reasonable power against a wide range of alternative distributions, showing itself to be competitive against the most recommended procedures for testing a multivariate hypothesis of normality. We further illustrate the use of the new test for the Fisher Iris data set.

KEYWORDS: Tests for multivariate normality; BHEP tests; Multiple testing; Monte Carlo power comparison.

AMS 2010 SUBJECT CLASSIFICATIONS: 62G10, 62H15

[*] This is an electronic version of an article accepted for publication in Communications in Statistics – Simulation and Computation (Vol. 46, 2017, 1746–1759). DOI:10.1080/03610918.2015.1011334.

[†] CMUC, Department of Mathematics, University of Coimbra, Apartado 3008, 3001–501 Coimbra, Portugal. E-mail: tenreiro@mat.uc.pt. URL: `http://www.mat.uc.pt/~tenreiro/`.

# 1 Introduction

If $X_1, \ldots, X_n, \ldots$ is a sequence of independent copies of a $d$-dimensional absolutely continuous random vector $X$ with unknown probability density function $f$, also denoted by $f_X$, the problem of assessing multivariate normality (MVN) is to test, on the basis of $X_1, \ldots, X_n$, the hypothesis

$$H_0 \,:\, f \in \mathcal{N}_d,$$

against a general alternative, where $\mathcal{N}_d$ is the family of normal probability density functions on $\mathbb{R}^d$. The multivariate normal distribution is widely used in many applications and several test procedures for this classical problem have been proposed in the literature showing a continued interest in this subject. Some of the work published in the last ten years include the papers of Liang et al. (2005), Mecklin and Mundfrom (2005), Székely and Rizzo (2005), Sürücü (2006), Arcones (2007), Farrel et al. (2007), Chiu and Liu (2009), Liang et al. (2009), Tenreiro (2009, 2011), Oliveira and Ferreira (2010), Ebner (2012) and Wang (2015). For some additional bibliography on this topic see Csörgő (1986), Rayner and Best (1989, p. 98–109), Thode (2002, p. 181–224) and the review articles of Henze (2002) and Mecklin and Mundfrom (2004).

An important class of test procedures for assessing MVN is the BHEP (Baringhaus-Henze-Epps-Pulley) family introduced by Baringhaus and Henze (1988) and Henze and Zirkler (1990), which extends the Epps and Pulley (1983) procedure to the multivariate context. In order to define this family of test statistics, let us denote by

$$Y_j = S_n^{-1/2}(X_j - \overline{X}_n), \quad j = 1, \ldots, n$$

the scaled residuals associated with the observations $X_1, \ldots, X_n$, where

$$\overline{X}_n = n^{-1} \sum_{j=1}^{n} X_j \quad \text{and} \quad S_n = n^{-1} \sum_{j=1}^{n} (X_j - \overline{X}_n)(X_j - \overline{X}_n)',$$

are the sample mean vector and the sample covariance matrix, respectively, and $S_n^{-1/2}$ is the symmetric positive definite square root of $S_n^{-1}$. We always assume that $S_n$ is nonsingular almost surely which, in accordance with Dykstra (1970), holds whenever $n \geq d+1$. The BHEP test statistic associated to the strictly positive real number $h$, is a weighted $L_2$-distance between the empirical characteristic function of the scaled residuals,

$$\Psi_n(t) = \frac{1}{n} \sum_{j=1}^{n} \exp\left(i\, t' Y_j\right), \quad t \in \mathbb{R}^d,$$

and the characteristic function $\Phi$ of the $d$-dimensional standard Gaussian density $\phi(x) = (2\pi)^{-d/2} \exp(-x'x/2)$, $x \in \mathbb{R}^d$, with weight function $t \to |\Phi_h(t)|^2 = \exp(-h^2 t' t)$, where $\Phi_h$

is the characteristic function of $\phi_h(\cdot) = \phi(\cdot/h)/h^d$. The BHEP test statistic is then defined as

$$\mathrm{B}(h) = n \int |\Psi_n(t) - \Phi(t)|^2 |\Phi_h(t)|^2 dt, \quad h > 0, \tag{1}$$

where the unspecified integral denotes integration over the whole space. The considered weight function is particularly useful because in such a case $\mathrm{B}(h)$ does not require any integration. In fact we can rewrite the BHEP test statistic as

$$\mathrm{B}(h) = (2\pi)^d \frac{1}{n} \sum_{i,j=1}^{n} Q(Y_i, Y_j; h),$$

with

$$Q(u, v; h) = \phi_{(2h^2)^{1/2}}(u - v) - \phi_{(1+2h^2)^{1/2}}(u) - \phi_{(1+2h^2)^{1/2}}(v) + \phi_{(2+2h^2)^{1/2}}(0),$$

for $u, v \in \mathbb{R}^d$ and $h > 0$.

The asymptotic behaviour of $\mathrm{B}(h)$ under the null hypothesis, a fixed alternative distribution and a sequence of local alternatives, can be obtained from the work of several authors such as Baringhaus and Henze (1988), Csörgő (1989), Henze and Zirkler (1990), Henze (1997), Henze and Wagner (1997) and Gürtler (2000). In particular, for each $0 < h < \infty$, $\mathrm{B}(h)$ has as limiting null distribution a weighted sum of $\chi^2$ independent random variables and, contrary to almost all MVN tests considered in the literature, the associated test procedure is consistent against each fixed alternative distribution.

It is worth mentioning that the statistic (1) can be interpreted as the $L_2$-distance between the Parzen-Rosenblatt kernel estimator based on the scaled residuals with kernel $K = \phi$ and smoothing parameter (bandwidth) $h$, and the convolution $K_h * \phi$, which can be seen as an approximation of the standardised null density when $h$ is close to zero (see Henze and Zirkler, 1990, Fan, 1998). In this form the statistic $\mathrm{B}(h)$ was firstly considered by Bowman and Foster (1993) and its asymptotic behaviour under the null hypothesis, a fixed alternative distribution and a sequence of local alternatives was first establish in Gürtler (2000). For an unifying treatment of the asymptotic behaviour of $\mathrm{B}(h)$ for the non-fixed ($h = h_n \to 0$) and fixed ($0 < h < \infty$) bandwidth cases, see also Tenreiro (2007).

From a practical point of view, it is well known that the finite sample power performance of the BHEP test is very sensitive to the choice of $h$ which acts as a tuning parameter (see Henze and Zirkler, 1990, Henze and Wagner, 1997, Tenreiro, 2009). In Tenreiro (2009) the choice of $h$ has been examined through a large-scale simulation study based on a set of meta-Gaussian distributions whose marginal distributions are members of the generalised lambda family discussed in Ramberg and Schmeiser (1974). Two distinct behaviour patterns for the

BHEP empirical power as a function of $h$ were identified which has led the author to propose two distinct choices of the bandwidth, depending on the data dimension ($2 \leq d \leq 15$):

$$h = h_{\mathrm{S}} := 0.448 + 0.026\, d \quad \text{and} \tag{2}$$

and

$$h = h_{\mathrm{L}} := 0.928 + 0.049\, d. \tag{3}$$

The bandwidth $h_{\mathrm{S}}$ has revealed to be suitable for short tailed or high moment alternatives while the bandwidth $h_{\mathrm{L}}$ has shown to be appropriate for long tailed or moderately skewed alternative distributions.

If there is no relevant information about the alternative distribution, which is the most common case in a real situation, the proposal of Tenreiro (2009) is to use

$$h = \bar{h} = (h_{\mathrm{S}} + h_{\mathrm{L}})/2, \tag{4}$$

because this value produces an omnibus test for normality. Despite this good property, for several alternative distributions the MVN test based on $\mathrm{B}(\bar{h})$ is outperformed by one of the classical Mardia (1970) tests, which are among the most recommended procedures for testing MVN. The Mardia tests are based on the test statistics MS (multivariate skewness) and MK (multivariate kurtosis) given by

$$\mathrm{MS} = n b_{1,d} \tag{5}$$

and

$$\mathrm{MK} = \sqrt{n}\, |\, b_{2,d} - d(d+2)|, \tag{6}$$

where

$$b_{1,d} = \frac{1}{n^2} \sum_{j,k=1}^{n} (Y_j' Y_k)^3 \quad \text{and} \quad b_{2,d} = \frac{1}{n} \sum_{j=1}^{n} (Y_j' Y_j)^2, \tag{7}$$

are the Mardia empirical measures of multivariate skewness and kurtosis, respectively. The Mardia skewness test performs well for skewed or long tailed alternatives and the Mardia kurtosis test is especially good for short tailed alternatives (cf. Henze and Zirkler, 1990, Baringhaus and Henze, 1992, Romeu and Ozturk, 1993).

Intending to propose a MVN test that could reveal a good empirical power for a wider range of alternative distributions than the BHEP test based on $\mathrm{B}(\bar{h})$, an improved Bonferroni method considered by Fromont and Laurent (2006) is used in Tenreiro (2011) in order to combine the previous BHEP tests based on $\mathrm{B}(h_S)$ and $\mathrm{B}(h_L)$, and the Mardia tests based on MS and MK. A simulation study carried out in Tenreiro (2011) for a wide range of alternative distributions, indicated that the resulting multiple test procedure, named

MB, presents a reasonable performance against a large set of alternative distributions and a good overall performance against other highly recommended MVN tests.

However, other combinations of affine invariant MVN tests are naturally possible. In this paper we consider one of such combinations which is exclusively based on the BHEP test statistic this being an interesting feature of the proposed multiple test procedure. Similarly to the MB test, the new multiple test combines four affine invariant MVN tests. Two of them are the BHEP tests based on $\mathrm{B}(h_S)$ and $\mathrm{B}(h_L)$, that were also included in the MB multiple test. The other two tests for MVN are based on the statistics derived in Henze (1997) by letting the bandwidth $h$ tend to zero and to infinity in $\mathrm{B}(h)$. Therefore the resulting multiple test procedure is based on the BHEP test statistic by combining extreme and non-extreme choices of the tuning parameter $h$.

The paper is planned as follows. In Section 2 we identify the two test statistics obtained in Henze (1997) by letting the parameter $h$ in the definition of the BHEP statistic tend to zero and to infinity. Two of the goodness-of-fit tests for MVN that can be associated to these statistics are combined with the tests based on $\mathrm{B}(h_S)$ and $\mathrm{B}(h_L)$ in order to propose a new multiple test for MVN. In Section 3 we define such a multiple test procedure and, as a consequence of the results in Tenreiro (2011), we describe its main properties. In Section 4 we report the results of a simulation study carried out to analyse the finite sample power performance of the new multiple test compared with the MB multiple test, that we take here as a benchmark MVN test. For the generality of the alternative distributions included in our Monte Carlo study, the two tests reveal quite similar results showing a good performance for a wide range of alternative distributions. Finally, in Section 5 the proposed test is illustrated using the Fisher Iris data set and in Section 6 we provide some overall conclusions. All the proofs are deferred to Section 7. The simulations and plots in this paper were carried out using the R software (R Development Core Team, 2011).

## 2 The extreme BHEP test statistics

Henze (1997) proposed and studied tests for multivariate normality whose test statistics are obtained by letting the bandwidth $h$ in the definition of the BHEP statistic tend to zero or to infinity. In this section we identify such test statistics and we describe the main properties of the associated MVN tests. Here, and throughout this article $||\cdot||$ denotes the Euclidean norm in $\mathbb{R}^d$ and $\xrightarrow{d}$ denotes convergence in distribution.

**Lemma 1** (Henze, 1997, Theorems 2.1 and 3.1)**.** *We have:*

*i)* $\underline{\textit{Limit of } \mathrm{B}(h) \textit{ as } h \to \infty}$*:*

$$\lim_{h\to\infty} (2\pi)^{-d/2}(h\sqrt{2}\,)^{d+6}\mathrm{B}(h) = n\big(b_{1,d}/6 + \widetilde{b}_{1,d}/4\big),$$

*where $b_{1,d}$ is the Mardia skewness measure given in (7) and*

$$\widetilde{b}_{1,d} = \frac{1}{n^2} \sum_{j,k=1}^{n} Y_j' Y_k ||Y_j||^2 ||Y_k||^2.$$

*ii)* <u>*Limit of* B$(h)$ *as* $h \to 0$</u>*:*

$$\lim_{h\to 0}\ 2^{-1} n^{-1/2}\big((2\pi)^{-d/2}\mathrm{B}(h) - 2^{-d/2}(h^{-d} - n)\big) = -\sqrt{n}\,\big(\widetilde{b}_{2,d} - 2^{-d/2}\big),$$

*where*

$$\widetilde{b}_{2,d} = \frac{1}{n}\sum_{j=1}^{n} \exp\big(-||Y_j||^2/2\big).$$

Under the null hypothesis of MVN, Henze (1997) established that $n\,(b_{1,d}/6 + \widetilde{b}_{1,d}/4) \xrightarrow{d} \frac{1}{2}(d+4)\chi^2_d + \chi^2_{d(d-1)(d+4)/6}$ and $\sqrt{n}\,\big(\widetilde{b}_{2,d} - 2^{-d/2}\big) \xrightarrow{d} N\big(0, 3^{-d/2} - 2^{-d} - d\,2^{-(d+3)}\big)$, which led him to propose two new MVN tests based on the affine invariant statistics defined as

$$\mathrm{B}(\infty) := n\,(b_{1,d}/6 + \widetilde{b}_{1,d}/4) \tag{8}$$

and

$$\mathrm{B}(0) := \sqrt{n}\,|\,\widetilde{b}_{2,d} - 2^{-d/2}|. \tag{9}$$

In both cases these tests reject $H_0$ for large values of the corresponding test statistics. As noticed by Henze (1997), B$(\infty)$ is based on a weighted sum of the empirical skewness measures $b_{1,d}$ and $\widetilde{b}_{1,d}$, the latter being a sample version of a measure of multivariate skewness introduced and studied by Móri et al. (1993) (see also Henze, 2002). In relation to the statistic $\widetilde{b}_{2,d}$ involved in B(0), Henze (1997) pointed out that it is similar to the Mardia kurtosis measure $b_{2,d}$ in the sense that it only uses information contained in the Mahalanobis distances $||Y_1||^2, \ldots, ||Y_n||^2$.

Taking into account these similarities, it will be not surprising if the tests based on the statistics (8) and (9) share some of the properties of the classical Mardia's tests based on the statistics MS and MK, respectively. As for these tests, the tests based on B$(\infty)$ and B(0) are not consistent against each alternative distribution. Therefore, the universal consistency of the BHEP test for each fixed $0 < h < \infty$ (Csörgő, 1989) is lost in the limit cases $h \to \infty$ and $h \to 0$. Denoting by $\widetilde{\beta}_{1,d} = ||\mathrm{E}\big(||W||^2 W\big)||^2$ and $\widetilde{\beta}_{2,d} = \mathrm{E}\big(\exp(-||W||^2/2)\big)$ the population counterparts to $\widetilde{b}_{1,d}$ and $\widetilde{b}_{2,d}$, respectively, where $W = \Sigma^{-1/2}(X - \mu)$, $\mu$ and $\Sigma$ are the mean vector and the covariance matrix of $X$, and $\Sigma^{-1/2}$ is the symmetric positive definite square root of $\Sigma^{-1}$, Henze (1997) showed that if $\mathrm{E}||X||^6 < \infty$ and the alternative distribution is supported by a set of positive $d$-dimensional Lebesgue measure, then the MVN test based on $\widetilde{b}_{1,d}$ is consistent if $\widetilde{\beta}_{1,d} > 0$. Additionally, he established that if $\mathrm{E}||X||^2 < \infty$, the MVN test based on $\widetilde{b}_{2,d}$ is consistent if and only if $\widetilde{\beta}_{2,d}$ differs from $2^{-d/2}$.

# 3 A new multiple test for assessing MVN

The new test for testing MVN proposed in this paper, labelled BB henceforth, is based on the combination of the extreme BHEP statistics given by (8) and (9), and the non-extreme BHEP statistics $\mathrm{B}(h)$ with $h = h_S$ and $h = h_L$ given by (2) and (3), respectively.

## 3.1 Definition and finite sample behaviour under $H_0$

For $u \in ]0,1[$ and

$$T_{n,1} = \mathrm{B}(0), \quad T_{n,2} = \mathrm{B}(h_\mathrm{S}), \quad T_{n,3} = \mathrm{B}(h_\mathrm{L}) \quad \text{and} \quad T_{n,4} = \mathrm{B}(\infty), \tag{10}$$

consider the corrected statistic

$$\mathbf{T}_n(u) = \max_{h \in H} \big(T_{n,h} - c_{n,h}(u)\big), \tag{11}$$

where $H = \{1,2,3,4\}$ and $c_{n,h}(u)$ is the quantile of order $1-u$ of the test statistic $T_{n,h}$ under the null hypothesis of MVN. As the test statistics $T_{n,h}$, $h \in H$, are affine invariant, that is, $T_{n,h}(AX_1 + b, \ldots, AX_n + b) = T_{n,h}(X_1, \ldots, X_n)$, for all $b \in \mathbb{R}^d$ and nonsingular matrix $A$, and $f_X \in \mathcal{N}_d$ if and only if $f_{AX+b} \in \mathcal{N}_d$, the quantile $c_{n,h}(u)$ does not depend on the distribution considered under the null hypothesis, and therefore $\mathbf{T}_n(u)$ is affine invariant for every $u \in ]0,1[$.

For a preassigned level of significance $\alpha \in ]0,1[$, the BB multiple test is defined as the test procedure that rejects the null hypothesis of MVN whenever

$$\mathbf{T}_n(u_{n,\alpha}) > 0$$

where

$$u_{n,\alpha} = \sup I_{n,\alpha} \tag{12}$$

with

$$I_{n,\alpha} = \big\{u \in ]0,1[ : \mathrm{P}_\phi(\mathbf{T}_n(u) > 0) \leq \alpha\big\},$$

and $\phi$ the $d$-dimensional standard Gaussian density.

Taking into account that $\alpha/4 \leq u_{n,\alpha}$, we conclude that the BB multiple test is at least as powerful as the Bonferroni procedure that leads to the rejection of $H_0$ if at least one of the test statistics $T_{n,h}$, for $h \in H$, is larger than its quantile of order $1 - \alpha/4$.

Similarly to the Bonferroni test procedure based on $T_{n,h}$, for $h \in H$, the next non-asymptotic result, which is a consequence of Theorem 1 of Tenreiro (2011), states that the BB multiple test has a level of significance that is at most equal to $\alpha$.

| Sample size | Data dimension | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 7 | 10 |
| | $\alpha = 0.01$ | | | | | |
| 20 | .0037 | .0032 | .0033 | .0031 | .0029 | .0032 |
| 40 | .0032 | .0032 | .0030 | .0029 | .0028 | .0026 |
| 60 | .0032 | .0032 | .0031 | .0027 | .0030 | .0025 |
| 80 | .0032 | .0029 | .0029 | .0029 | .0028 | .0028 |
| 100 | .0036 | .0030 | .0030 | .0025 | .0030 | .0027 |
| 200 | .0030 | .0030 | .0033 | .0029 | .0030 | .0029 |
| 400 | .0033 | .0026 | .0031 | .0029 | .0032 | .0030 |
| | $\alpha = 0.05$ | | | | | |
| 20 | .0187 | .0169 | .0167 | .0167 | .0159 | .0159 |
| 40 | .0173 | .0176 | .0167 | .0161 | .0162 | .0145 |
| 60 | .0182 | .0163 | .0166 | .0162 | .0160 | .0152 |
| 80 | .0186 | .0170 | .0168 | .0158 | .0152 | .0149 |
| 100 | .0173 | .0163 | .0159 | .0152 | .0152 | .0150 |
| 200 | .0171 | .0165 | .0171 | .0159 | .0157 | .0154 |
| 400 | .0179 | .0161 | .0170 | .0164 | .0163 | .0158 |

Table 1: Estimates of $u_{n,\alpha}$ for $\alpha = 0.01, 0.05$ based on a regular grid of size 0.0001 on the interval $]0,1[$. The number of replications for each stage of the estimation process is 50,000.

**Theorem 1.** *For $n > d$ and $0 < \alpha < 1$ we have $\mathrm{P}_f(\mathbf{T}_n(u_{n,\alpha}) > 0) \leq \alpha$, for all $f \in \mathcal{N}_d$.*

In practice, the value $u_{n,\alpha}$, the level at which each one of the tests $T_{n,h}$, $h \in H$, is performed, is estimated by Monte Carlo experiments under the null hypothesis as described in Fromont and Laurent (2006). We have used 50,000 simulations under the null hypothesis of the involved test statistics and the R function quantile(·,type=7) for estimating the $1-u$ quantiles $c_{n,h}(u)$ for $u$ varying on a regular grid, $u_{i+1} = u_i + p$ with $u_1 = p$ and $p = 0.0001$, on the interval $]0,1[$, and further 50,000 simulations were used for estimating the probabilities $\mathrm{P}_\phi(\, \mathbf{T}_n(u) > 0\,)$. Finally, we have taken the largest value of $u$ that satisfies $\mathrm{P}_\phi(\mathbf{T}_n(u) > 0) \leq \alpha$ as an approximation for $u_{n,\alpha}$ defined by (12).

For $\alpha = 0.01$ and $\alpha = 0.05$, and several sample sizes $n$ and data dimensions $d$, we present in Table 1 the estimated levels $u_{n,\alpha}$ based on a regular grid of size $p = 0.0001$. For the large majority of the considered combinations, the estimated level $u_{n,\alpha}$ is clearly larger than $\alpha/4$, the level at which each one of the tests $T_{n,h}$, $h \in H$, is performed whenever a Bonferroni multiple test based on these statistics is used. However, for $\alpha = 0.01$ and for some of the considered sample size and data dimension combinations, the estimated level $u_{n,\alpha}$ is close to $\alpha/4$, which means that the considered multiple test BB is, in those cases, close to the Bonferroni test procedure.

| Sample size | Data dimension | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 7 | 10 |
| | $\alpha = 0.01$ | | | | | |
| 20 | .0098 | .0099 | .0095 | .0087 | .0099 | .0095 |
| 40 | .0092 | .0095 | .0096 | .0094 | .0093 | .0089 |
| 60 | .0093 | .0096 | .0092 | .0083 | .0097 | .0082 |
| 80 | .0097 | .0010 | .0087 | .0091 | .0088 | .0098 |
| 100 | .0011 | .0098 | .0089 | .0082 | .0098 | .0092 |
| 200 | .0087 | .0096 | .0095 | .0091 | .0099 | .0087 |
| 400 | .0098 | .0087 | .0091 | .0092 | .0093 | .0094 |
| | $\alpha = 0.05$ | | | | | |
| 20 | .0504 | .0471 | .0499 | .0482 | .0498 | .0496 |
| 40 | .0499 | .0512 | .0508 | .0509 | .0515 | .0485 |
| 60 | .0512 | .0466 | .0473 | .0462 | .0497 | .0479 |
| 80 | .0533 | .0496 | .0494 | .0490 | .0479 | .0481 |
| 100 | .0502 | .0490 | .0477 | .0477 | .0481 | .0482 |
| 200 | .0488 | .0481 | .0496 | .0487 | .0494 | .0477 |
| 400 | .0494 | .0487 | .0492 | .0496 | .0506 | .0491 |

Table 2: Estimates of the nominal level of significance of the multiple test BB for a preassigned level $\alpha$. The number of replications for each case is 100,000.

For the previously considered values of $\alpha$, $n$ and $d$, Table 2 shows estimates for the nominal levels of significance of the BB test based on 100,000 simulations under the null hypothesis. These estimates were evaluated by using an approximation of the $p$-value of the BB test that can be obtained along the lines described in Tenreiro (2011, p. 1991). The R program for computing the $p$-value may be obtained from the author. Although we were not able to prove that the BB test has an exact $\alpha$-level of significance, a sufficient condition for which being the continuity of the distribution function of the statistics $T_{n,h}$, for all $h \in H$ (see Theorem 1 of Tenreiro, 2011), we conclude from Table 2 that the previous implementation enables us to obtain a multiple test procedure with an attained level of significance very close to $\alpha$. With some few exceptions the estimated levels are inside the approximate 95% confidence interval for the preassigned level $\alpha$.

## 3.2 Consistency against fixed and local alternatives

The next result, which is a consequence of Theorem 1 of Tenreiro (2011), states that the BB multiple test is consistent against each fixed alternative.

**Theorem 2.** *For* $0 < \alpha < 1$ *we have* $\mathrm{P}_f(\mathbf{T}_n(u_{n,\alpha}) > 0) \to 1,$ *as* $n \to +\infty$, *for all* $f \notin \mathcal{N}_d$.

A similar result is valid for a sequence of local alternatives converging to the null density function at a rate slower than $n^{-1/2}$. To define local alternatives we consider $X_{n1}, \ldots, X_{nn}, \ldots$ a sequence of independent and identically distributed $d$-dimensional absolutely continuous random vectors with mean $\mu_n$ and nonsingular covariance matrix $\Sigma_n$, whose probability density function $f_n$ satisfies

$$f_n^*(x) = \phi(x)\big(1 + \gamma_n \eta(x) + o(\gamma_n)\eta_n(x)\big),$$

for $x \in \mathbb{R}^d$, with $f_n^*(x) = |\Sigma_n^{1/2}| f_n\big(\mu_n + \Sigma_n^{1/2} x\big)$, $\eta$ an a.e. non-identically null function, $(\gamma_n)$ a sequence of positive real numbers tending to zero as $n$ tends to infinity, and the functions $\eta$ and $\eta_n$ satisfy

$$\sup_{x \in \mathbb{R}^d} |\eta(x)| < \infty, \quad \sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} |\eta_n(x)| < \infty.$$

**Theorem 3.** *For $0 < \alpha < 1$ we have $\mathrm{P}_{f_n}(\mathbf{T}_n(u_{n,\alpha}) > 0) \to 1,$ as $n \to +\infty$, for a sequence of local alternatives with $n^{-1/2} = o(\gamma_n)$.*

# 4 Finite sample power analysis

In this section we present the results of a simulation study that was conducted to compare the empirical power performance of the new BB multiple test against the MB multiple test proposed by Tenreiro (2011). We recall that the latter test is defined similarly to the BB multiple test with $T_{n1} = \mathrm{MK}$, $T_{n2} = \mathrm{B}(h_\mathrm{S}), T_{n,3} = \mathrm{B}(h_\mathrm{L})$ and $T_{n,4} = \mathrm{MS}$. Based on the Monte Carlo results presented in Tenreiro (2011), we know that the MB test procedure reveals a good empirical power for a wide range of alternative distributions, and shows an overall good performance against the most recommended procedures for testing MVN such as the Henze and Zirkler (1990) test which is based on $\mathrm{B}(h_\mathrm{HZ})$ with $h_\mathrm{HZ} = 1.41$, the BHEP test based on $\mathrm{B}(\bar{h})$ with $\bar{h}$ given by (4), and the test proposed by Székely and Rizzo (2005), among others (see Tenreiro, 2011, p. 1986). For this very reason the MB test is considered here as a benchmark test for testing MVN against which we will compare the new multiple test proposed in this paper.

## 4.1 The alternative distributions

A wide set of alternative distributions, including all the distributions considered in Tenreiro (2009, 2011), was selected to our study. This set includes distributions previously considered in other simulations studies such as those of Henze and Zirkler (1990), Romeu and Ozturk (1993), Mecklin and Mundfrom (2005) and Székely and Rizzo (2005). We investigate: i) some symmetric distributions from Pearson's Types II and VII families (see Johnson, 1987, p. 110–121); ii) some heavily skewed distributions such as the multivariate $\chi^2_1$ and the multivariate lognormal with independent marginals, and some members of the multivariate asymmetric Laplace family (see Kotz et al., 2001, chapter 6); iii) some distributions with some characteristics identical to MVN such as (meta-)Burr-Pareto-Logistic distributions with normal marginals (see Johnson, 1987, chapter 9) and Khintchine distributions with generalised exponential power marginal distributions (see Johnson, 1987, chapter 8 and paragraph 2.4); iv) some mixtures of two multivariate normals (location and scale mixtures) in order to assess the effect of data contamination; and finally v) a set of meta-Gaussian distributions whose marginal distributions, given in Table 2 of Tenreiro (2009, p. 1043), are members of the generalised lambda family discussed in Ramberg and Schmeiser (1974). For a detailed description of all these alternatives see Tenreiro (2009, p. 1045; 2011, p. 1986).

## 4.2 Empirical power comparisons

The Monte Carlo results presented in this section are based on 10,000 samples of different sizes ($n = 20, 40, 60, 80, 100, 200, 400$) and data dimensions ($d = 2, 3, 4, 5, 7, 10$) from the considered set of alternative distributions. With this number of repetitions the margin of error for approximate 95% confidence intervals for the true power does not exceed 0.01. The standard level of significance $\alpha = 0.05$ was used.

The observed numerical results indicate that the tests MB and BB exhibit a similar behaviour for the large majority of the considered alternative distributions. This is particularly clear when one of the considered non-extreme BHEP tests, $\mathrm{B}(h_\mathrm{S})$ or $\mathrm{B}(h_\mathrm{L})$, is, by a wide margin, the best of the tests involved in both multiple test procedures for a given alternative distribution. As these tests are included in both multiple tests, the power performances of MB and BB are quite similar for such alternatives. This is illustrated in Figures 1 and 2, for two normal location mixture distributions of the form $pN_d(0, I) + (1 - p)N_d(\mu, I)$ with $p = 0.5$ (centrally symmetric with tails lighter than MVN) and $p = 0.9$ (asymmetric with tails heavier than MVN), where $I$ is the $d$-dimensional identity matrix and $\mu = (3, \ldots, 3)$. Besides the empirical power of the two multiple tests we want to compare, we also present the empirical power of each one of the tests involved in both multiple test procedures. The same situation is reported in Figure 3 where we consider a Khintchine alternative whose values of the Mardia skewness and kurtosis are equal to the MVN ones (high moment alternative), which explains the poor performance of the tests MK, MS, B(0) and B($\infty$) for this alternative.

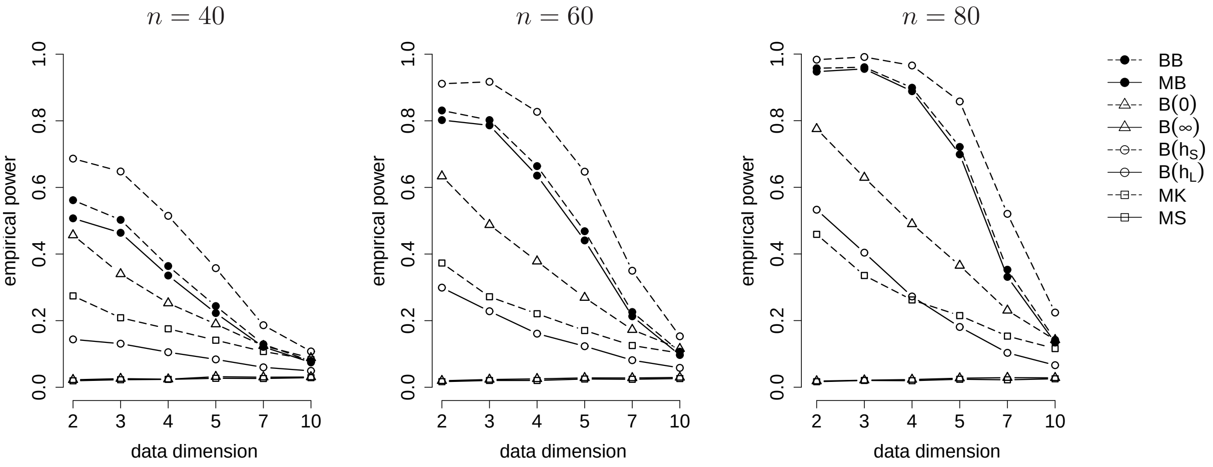

Figure 1: Normal location mixture distribution with $p = 0.5$.

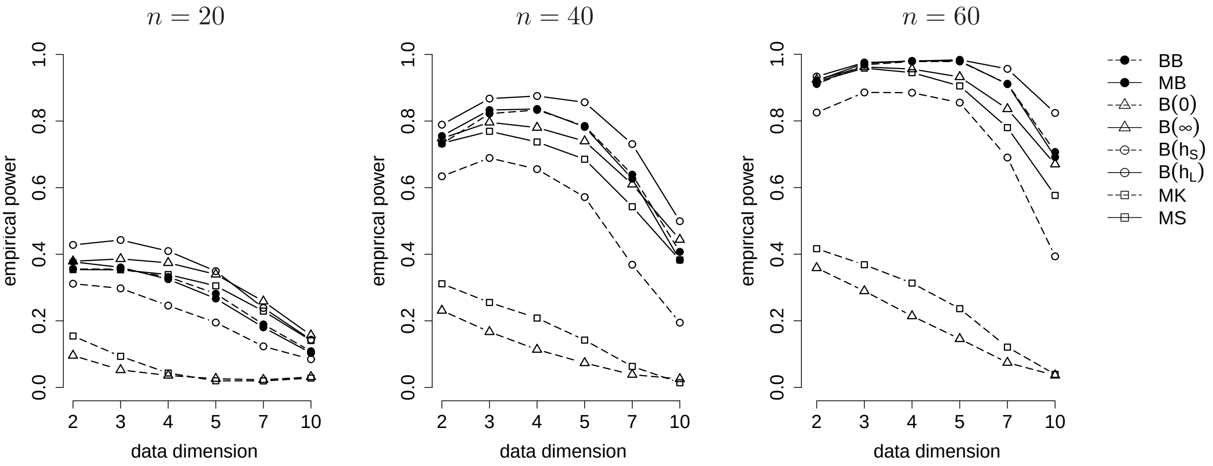

Figure 2: Normal location mixture distribution with $p = 0.9$.

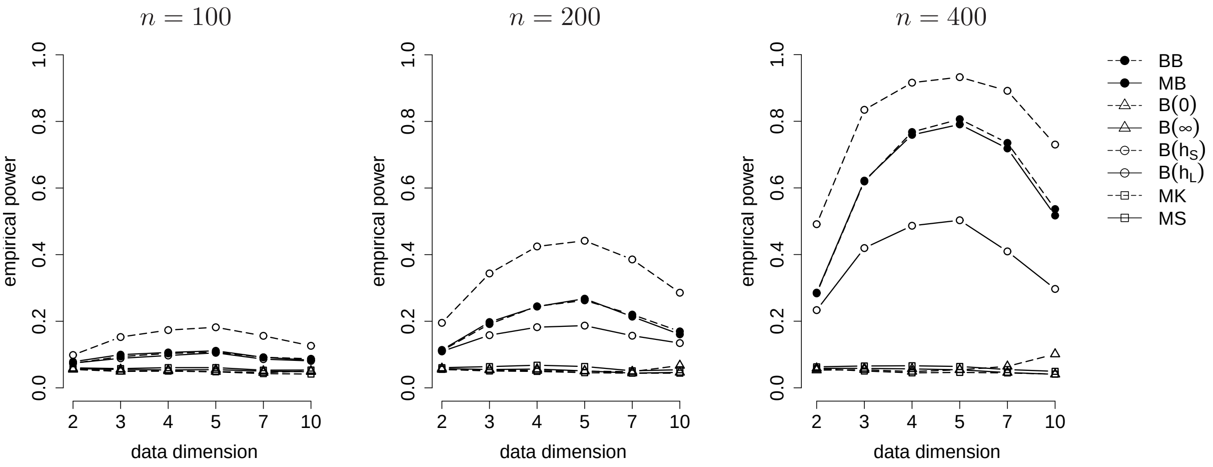

Figure 3: High moment Khintchine distribution with GEP marginals.

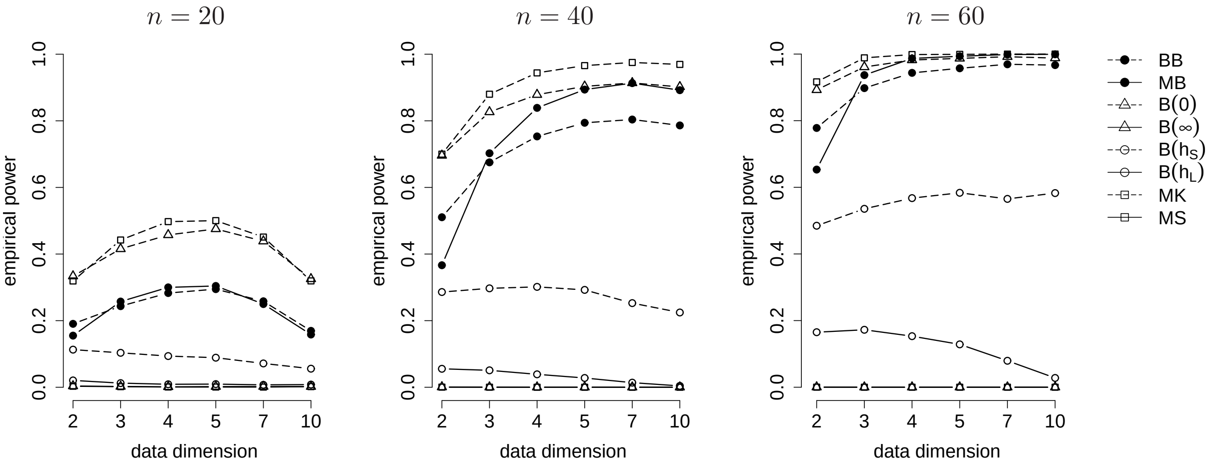

Figure 4: Pearson Type II distribution with $m = 0.5$.

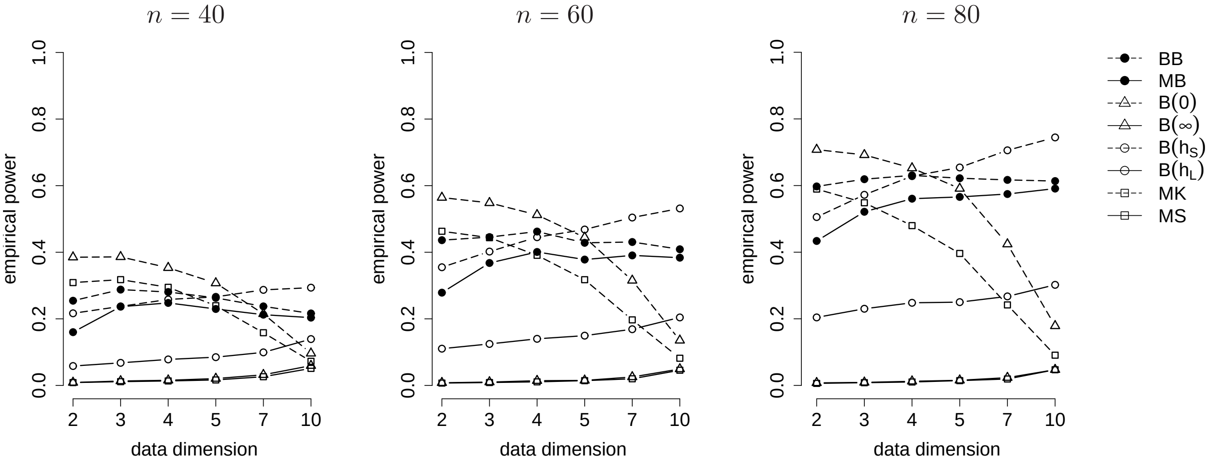

Figure 5: Meta-Gaussian distribution with marginals S.$2_d(0.7)$.

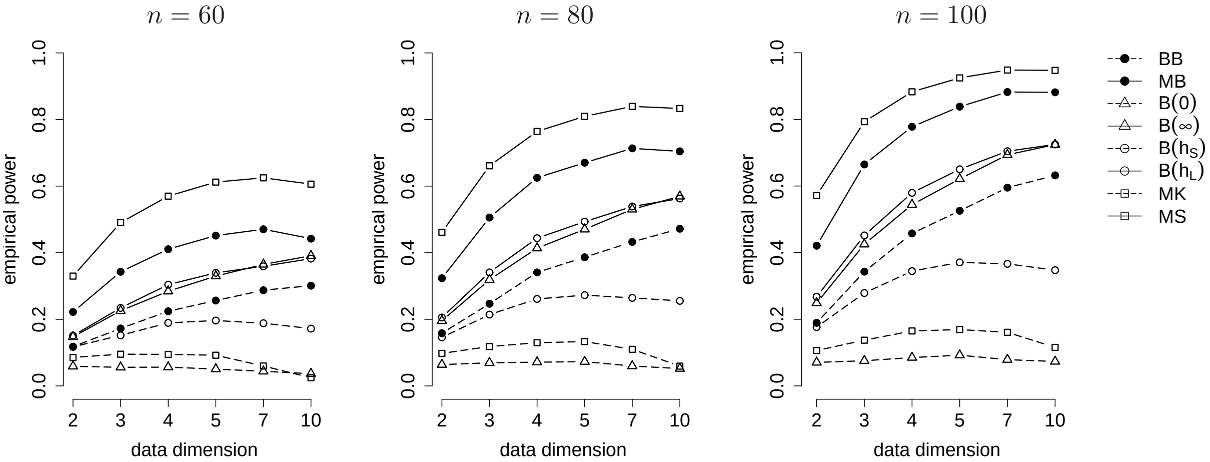

Figure 6: Burr-Pareto-Logistic distribution with normal marginals and $\alpha = 1$.

Next we report three other situations where the empirical powers of the two multiple tests are not as close as before, which can be mainly explained by the distinct power performances of the tests B(0) and MK, or B($\infty$) and MS, for the considered alternatives. This is illustrated in Figures 4 and 5 for a Pearson Type II distribution (elliptically symmetric with tails lighter than MVN) and for a centrally symmetric meta-Gaussian distribution with generalised lambda marginals (tails lighter than MVN), and in Figure 6 for a Burr-Pareto-Logistic distribution with normal marginals (asymmetric with kurtosis close to the MVN one). For this latter alternative we observed the greatest difference between the power of the tests MB and BB among the considered set of alternative distributions. Due to the exceptional good performance of the Mardia MS test in relation to B($\infty$), the MB test outperforms the BB test by a wide margin, and this occurs uniformly in relation to the considered sample sizes and data dimensions.

Taking into account the power performance of the BB test for all the considered alternative distributions, sample sizes and data dimensions (not presented here for brevity's sake), we conclude that the new test reveals a reasonable performance for a wide range of alternative distributions, demonstrating itself to be competitive against the multiple MB test which has shown an overall good performance against the most recommended procedures for testing MVN (Tenreiro, 2011, p. 1986).

# 5 An example: the Fisher Iris data

We consider in this section the well-known iris data of Fisher (1936). This data comprises flower measurements from three iris species of fifty plants each: iris setosa, iris versicolor and iris virginica. For each plant four measurements in centimeters were taken: sepal length, sepal width, petal length and petal width. We consider testing for multivariate normality of the four measurements for each of the considered species.

Approximations of the $p$-values of the multiple tests BB and MB are reported in Table 3. At level $\alpha = 0.05$, multivariate normality is not rejected for iris versicolor and iris virginica data which is compatible with the results obtained by Beirlant et al. (1999, p. 124). In relation to the iris setosa data set, the normality assumption was rejected by Small (1980) through a test statistic that combines marginal skewness and kurtosis, and by two of the three test statistics considered by Beirlant et al. (1999, p. 124–125). Although smaller $p$-values were observed for this data set when compared with the $p$-values observed for the other two iris species, the MVN hypothesis was not rejected by any of the multiple tests BB or MB, which agrees with the result obtained by the third MVN test considered by

| Data set | Multiple tests | |
|---|---|---|
| | BB | MB |
| Iris setosa | $0.1385 < p\text{-value} < 0.139$ | $0.1405 < p\text{-value} < 0.141$ |
| Iris versicolor | $0.3 < p\text{-value} < 0.35$ | $0.3 < p\text{-value} < 0.35$ |
| Iris virginica | $0.3 < p\text{-value} < 0.35$ | $0.3 < p\text{-value} < 0.35$ |

Table 3: Approximations of the $p$-values of the multiple tests BB and MB for the three iris species measurements of fifty observations each of the Fisher Iris data set.

Beirlant et al. (1999, p. 125). Of the six tests involved in the two multiple test procedures, only the BHEP tests $\mathrm{B}(h_\mathrm{S})$ and $\mathrm{B}(h_\mathrm{L})$ reject, and by a small margin, the MVN hypothesis for the iris setosa data. In fact, we obtain for both tests approximate $p$-values between 0.049 and 0.0495.

# 6 Conclusions

In this paper we propose a new multiple test procedure for assessing MVN which combines tests from the BHEP family by considering extreme and non-extreme choices of the tuning parameter figuring in the definition of the BHEP test statistic. Contrary to the multiple test MB previously proposed by the author, which combines the Mardia and non-extreme BHEP tests (Tenreiro, 2011), the new test exclusively combines test statistics based on the BHEP family, this being an interesting feature of the proposed test procedure. The Monte Carlo study indicates that the new test presents a reasonable performance for a wide range of alternative distributions, which is a desirable feature particularly when no information about the alternative hypothesis is available.

# 7 Proofs

**Proof of Theorem 1:** Following closely the proof of Theorem 3 in Tenreiro (2011, p. 1992) we conclude that the null distribution function $F_{T_{n,h}}$ of each one of the statistics (10) is strictly increasing. Thus, from Theorem 1 of Tenreiro (2011) we conclude that the BB multiple test $I(\mathbf{T}_n(u_{n,\alpha}) > 0)$ has a level of significance less than or equal to $\alpha$.

□

**Proof of Theorem 2:** Given $f$ a non-normal density, we have

$$T_{n,2} = \mathrm{B}(h_\mathrm{S}) \xrightarrow{p} +\infty \quad \text{under } f, \tag{13}$$

(see Csörgő, 1989), where $\xrightarrow{p}$ denotes convergence in probability, and

$$\mathrm{P}_f\big(\mathbf{T}_n(u_{n,\alpha}) > 0\big) \geq \mathrm{P}_f\big(T_{n,2} > c_{n,2}(u_{n,\alpha})\big) \geq \mathrm{P}_f\big(T_{n,2} > c_{n,2}(\alpha/4)\big) \tag{14}$$

(the same reasoning could be based on $T_{n,3} = \mathrm{B}(h_{\mathrm{L}})$).

Moreover, $T_{n,2}$ has a weighted sum of $\chi^2$ independent random variables as limiting null distribution (see Baringhaus and Henze, 1988). Denoting this limiting random variable by $T_{\infty,2}$, from the continuity of $F^{-1}_{T_{\infty,2}}$ and the convergence $F^{-1}_{T_{n,2}}(t) \longrightarrow F^{-1}_{T_{\infty,2}}(t)$, for all $0 < t < 1$ (see Shorack and Wellner, 1986, p. 10), we get

$$c_{n,2}(\alpha/4) = F^{-1}_{T_{n,2}}(1-\alpha/4) \longrightarrow F^{-1}_{T_{\infty,2}}(1-\alpha/4). \tag{15}$$

Finally, from (13), (14) and (15) we deduce that

$$\mathrm{P}_f\big(\mathbf{T}_n(u_{n,\alpha}) > 0\big) \geq \mathrm{P}_f\big(T_{n,2} > \sup_{n\in\mathbb{N}} c_{n,2}(\alpha/4)\big) \longrightarrow 1.$$

□

**Proof of Theorem 3:** Following the proof of Theorem 2, for a sequence $f_n$ of local alternatives we have

$$\mathrm{P}_{f_n}\big(\mathbf{T}_n(u_{n,\alpha}) > 0\big) \geq \mathrm{P}_{f_n}\big(T_{n,2} > \sup_{n\in\mathbb{N}} c_{n,2}(\alpha/4)\big).$$

The stated result follows now from the fact that $T_{n,2} = \mathrm{B}(h_{\mathrm{S}}) \xrightarrow{p} +\infty$ under $f_n$ whenever $n^{-1/2} = o(\gamma_n)$ (see Tenreiro, 2007, p. 115).

□

**Acknowledgments.** The author expresses his thanks to Professor Norbert Henze who has suggested replacing the Mardia tests used in the MB multiple test by the extreme BHEP statistics, and to the reviewers for their comments and suggestions. This research was partially supported by the Centro de Matemática da Universidade de Coimbra (funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT–Fundação para a Ciência e Tecnologia under the project PEst-C/MAT/UI0324/2013).

# References

Arcones, M.A., 2007. Two tests for multivariate normality based on the characteristic function. Math. Methods Statist. 16, 177–201.

Baringhaus, L., Henze, N., 1988. A consistent test for multivariate normality based on the empirical characteristic function. Metrika 35, 339–348.

Baringhaus, L., Henze, N., 1992. Limit distributions for Mardia's measure of multivariate skewness. Ann. Statist. 20, 1889–1902.

Beirlant, J., Mason, D.M., Vynckier, C., 1999. Goodness-of-fit analysis for multivariate normality based on generalized quantiles. Comput. Statist. Data Anal. 30, 119–142.

Bowman, A.W., Foster, P.J., 1993. Adaptive smoothing and density-based tests of multivariate normality. J. Amer. Statist. Assoc. 88, 529–537.

Chiu, S.N., Liu, K.I., 2009. Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions. Comput. Statist. Data Anal. 53, 3817–3834.

Csörgő, S., 1986. Testing for normality in arbitrary dimension. Ann. Statist. 14, 708–723.

Csörgő, S., 1989. Consistency of some tests for multivariate normality. Metrika 36, 107–116.

Dykstra, R.L., 1970. Establishing the positive definiteness of the sample covariance matrix. Ann. Math. Statist. 41, 2153–2154.

Ebner, B., 2012. Asymptotic theory for the test for multivariate normality by Cox and Small. J. Multivariate Anal. 111, 368–379.

Epps, T.W., Pulley, L.B., 1983. A test for normality based on the empirical characteristic function. Biometrika 70, 723–726.

Fan, Y., 1998. Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. Econometric Theory 14, 604–621.

Farrel, P.J., Salibian-Barrera, M., Naczk, K., 2007. On tests for multivariate normality and associated simulation studies. J. Stat. Comput. Simul. 77, 1065–1080.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics 7, 179–188.

Fromont, M., Laurent, B., 2006. Adaptive goodness-of-fit tests in a density model. Ann. Statist. 34, 680–720.

Gürtler, N. (2000). Asymptotic theorems for the class of BHEP-tests for multivariate normality with fixed and variable smoothing parameter (in German). Doctoral dissertation, University of Karlsruhe, Germany.

Henze, N., 2002. Invariant tests for multivariate normality: a critical review. Statist. Papers 43, 467–506.

Henze, N., Zirkler, B., 1990. A class of invariante consistent tests for multivariate normality. Comm. Stat. Theory Methods 19, 3595–3617.

Henze, N., 1997. Extreme smoothing and testing for multivariate normality. Statist. Probab. Lett. 35, 203–213.

Henze, N., Wagner, T., 1997. A new approach to the BHEP tests for multivariate normality. J. Multivariate Anal. 62, 1–23.

Johnson, M.E., 1987. Multivariate Statistical Simulation, Wiley, New York.

Kotz, S., Kozubowski, T., Podgorski, K., 2001. The Laplace Distribution and Generalizations, Birkhauser, Boston.

Liang, J., Pan, W.S.Y., Yang, Z.-H., 2005. Characterization-based Q-Q plots for testing multinormality. Statis. Probab. Lett. 70, 183–190.

Liang, J., Tang, M.-L., Chan, P.S., 2009. A generalized Shapiro–Wilk W statistic for testing high-dimensional normality. Comput. Statist. Data Anal. 53, 3883–3891.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530.

Mecklin, C.J., Mundfrom, D.J., 2004. An appraisal and bibliography of tests for multivariate normality. Int. Stat. Rev. 72, 123–138.

Mecklin, C.J., Mundfrom, D.J., 2005. A Monte Carlo comparison of Type I and Type II error rates of tests of multivariate normality. J. Stat. Comput. Simul. 75, 93–107.

Móri, T.F., Rohatgi, V.K., Székely, G.J., 1993. On multivariate skeweness and kurtosis. Theory Probab. Appl. 38, 547–551.

Oliveira, I.R.C., Ferreira, D.F., 2010. Multivariate extension of chi-squared univariate normality test. J. Stat. Comput. Simul. 80, 513–526.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org

Ramberg, J.S., Schmeiser, B.W., 1974. An approximate method for generating asymmetric random variables. Commun. ACM 17, 78–82.

Rayner, J.C.W., Best, D.J., 1989. *Smooth tests of goodness of fit.* New York: Oxford University Press.

Romeu, J.L., Ozturk, A., 1993. A comparative study of goodness-of-fit tests for multivariate normality. J. Multivariate Anal. 46, 309–334.

Shorack, G.R., Wellner, J.A., 1986. Empirical Processes with Applications to Statistics, Wiley, New York.

Small, N.J.H., 1980. Marginal skewness and kurtosis in testing multivariate normality. J. Roy. Statist. Soc. Ser. C 29, 85–87.

Sürücü, B., 2006. Goodness-of-fit tests for multivariate distributions. Comm. Statist. Theory Methods 35, 1319–1331.

Székely, G.J., Rizzo, M.L., 2005. A new test for multivariate normality. J. Multivariate Anal. 93, 58–80.

Tenreiro, C., 2007. On the asymptotic behaviour of location-scale invariant Bickel-Rosenblatt tests. J. Statist. Plann. Inference 137, 103–116. Erratum: 139, 2115, 2009.

Tenreiro, C., 2009. On the choice of the smoothing parameter for the BHEP goodness-of-fit test. Comput. Statist. Data Anal. 53, 1038–1053.

Tenreiro, C., 2011. An affine invariant multiple test procedure for assessing multivariate normality. Comput. Statist. Data Anal. 55, 1980–1992.

Thode, Jr., H.C., 2002. *Testing for normality.* New York: Marcel Dekker.

Wang, C.-C., 2015. A MATLAB package for multivariate normality test. J. Stat. Comput. Simul. 85, 166–188.