Are You All Normal? It Depends!

Wanfang Chen[1], Marc G. Genton[2]

August 26, 2020

Abstract

The assumption of normality has underlain much of the development of statistics, including spatial statistics, and many tests have been proposed. In this work, we focus on the multivariate setting and we first provide a synopsis of the recent advances in multivariate normality tests for i.i.d. data, with emphasis on the skewness and kurtosis approaches. We show through simulation studies that some of these tests cannot be used directly for testing normality of spatial data, since the multivariate sample skewness and kurtosis measures, such as the Mardia's measures, deviate from their theoretical values under Gaussianity due to dependence, and some related tests exhibit inflated type I error, especially when the spatial dependence gets stronger. We review briefly the few existing tests under dependence (time or space), and then propose a new multivariate normality test for spatial data by accounting for the spatial dependence of the observations in the test statistic. The new test aggregates univariate Jarque-Bera (JB) statistics, which combine skewness and kurtosis measures, for individual variables. The asymptotic variances of sample skewness and kurtosis for standardized observations are derived given the dependence structure of the spatial data. Consistent estimators of the asymptotic variances are then constructed for finite samples. The test statistic is easy to compute, without any smoothing involved, and it is asymptotically χ^2_{2p} under normality, where p is the number of variables. The new test has a good control of the type I error and a high empirical power, especially for large sample sizes.

Key words: Gaussian process, Jarque-Bera test, Skewness and Kurtosis, Spatial Dependence, Spatial Statistics, Tests for Multivariate Normality

---

[1](Corresponding author) Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail: wanfang.chen@kaust.edu.sa

[2]Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail: marc.genton@kaust.edu.sa

1   Introduction

Normality is one of the most commonly made assumptions in the development and use of statistical procedures, such as t-tests, tests for regression coefficients, the F-test of homogeneity of variance, discriminant analysis and analysis of variance (ANOVA). The performance of these procedures can be affected to various extents if the normality assumption is violated (see, e.g., Pitman (1938), Geary (1947), Box (1953), Tukey (1960), Subrahmaniam et al. (1975), D'Agostino and Lee (1977), and Looney (1995)). Hence, the problem of testing whether a sample of observations comes from a normal distribution or not has received much attention, and numerous methods for testing for normality have been developed. There is now a very large body of literature on tests for univariate normality; for a review of classical tests, see, e.g., Mardia (1980), D'Agostino and Stephens (1986) and Thode (2002), and for comparative studies on the power of selected normality tests, see, e.g., Shapiro et al. (1968), Pearson et al. (1977), Keskin (2006), Öztuna et al. (2006), Farrell and Rogers-Stewart (2006), Thadewald and Büning (2007), Yazici and Yolacan (2007), Romao et al. (2010), Yap and Sim (2011), Noughabi and Arghami (2011), Ahmad and Khan (2015), Islam (2017) and Sánchez-Espigares et al. (2019).

Relatively less work has been done in the field of testing for multivariate normality (MVN) compared to that done for the univariate case, since there can be many difficult cases for MVN. For instance, non-normal distributions that have all lower-dimensional marginals being normal (see, e.g., Dutta and Genton (2014)), and classical univariate normality tests, such as the χ^2-test, have limited applicability in higher dimensions. Reviews on the tests for MVN have been given by Thode (2002), Henze (2002) and Ebner and Henze (2020), with the last one emphasizing on several classes of the weighted L^2-statistics. Evaluation on the power of various tests for MVN is quite sparse, and among the more comprehensive studies are those of Horswell and Looney (1992), Romeu and Ozturk (1993), Mecklin and Mundfrom (2005), Farrell et al. (2007), Joenssen and Vogel (2014) and Hanusz et al. (2018). The Jarque-Bera (JB) type test (Jarque and Bera, 1981), which combines the sample skewness and kurtosis measures, is among one of the most commonly used tests due to its simplicity and good power properties.

In spatial statistics applications, the Gaussian assumption is also widely used to improve

finite-sample inference and effectively employ Bayesian methods. Zimmerman and Stein (2010) and Gelfand and Schliep (2016) provided surveys of Gaussian modeling in spatial statistics. Recent research has focused on applying spatial statistical methods based on the Gaussian assumption to large datasets and advancing computational approaches; see, e.g., Nychka et al. (2015), Paciorek et al. (2015), Katzfuss (2017) and Guhaniyogi and Banerjee (2018). Despite the prevalence of the Gaussian assumption made in spatial statistics, there appears to be very few significance tests that could be used to assess if it is reasonable to assume that a given spatial dataset can be treated as a realization of a Gaussian random field. All the aforementioned tests cannot be directly used for spatial data, since they are designed for examining the normality in a random sample (i.e., i.i.d. observations), so that the conventional large-sample approximations to the null distributions of the test statistics are either unknown or inaccurate under spatial dependence. In this work, we show that the sample skewness and kurtosis deviate from their theoretical values in the i.i.d. case as the degree of spatial dependence increases. Hence, the usual test of normality based on the sample skewness and kurtosis may be misleading if the observations in the sample are actually dependent, as also indicated by the inflated type I error from a further simulation study.

A review on univariate normality tests for data with serial dependence in time series is given by Psaradakis and Vávra (2020), but these tests need to be justified, extended or modified if they are to be applied to spatial data, and further generalized to the multivariate setting, which is not always possible. Pardo-Igúzquiza and Dowd (2004) demonstrated a methodology for the application of standard univariate normality tests, such as the Kolmogorov–Smirnov test, the chi-square test, and the Shapiro–Wilks test, to spatially correlated data, using block kriging in de-clustering to obtain unbiased estimates of the probability density function or the cumulative density function. Olea and Pawlowsky-Glahn (2009) and Zheng (2019) investigated the Kolmogorov-Smirnov test under spatial correlations, using bootstrap methods or Monte Carlo procedures. However, these tests are either difficult to implement or computationally intensive. Horváth et al. (2020) developed a JB-type test for spatial data defined on a grid under the assumption of stationarity by accounting for the spatial dependence of the observations. The

test is easy to implement, shown to have good empirical size and power, and can be justified asymptotically. To our knowledge, no normality test for multivariate spatial data has been proposed yet.

The goal of this study is twofold. First, we aim at providing a comprehensive review on recent MVN tests for i.i.d. data based on skewness and kurtosis approaches, proposed since the review works by Thode (2002) and Henze (2002). Second, we propose a MVN test for spatially correlated data by extending the test of Horváth et al. (2020) to the multivariate setting to assess if a multivariate spatial dataset can be assumed to be a realization from a multivariate Gaussian random field. The type I error and empirical power of the new test are assessed by simulation studies.

2   Preliminaries, Terminologies and Notations

In this section, we describe the preliminaries, terminologies and notations that will be used throughout this paper.

The significance testing problem is formulated as follows. Let $\boldsymbol{X}_i \in \mathbb{R}^p, i = 1, \ldots, n$, be observations (a random sample or spatially correlated data) from a p-variate distribution with cumulative distribution function (CDF) $F_{\boldsymbol{X}}$. Let $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the p-variate normal distribution with expectation $\boldsymbol{\mu}$ and nonsingular covariance matrix $\boldsymbol{\Sigma}$, and let $\mathcal{N}_p$ denote the class of all non-degenerate p-variate normal distributions. Our interest is to test, based on the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, the hypothesis

$$H_0 : F_{\boldsymbol{X}} \in \mathcal{N}_p,$$

against general alternatives.

It is usually desired that the tests for MVN possess the properties of (a) affine invariance and (b) universal consistency. Since the class $\mathcal{N}_p$ is closed with respect to full rank affine transformations, in order to ensure the same conclusion regarding rejection or acceptance of H_0 given the original data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and the transformed data $\boldsymbol{A}\boldsymbol{X}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{X}_n + \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ is

nonsingular and $\boldsymbol{b} \in \mathbb{R}^p$, any test statistic $T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ should be affine invariant, i.e.,

$$T_n(\boldsymbol{A}\boldsymbol{X}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{X}_n + \boldsymbol{b}) = T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).$$

The consistency class of a test statistic T_n for H_0 is the set of probability distributions P over $\mathbb{R}^p$ such that, if the underlying distribution is P, the probability of rejecting H_0 tends to one as the sample size n goes to infinity, when using the test statistic T_n. As the alternatives to normality are rarely known in practice, it is important that the consistency class of a test for MVN is the set of all $P \notin \mathcal{N}_p$, which implies that the test is able to detect any non-normal alternative distribution, at least for large samples. Here, we call a test to be universally consistent if it is consistent against any fixed non-normal alternative distributions.

Since there are, in principle, an infinite number of alternatives to normal distributions, no uniformly most powerful test exists for MVN. Therefore, two types of tests are developed tailored to the problem of interest. One type consists of *omnibus* tests that are designed to cover all possible alternatives, usually with only reasonably high and generally suboptimal powers. Most of the tests in the literature are omnibus tests. The other type refers to *directed* tests that are highly powerful for some specific classes of alternatives, at the cost of being blind to other types of alternatives. Combinations of directed tests have also been suggested as omnibus tests. Tests based on measures of multivariate skewness and kurtosis are typically directed tests, and they have certain diagnostic limitations as clarified by Henze (2002) and also mentioned in Section 3. Nevertheless, one important role of directed tests is that they can be used to detect types of departures from normality that are most dangerous in the underlying problem. For example, the size of the Hotelling T^2 test (Hotelling, 1931) is much influenced by the asymmetry of the distribution, while symmetric departures from normality are not so crucial (Mardia, 1970). In addition, for restricted families of alternatives that are closed under the action of some groups of transformations, it may be possible to construct most powerful invariant (MPI) tests and thus set benchmarks for assessing the performance of other invariant tests.

In what follows, let $\boldsymbol{0}$ denote the null vector of length p, $\boldsymbol{I}_p$ denote the identity matrix of size $p \times p$, $\|\cdot\|$ denote the Euclidean norm in $\mathbb{R}^p$, and a superscript $\top$ denote a transpose. Also, denote the sample mean vector and sample covariance matrix for the p-variate observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$

as

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\top},$$

respectively, and $\widetilde{\boldsymbol{S}} = \frac{n}{n-1}\boldsymbol{S}$ is the unbiased sample covariance matrix. In addition, assume that $n \geq p+1$ so that $\boldsymbol{S}$ is invertible with probability one (Eaton and Perlman, 1973). Denote by $\boldsymbol{S}^{-1/2}$ the unique symmetric square root of $\boldsymbol{S}$, and define the scaled residuals as $\boldsymbol{Y}_i = \boldsymbol{S}^{-1/2}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}), i = 1, \ldots, n$, which are asymptotically $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$ under H_0.

3 Recent Advances of MVN Tests Based on Skewness and Kurtosis Approaches for I.I.D. Data

Recent work on MVN tests for i.i.d. data can be classified into five categories: 1) skewness and kurtosis approaches, 2) chi-squared type tests, 3) BHEP-type tests based on the empirical characteristic function, 4) other generalizations of univariate normality tests, and 5) multiple testing procedures that combine multiple tests for MVN. In this section, we review the first category, i.e., tests based on skewness and kurtosis measures, and also present the review for the remaining four categories in the Appendix for readers' reference. We summarize some important properties (affine invariance, universal consistency, explicit null distribution) for all the reviewed tests as well as one classical test selected from each of the first three categories in Table A1 in the Appendix.

In univariate statistics, the skewness and kurtosis of a random variable X, with mean μ and variance σ^2, are defined as

$$\beta_1 = \mathbb{E}\left\{\left(\frac{X-\mu}{\sigma}\right)^3\right\} = \frac{\mu_3}{\mu_2^{3/2}}, \text{ and } \beta_2 = \mathbb{E}\left\{\left(\frac{X-\mu}{\sigma}\right)^4\right\} = \frac{\mu_4}{\mu_2^2},$$

respectively, where μ_i is the ith central moment of X. For a normal distribution, $\beta_1 = 0$ and $\beta_2 = 3$. Hence, $\beta_2 - 3$ is called excess kurtosis with respect to a normal distribution. The skewness $\beta_1 = 0$ for symmetric distributions and $\beta_1 > 0$ (< 0) for right (left)-asymmetric distributions, while the kurtosis $\beta_2 = 3$ for the normal distribution, and $\beta_2 > 3$ (< 3) for distributions that are heavier-tailed (lighter-tailed) than the normal one.

Tests based on the univariate sample skewness and kurtosis are among the earliest procedures for assessing univariate normality. Due to their popularity and good power properties, some of the first tests for MVN are based on extensions of the notion of skewness and kurtosis to the multivariate setting. The Mardia's tests (Mardia, 1970, 1974) are perhaps the most often referenced tests for MVN. Mardia (1970) firstly extended the measures of skewness and kurtosis of a p-dimensional random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, as

$$\beta_{1,p} = \mathbb{E}\left[\left\{(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})\right\}^3\right],$$

$$\beta_{2,p} = \mathbb{E}\left[\left\{(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right\}^2\right],$$

respectively, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independently and identically distributed random vectors. For a p-variate normal distribution, $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$. For all distributions, $\beta_{1,p} \geq 0$, and for $p = 1$, $\beta_{1,p}$ reduces to the square of the univariate skewness. The sample measures are also defined for i.i.d. samples, $\boldsymbol{X}_i$, $i = 1, \ldots, n$, as

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})\right\}^3 = \frac{1}{n^2}\sum_{j,k=1}^{n}(\boldsymbol{Y}_j^\top \boldsymbol{Y}_k)^3,$$

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\left\{(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})\right\}^2 = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{Y}_j^\top \boldsymbol{Y}_j)^2.$$

Mardia (1970) then proposed tests based on $b_{1,p}$ and $b_{2,p}$ as:

$$\text{MS} = nb_{1,p}/6, \quad \text{MK} = \{b_{2,p} - p(p+2)\}/\{8p(p+2)/n\}^{1/2}, \tag{1}$$

which are asymptotically $\chi^2_{p(p+1)(p+2)/6}$ and $\mathcal{N}(0,1)$ under H_0. Other classical measures of multivariate skewness and kurtosis and related tests for MVN have been proposed by, for example, Malkovich and Afifi (1973), Isogai (1982), Srivastava (1984), Koziol (1987) and Móri et al. (1994).

Univariate normality tests often use classical measures of asymmetry based on the standardized distance between two separate location parameters, and measures of kurtosis based on the ratios of two scale measures, such as the classical standardized fourth moment. Motivated by these facts, Kankainen et al. (2007) proposed a measure of multivariate skewness based on the Mahalanobis distance between two multivariate location vector estimates, and a measure of multivariate kurtosis based on the (matrix) distance between two scatter matrix estimates (see

Section 2 in their paper for the definitions of location vectors and scatter matrices). Then, the test statistic for MVN (to detect skewness) is given by

$$U = (\boldsymbol{T}_1 - \boldsymbol{T}_2)^\top \boldsymbol{C}^{-1}(\boldsymbol{T}_1 - \boldsymbol{T}_2),$$

where $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are two separate location vectors and $\boldsymbol{C}$ is a scatter matrix, and the kurtosis test statistic is given by

$$W = \|\boldsymbol{C}_1^{-1}\boldsymbol{C}_2 - \boldsymbol{I}_p\|^2 = \left[\mathrm{tr}\{(\boldsymbol{C}_1^{-1}\boldsymbol{C}_2)^2\} - \frac{1}{p}\mathrm{tr}^2(\boldsymbol{C}_1^{-1}\boldsymbol{C}_2)\right] + \frac{1}{p}\left\{\mathrm{tr}(\boldsymbol{C}_1^{-1}\boldsymbol{C}_2) - p\right\}^2,$$

where $\| \cdot \|^2 = \mathrm{tr}(\cdot^\top \cdot)$, and $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are two separate scatter matrices. Using special choices of location and scatter estimators, it is possible to obtain generalizations of classical Mardia's measures of multivariate skewness and kurtosis.

Thulin (2014) proposed a measure of multivariate skewness in a way that resembles the construction in Mardia (1970). For the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, write $\overline{\boldsymbol{X}} = (\overline{X}_1, \ldots, \overline{X}_p)^\top$, $\boldsymbol{S} = \{S_{ij}\}$, and $\boldsymbol{u} = (S_{11}, \ldots, S_{pp}, S_{12}, \ldots, S_{1p}, \ldots, S_{2p}, \ldots, S_{p-1,p})^\top$. It is well known that $\overline{\boldsymbol{X}}$ and $\boldsymbol{u}$ are independent under H_0. Denote the covariance matrix of $\overline{\boldsymbol{X}}$ and $\boldsymbol{u}$ by

$$\mathrm{Cov}(\overline{\boldsymbol{X}}, \boldsymbol{u}) = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix},$$

where $\boldsymbol{\Lambda}_{11}$ is the covariance matrix of $\overline{\boldsymbol{X}}$ and so on. The canonical correlations, $\lambda_1, \ldots, \lambda_p$, of $\overline{\boldsymbol{X}}$ and $\boldsymbol{u}$ are the square roots of the eigenvalues of $\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}$, and they are all equal to zero under H_0. The measure of multivariate skewness proposed by Mardia (1970) is based on the sum of the squared canonical correlations:

$$\beta_{1,p} = 2\sum_{i=1}^{p}\lambda_i^2 = 2\,\mathrm{tr}(\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}), \tag{2}$$

under the assumption that the cumulants of order higher than 3 of $\boldsymbol{X}$ are negligible. The sample counterpart of $\beta_{1,p}$ can be used to construct tests for MVN. Thulin (2014) derived explicit expressions for the elements of $\mathrm{Cov}(\overline{\boldsymbol{X}}, \boldsymbol{u})$ in terms of the moments of $(X_1, \ldots, X_p)$ (see his Theorem 1), and proposed a new test, $Z_{2,p}^{HL}$, based on the sample counterpart of $\mathrm{Cov}(\overline{\boldsymbol{X}}, \boldsymbol{u})$ (see his Equation (12)). The author constructed another test based on the fact that $\overline{\boldsymbol{X}}$ and

$\boldsymbol{v} = (S_{111}, S_{112}, \ldots, S_{p,p,(p-1)}, S_{ppp})^\top$ are also independent under H_0, where

$$S_{ijk} = \frac{n}{(n-1)(n-2)} \sum_{r=1}^{n} (X_{r,i} - \overline{X}_i)(X_{r,j} - \overline{X}_j)(X_{r,k} - \overline{X}_k).$$

He further constructed tests based on three other functions of the squared canonical correlations. Yamada et al. (2015) generalized Mardia's multivariate kurtosis for testing MVN for the case when the data consist of a random sample of two-step monotone incomplete observations.

One disadvantage of the above tests is that they only consider departures from multivariate normality revealed by skewness and kurtosis, and failure to reject the null hypothesis leaves open the question of whether there are departures from normality in other ways. Consequently, these tests are not universally consistent. For example, the test based on multivariate kurtosis in the sense of Malkovich and Afifi (1973) is inconsistent against spherically symmetric alternative distributions with normal marginal kurtosis, 3. Furthermore, these tests rely only on asymptotic properties, that is, they require large samples to achieve both reasonably accurate control of type I error and high power.

The omnibus Jarque-Bera (JB)-type tests address the above issue by combining the skewness and kurtosis measures. The univariate JB test (Jarque and Bera, 1981), based on a univariate random sample $X_i \in \mathbb{R}, i = 1, \ldots, n$, is given by

$$\text{JB} = \frac{nb_1^2}{6} + \frac{n(b_2 - 3)^2}{24},$$

where b_1 and b_2 are the sample skewness and kurtosis, respectively, given by $b_1 = \frac{\sqrt{n(n-1)}}{n-2} \frac{m_3}{m_2^{3/2}}$ and $b_2 = \frac{m_4}{m_2^2}$, where $m_k = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \frac{1}{n} \sum_{j=1}^{n} X_j \right)^k$. Under univariate normality, the JB statistic is asymptotically χ_2^2. The simplest way to construct multivariate JB-type tests, based on the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, is to aggregate individual (univariate) skewness and kurtosis as

$$\text{LM} = \sum_{i=1}^{p} \frac{nb_{1(i)}^2}{6} + \sum_{i=1}^{p} \frac{n(b_{2(i)} - 3)^2}{24},$$

where $b_{1(i)}$ and $b_{2(i)}$ denote the sample skewness and kurtosis of component i, respectively. LM is asymptotically distributed as χ_{2p}^2 under H_0 (see, e.g., Lütkepohl (2005)). However, for both JB and LM, the sample skewness and kurtosis are not independent in finite samples, and using the asymptotic distribution leads to under-rejection. To remedy this problem, Doornik and Hansen

([2008](#)) proposed to use transformed skewness and kurtosis, where the transformation creates statistics that are much closer to standard normal, based on the work of Bowman and Shenton ([1975](#)). Specifically, the test statistic is

$$\text{JB}_{\text{DH}} = \boldsymbol{B}_1^\top \boldsymbol{B}_1 + \boldsymbol{B}_2^\top \boldsymbol{B}_2, \tag{3}$$

where $\boldsymbol{B}_1 = (b_{11}, \ldots, b_{1p})^\top$ and $\boldsymbol{B}_2 = (b_{21}, \ldots, b_{2p})^\top$ are the transformed vectors of skewness and kurtosis, respectively. JB_{DH} is asymptotically χ^2_{2p} under H_0. Jönsson ([2011](#)) further noticed that there is a pattern of downward size distortions to the test based on LM; see his Figure 1. He suggested using the test statistic that pools the individual p-values:

$$\widetilde{\text{LM}} = -2 \sum_{i=1}^p \ln(\pi_i),$$

where π_i is the p-value of the univariate JB test for the ith component. $\widetilde{\text{LM}}$ has an asymptotic χ^2_{2p} distribution under H_0, and simulation studies showed that the previous poor size properties are eliminated (see his Figure 2) without loss of power. The calculation of $\widetilde{\text{LM}}$ is somewhat more convenient than using the transformation approach proposed by Doornik and Hansen ([2008](#)). Kim ([2016](#)) proposed to aggregate the univariate JB-type statistics based on transformed data. Suppose the random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the standardized data,

$$\boldsymbol{Z}_i = \boldsymbol{S}^{*\top}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}), \ i = 1, \ldots, n,$$

follow a $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I})$ asymptotically under H_0, where $\boldsymbol{S}^*$ is defined by $\boldsymbol{S}^{*\top}\boldsymbol{S}\boldsymbol{S}^* = \boldsymbol{I}_p$. The multivariate test statistics are then formed by adding up the univariate JB-type statistics for each coordinate of the transformed vectors.

Another way to construct multivariate JB-type tests is to combine multivariate skewness and kurtosis measures (see, e.g., Mardia and Foster ([1983](#)), Bera and John ([1983](#)) and Mardia and Kent ([1991](#))). Koizumi et al. ([2009](#)) proposed two JB-type tests based on the sample measures of multivariate skewness and kurtosis of Srivastava ([1984](#)). For the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, let $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{D}_\omega\boldsymbol{H}^\top$, where $\boldsymbol{H} = (\boldsymbol{h}_1 \ldots \boldsymbol{h}_p)$ is an orthogonal matrix and $\boldsymbol{D}_\omega = \text{diag}(\omega_1, \ldots, \omega_p)$. The

sample measures of multivariate skewness and kurtosis given by Srivastava (1984) are:

$$\widetilde{b}_{1,p} = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{m_{3i}}{m_{2i}^{3/2}}\right)^2, \quad \widetilde{b}_{2,p} = \frac{1}{p}\sum_{i=1}^{p}\frac{m_{4i}}{m_{2i}^2}, \tag{4}$$

respectively, where $m_{ki} = \frac{1}{n}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_i)^k$, with $Y_{ij} = \boldsymbol{h}_i^\top \boldsymbol{X}_j$ and $\overline{Y}_i = \frac{1}{n}\sum_{j=1}^{n}Y_{ij}$, $i = 1,\ldots,p$, $j = 1,\ldots,n$. The two JB-type statistics based on $\widetilde{b}_{1,p}$ and $\widetilde{b}_{2,p}$ are:

$$M_1 = np\left\{\frac{\widetilde{b}_{1,p}}{6} + \frac{(\widetilde{b}_{2,p}-3)^2}{24}\right\}, \tag{5}$$

$$M_2 = \frac{p\widetilde{b}_{1,p}}{\mathrm{E}(\widetilde{b}_{1,p})} + \frac{\{\widetilde{b}_{2,p} - \mathrm{E}(\widetilde{b}_{2,p})\}^2}{\mathrm{Var}(\widetilde{b}_{2,p})}, \tag{6}$$

which are both asymptotically χ^2_{p+1} under H_0, with $\mathrm{E}(\widetilde{b}_{1,p}) = \frac{6(n-2)}{(n+1)(n+3)}$, $\mathrm{E}(\widetilde{b}_{2,p}) = \frac{3(n-1)}{n+1}$, and $\mathrm{Var}(\widetilde{b}_{2,p}) = \frac{24n(n-2)(n-3)}{p(n+1)^2(n+3)(n+5)}$ under H_0. Enomoto et al. (2012) noticed a difference between the upper percentiles of the distributions of M_2 and the χ^2 distribution for small n. To mitigate the difference, they proposed a new test statistic by using the variance of M_2:

$$M_3 = cM_2 + (1-c)(p+1),$$

which is also asymptotically χ^2_{p+1} under H_0, with $c = \left\{\frac{2p(p+1)}{\mathrm{Var}(M_2)}\right\}^2$, and $\mathrm{Var}(M_2)$ is derived as their Equation (3.1). Koizumi et al. (2014) suggested two other improved tests of M_1 and M_2. First, they noticed that in M_1, the skewness term asymptotically dominates the kurtosis term for large p, so that the omnibus test becomes a directional test for the skewness only. Therefore, they proposed the following test statistic:

$$\mathrm{MJB}_2 = z_{WH}^2 + \frac{np}{24}(\widetilde{b}_{2,p}-3)^2,$$

where $z_{WH} = \frac{(z_1/p)^{1/3}-1+2/(9p)}{\sqrt{2/(9p)}}$ is the Wilson-Hilferty transform (Wilson and Hilferty, 1931) of $z_1 = np\widetilde{b}_{1,p}/6$. When both p and n go to infinity, MJB_2 is asymptotically χ^2_2 under H_0, which does not depend on the dimensionality p, and hence the omnibus property of the test is maintained even for large p. However, their simulation study showed that the MJB_2 test has poor performance in terms of type I error. They further improved MJB_2 by a normalizing transform of the sample kurtosis as suggested in Seo and Ariga (2011):

$$\mathrm{mMJB} = z_{WH}^2 + z_{NT}^2,$$

where $z_{NT} = \sqrt{\frac{np}{24}}\{-e^{-(\widetilde{b}_{2,p}-3)} + 1 + \frac{6}{n} + \frac{12}{np}\}$. mMJB is asymptotically χ^2_2 under H_0, and proved

to have a more stable behavior in small samples. They further studied the F-approximation for mMJB which is shown to be better than the χ^2 approximation, and therefore can be recommended for testing MVN in both small and large samples.

4 Simulation Study

In this section, we investigate the influence of spatial dependence on the measures of skewness and kurtosis for multivariate Gaussian random fields through Monte Carlo simulation studies. The results reveal that the sample skewness and kurtosis deviate from their theoretical values in the i.i.d. case as the degree of spatial dependence increases. This indicates that the usual test of normality based on the sample skewness and kurtosis may be misleading if the observations in the sample are actually dependent.

For a multivariate random field, the cross-covariances measure the spatial dependences within individual variables as well as between distinct variables. For a p-variate random field $\boldsymbol{Z}(\boldsymbol{s}) = (Z_1(\boldsymbol{s}), Z_2(\boldsymbol{s}), \ldots, Z_p(\boldsymbol{s}))^\top$, $\boldsymbol{s} \in \mathbb{R}^d$, the matrix-valued cross-covariance function of $\boldsymbol{Z}(\boldsymbol{s})$ at two locations, $\boldsymbol{s}_1 \in \mathbb{R}^d$ and $\boldsymbol{s}_2 \in \mathbb{R}^d$, is defined as $\boldsymbol{C}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \{C_{ij}(\boldsymbol{s}_1, \boldsymbol{s}_2)\}_{i,j=1}^p$, where $C_{ij}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{cov}\{Z_i(\boldsymbol{s}_1), Z_j(\boldsymbol{s}_2)\}$, $i, j = 1, \ldots, p$. The covariance matrix $\boldsymbol{\Sigma} = \{\boldsymbol{C}(\boldsymbol{s}_i, \boldsymbol{s}_j)\}_{i,j=1}^n$ should satisfy the nonnegative definite condition: $\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a} \geq 0$ for any vector $\boldsymbol{a} \in \mathbb{R}^{np}$, any spatial locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, and any integer n. Various valid cross-covariance models have been built (see Genton and Kleiber (2015) for a review), and the multivariate Matérn model (Gneiting et al., 2010) has received a great deal of attention. In particular, the parsimonious Matérn model for a stationary bivariate random field, where the cross-covariances depend on the spatial lags only, is given by

$$C_{11}(\boldsymbol{h}) = \sigma_1^2 M(\boldsymbol{h}|\nu_1, \beta), \quad C_{22}(\boldsymbol{h}) = \sigma_2^2 M(\boldsymbol{h}|\nu_2, \beta), \tag{7}$$

and

$$C_{12}(\boldsymbol{h}) = C_{21}(\boldsymbol{h}) = \rho_{12}\sigma_1\sigma_2 M\left(\boldsymbol{h}|\frac{1}{2}(\nu_1 + \nu_2), \beta\right), \tag{8}$$

where σ_1^2 and σ_2^2 are the marginal variances, and $M(\boldsymbol{h}|\nu, \beta) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|\boldsymbol{h}\|}{\beta}\right)^\nu \mathcal{K}_\nu \left(\frac{\|\boldsymbol{h}\|}{\beta}\right)$, with $\nu > 0$ is the smoothness parameter, $\beta > 0$ is the spatial range parameter, and $\mathcal{K}_\nu$ is a modified Bessel

function of the second kind of order ν. The colocated correlation coefficient ρ_{12} should satisfy the following condition for the model to be valid:

$$|\rho_{12}| \leq \frac{\Gamma(\nu_1 + \frac{d}{2})^{1/2}}{\Gamma(\nu_1)^{1/2}} \frac{\Gamma(\nu_2 + \frac{d}{2})^{1/2}}{\Gamma(\nu_2)^{1/2}} \frac{\Gamma\{\frac{1}{2}(\nu_1 + \nu_2)\}}{\Gamma\{\frac{1}{2}(\nu_1 + \nu_2) + \frac{d}{2}\}}. \tag{9}$$

In this section, we simulate bivariate random fields defined on $[0,1] \times [0,1] \subset \mathbb{R}^2$ with certain cross-covariance structures, and examine the behaviors of sample skewness and kurtosis as a function of the degree of spatial dependence specified in the cross-covariance function. Specifically, we use the bivariate Matérn model (7) and (8) with smoothness parameters $\nu_1 = \nu_2 = 0.5$ (Exponential) or $\nu_1 = \nu_2 = 1$ (Whittle), and the colocated correlation coefficient ρ can be either positive (e.g., 0.5) or negative (e.g., -0.5) as long as it satisfies the inequality (9). Both marginal variances are set to 1 for simplicity. Further, the spatial dependence can be characterized by the effective range h^*, which is defined as the distance beyond which the correlation between observations is less than or equal to 0.05 (Irvine et al., 2007). We simulate the random fields at 15×15 regular grid of locations over the unit square $[0,1] \times [0,1] \subset \mathbb{R}^2$, set the effective range $h^* \in \{0.1, 0.12, 0.14, \ldots, 0.88, 0.9\}$, which implies an increasing degree of spatial dependence of the random field, and solve the following equations:

$$R(h^*) = \exp\left(\frac{h^*}{\beta}\right) = 0.05 \quad \text{(Exponential)} \quad \text{or} \quad R(h^*) = \frac{h^*}{\beta}\mathcal{K}_1\left(\frac{h^*}{\beta}\right) = 0.05 \quad \text{(Whittle)} \tag{10}$$

to get the values of the spatial range parameter β. We simulate 200 times for each combination of parameters. In order to see the pure effect of spatial dependence determined by h^* or the induced parameter β, in each simulation we simulate a standard multi-normal random vector $\mathbf{e}$ and fix it, and then impose the covariance matrix on it. Specifically, to simulate a bivariate random field $\boldsymbol{Z}(\boldsymbol{s}) = (Z_1(\boldsymbol{s}), Z_2(\boldsymbol{s}))^\top$ at a regular grid of n locations, we first stack the variables in a long vector $\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top)^\top = (Z_1(\boldsymbol{s}_1), \ldots, Z_1(\boldsymbol{s}_n), Z_2(\boldsymbol{s}_1), \ldots, Z_2(\boldsymbol{s}_n))^\top$, then simulate and fix a standard multi-normal random vector $\mathbf{e} \in \mathbb{R}^{2n}$, and get the values of $\boldsymbol{Z}$ by

$$\boldsymbol{Z} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}(h^*))\mathbf{e} \in \mathbb{R}^{2n},$$

for each combination of parameters $\boldsymbol{\theta}$ that depends on the effective range h^*, where $\boldsymbol{\Sigma}^{1/2}$ is the square root of $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, the covariance matrix of $\boldsymbol{Z}$, with $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ being the auto-covariance matrices for $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, respectively, and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top$ being the cross-covariance

matrix between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. By doing this, we can eliminate the effect of randomness coming from $\mathbf{e}$ and isolate the effect of changing the parameters, particularly changing the degree of spatial dependence, in the covariance function.

Following these procedures, we thus have 200 sample skewness and kurtosis for each level of spatial dependence (i.e., the effective range h^* or the correlation parameter ρ) that is specified in the covariance structure. We then summarize the 200 curves of sample skewness and kurtosis as a function of h^* or ρ by functional boxplot (Sun and Genton, 2011), which is an extension of the classical boxplot for visualizing data that continuously vary in space and time. The classical boxplot can be created by simply ordering one-dimensional observations from the smallest to the largest value. For functional data, each observation is a function (e.g., a curve or an image), and all the observations are center-outward ordered based on the concept of band depth (López-Pintado and Romo, 2009) or other notions of depth. Based on the ranking, a functional boxplot is able to display three descriptive statistics: the median curve, the envelope of the 50% central region, and the maximum non-outlying envelope (Sun and Genton, 2011). Outliers are detected as exceeding 1.5 times the 50% central region, similarly to classical boxplots.

Figure 1 shows the functional boxplots of the Mardia's sample skewness and kurtosis of the bivariate Gaussian random field on $\mathbb{R}^2$ as a function of the effective range h^*. Recall that Mardia's measure of multivariate skewness is always positive. We can find that the sample skewness and kurtosis increase as the effective range increases, and the smoother the field, the larger the influence from spatial dependence. The difference between the cases where $\rho_{12} = 0.5 > 0$ and where $\rho_{12} = -0.5 < 0$ is small if we compare, e.g., (a) with (c) or (b) with (d).

5 The New Test for MVN Under Spatial Dependence

5.1 Construction of the new test

The results from the simulation study in the previous section suggest that the dependence in spatial data should be appropriately accounted for in the tests for MVN based on sample skewness and kurtosis measures; otherwise, the un-adjusted tests may lead to conservative decisions on
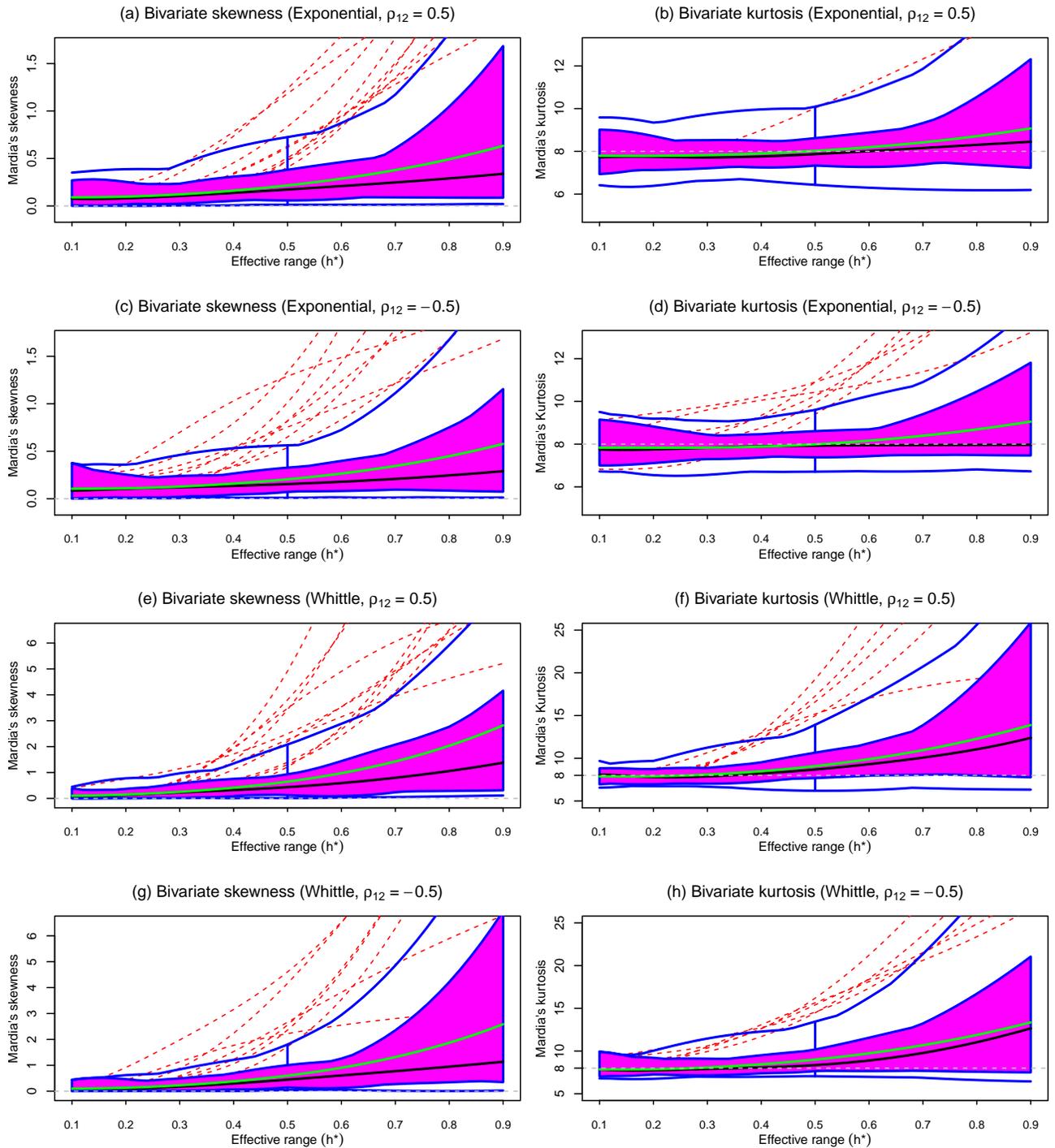
Figure 1: Functional boxplot of the Mardia's sample skewness and kurtosis of the bivariate Gaussian random field in $[0,1] \times [0,1]$ as a function of the effective range h^* for (a)-(d) the Exponential and (e)-(h) the Whittle covariance functions. The green line is the point-wise mean curve, the black line is the median curve, the purple shaded region is the envelope of the 50% central region, the outer blue lines represent the maximum non-outlying envelope, and the red dashed lines are detected outliers. The theoretical values of Mardia's measures of skewness (i.e., $\beta_{1,2} = 0$) and kurtosis (i.e., $\beta_{2,2} = 8$) for a bivariate normal distribution are indicated by gray dashed lines.

assessing the Gaussianity in the data. Horváth et al. (2020) proposed a JB-type test to address this problem for the univariate case. Assume that the spatial dataset $\{X(\mathbf{s}_1), X(\mathbf{s}_2), \ldots, X(\mathbf{s}_n)\}$, where $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\} \in \mathbb{Z}^d$ are locations in the d-dimensional space with integer coordinates, is from a strictly stationary Gaussian spatial moving average process under the H_0:

$$X(\boldsymbol{s}) = \mu + \sum_{\boldsymbol{t} \in \mathbb{Z}^d} a(\boldsymbol{t}) \varepsilon(\boldsymbol{s} - \boldsymbol{t}), \quad \boldsymbol{s} \in \mathbb{Z}^d,$$

where the innovations $\varepsilon(\boldsymbol{s}), \boldsymbol{s} \in \mathbb{Z}^d$ are i.i.d. from $\mathcal{N}(0,1)$, and the constants $a(\boldsymbol{s}), \boldsymbol{s} \in \mathbb{Z}^d$, satisfy $\sum_{\boldsymbol{s} \in \mathbb{Z}^d} |a(\boldsymbol{s})|^2 < \infty$. The JB-type test statistic is

$$\mathrm{JB} = \frac{S_n^2}{\hat{\phi}_{\mathrm{S}}^2} + \frac{K_n^2}{\hat{\phi}_{\mathrm{K}}^2},$$

where S_n and K_n are sample skewness and kurtosis of the standardized observations, respectively, and $\hat{\phi}_{\mathrm{S}}^2$ and $\hat{\phi}_{\mathrm{K}}^2$ are consistent estimators of the asymptotic variances of S_n and K_n, respectively. Horváth et al. (2020) defined the non-parametric kernel estimators, $\hat{\phi}_{\mathrm{S}}^2$ and $\hat{\phi}_{\mathrm{K}}^2$, as

$$\hat{\phi}_{\mathrm{S}}^2 = 6 \sum_{\boldsymbol{h}} \omega_{\boldsymbol{b}}(\boldsymbol{h}) \hat{\gamma}^3(\boldsymbol{h}) := 6 \sum_{l=1}^{d} \sum_{|h_l| \leq b_l} \left\{ \prod_{l=1}^{d} K\left(\frac{h_l}{b_l}\right) \right\} \hat{\gamma}^3(h_1, \ldots, h_d),$$

$$\hat{\phi}_{\mathrm{K}}^2 = 24 \sum_{\boldsymbol{h}} \omega_{\boldsymbol{b}}(\boldsymbol{h}) \hat{\gamma}^4(\boldsymbol{h}) := 24 \sum_{l=1}^{d} \sum_{|h_l| \leq b_l} \left\{ \prod_{l=1}^{d} K\left(\frac{h_l}{b_l}\right) \right\} \hat{\gamma}^4(h_1, \ldots, h_d),$$

where $\hat{\gamma}(\boldsymbol{h})$ is the sample auto-covariance function for the standardized observations with spatial lag $\boldsymbol{h} = (h_1, \ldots, h_d)^\top$; K is a univariate kernel and $\{b_1, \ldots, b_d\}$ are smoothing bandwidths, satisfying some regularity conditions. The spatial dependence in the data is accounted for in $\hat{\gamma}(\boldsymbol{h})$, and the kernel smoothing method is used to establish consistency of the asymptotic variance estimators. Under H_0, the statistic JB is asymptotically χ_2^2.

To generalize the univariate test of Horváth et al. (2020) to the multivariate setting, we adopt the idea in Kim (2016) of aggregating the univariate JB-type statistics for individual variables based on transformed multivariate data. Specifically, suppose we have a p-variate spatial dataset $\boldsymbol{X} = \{\boldsymbol{X}(\mathbf{s}_1), \boldsymbol{X}(\mathbf{s}_2), \ldots, \boldsymbol{X}(\mathbf{s}_n)\}$, where $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\} \in \mathbb{Z}^d$ are n spatial locations, $\boldsymbol{X}(\boldsymbol{s}_i) = (X_1(\boldsymbol{s}_i), X_2(\boldsymbol{s}_i), \ldots, X_p(\boldsymbol{s}_i))^\top$ is the vector of p variables at location $\mathbf{s}_i, i = 1, \ldots, n$. We assume that under H_0, the observations follow a multivariate Gaussian spatial moving average

(or kernel convolution) process (Gelfand and Banerjee, 2010):

$$X_l(\boldsymbol{s}) = \mu_l + \sigma_l \sum_{\boldsymbol{t} \in \mathbb{Z}^d} k_l(\boldsymbol{s} - \boldsymbol{t})\omega(\boldsymbol{t}), \quad \boldsymbol{s} \in \mathbb{Z}^d, \ l = 1, \ldots, p, \tag{11}$$

where μ_l is the unknown mean, $k_l(\cdot), l = 1, \ldots, p$, is a set of p square integrable kernel functions on $\mathbb{Z}^d$ with $k_l(\boldsymbol{0}) = 1$, and $\omega(\cdot)$ is a mean 0, variance 1 Gaussian random field on $\mathbb{Z}^d$ with certain correlation function ρ. The kernel convolution technique is a well-known approach for creating rich classes of stationary processes (Bernardo et al., 2003). Under H_0, $\boldsymbol{X}$ is thus from a stationary multivariate Gaussian random field with the associated $p \times p$ matrix-valued cross-covariance function $C(\boldsymbol{s}, \boldsymbol{s}')$ having (l, l') entry

$$(C(\boldsymbol{s}, \boldsymbol{s}'))_{ll'} = \sigma_l \sigma_{l'} \sum_{\boldsymbol{t} \in \mathbb{Z}^d} \sum_{\boldsymbol{t}' \in \mathbb{Z}^d} k_l(\boldsymbol{s} - \boldsymbol{t}) k_{l'}(\boldsymbol{s}' - \boldsymbol{t}') \rho(\boldsymbol{t} - \boldsymbol{t}').$$

We then standardize the observations to obtain the scaled residuals as

$$\boldsymbol{Y}(\boldsymbol{s}_i) = \boldsymbol{S}^{-1/2}\{\boldsymbol{X}(\boldsymbol{s}_i) - \overline{\boldsymbol{X}}\}, \ i = 1, \ldots, n,$$

where $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)$ is the sample mean and $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{X}(\boldsymbol{s}_i) - \overline{\boldsymbol{X}}\}\{\boldsymbol{X}(\boldsymbol{s}_i) - \overline{\boldsymbol{X}}\}^{\top}$ is the sample covariance matrix. For each $\boldsymbol{s}_i$, $i = 1, \ldots, n$, $\boldsymbol{Y}(\boldsymbol{s}_i)$ is distributed asymptotically as $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$ under H_0. For each variable, we compute the sample skewness and excess kurtosis for the scaled residuals as

$$S_{n,l} = \hat{\mu}_{3,l}\hat{\mu}_{2,l}^{-3/2}, \quad K_{n,l} = \hat{\mu}_{4,l}\hat{\mu}_{2,l}^{-2} - 3, \quad l = 1, \ldots, p,$$

where $\hat{\mu}_{k,l} = \frac{1}{n}\sum_{i=1}^{n}(Y_l(\boldsymbol{s}_i) - \overline{Y}_l)^k$ is the kth sample moment with $\overline{Y}_l = \frac{1}{n}\sum_{i=1}^{n} Y_l(\boldsymbol{s}_i)$ being the sample mean of the lth variable. The consistent estimators, $\hat{\phi}_{S_{n,l}}^2$ and $\hat{\phi}_{K_{n,l}}^2$ of $S_{n,l}$ and $K_{n,l}$, respectively, can be obtained using the kernel smoothing method in Horváth et al. (2020), but the statistical inference can be very sensitive to the selection of the user-chosen kernels and bandwidth. Therefore, we adopt the approximation method in Lobato and Velasco (2004) that does not require any smoothing and obtain the estimators:

$$\hat{\phi}_{S_{n,l}}^2 = \sum_{\boldsymbol{h}} \hat{\gamma}_l^3(\boldsymbol{h}), \quad \hat{\phi}_{K_{n,l}}^2 = \sum_{\boldsymbol{h}} \hat{\gamma}_l^4(\boldsymbol{h}), \quad l = 1, \ldots, p, \tag{12}$$

where $\hat{\gamma}_l(\cdot)$ is the sample auto-covariance function for the lth variable based on the scaled resid-

uals, i.e.,

$$\hat{\gamma}_l(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}\in D} \{Y_l(\boldsymbol{s}) - \overline{Y}_l\}\{Y_l(\boldsymbol{s}+\boldsymbol{h}) - \overline{Y}_l\}, \;\; l = 1, \ldots, p,$$

where $N(\boldsymbol{h})$ is the number of pairs of locations with spatial lag $\boldsymbol{h}$, and D is the spatial domain of the observations. The estimators in Equation (12) are consistent based on the same argument in Lobato and Velasco (2004) that the powers of the sample auto-covariances provide the stochastic dampening factors. To further reduce the computational burden, the auto-covariances can be estimated using the truncation method:

$$\hat{\gamma}_l(\boldsymbol{h}) = \frac{1}{N^*(\boldsymbol{h})} \sum_{\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}\in D, \, \|\boldsymbol{h}\|\leq h^*} \{Y_l(\boldsymbol{s}) - \overline{Y}_l\}\{Y_l(\boldsymbol{s}+\boldsymbol{h}) - \overline{Y}_l\}, \;\; l = 1, \ldots, p,$$

where h^* is the effective range, and $N^*(\boldsymbol{h})$ is the number of pairs of locations separated by $\boldsymbol{h}$ such that $\|\boldsymbol{h}\| \leq h^*$. The corresponding estimators in Equation (12) are still consistent since the correlations of the observations with distance beyond h^* are restricted to zero in the sample auto-covariances. Finally, the new JB-type test for MVN based on the multivariate spatial dataset $\boldsymbol{X} = \{\boldsymbol{X}(\mathbf{s}_1), \boldsymbol{X}(\mathbf{s}_2), \ldots, \boldsymbol{X}(\mathbf{s}_n)\}$ is given by

$$JB^* = \sum_{l=1}^{p} \frac{S_{n,l}^2}{\hat{\phi}_{S_{n,l}}^2} + \sum_{l=1}^{p} \frac{K_{n,l}^2}{\hat{\phi}_{K_{n,l}}^2}. \tag{13}$$

Since the scaled residuals are independent under H_0, and both $S_{n,l}/\hat{\phi}_{S_{n,l}}$ and $K_{n,l}/\hat{\phi}_{K_{n,l}}$ converge to a standard normal distribution according to Horváth et al. (2020), the test statistic JB^* is asymptotically χ_{2p}^2 under H_0. In addition, since the new test is a JB-type test, it is affine invariant and universally consistent.

5.2 Type I error and empirical power of the new test

In this section, we assess the type I error and empirical power of the new test via Monte-Carlo simulations for various configurations of the sample size and the degree of spatial dependence.

To assess the type I error (or empirical size) of the new test, we first simulate a zero-mean p-variate Gaussian random field on $\mathbb{Z}^2$ (i.e., $d = 2$, most commonly encountered in applications) from the spatial moving average (kernel convolution) process of Equation (11). Specifically, each variable is generated from the spatial moving average model defined in Haining (1978), located

on the points of a rectangular square lattice $\mathbb{Z}^2$:

$$X_l(i,j) = \theta_l\{e(i-1,j) + e(i+1,j) + e(i,j-1) + e(i,j+1)\} + e(i,j), \quad l = 1,\ldots,p, \quad (14)$$

where i and j are integers satisfying $1 \leq i \leq M$ and $1 \leq j \leq N$, $e(\cdot,\cdot)$ is a zero-mean, unit-variance Gaussian process on $\mathbb{Z}^2$ with some correlation function ρ, and $e(i,0) = e(0,j) = e(0,0) = 0$ for all $1 \leq i \leq M$ and $1 \leq j \leq N$. When $|\theta_l| \leq 1/4$, this model is invertible to the following first-order quadrilateral autoregressive random field:

$$X_l(i,j) = \theta_l\{X(i-1,j) + X(i+1,j) + X(i,j-1) + X(i,j+1)\} + e(i,j), \quad l = 1,\ldots,p,$$

which has been a preoccupation for the study of finite random fields within geography as a model for spatial dependence (Haining, 1978). Equation (14) is a special case of the spatial kernel convolution process of Equation (11), where the kernels are functions taking the form of a constant height over a bounded rectangle and zero outside. To investigate the performance of the new test for different degrees of spatial dependence, we set the correlation function ρ of the process $e(\cdot,\cdot)$ as the exponential correlation that has been used in Section 4, with varying effective ranges.

Based on the above settings, we first consider the bivariate case (i.e., $p = 2$), set $\theta_1 = 1/5$, $\theta_2 = -1/5$, simulate the random field at an $M \times N$ regular grid of locations over the unit square $[0,1]^2$, and vary the effective ranges, h^*, of the process $e(\cdot,\cdot)$ in $[0.1, 0.9]$ by steps of 0.02. For each level of the spatial dependence indicated by h^*, we use $1,000$ replications for the data generating and testing procedure, and the type I error is approximated by the relative frequency of null hypothesis rejection. The null hypothesis, H_0, is rejected when the p-value given by the test is smaller than the nominal significance level, α. For comparison, we also apply several tests for MVN that do not account for the spatial dependence in the data, i.e., 1) Mardia's tests, MS and MK, defined in Equation (1), 2) the test of Doornik and Hansen (2008), $\mathrm{JB_{DH}}$, defined in Equation (3), and 3) a JB-type test based on Bowman and Shenton (1975), $\mathrm{JB_{BS}}$, defined by replacing the estimators $\hat{\phi}^2_{S_{n,l}}$ and $\hat{\phi}^2_{K_{n,l}}$, $l = 1,\ldots,p$, in our new test of Equation (13) with un-adjusted variances, 6 and 24, respectively.

To assess the empirical power of the new test, we simulate data from the non-Gaussian

sinh-arcsinh (SAS) transformed multivariate Matérn random field defined in Yan et al. (2020). Specifically, we obtain the non-Gaussian data using the element-wise and inverse SAS transformation (Jones and Pewsey, 2009) on the data from Gaussian random fields, i.e., the data used above for assessing the type I error. The corresponding transformation parameter setting used for the first variable is $(0.5, 0.5)$, which indicates positive skewness and lighter tail than the normal distribution. The parameter setting used for the second variable is $(0, 1)$, meaning that no transformation is performed, i.e., the second marginal distribution of the bivariate non-Gaussian random field is Gaussian. Again, we use $1,000$ replications for the data generating and testing procedure, and the empirical power is approximated by the proportion of null hypothesis rejection.
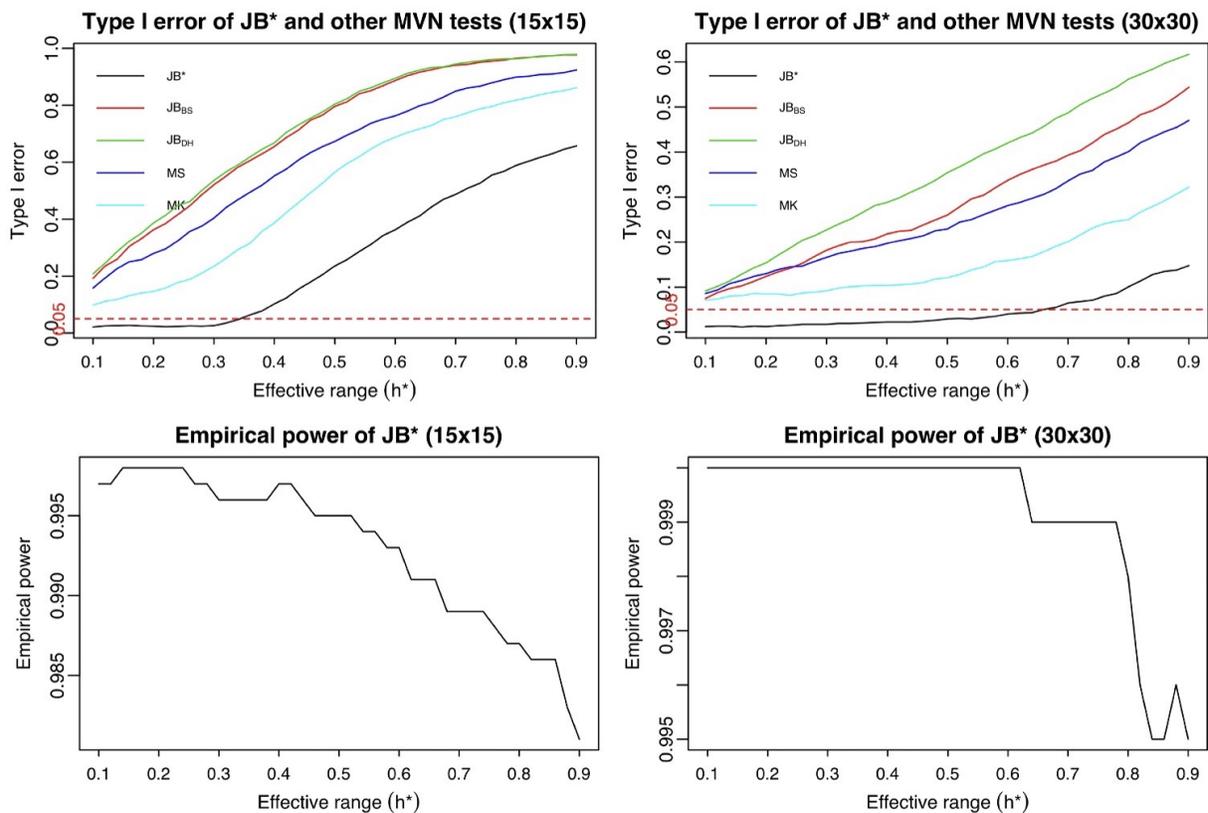

Figure 2: Top: Type I error (empirical size) of the new test for MVN under spatial dependence and four MVN tests for i.i.d. data and bottom: empirical power of the new test for the nominal significance level of $\alpha = 5\%$. JB* represents our new test, JB_{BS} represents the modified test of Bowman and Shenton (1975), JB_{DH} represents the test of Doornik and Hansen (2008), and MS and MK represent the tests of Mardia (1970).

The results for $M = N = 15$ and $M = N = 30$ are shown in Figure 2. The top panels show

the type I error (empirical size) of the new test for MVN under spatial dependence, compared with four MVN tests for i.i.d. data, for the nominal significance level of $\alpha = 5\%$. The probability of the type I error should, by any statistical test, be bounded upwards by the nominal level of significance; otherwise, the test cannot be used for the given purpose. On the other hand, a type I error far smaller than a chosen α is indicative of a test with low power, but does not disqualify the procedure for testing. The simulation results imply that when the sample size is small, the type I error of our new test is bounded by and not too far from $\alpha = 0.05$ until the effective range gets beyond 0.3, i.e., when the data exhibit medium to strong spatial dependence. The performance of the new test improves as the sample size increases; when $M = N = 30$, our test can be used for data with weak to medium spatial dependence. This indicates that our new test has a good large-sample empirical size, and it may become problematic only when dealing with data with strong dependence. All four MVN tests for i.i.d data have inflated type I error for all levels of spatial dependence, providing evidence that these tests cannot be used for spatially correlated data. The bottom panels of Figure 2 show the empirical power of the new test for $\alpha = 5\%$. We can see that the empirical power decreases as the spatial dependence increases, but all at a high level that is near 100%; again, the performance improves as the sample size increases.

6   Discussion

In this work, we reviewed the recent development of tests for multivariate normality for i.i.d. data, with emphasis on the skewness and kurtosis approaches. Based on simulation studies, we showed that when there exists spatial dependence in the data, the multivariate sample skewness and kurtosis measures proposed by Mardia (1970) deviate from their theoretical values under Gaussianity due to dependence, and some of the tests designed for i.i.d. data exhibit inflated type I error; the deviation and type I error increases as the spatial dependence increases. Extending the work of Horváth et al. (2020) to the multivariate case, we then proposed a new JB-type test for multivariate normality for spatially correlated data, by aggregating univariate JB statistics for individual variables. The new test statistic is asymptotically χ^2_{2p} under H_0, where p is the

number of variables. The spatial dependence is accounted for in the asymptotic variances of skewness and kurtosis in the JB test statistic. Easy-to-compute and consistent estimators of the asymptotic variances are constructed for finite samples. The new test has a good control of the type I error, especially for large sample sizes, and it is inappropriate only when the spatial dependence in the data is very strong. In addition, the new test has a high power for finite samples at all levels of spatial dependence, and it has very high power for large sample sizes.

One limitation of the new test is that it can only be used for multivariate spatial data on a regular grid. Tests for data at irregular spatial locations need to be developed, but this can be challenging because the tests would be difficult to be justified asymptotically. Our proposed test can be used in various applications based on the abundant gridded data simulated from reanalysis products, General Circulation Model (GCM) experiments, Regional Climate Model (RCM) experiments or Numerical Weather Prediction (NWP) models. As we have mentioned in Section 3, besides aggregating individual JB test statistics, another way to construct multivariate JB-type tests is to combine multivariate skewness and kurtosis measures. Therefore, it would be an interesting topic to propose a JB-type test for MVN under spatial dependence that combines Mardia's multivariate skewness and kurtosis measures. Simulations in this study show that the un-adjusted tests based on Mardia's measures are misleading if applied to a spatial dataset. To account for the spatial dependence, we need to derive the asymptotic variances of the multivariate skewness and kurtosis of the scaled residuals under some kind of dependence structure, which is a non-trivial task. In addition, we need to construct consistent estimators of the asymptotic variances, and establish the asymptotic properties (limiting null distribution, etc.) of the new test. These are left for our future work.

7 Appendix: Review of Other Recent Tests for MVN

7.1 Chi-squared type tests

The χ^2 test, proposed by Karl Pearson in 1900 (Pearson, 1900), is among the most useful goodness-of-fit tests. For the univariate case, the range of the n observations is divided into

k mutually exclusive classes; $O_i = n_i$ is the observed frequency in class i, and p_i is the probability that an observation will fall into class i under the null hypothesis, so that $E_i = np_i$ is the expected frequency in class i. The χ^2 statistic is then given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{A1}$$

which is asymptotically χ^2_{k-1} under any null distribution. One disadvantage of the χ^2 test is that the testing results can be substantially affected by the number and size of the k classes chosen (see Section 5.2 in Thode (2002) for more details). The χ^2 test is, however, not recommended as a test for univariate normality (Moore, 1986), mostly because of its lack of power relative to other tests for normality. However, the test is easily adaptable to any null distribution, including those that are multivariate in nature, so that it can be used for testing MVN rather than other tests that are much more difficult to implement. As in the univariate case, the sample space is required to be partitioned into mutually exclusive classes; hence, the same problem must still be addressed, i.e., the class size and number of classes. In addition, the problem of choosing class intervals becomes much more difficult as the dimension of the sample space increases, and even in the multivariate normal case, calculating expected frequencies can be extremely difficult. Early attempts to develop extensions of χ^2 test for MVN include Kowalski (1970), Moore and Stubblebine (1981) and Mason and Young (1985), and a few recent studies, presented below, also focused on the chi-squared type tests for MVN.

Cardoso de Oliveira and Ferreira (2010) proposed a multivariate χ^2 test for MVN based on the fact that the statistics

$$B_i = \frac{n}{(n-1)^2}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\top \widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}), \ i = 1, \ldots, n, \tag{A2}$$

where $\widetilde{\boldsymbol{S}}$ is the unbiased sample covariance matrix, are each distributed exactly as $\text{Beta}(p/2, (n-p-1)/2)$ under H_0 (Gnanadesikan and Kettenring, 1972). The authors defined k equal-sized classes based on the empirical rule

$$k \approx \begin{cases} \sqrt{n}, & \text{if } n \leq 100, \\ 5\log_{10}(n), & \text{if } n > 100. \end{cases}$$

The class intervals in the sample space of $B_1, \ldots, B_n$ correspond to regions partitioned from the original p-dimensional sample space of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Now, let q_i be the upper $(k - i)/k \times 100\%$ quantile of the Beta$(p/2, (n - p - 1)/2)$ distribution, then the ith class is defined by $\{q|q_{i-1} < q \leq q_i\}$ for $i = 1, \ldots, k$, where $q_0 = 0$ and $q_k = 1$. The observed frequency O_i of the ith class is the number of values for $B_1, \ldots, B_n$ that fall within the class limit $(q_{i-1}, q_i]$, and the expected frequency is simply $E_i = n/k$, $i = 1, \ldots, k$. The test statistic is then calculated using Equation (A1), which is asymptotically distributed as χ^2_{k-1} under H_0.

Noticing that the above testing procedure was in fact a k-dimensional multinomial goodness-of-fit test, and Pearson's χ^2 statistic was used to measure the discrepancy between the observed and expected proportions, Batsidis et al. (2013) proposed a broader class of tests based on the power divergence family of statistics (Cressie and Read, 1984; Read and Cressie, 2012):

$$
Z_{(\lambda)} = \begin{cases}
\frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} O_i \left\{ \left(\frac{O_i}{E_i}\right)^{\lambda} - 1 \right\}, & \text{when } \lambda \in \mathbb{R}, \lambda \neq -1, 0, \\
2 \sum_{i=1}^{k} E_i \log \frac{E_i}{O_i}, & \text{when } \lambda = -1, \\
2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i}, & \text{when } \lambda = 0,
\end{cases}
$$

which includes as a specific case the Pearson's χ^2 statistic, Equation (A1), when $\lambda = 1$. $Z_{(\lambda)}$ is also aymptotically χ^2_{k-1} under H_0, where O_i and E_i are calculated in the same way as in Cardoso de Oliveira and Ferreira (2010).

Apart from formal testing procedures for MVN with explicitly defined test statistics, subjective graphical methods based on quantiles have also been proposed, such as Small (1978), who assessed MVN based on the plot of the points $(B_{(i)}, D_i), i = 1, \ldots, n$ with the line $y = x$, where $B_{(i)}$'s are the ordered statistics of B_i's defined in Equation (A2), and D_i's are Beta order statistics using Blom's general plotting position (Blom, 1958):

$$
\frac{i - \alpha}{n - \alpha - \beta + 1}, \quad i = 1, \ldots, n,
$$

with $\alpha = (p - 2)/(2p)$ and $\beta = 0.5 - (n - p - 1)^{-1}$. Another graphical method was proposed by Srivastava (1984). Hanusz and Tarasińska (2012) formalized both graphical methods using explicit test statistics. For example, they formalized the testing procedure of Small (1978) by constructing a geometric test statistic, SmG, that measures the departure of empirical points from the line $y = x$, i.e., the sum of the areas between the points $(B_{(i)}, D_i), i = 1, \ldots, n$ and the

line $y = x$, as shown in their Figure 1. Large values of the statistic lead to rejection of MVN of the data. Madukaife and Okafor (2019) pointed out that some areas in the above test statistic may be irregular in shape, and thus may not be easily computed without the use of special computer programs. They therefore proposed a more tractable statistic based on the distances between an ordered set of the transformed observations

$$Z_i = (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\top \widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}), \; i = 1, \ldots, n,$$

which are asymptotically distributed as χ_p^2 under H_0, and the set of the population quantiles of the χ_p^2 distribution. Specifically, the test statistic is

$$G = \sum_{i=1}^{n} (Z_{(i)} - C_i)^2,$$

where $Z_{(i)}$'s are the ordered statistics of Z_i's, and C_i's are the corresponding approximate expected order statistics, i.e., the quantiles of the χ_p^2 distribution. Again, large values of G will lead to rejection of MVN of the data.

Voinov et al. (2016) found that the χ^2 test statistic for MVN, i.e., the Nikulin-Rao-Robson (NRR) statistic, proposed in Moore and Stubblebine (1981) is asymptotically chi-square distributed under H_0 if and only if the covariance matrix $\boldsymbol{\Sigma}$ is a diagonal matrix. They derived the forms of the NRR statistic, Y_n^2, as well as its decomposition, $Y_n^2 = U_n^2 + S_n^2$, for any diagonal covariance matrix of any dimensionality p (see their equations (6), (9) and (10)) and suggested a procedure for testing MVN: 1) produce the Karhunen-Loève transformation of the sample data, which will diagonalize the sample covariance matrix, and 2) compute the statistics Y_n^2, U_n^2 and S_n^2 according to their equations (6), (9) and (10), respectively, based on the transformed data. Since U_n^2 and S_n^2 are asymptotically independent under H_0, they can be used as test statistics independently from each other.

7.2 BHEP-type tests

The BHEP (Baringhaus-Henze-Epps-Pulley) tests, coined by Csörgő (1989), is a class of affine invariant and universally consistent tests for MVN based on the empirical characteristic function (CF). Epps and Pulley (1983) provided a test for univariate normality based on the empirical

CF, and Baringhaus and Henze (1988) generalized their idea to the multivariate case. Henze and Zirkler (1990) studied the test in a more general setting to gain more flexibility with respect to the power of the test against specific alternatives. The BHEP statistic is given by

$$T_{n,\beta} = n \int_{\mathbb{R}^p} |\Psi_n(\boldsymbol{t}) - \Psi(\boldsymbol{t})|^2 \, \psi_\beta(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}, \tag{A3}$$

where $\beta > 0$ is the smoothing parameter, $\Psi_n(\boldsymbol{t}) = \frac{1}{n} \sum_{j=1}^n \exp(i\boldsymbol{t}^\top \boldsymbol{Y}_j)$ is the empirical CF of the scaled residuals $\boldsymbol{Y}_j, j = 1, \ldots, n$, $\Psi(\boldsymbol{t}) = \exp\left(-\|\boldsymbol{t}\|^2/2\right)$ is the CF of $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, and the weighting function $\psi_\beta(\boldsymbol{t}) = (2\pi\beta^2)^{-p/2} \exp\left(-\frac{\|\boldsymbol{t}\|^2}{2\beta^2}\right)$ is the density of $\mathcal{N}_p(\boldsymbol{0}, \beta^2 \boldsymbol{I}_p)$. Theoretical properties of the statistic $T_{n,\beta}$ and alternative test statistics based on the empirical CF using other functional distances have been studied by Baringhaus and Henze (1988), Csörgő (1989), Henze and Zirkler (1990), Henze (1990), Henze (1997), Henze and Wagner (1997) and Epps (1999) (see Section 6 in Henze (2002) and the references therein). Continuous interest has been shown in developing BHEP-type tests since the review paper of Henze (2002), as discussed below.

Pudełko (2005) proposed a test statistic based on the weighted supremum distance:

$$T_{n,r} = \sqrt{n} \sup_{\|\boldsymbol{t}\| < r} |W_n(\boldsymbol{t})|,$$

where $r > 0$ and

$$W_n(\boldsymbol{t}) = \begin{cases} \frac{\Psi_n(\boldsymbol{t}) - \Psi(\boldsymbol{t})}{\|\boldsymbol{t}\|}, & \boldsymbol{t} \neq \boldsymbol{0}, \\ 0, & \boldsymbol{t} = \boldsymbol{0}, \end{cases}$$

where $\Psi_n(\boldsymbol{t})$ and $\Psi(\boldsymbol{t})$ are defined as above. The asymptotic null distribution is derived as the distribution of the supremum norm of a non-stationary complex-valued d-dimensional Gaussian random process.

Arcones (2007) proposed two BHEP-type tests based on the Lévy characterization of the normal distribution (Loève, 1977) and its variant. The test statistics, however, are rather complicated to compute. For example, the first test statistic is given by

$$\widehat{D}_{n,m} = \int_{\mathbb{R}^p} \left|\widehat{\psi}_{n,m}(\boldsymbol{t}) - \Psi(\boldsymbol{t})\right|^2 \psi_\beta(\boldsymbol{t}) \mathrm{d}\boldsymbol{t},$$

where

$$\widehat{\psi}_{n,m}(\boldsymbol{t}) := \frac{(n-m)!}{n!} \sum_{(j_1,\ldots,j_m) \in I_m^n} \exp\left[im^{-1/2}\boldsymbol{t}^\top \left\{\sum_{k=1}^m \widehat{\boldsymbol{\Sigma}}_n^{-1/2}(\boldsymbol{X}_{j_k} - \widehat{\boldsymbol{\mu}}_n)\right\}\right],$$

25

$\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ are estimators of $\boldsymbol{\mu}_{F_X}$ and $\boldsymbol{\Sigma}_{F_X}$, respectively, and $I_m^n = \{(j_1, \ldots, j_m) \in \mathbb{N}^m : 1 \leq j_k \leq n, j_k \neq j_l$ if $k \neq l\}$. If $m = 1$, $\widehat{\boldsymbol{\mu}}_n = \overline{\boldsymbol{X}}$, and $\widehat{\boldsymbol{\Sigma}}_n = \boldsymbol{S}$, then $\widehat{D}_{n,m}$ agrees with $T_{n,\beta}$ in Equation (A3).

Henze and Jiménez-Gamero (2018) constructed a "moment generating function (MGF) analogue" to the BHEP statistic $T_{n,\beta}$. The test statistic is given by

$$\widetilde{T}_{n,\beta} = n \int_{\mathbb{R}^p} \{M_n(\boldsymbol{t}) - m(\boldsymbol{t})\}^2 \omega_\beta(\boldsymbol{t}) \mathrm{d}\boldsymbol{t},$$

where $M_n(\boldsymbol{t}) = \frac{1}{n} \sum_{j=1}^n \exp(\boldsymbol{t}^\top \boldsymbol{Y}_j)$ is the empirical MGF of the scaled residuals $\boldsymbol{Y}_j, j = 1, \ldots, n$, $m(\boldsymbol{t}) = \exp(\|\boldsymbol{t}\|^2/2)$ is the MGF of $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, and $\omega_\beta(\boldsymbol{t}) = \exp(-\beta\|\boldsymbol{t}\|^2)$ with $\beta > 1$ is the weighting function, which leads to a representation of $\widetilde{T}_{n,\beta}$ (see their Equation (1.4)) that is amendable to computational purposes. The authors showed that after a suitable scaling, $\widetilde{T}_{n,\beta}$ approaches a linear combination of sample measures of multivariate skewness in the sense of Mardia (1970) and Móri et al. (1994), as $\beta \to \infty$ (see their Theorem 2.1). They also showed that $\widetilde{T}_{n,\beta}$ has a non-degenerate asymptotic null distribution only when $\beta > 2$.

Henze et al. (2019) constructed a class of tests based on both the CF and the MGF. The authors generalized a characterization of univariate normal distributions in Volkmer (2014) to the multivariate case (see their Proposition 2.1), and showed that $\boldsymbol{X} \in \mathbb{R}^p$ is zero-mean normal distributed if and only if $R_{\boldsymbol{X}}(\boldsymbol{t})M_{\boldsymbol{X}}(\boldsymbol{t}) - 1 = 0$, where $R_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{Re}\{\phi_{\boldsymbol{X}}(\boldsymbol{t})\}$ is the real part of the CF, $\phi_{\boldsymbol{X}}(\boldsymbol{t})$, and $M_{\boldsymbol{X}}(\boldsymbol{t})$ is the MGF of $\boldsymbol{X}$. Let $R_n(\boldsymbol{t}) = \frac{1}{n}\sum_{j=1}^n \cos(\boldsymbol{t}^\top \boldsymbol{Y}_j)$ be the empirical cosine transform, $M_n(\boldsymbol{t}) = \frac{1}{n}\sum_{j=1}^n \exp(\boldsymbol{t}^\top \boldsymbol{Y}_j)$ be the empirical MGF of the scaled residuals $\boldsymbol{Y}_j, j = 1, \ldots, n$, and $U_n(\boldsymbol{t}) = \sqrt{n}\{R_n(\boldsymbol{t})M_n(\boldsymbol{t}) - 1\}$. The test statistic is given by

$$T_{n,\gamma} = \int_{\mathbb{R}^p} U_n^2(\boldsymbol{t})\omega_\gamma(\boldsymbol{t})\mathrm{d}\boldsymbol{t} = n \int_{\mathbb{R}^p} \{R_n(\boldsymbol{t})M_n(\boldsymbol{t}) - 1\}^2 \omega_\gamma(\boldsymbol{t})\mathrm{d}\boldsymbol{t},$$

where $\omega_\gamma(\boldsymbol{t}) = \exp(-\gamma\|\boldsymbol{t}\|^2)$ with $\gamma > 0$ is the weighting function, which leads to a computationally feasible form of $T_{n,\gamma}$ (see their Equation (3.7)). They found a simpler form if the test statistic if defined by $\widetilde{T}_{n,\gamma} = \int_{\mathbb{R}^p} U_n(\boldsymbol{t})\omega_\gamma(\boldsymbol{t})\mathrm{d}\boldsymbol{t}$:

$$\widetilde{T}_{n,\gamma} = \left(\frac{\pi}{\gamma}\right)^{p/2} \sqrt{n} \left\{ \frac{1}{n^2} \sum_{j,k=1}^n \exp\left(\frac{\|\boldsymbol{Y}_j\|^2 - \|\boldsymbol{Y}_k\|^2}{4\gamma}\right) \cos\left(\frac{\boldsymbol{Y}_j^\top \boldsymbol{Y}_k}{2\gamma}\right) - 1 \right\}.$$

The asymptotic null distribution of $\widetilde{T}_{n,\gamma}$ is $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 = 2\pi^p(\gamma^2 - 0.25)^{-p/2} + 2\pi^p(\gamma^2 + 0.25)^{-p/2} - 4\pi^p\gamma^{-p}$.

7.3 Other generalizations of univariate normality tests

The above testing procedures for MVN are all extensions of univariate techniques. In this section, we present four other recent generalizations that cannot be classified into any of the above groups.

Székely and Rizzo (2005) proposed a class of multivariate goodness-of-fit tests based on Euclidean distance between sample elements, and applied the tests for assessing MVN. The goodness-of-fit test statistic is defined by

$$\delta_{n,p} = n\left(\frac{2}{n}\sum_{j=1}^{n}\mathrm{E}(\|\boldsymbol{X}_j - \boldsymbol{X}\|) - \mathrm{E}(\|\boldsymbol{X} - \boldsymbol{X}'\|) - \frac{1}{n^2}\sum_{j,k=1}^{n}\|\boldsymbol{X}_j - \boldsymbol{X}_k\|\right),$$

where $\boldsymbol{X}$ and $\boldsymbol{X}'$ are i.i.d. random vectors from the null distribution. $\delta_{n,p}/n$ is actually a von-Mises-statistic, or simply V-statistic (von Mises, 1947; Hoeffding et al., 1948): $\delta_{n,p}/n = \frac{1}{n^2}\sum_{j,k=1}^{n}h(\boldsymbol{X}_j, \boldsymbol{X}_k)$, with kernel

$$h(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{E}(\|\boldsymbol{x} - \boldsymbol{X}\|) + \mathrm{E}(\|\boldsymbol{y} - \boldsymbol{X}\|) - \mathrm{E}(\|\boldsymbol{X} - \boldsymbol{X}'\|) - \|\boldsymbol{x} - \boldsymbol{y}\|, \ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p.$$

Since the kernel $h(\boldsymbol{x}, \boldsymbol{y})$ for $p = 1$ is closely related to the Cramér-von Mises distance (see their Equation (17)), this test is a multivariate version of a Cramér-von Mises type test. If the null distribution is $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the test statistic for MVN is given by

$$\delta_{n,p} = n\left(\frac{2}{n}\sum_{j=1}^{n}\mathrm{E}(\|\boldsymbol{Y}_j - \boldsymbol{Z}\|) - \mathrm{E}(\|\boldsymbol{Z} - \boldsymbol{Z}'\|) - \frac{1}{n^2}\sum_{j,k=1}^{n}\|\boldsymbol{Y}_j - \boldsymbol{Y}_k\|\right),$$

where $\boldsymbol{Y}_j, j = 1, \ldots, n$ are the scaled residuals, and $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ are i.i.d. random vectors from $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. The explicit form of $\delta_{n,p}$ is given by their Equation (8).

Villasenor et al. (2009) proposed a generalization of Shapiro–Wilk's test for MVN. Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$ is the set of order statistics from a standard normal random sample, and $\mathrm{E}(\boldsymbol{Z}) = \boldsymbol{m}$ and $\mathrm{cov}(\boldsymbol{Z}) = \boldsymbol{V}$. If a set of ordered sample, $\boldsymbol{X}^* = (X_1, \ldots, X_n)^\top$, comes from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, then on a normal probability plot, $X_i = \mu + \sigma Z_i, i = 1, \ldots, n$. Hence, the best linear unbiased estimates of μ and σ are the generalized least square estimates that minimize the quadratic form $(\boldsymbol{X}^* - \mu\boldsymbol{1} - \sigma\boldsymbol{m})^\top \boldsymbol{V}^{-1}(\boldsymbol{X}^* - \mu\boldsymbol{1} - \sigma\boldsymbol{m})$, where $\boldsymbol{1}$ is the vector of

ones of length n; that is, $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = (\boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{m})^{-1} \boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{X}^*$. Let $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$ be the unbiased estimate of σ^2. The Shapiro-Wilk test statistic is defined by

$$W = \frac{R^4 \widehat{\sigma}^2}{(n-1)C^2 s^2} = \frac{b^2}{(n-1)s^2} = \frac{(\boldsymbol{a}^\top \boldsymbol{X}^*)^2}{(n-1)s^2} = \frac{(\sum_{i=1}^n a_i X_i)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}, \tag{A4}$$

where $R^2 = \boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{m}$, $C^2 = \boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m}$, $\boldsymbol{a}^\top = (\boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m})^{-1/2} \boldsymbol{m}^\top \boldsymbol{V}^{-1}$ and $b = R^2 \widehat{\sigma}/C$. W is close to one under normality. For the multivariate random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, using the fact that under H_0, the scaled residuals $\boldsymbol{Y}_j, j = 1, \ldots, n$ have a distribution close to $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, which means that the coordinates of $\boldsymbol{Y}_j$, denoted by $Y_{1j}, \ldots, Y_{pj}$, are approximately i.i.d. random variables from $\mathcal{N}(0, 1)$, Villasenor et al. (2009) proposed a test statistic for assessing MVN:

$$W^* = \frac{1}{p} \sum_{i=1}^p W_{Y_i},$$

where W_{Y_i} is the Shapiro–Wilk's test statistic evaluated on the ith coordinate of the transformed observations, $Y_{i1}, \ldots, Y_{in}$, $i = 1, \ldots, p$.

Majerski and Szkutnik (2010) derived some approximations to the most powerful invariant (MPI) tests for MVN. Exact MPI tests for univariate normality have been well studied for some specific alternatives, such as uniform, double exponential, exponential, and Cauchy (Uthoff, 1970; Uthoff et al., 1973; Franck, 1981). Exact, but computationally cumbersome, MPI tests for binomality have been developed by Szkutnik et al. (1988) for two specific alternatives only, i.e., bivariate uniform and bivariate exponential, and MPI tests for $p > 2$ have not been studied so far. Majerski and Szkutnik (2010) constructed approximations to the tests presented by Szkutnik et al. (1988) using the Laplace expansion for integrals, and showed that the approximations are asymptotically equivalent to the likelihood ratio (LR) tests, as is the case in the univariate setting. Furthermore, the authors extended their results to the cases of $p > 2$, which are, however, limited to low-dimensional cases, due to the computational accuracy and complexity of numerical integration approximations for high dimensions. By showing in simulation studies that the MPI tests have practically the same powers as the LR tests, they provided a strong motivation for using the simple and fast LR test procedures for higher dimensions.

Kim and Park (2018) presented extensions of the univariate omnibus LR tests, which are based on empirical distribution functions (EDF), to the tests for MVN. Zhang (2002) proposed a goodness-of-fit LR test statistic based on the univariate observations $X_1, \ldots, X_n$:

$$Z_A = -\sum_{i=1}^{n} \left[\frac{\log F_0(X_{(i)})}{n - i + 1/2} + \frac{\log\{1 - F_0(X_{(i)})\}}{i - 1/2} \right],$$

where $F_0(\cdot)$ is the null distribution function, and $X_{(i)}$'s are the order statistics. For the multivariate sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, the scaled residuals $\boldsymbol{Y}_j, j = 1, \ldots, n$, are approximately $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, which indicates that the coordinates of $\boldsymbol{Y}_j$, denoted by $Y_{1j}, \ldots, Y_{pj}$, and furthermore, all the elements $Y_{ij}, i = 1, \ldots, p, j = 1, \ldots, n$, are approximately i.i.d. random variables from $\mathcal{N}(0, 1)$. Kim and Park (2018) thus suggested the test statistic using the coordinate-wise characterization as

$$Z_A^* = \frac{1}{p} \sum_{i=1}^{p} Z_A^{(i)} = -\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{n} \left[\frac{\log \Phi(Y_{i(j)})}{n - i + 1/2} + \frac{\log\{1 - \Phi(Y_{i(j)})\}}{i - 1/2} \right],$$

where $Z_A^{(i)}, i = 1, \ldots, p$, is the univariate LR statistic for the ith component, $Y_{i(j)}$ is the jth order statistic of $Y_{i1}, \ldots, Y_{in}$, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of $\mathcal{N}(0, 1)$. The second test statistic based on the element-wise characterization is given by

$$Z_A^{**} = -\sum_{i=1}^{m} \left[\frac{\log \Phi(Y_{(i)})}{m - i + 1/2} + \frac{\log\{1 - \Phi(Y_{(i)})\}}{i - 1/2} \right],$$

where $Y_{(i)}$ is the ith order statistic of $\text{vec}\{(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^\top\}$, and $m = np$.

7.4 Multiple test procedures

In this section, we present two recent testing procedures that combine multiple tests for MVN.

Tenreiro (2011) proposed a multiple test procedure that combines a finite set of affine invariant test statistics for MVN through an improved Bonferroni method. The test statistic is

$$T_n(u) = \max_{h \in H} \{T_h - c_{n,h}(u)\}, \tag{A5}$$

where $u \in [0, 1]$, $T_h, h \in H$, is any finite family of affine invariant test statistics for MVN, and $c_{n,h}(u)$ is the quantile of order $1 - u$ of T_h under H_0. For a significance level $\alpha \in [0, 1]$, the multiple test procedure rejects the null hypothesis of MVN whenever $T_n(u_{n,\alpha}) > 0$, where $u_{n,\alpha} = \sup\{u \in [0, 1] : P_\phi\{T_n(u)\} > 0 \le \alpha\}$, and ϕ is the density for $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. The usefulness of such an approach is illustrated by a multiple test combining some of the most recommended

tests, i.e., the Mardia's skewness and kurtosis tests (Mardia, 1970, 1974) (the former performs well for skewed or long-tailed alternatives, and the latter for short-tailed alternatives), and the BHEP tests with two choices of the tuning parameter β in the statistic $T_{n,\beta}$:

$$\beta_S = 0.448 + 0.026p \quad \text{and} \quad \beta_L = 0.928 + 0.049p, \tag{A6}$$

which depend on the dimension p for $2 \le p \le 15$, and are identified from simulation studies by Tenreiro (2009) based on their distinct behavior patterns for the empirical power of BHEP tests as a function of β. β_S is shown to be suitable for short-tailed or high-moment alternatives, while β_L is appropriate for long-tailed or moderately skewed alternative distributions. The multiple test procedure was further studied in Tenreiro (2017), who combined BHEP tests with four different values of β in the statistic $T_{n,\beta}$: two non-extreme choices, where $\beta = \beta_S$ and $\beta = \beta_L$ as defined in Equation (A6), and two extreme cases, where $\beta \to 0$ and $\beta \to \infty$.

Zhou and Shao (2014) proposed a test that combines the univariate Shapiro-Wilk test for projected data and Mardia's multivariate kurtosis test. The Shapiro-Wilk test statistic W, given by Equation (A4), can be used to detect non-normality in univariate projections of the scaled residuals $\boldsymbol{Y}_j, j = 1, \ldots, n$, in the direction $\boldsymbol{\theta}$:

$$G_n(\boldsymbol{\theta}) = W(\boldsymbol{\theta}^\top \boldsymbol{Y}_1, \ldots, \boldsymbol{\theta}^\top \boldsymbol{Y}_n).$$

While Fattorini (1986) considered a test for detecting non-normality of multivariate data projected in the most "extreme" direction among $\|\boldsymbol{Y}_j\|^{-1}\boldsymbol{Y}_j$, $j = 1, \ldots, n$, corresponding to the smallest G_n value, Zhou and Shao (2014) considered the p most "extreme" directions corresponding to the p smallest G_n values evaluated at the same random directions, denoted by Θ_1. They also consider the p unit vector directions $\boldsymbol{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$, $j = 1, \ldots, p$, denoted by Θ_2, which project the multivariate data to the p marginal variates that are also normal under H_0. The new test statistic, incorperating Mardia's multivariate kurtosis statistic, is defined as

$$T_{n,c} = 1 - \frac{1}{2p} \sum_{\boldsymbol{\theta} \in \Theta_1 \cup \Theta_2} G_n(\boldsymbol{\theta}) I_A$$

where $A = \{c_1 \le MK \le c_2\}$ with c_1 and c_2 being certain percentiles (e.g., 1% and 99%, respectively) of the MK statistic given by Equation (1), and I_A is the indicator function with a value of 1 if A is true and 0 otherwise.

Table A1: Properties of the recent tests and classical tests for MVN for i.i.d. data.

Test	Affine invariance	Universal consistency	Known null distribution	Reference
1. Skewness and kurtosis approaches				
MS, MK	✓	✗	✓	Mardia (1974)
U, W	✓	✗	✓	Kankainen et al. (2007)
$Z_{2,p}^{HL}$	✓	✗	✗	Thulin (2014)
$b_{2,p,q}$	✓	✗	✓	Yamada et al. (2015)
JB_{BS}	✓	✓	✓	Bowman and Shenton (1975)
JB_{BS}	✓	✓	✓	Doornik and Hansen (2008)
$\widetilde{\text{LM}}$	✓	✓	✓	Jönsson (2011)
$\text{JB}_M, \text{RJB}_M, \text{RT}_M, \text{JBT}_M$	✓	✓	✓	Kim (2016)
M_1, M_2	✓	✓	✓	Koizumi et al. (2009)
M_3	✓	✓	✓	Enomoto et al. (2012)
$\text{MJB}_2, mMJB$	✓	✓	✓	Koizumi et al. (2014)
2. Chi-squared type tests				
NRR	✓	✗	✓	Moore and Stubblebine (1981)
χ^2	✓	✗	✓	Cardoso de Oliveira and Ferreira (2010)
$Z_{(\lambda)}$	✓	✗	✓	Batsidis et al. (2013)
SmG	✓	✗	✗	Hanusz and Tarasińska (2012)
G	✓	✓	✗	Madukaife and Okafor (2019)
Y_n^2, U_n^2, S_n^2	✓	✓	✓	Voinov et al. (2016)
3. BHEP-type tests				
$T_{n,\beta}$	✓	✓	✓	Henze and Zirkler (1990)
$T_{n,r}$	✓	✓	✓	Pudełko (2005)
$\widehat{D}_{n,m}$	✓	✓	✓	Arcones (2007)
$\check{T}_{n,\beta}$	✓	✓	✓	Henze and Jiménez-Gamero (2018)
$T_{n,\gamma}, \widetilde{T}_{n,\gamma}$	✓	✓	✓	Henze et al. (2019)
4. Other generalizations of univariate normality test				
$\delta_{n,p}$	✓	✓	✗	Székely and Rizzo (2005)
W^*	✓	✓	✗	Villasenor et al. (2009)
MPI	✗	✗	✗	Majerski and Szkutnik (2010)
Z_A^*, Z_A^{**}	✓	✓	✗	Kim and Park (2018)
5. Multiple test procedures				
$T_n(u)$	✓	✓	✗	Tenreiro (2011, 2017)
$T_{n,c}$	✓	✓	✗	Zhou and Shao (2014)

References

Ahmad, F. and Khan, R. A. (2015). A power comparison of various normality tests. *Pakistan Journal of Statistics and Operation Research*, 11(3):331–345.

Arcones, M. A. (2007). Two tests for multivariate normality based on the characteristic function. *Mathematical Methods of Statistics*, 16(3):177–201.

Baringhaus, L. and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348.

Batsidis, A., Martin, N., Pardo, L., and Zografos, K. (2013). A necessary power divergence type family tests of multivariate normality. *Communications in Statistics-Simulation and Computation*, 42(10):2253–2271.

Bera, A. and John, S. (1983). Tests for multivariate normality with Pearson alternatives. *Communications in Statistics-Theory and Methods*, 12(1):103–117.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, page 181. Oxford University Press, USA.

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. New York: John Wiley and Sons.

Bowman, K. and Shenton, L. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, 62(2):243–250.

Box, G. (1953). Non-normality and tests on variances. *Biometrika*, 40:318–335.

Cardoso de Oliveira, I. and Ferreira, D. (2010). Multivariate extension of chi-squared univariate normality test. *Journal of Statistical Computation and Simulation*, 80(5):513–526.

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464.

Csörgő, S. (1989). Consistency of some tests for multivariate normality. *Metrika*, 36(1):107–116.

D'Agostino, R. B. and Lee, A. F. (1977). Robustness of location estimators under changes of population kurtosis. *Journal of the American Statistical Association*, 72(358):393–396.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques*, volume 68 of *Statistics: A Series of Textbooks and Monographs*. New York: Marcel Dekker.

Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70:927–939.

Dutta, S. and Genton, M. G. (2014). A non-Gaussian multivariate distribution with all lower-dimensional Gaussians and related families. *Journal of Multivariate Analysis*, 132:82–93.

Eaton, M. L. and Perlman, M. D. (1973). The non-singularity of generalized sample covariance matrices. *The Annals of Statistics*, 1(4):710–717.

Ebner, B. and Henze, N. (2020). Tests for multivariate normality–a critical review with emphasis on weighted L^2-statistics. *arXiv preprint arXiv:2004.07332*.

Enomoto, R., Okamoto, N., and Seo, T. (2012). Multivariate normality test using Srivastava's skewness and kurtosis. *SUT Journal of Mathematics*, 48(1):103–115.

Epps, T. (1999). Limiting behavior of the ICF test for normality under Gram–Charlier alternatives. *Statistics & Probability Letters*, 42(2):175–184.

Epps, T. W. and Pulley, L. B. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726.

Farrell, P. J. and Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, 76(9):803–816.

Farrell, P. J., Salibian-Barrera, M., and Naczk, K. (2007). On tests for multivariate normality and associated simulation studies. *Journal of Statistical Computation and Simulation*, 77(12):1065–1080.

Fattorini, L. (1986). Remarks on the use of Shapiro-Wilk statistic for testing multivariate normality. *Statistica*, 46(2):209–217.

Franck, W. E. (1981). The most powerful invariant test of normal versus Cauchy with applications to stable alternatives. *Journal of the American Statistical Association*, 76(376):1002–1005.

Geary, R. C. (1947). Testing for normality. *Biometrika*, 34(3/4):209–242.

Gelfand, A. E. and Banerjee, S. (2010). Multivariate spatial process models. In *Handbook of Spatial Statistics (eds A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp)*, pages 495–515. Boca Raton: Chapman and Hall–CRC.

Gelfand, A. E. and Schliep, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, 18:86–104.

Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30(2):147–163.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

Guhaniyogi, R. and Banerjee, S. (2018). Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets. *Technometrics*, 60(4):430–444.

Haining, R. (1978). The moving average model for spatial interaction. *Transactions of the Institute of British Geographers*, 3(2):202–225.

Hanusz, Z., Enomoto, R., Seo, T., and Koizumi, K. (2018). A Monte Carlo comparison of Jarque–Bera type tests and Henze–Zirkler test of multivariate normality. *Communications in Statistics-Simulation and Computation*, 47(5):1439–1452.

Hanusz, Z. and Tarasińska, J. (2012). New tests for multivariate normality based on Small's and Srivastava's graphical methods. *Journal of Statistical Computation and Simulation*, 82(12):1743–1752.

Henze, N. (1990). An approximation to the limit distribution of the Epps-Pulley test statistic for normality. *Metrika*, 37(1):7–18.

Henze, N. (1997). Extreme smoothing and testing for multivariate normality. *Statistics & Probability Letters*, 35(3):203–213.

Henze, N. (2002). Invariant tests for multivariate normality: a critical review. *Statistical Papers*, 43(4):467–506.

Henze, N. and Jiménez-Gamero, M. D. (2018). A new class of tests for multinormality with iid and GARCH data based on the empirical moment generating function. *TEST*, 28(2):499–521.

Henze, N., Jiménez-Gamero, M. D., and Meintanis, S. G. (2019). Characterizations of multi-normality and corresponding tests of fit, including for GARCH models. *Econometric Theory*, 35(3):510–546.

Henze, N. and Wagner, T. (1997). A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis*, 62(1):1–23.

Henze, N. and Zirkler, B. (1990). A class of invariant and consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617.

Hoeffding, W. et al. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Horswell, R. L. and Looney, S. W. (1992). A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. *Journal of Statistical Computation and Simulation*, 42(1-2):21–38.

Horváth, L., Kokoszka, P., and Wang, S. (2020). Testing normality of data on a multivariate grid. *Journal of Multivariate Analysis*, 179:104640.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2(3):360–378.

Irvine, K. M., Gitelman, A. I., and Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: effects on covariance estimation. *Journal of Agricultural, Biological and Environmental Statistics*, 12(4):450–469.

Islam, T. U. (2017). Stringency-based ranking of normality tests. *Communications in Statistics–Simulation and Computation*, 46(1):655–668.

Isogai, T. (1982). On a measure of multivariate skewness and a test for multivariate normality. *Annals of the Institute of Statistical Mathematics*, 34(1):531–541.

Jarque, C. M. and Bera, A. K. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, 7(4):313–318.

Joenssen, D. W. and Vogel, J. (2014). A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, 84(5):1055–1078.

Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780.

Jönsson, K. (2011). A robust test for multivariate normality. *Economics Letters*, 113(2):199–201.

Kankainen, A., Taskinen, S., and Oja, H. (2007). Tests of multinormality based on location vectors and scatter matrices. *Statistical Methods and Applications*, 16(3):357–379.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214.

Keskin, S. (2006). Comparison of several univariate normality tests regarding type I error rate and power of the test in simulation based small samples. *Journal of Applied Science Research*, 2(5):296–300.

Kim, I. and Park, S. (2018). Likelihood ratio tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 47(8):1923–1934.

Kim, N. (2016). A robustified Jarque–Bera test for multivariate normality. *Economics Letters*, 140:48–52.

Koizumi, K., Hyodo, M., and Pavlenko, T. (2014). Modified Jarque–Bera type tests for multivariate normality in a high-dimensional framework. *Journal of Statistical Theory and Practice*, 8(2):382–399.

Koizumi, K., Okamoto, N., and Seo, T. (2009). On Jarque-Bera tests for assessing multivariate normality. *Journal of Statistics: Advances in Theory and Applications*, 1(2):207–220.

Kowalski, C. J. (1970). The performance of some rough tests for bivariate normality before and after coordinate transformations to normality. *Technometrics*, 12(3):517–544.

Koziol, J. (1987). An alternative formulation of Neyman's smooth goodness of fit tests under composite alternatives. *Metrika*, 34(1):17–24.

Lobato, I. N. and Velasco, C. (2004). A simple test of normality for time series. *Econometric Theory*, 20(4):671–689.

Loève, M. (1977). *Probability Theory, Vol. 1, 4th ed.* New York: Springer.

Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality. *The American Statistician*, 49(1):64–70.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. New York: Springer.

Madukaife, M. S. and Okafor, F. C. (2019). A new large sample goodness of fit test for multivariate normality based on chi-squared probability plots. *Communications in Statistics-Simulation and Computation*, 48(6):1651–1664.

Majerski, P. and Szkutnik, Z. (2010). Approximations to most powerful invariant tests for multinormality against some irregular alternatives. *TEST*, 19(1):113–130.

Malkovich, J. F. and Afifi, A. (1973). On tests for multivariate normality. *Journal of the American Statistical Association*, 68(341):176–179.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, 36:115–128.

Mardia, K. V. (1980). 9 tests of unvariate and multivariate normality. In *Handbook of Statistics (eds P. R. Krishnaiah)*, volume 1, pages 279–320. New York: Elsevier.

Mardia, K. V. and Foster, K. (1983). Omnibus tests of multinormality based on skewness and kurtosis. *Communications in Statistics-Theory and Methods*, 12(2):207–221.

Mardia, K. V. and Kent, J. (1991). Rao score tests for goodness of fit and independence. *Biometrika*, 78(2):355–363.

Mason, R. L. and Young, J. C. (1985). Re-examining two tests for bivariate normality. *Communications in Statistics-Theory and Methods*, 14(7):1531–1546.

Mecklin, C. J. and Mundfrom, D. J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75(2):93–107.

Moore, D. S. (1986). Tests of chi-squared type. In *Goodness-of-Fit Techniques (eds R. B. D'Agostino and M. A. Stephens)*, pages 63–96. New York: Marcel Dekker.

Moore, D. S. and Stubblebine, J. B. (1981). Chi-square tests for multivariate normality with application to common stock prices. *Communications in Statistics-Theory and Methods*, 10(8):713–738.

Móri, T. F., Rohatgi, V. K., and Székely, G. (1994). On multivariate skewness and kurtosis. *Theory of Probability & Its Applications*, 38(3):547–551.

Noughabi, H. A. and Arghami, N. R. (2011). Monte carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation*, 81(8):965–972.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599.

Olea, R. A. and Pawlowsky-Glahn, V. (2009). Kolmogorov–Smirnov test for spatially correlated data. *Stochastic Environmental Research and Risk Assessment*, 23(6):749–757.

Öztuna, D., Elhan, A. H., and Tüccar, E. (2006). Investigation of four different normality tests in terms of type I error rate and power under different distributions. *Turkish Journal of Medical Sciences*, 36(3):171–176.

Paciorek, C. J., Lipshitz, B., Zhuo, W., Kaufman, C. G., and Thomas, R. C. (2015). Parallelizing Gaussian process calculations in R. *Journal of Statistical Software*, 63(i10).

Pardo-Igúzquiza, E. and Dowd, P. A. (2004). Normality tests for spatially correlated data. *Mathematical Geology*, 36(6):659–681.

Pearson, E. S., D'Agostino, R. B., and Bowman, K. O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, 64(2):231–246.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, 29(3/4):322–335.

Psaradakis, Z. and Vávra, M. (2020). Normality tests for dependent data: large-sample and bootstrap approaches. *Communications in Statistics-Simulation and Computation*, 49(2):283–304.

Pudełko, J. (2005). On a new affine invariant and consistent test for multivariate normality. *Probability and Mathematical Statistics*, 25:43–54.

Read, T. R. and Cressie, N. A. (2012). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

Romao, X., Delgado, R., and Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation*, 80(5):545–591.

Romeu, J. L. and Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality. *Journal of Multivariate Analysis*, 46(2):309–334.

Sánchez-Espigares, J. A., Grima, P., and Marco-Almagro, L. (2019). Graphical comparison of normality tests for unimodal distribution data. *Journal of Statistical Computation and Simulation*, 89(1):145–154.

Seo, T. and Ariga, M. (2011). On the distribution of sample measure of multivariate kurtosis. *Journal of Combinatorics, Information & System Sciences*, 36(1-4):179.

Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63(324):1343–1372.

Small, N. (1978). Plotting squared radii. *Biometrika*, 65(3):657–658.

Srivastava, M. S. (1984). A measure of skewness and kurtosis and a graphical method for assessing multivariate normality. *Statistics & Probability Letters*, 2(5):263–267.

Subrahmaniam, K., Subrahmaniam, K., and Messeri, J. (1975). On the robustness of some tests of significance in sampling from a compound normal population. *Journal of the American Statistical Association*, 70(350):435–438.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334.

Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80.

Szkutnik, Z. et al. (1988). Most powerful invariant tests for binormality. *The Annals of Statistics*, 16(1):292–301.

Tenreiro, C. (2009). On the choice of the smoothing parameter for the BHEP goodness-of-fit test. *Computational Statistics & Data Analysis*, 53(4):1038–1053.

Tenreiro, C. (2011). An affine invariant multiple test procedure for assessing multivariate normality. *Computational Statistics & Data Analysis*, 55(5):1980–1992.

Tenreiro, C. (2017). A new test for multivariate normality by combining extreme and nonextreme BHEP tests. *Communications in Statistics-Simulation and Computation*, 46(3):1746–1759.

Thadewald, T. and Büning, H. (2007). Jarque–Bera test and its competitors for testing normality–a power comparison. *Journal of Applied Statistics*, 34(1):87–105.

Thode, H. C. (2002). *Testing for normality*, volume 164. Boca Raton: Chapman and Hall–CRC.

Thulin, M. (2014). Tests for multivariate normality based on canonical correlations. *Statistical Methods & Applications*, 23(2):189–208.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics (eds I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann)*, pages 448–485. CA: Stanford University Press.

Uthoff, V. A. (1970). An optimum test property of two weil-known statistics. *Journal of the American Statistical Association*, 65(332):1597–1600.

Uthoff, V. A. et al. (1973). The most powerful scale and location invariant test of the normal versus the double exponential. *The Annals of Statistics*, 1(1):170–174.

Villasenor, A., José, A., and Estrada, E. G. (2009). A generalization of Shapiro–Wilk's test for multivariate normality. *Communications in Statistics–Theory and Methods*, 38(11):1870–1883.

Voinov, V., Pya, N., Makarov, R., and Voinov, Y. (2016). New invariant and consistent chi-squared type goodness-of-fit tests for multivariate normality and a related comparative simulation study. *Communications in Statistics-Theory and Methods*, 45(11):3249–3263.

Volkmer, H. (2014). A characterization of the normal distribution. *Journal of Statistical Theory and Applications*, 13(1):83–85.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348.

Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America*, 17(12):684.

Yamada, T., Romer, M. M., and Richards, D. S. P. (2015). Kurtosis tests for multivariate normality with monotone incomplete data. *TEST*, 24(3):532–557.

Yan, Y., Jeong, J., and Genton, M. G. (2020). Multivariate transformed Gaussian processes. *Japanese Journal of Statistics and Data Science*, 3(1):129–152.

Yap, B. W. and Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155.

Yazici, B. and Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2):175–183.

Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):281–294.

Zheng, W. (2019). *Kolmogorov–Smirnov type tests under spatial correlations*. PhD thesis, UT School of Public Health Dissertations (Open Access).

Zhou, M. and Shao, Y. (2014). A powerful test for multivariate normality. *Journal of Applied Statistics*, 41(2):351–363.

Zimmerman, D. and Stein, M. (2010). Classical geostatistical methods. In *Handbook of Spatial Statistics (eds A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp)*, pages 29–44. Boca Raton: Chapman and Hall–CRC.