International Statistical Review

International Statistical Review (2023), **91**, 1, 114–139 doi: 10.1111/insr.12512

Are You All Normal? It Depends!

Wanfang Chen[1] and **Marc G. Genton**[2]

[1]*Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai, China*
[2]*Statistics Program, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*
Correspondence to: Wanfang Chen, Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai, China. Email: wfchen@fem.ecnu.edu.cn

Summary

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 first review 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. We further review briefly the few existing univariate tests under dependence (time or space), and then propose a new multivariate normality test for spatial data by accounting for the spatial dependence. The new test utilises the union-intersection principle to decompose the null hypothesis into intersections of univariate normality hypotheses for projection data, and it rejects the multivariate normality if any individual hypothesis is rejected. The individual hypotheses for univariate normality are conducted using a Jarque–Bera type test statistic that accounts for the spatial dependence in the data. We also show in simulation studies that the new test has a good control of the type I error and a high empirical power, especially for large sample sizes. We further illustrate our test on bivariate wind data over the Arabian Peninsula.

Key words: Gaussian process; Jarque–Bera test; skewness and kurtosis; spatial dependence; spatial statistics; test for multivariate normality.

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 & 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, for example, Mardia (1980), D'Agostino & 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 & Rogers-Stewart (2006),

Thadewald & Büning (2007), Yazici & Yolacan (2007), Romao *et al.* (2010), Yap & Sim (2011), Noughabi & Arghami (2011), Ahmad & Khan (2015), Islam (2017) and Sánchez-Espigares *et al.* (2019).

Relatively less work has been carried out in the field of testing for multivariate normality (MVN) compared with that carried out for the univariate case, because 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 & Genton, 2014). In addition, classical univariate normality tests, such as the chi-squared test, have limited applicability in higher dimensions. Reviews on the tests for MVN have been given by Thode (2002), Henze (2002) and Ebner & Henze (2020), with the last one emphasising 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 & Looney (1992), Romeu & Ozturk (1993), Mecklin & Mundfrom (2005), Farrell *et al.* (2007), Joenssen & Vogel (2014) and Hanusz *et al.* (2018). The Jarque–Bera (JB) type test (Jarque & 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 & Stein (2010) and Gelfand & Schliep (2016) provided surveys of Gaussian modelling 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, for example, Nychka *et al.* (2015), Paciorek *et al.* (2015), Katzfuss (2017) and Guhaniyogi & 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 realisation of a Gaussian random field. All the aforementioned tests cannot be directly used for spatial data, because 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 by simulation studies in Section 5 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 dependent, as also indicated by the severely inflated type I error from our simulation study in Section 6.

A review on univariate normality tests for data with serial dependence in time series is given by Psaradakis & Vávra (2020), but these tests need to be justified, extended or modified if they are to be applied to spatial data, and further generalised to the multivariate setting, which is not always possible. Pardo-Igúzquiza & Dowd (2004) demonstrated a methodology for the application of standard univariate normality tests, such as the Kolmogorov–Smirnov test, the chi-squared 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 & 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 ever 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. We consider the practically common case where the data to be tested are the zero-mean residuals of regression and spatial models. The type I error and empirical power of the new test are assessed by simulation studies. In the title of our paper, the 'All' represents 'multivariate', and the answer to the question of testing multivariate normality 'depends' on the underlying dependence (in space or time).

The rest of this paper is organised as follows. Section 2 introduces some useful preliminaries, terminologies and notations. Section 3 reviews the recent developments of MVN tests based on the skewness and kurtosis approaches in the i.i.d. setting; Section 4 reviews the chi-squared type and BHEP-type tests, and the other types of MVN test for i.i.d. data are presented in the supporting information. Section 5 demonstrates a simulation study to investigate the influence of spatial dependence on the measures of skewness and kurtosis for multivariate Gaussian random fields. Section 6 describes our new test for MVN under spatial dependence and its performance based on the type I error and empirical power. Section 7 describes a data application based on bivariate wind data over the Arabian Peninsula. Section 8 concludes and discusses future work directions.

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 affine invariance and universal consistency. Because 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 $\mathbf{A}\boldsymbol{X}_1 + \boldsymbol{b}, \ldots, \mathbf{A}\boldsymbol{X}_n + \boldsymbol{b}$, where $\mathbf{A} \in \mathbb{R}^{p \times p}$ 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, that is, $T_n(\mathbf{A}\boldsymbol{X}_1 + \boldsymbol{b}, \ldots, \mathbf{A}\boldsymbol{X}_n + \boldsymbol{b}) = T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. 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.

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

3 Recent Advances of Multivariate Normality 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: (i) skewness and kurtosis approaches; (ii) chi-squared type tests; (iii) BHEP-type tests based on the empirical characteristic function; (iv) other generalisations of univariate normality tests; and (v) multiple testing procedures that combine multiple tests for MVN. In this section, we review the first category, that is, tests based on skewness and kurtosis measures. In the next section, we review the chi-squared type and BHEP-type tests, and also present the review for the remaining two categories in the supporting information for readers' reference.

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

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

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

$$\beta_{1,p} = \mathrm{E}\left[\left\{(\boldsymbol{X} - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})\right\}^3\right] \text{ and } \beta_{2,p} = \mathrm{E}\left[\left\{(\boldsymbol{X} - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right\}^2\right],$$

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

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

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

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

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

which are asymptotically $\chi^2_{p(p+1)(p+2)/6}$ and $\mathcal{N}(0,1)$, respectively, under H_0. Other classical measures of multivariate skewness and kurtosis and related tests for MVN have been proposed by, for example, Malkovich & 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 standardised distance between two separate location parameters, and measures of kurtosis based on the ratios of two scale measures, such as the classical standardised 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. A vector-valued (matrix-valued) statistic is called a location vector (a scatter matrix) if it is affine equivariant (see Section 2 in Kankainen *et al.* (2007)). Then, the test statistic for MVN (to detect skewness) is given by $U = (\boldsymbol{T}_1 - \boldsymbol{T}_2)^{\top}\mathbf{C}^{-1}(\boldsymbol{T}_1 - \boldsymbol{T}_2)$, where $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are two separate location vectors and $\mathbf{C}$ is a scatter matrix, and the kurtosis test statistic is given by

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

where $\|\cdot\|^2 = \mathrm{tr}(\cdot^{\top}\cdot)$, and $\mathbf{C}_1$ and $\mathbf{C}_2$ are two separate scatter matrices. Using special choices of location and scatter estimators, it is possible to obtain generalisations 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 $\overline{\boldsymbol{X}} = (\overline{X}_1, \ldots, \overline{X}_p)^{\top}$, $\mathbf{S} = \{S_{ij}\}$, 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 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, $Z^{HL}_{2,p}$, 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

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

Yamada *et al.* (2015) generalised Mardia's multivariate kurtosis for testing MVN when the data consist of a random sample of two-step monotone incomplete observations.

One disadvantage of the aforementioned 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 & 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 aforementioned issue by combining the skewness and kurtosis measures. The univariate JB test (Jarque & Bera, 1981), based on a univariate random sample $X_i \in \mathbb{R}$, $i = 1, \ldots, n$, is given by $\mathrm{JB} = \frac{nb_1^2}{6} + \frac{n(b_2-3)^2}{24}$, 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, JB 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 $\mathrm{JB_M} = \sum_{i=1}^{p}\frac{nb_{1(i)}^2}{6} + \sum_{i=1}^{p}\frac{n(b_{2(i)}-3)^2}{24}$, where $b_{1(i)}$ and $b_{2(i)}$ denote the sample skewness and kurtosis, respectively, of component i. $\mathrm{JB_M}$ is asymptotically distributed as χ^2_{2p} under H_0 (see, e.g. Lütkepohl, 2005). However, for both JB and $\mathrm{JB_M}$, 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 & Hansen (2008) proposed to use transformed skewness and kurtosis, which creates statistics that are much closer to standard normal, based on the work of Bowman & Shenton (1975). Specifically, the test statistic is

$$\mathrm{JB_{DH}} = \boldsymbol{B}_1^{\top}\boldsymbol{B}_1 + \boldsymbol{B}_2^{\top}\boldsymbol{B}_2, \tag{3}$$

where $\boldsymbol{B}_1 = (b^*_{1(1)}, \ldots, b^*_{1(p)})^{\top}$ and $\boldsymbol{B}_2 = (b^*_{2(1)}, \ldots, b^*_{2(p)})^{\top}$ are the transformed vectors of skewness and kurtosis, respectively. $\mathrm{JB_{DH}}$ 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 $\mathrm{JB_M}$; see his Figure 1. He suggested using the test statistic that pools the individual p-values: $\widetilde{\mathrm{LM}} = -2\sum_{i=1}^{p}\ln(\pi_i)$, where π_i is the p-value of the univariate JB test for the ith component. $\widetilde{\mathrm{LM}}$

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 $\widetilde{\mathrm{LM}}$ is somewhat more convenient than using the transformation approach proposed by Doornik & 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 standardised data, $\boldsymbol{Z}_i = \mathbf{S}^{*\top}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})$, $i = 1, \ldots, n$, follow a $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ asymptotically under H_0, where $\mathbf{S}^*$ is defined by $\mathbf{S}^{*\top}\mathbf{S}\mathbf{S}^* = \mathbf{I}_p$. 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. Bera & John, 1983; Mardia & Foster, 1983; and Mardia & Kent, 1991). Koizumi *et al.* (2009) proposed two JB-type tests based on multivariate sample skewness and kurtosis of Srivastava (1984). For the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, let $\mathbf{S} = \mathbf{H}\mathbf{D}_\omega\mathbf{H}^\top$, where $\mathbf{H} = (\boldsymbol{h}_1 \ldots \boldsymbol{h}_p)$ is an orthogonal matrix and $\mathbf{D}_\omega = \mathrm{diag}(\omega_1, \ldots, \omega_p)$. The sample measures of multivariate skewness and kurtosis given by Srivastava (1984) are

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

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 $\tilde{b}_{1,p}$ and $\tilde{b}_{2,p}$ are:

$$M_1 = np\left\{\frac{\tilde{b}_{1,p}}{6} + \frac{(\tilde{b}_{2,p} - 3)^2}{24}\right\} \text{ and } M_2 = \frac{p\tilde{b}_{1,p}}{\mathrm{E}(\tilde{b}_{1,p})} + \frac{\{\tilde{b}_{2,p} - \mathrm{E}(\tilde{b}_{2,p})\}^2}{\mathrm{Var}(\tilde{b}_{2,p})},$$

both asymptotically χ^2_{p+1} under H_0, with $\mathrm{E}(\tilde{b}_{1,p}) = \frac{6(n-2)}{(n+1)(n+3)}$, $\mathrm{E}(\tilde{b}_{2,p}) = \frac{3(n-1)}{n+1}$, and $\mathrm{Var}(\tilde{b}_{2,p}) = \frac{24n(n-2)(n-3)}{p(n+1)^2(n+3)(n+5)}$ under H_0. Enomoto *et al.* (2012) noticed a difference between the upper percentiles of the distributions of M_2 and the chi-squared 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:

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

where $z_{WH} = \frac{(z_1/p)^{1/3} - 1 + 2/(9p)}{\sqrt{2/(9p)}}$ is the Wilson–Hilferty transform (Wilson & Hilferty, 1931) of $z_1 = np\tilde{b}_{1,p}/6$. 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 normalising transform of the sample kurtosis as suggested in Seo & Ariga (2011):

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

where $z_{NT} = \sqrt{\frac{np}{24}}\left\{-e^{-(\tilde{b}_{2,p}-3)} + 1 + \frac{6}{n} + \frac{12}{np}\right\}$. The statistic mMJB is asymptotically χ_2^2 under H_0, and proved to have a more stable behaviour in small samples. They further studied the F -approximation for mMJB which is shown to be better than the chi-squared approximation, and therefore can be recommended for testing MVN in both small and large samples.

4 Review of Other Recent Multivariate Normality Tests for i.i.d. Data

In this section, we review the chi-squared and BHEP-type tests. The remaining types of MVN tests for i.i.d. data (i.e. other generalisations of univariate normality tests and multiple testing procedures that combine multiple tests for MVN) are presented in the supporting information. We summarise some important properties (affine invariance, universal consistency, explicit null distribution) for all the reviewed tests in Table 1.

4.1 Chi-squared Type Tests

The chi-squared 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 chi-squared statistic is then given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{5}$$

which is asymptotically χ_{k-1}^2 under any null distribution. One disadvantage of the chi-squared 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 chi-squared 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, that is, 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 chi-squared test for MVN include Kowalski (1970), Moore & Stubblebine (1981) and Mason & Young (1985), and a few recent studies, presented below, also focused on the chi-squared type tests for MVN.

Cardoso de Oliveira & Ferreira (2010) proposed a multivariate chi-square test for MVN based on the fact that the statistics

$$B_i = \frac{n}{(n-1)^2}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\top} \tilde{\mathbf{S}}^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}),\ i = 1, \ldots, n, \tag{6}$$

Table 1. *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)
$\mathrm{JB}_{\mathrm{BS}}$	✓	✓	✓	Bowman & Shenton (1975)
$\mathrm{JB}_{\mathrm{DH}}$	✓	✓	✓	Doornik & Hansen (2008)
LM	✓	✓	✓	Jönsson (2011)
JB_M, RJB_M, RT_M, JBT_M	✓	✓	✓	Kim (2016)
M_1, M_2	✓	✓	✓	Koizumi *et al.* (2009)
M_3	✓	✓	✓	Enomoto *et al.* (2012)
MJB_2, *mMJB*	✓	✓	✓	Koizumi *et al.* (2014)
2. Chi-squared type tests				
NRR	✓	×	✓	Moore & Stubblebine (1981)
χ^2	✓	×	✓	Cardoso de Oliveira & Ferreira (2010)
$Z_{(\lambda)}$	✓	×	✓	Batsidis *et al.* (2013)
SmG	✓	×	×	Hanusz & Tarasińska (2012)
G	✓	✓	×	Madukaife & Okafor (2019)
Y_n^2, U_n^2, S_n^2	✓	✓	✓	Voinov *et al.* (2016)
3. BHEP-type tests				
$T_{n,\beta}$	✓	✓	✓	Henze & Zirkler (1990)
$T_{n,r}$	✓	✓	✓	Pudełko (2005)
$\hat{D}_{n,m}$	✓	✓	✓	Arcones (2007)
$\tilde{T}_{n,\beta}$	✓	✓	✓	Henze & Jiménez-Gamero (2018)
$T_{n,\gamma}$, $\tilde{T}_{n,\gamma}$	✓	✓	✓	Henze *et al.* (2019)
4. Other generalisations of univariate normality test				
$\delta_{n,p}$	✓	✓	×	Székely & Rizzo (2005)
W^*	✓	✓	×	Villasenor *et al.* (2009)
MPI	×	×	×	Majerski & Szkutnik (2010)
Z_A^*, Z_A^{**}	✓	✓	×	Kim & Park (2018)
5. Multiple test procedures				
$T_n(u)$	✓	✓	×	Tenreiro (2011), Tenreiro (2017)
$T_{n,c}$	✓	✓	×	Zhou & Shao (2014)

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

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

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 $\mathrm{Beta}(p/2, (n-p-1)/2)$ distribution, then the ith class is defined by $\{q | q_{i-1} < q \le q_i\}$ 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 (5), which is asymptotically distributed as χ^2_{k-1} under H_0.

Noticing that the aforementioned testing procedure was a k-dimensional multinomial goodness-of-fit test, and Pearson's chi-squared 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 & Read, 1984; Read & Cressie, 2012):

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

which includes as a specific case the Pearson's chi-squared statistic, Equation (5), when $\lambda = 1$. Here $Z_{(\lambda)}$ 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 & 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 (6), and D_i's are Beta order statistics using Blom's general plotting position (Blom, 1958): $\dfrac{i - \alpha}{n - \alpha - \beta + 1}$, $i = 1, \ldots, n$, with $\alpha = (p - 2)/(2p)$ and $\beta = 0.5 - (n - p - 1)^{-1}$. Another graphical method was proposed by Srivastava (1984). Hanusz & Tarasińska (2012) formalised both graphical methods using explicit test statistics. For example, they formalised the testing procedure of Small (1978) by constructing a geometric test statistic, SmG, that measures the departure of empirical points from the line $y = x$, that is, 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 & Okafor (2019) pointed out that some areas in the aforementioned 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

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

which are asymptotically distributed as χ^2_p under H_0, and the set of the population quantiles of the χ^2_p 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, that is, the quantiles of the χ^2_p distribution. Again, large values of G will lead to rejection of MVN of the data.

Voinov *et al.* (2016) found that the chi-squared test statistic for MVN, that is, the Nikulin–Rao–Robson (NRR) statistic, proposed in Moore & Stubblebine (1981), is asymptotically chi-squared 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^2_n, as well as its decomposition, $Y^2_n = U^2_n + S^2_n$, for any diagonal covariance matrix of any dimensionality p (see their equations (6), (9) and (10)) and suggested a procedure for testing MVN: (i) produce the Karhunen–Loève transformation of the sample data, which will diagonalise the sample covariance matrix and (ii) compute the statistics Y^2_n, U^2_n and S^2_n according to their equations (6), (9) and (10), respectively, based on

the transformed data. Because U_n^2 and S_n^2 are asymptotically independent under H_0, they can be used as test statistics independently from each other.

4.2 Baringhaus–Henze–Epps–Pulley 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 & Pulley (1983) provided a test for univariate normality based on the empirical CF, and Baringhaus & Henze (1988) generalised their idea to the multivariate case. Henze & 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

$$T_{n,\beta} = n\int_{\mathbb{R}^p} |\Psi_n(\boldsymbol{t}) - \Psi(\boldsymbol{t})|^2 \psi_\beta(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}, \tag{7}$$

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 $\mathcal{N}_p(\boldsymbol{0}, \mathbf{I}_p)$, 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 $\mathcal{N}_p(\boldsymbol{0}, \beta^2\mathbf{I}_p)$. 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 & Henze (1988), Csörgő (1989), Henze (1990), Henze (1997), Henze & 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 ever 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

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

with $\Psi_n(\boldsymbol{t})$ and $\Psi(\boldsymbol{t})$ defined as earlier. 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 characterisation 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

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

$\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ are estimators of $\boldsymbol{\mu}_{F_X}$ and $\boldsymbol{\Sigma}_{F_X}$, respectively, and $I_m^n = \{(j_1, \ldots, j_m) \in \mathbb{N}^m : 1 \leq j_k \leq n, j_k \neq j_l \text{ if } k \neq l\}$. If $m = 1$, $\widehat{\boldsymbol{\mu}}_n = \overline{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Sigma}}_n = \mathbf{S}$, then $\widehat{D}_{n,m}$ agrees with $T_{n,\beta}$ in Equation (7).

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

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

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 $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$, and $\omega_\beta(\boldsymbol{t}) = \exp(-\beta\|\boldsymbol{t}\|^2)$ with $\beta > 1$ is the weighting function, which leads to a representation of $\tilde{T}_{n,\beta}$ (see their Equation (1.4)) that is amenable to computational purposes. The authors showed that after a suitable scaling, $\tilde{T}_{n,\beta}$ 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 $\tilde{T}_{n,\beta}$ 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 generalised a characterisation 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 is defined by $\tilde{T}_{n,\gamma} = \int_{\mathbb{R}^p} U_n(\boldsymbol{t})\omega_\gamma(\boldsymbol{t})\mathrm{d}\boldsymbol{t}$:

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

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

5 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. Due to these deviations, the usual test statistics based on sample skewness and kurtosis can have a quite different asymptotic behaviour under spatial dependence, so that the usual test of normality, which depends on the asymptotic property derived under the i.i.d. assumption, may be misleading for spatially

correlated data. Therefore, there is a need to construct a new MVN test under spatial dependence, which is the focus of the next section.

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 $\mathbf{C}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \{C_{ij}(\boldsymbol{s}_1, \boldsymbol{s}_2)\}_{i,j=1}^p$, 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 $\boldsymbol{\Sigma} = \{\mathbf{C}(\boldsymbol{s}_i, \boldsymbol{s}_j)\}_{i,j=1}^n$ 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 & 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

$$C_{11}(\boldsymbol{h}) = \sigma_1^2 \mathcal{M}(\boldsymbol{h}|\nu_1, \beta), \quad C_{22}(\boldsymbol{h}) = \sigma_2^2 \mathcal{M}(\boldsymbol{h}|\nu_2, \beta), \tag{8}$$

$$C_{12}(\boldsymbol{h}) = C_{21}(\boldsymbol{h}) = \rho_{12}\sigma_1\sigma_2 \mathcal{M}\left(\boldsymbol{h}|\frac{1}{2}(\nu_1 + \nu_2), \beta\right), \tag{9}$$

where σ_1^2 and σ_2^2 are the marginal variances, $\mathcal{M}(\boldsymbol{h}|\nu, \beta) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\|\boldsymbol{h}\|/\beta)^\nu \mathcal{K}_\nu(\|\boldsymbol{h}\|/\beta)$, $\nu > 0$ 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:

$$|\rho_{12}| \leq \frac{\Gamma\left(\nu_1 + \frac{d}{2}\right)^{1/2}}{\Gamma(\nu_1)^{1/2}} \frac{\Gamma\left(\nu_2 + \frac{d}{2}\right)^{1/2}}{\Gamma(\nu_2)^{1/2}} \frac{\Gamma\left\{\frac{1}{2}(\nu_1 + \nu_2)\right\}}{\Gamma\left\{\frac{1}{2}(\nu_1 + \nu_2) + \frac{d}{2}\right\}}. \tag{10}$$

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 behaviours 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 (8) and (9) with smoothness parameters $\nu_1 = \nu_2 = 0.5$ (Exponential) or $\nu_1 = \nu_2 = 1$ (Whittle), and the colocated correlation coefficient ρ_{12} can be either positive (e.g. 0.5) or negative (e.g. -0.5) as long as it satisfies the inequality (10). Both marginal variances are set to 1 for simplicity. Further, the spatial dependence can be characterised 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 a 15×15 regular grid of locations over the unit square, set the effective range $h^* \in \{0.1, 0.12, 0.14, \ldots, 0.88, 0.9\}$, implying an increasing degree of spatial dependence, and solve the following equations:

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

to obtain the values of the spatial range parameter β. We simulate 200 replicates 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

$\boldsymbol{e}$ 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 $\boldsymbol{e} \in \mathbb{R}^{2n}$, and obtain the values of $\boldsymbol{Z}$ by $\boldsymbol{Z} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}(h^*))\boldsymbol{e} \in \mathbb{R}^{2n}$, 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

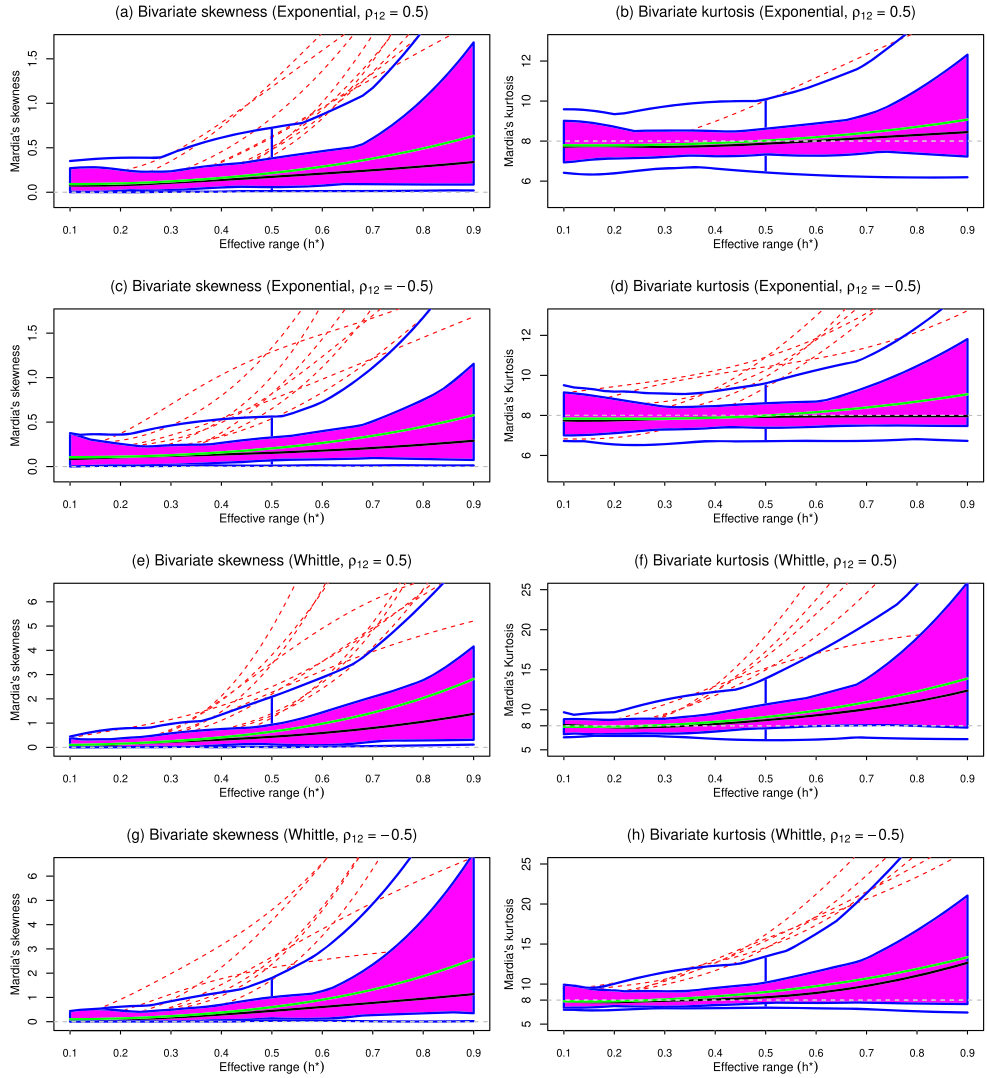

Figure 1. *Functional boxplot of the Mardia's sample skewness and kurtosis of the bivariate Gaussian random field in* $[0, 1] \times [0, 1]$ *as a function of the effective range* h^* *for (a)–(d) the Exponential and (e)-(h) the Whittle covariance functions. The green curve is the point-wise mean curve, the black curve is the median curve, the purple shaded region is the envelope of the 50% central region, the outer blue curves represent the maximum non-outlying envelope, and the red dashed curves 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 grey dashed lines*

$\boldsymbol{\Sigma}_{22}$ 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 $\boldsymbol{e}$ and isolate the effect of changing the parameters, particularly 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_{12}) that is specified in the covariance structure. We then summarise the 200 curves of sample skewness and kurtosis as a function of h^* or ρ_{12} by functional boxplot (Sun & Genton, 2011), which is an extension of the classical boxplot for visualising functional data. A functional boxplot displays three descriptive statistics: the median curve, the envelope of the 50% central region, and the maximum non-outlying envelope (Sun & 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 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 $\rho_{12} > 0$ and where $\rho_{12} < 0$ is small if we compare, for example, (a) with (c) or (b) with (d).

6 The New Test for Multivariate Normality Under Spatial Dependence

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

$$X(\boldsymbol{s}) = \mu + \sum_{\boldsymbol{t} \in \mathbb{Z}^d} a(\boldsymbol{t})\varepsilon(\boldsymbol{s} - \boldsymbol{t}), \quad \boldsymbol{s} \in \mathbb{Z}^d, \tag{11}$$

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

$$\mathrm{JB}^* = \frac{S_n^2}{\hat{\phi}_{\mathrm{S}}^2} + \frac{K_n^2}{\hat{\phi}_{\mathrm{K}}^2}, \tag{12}$$

where S_n and K_n are sample skewness and kurtosis of the standardised 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 kernel estimators, $\hat{\phi}_{\mathrm{S}}^2$ and $\hat{\phi}_{\mathrm{K}}^2$, as

$$\hat{\phi}_{\mathrm{S}}^2 = 6\sum_{\boldsymbol{h}} \omega_{\boldsymbol{b}}(\boldsymbol{h})\widehat{C}^3(\boldsymbol{h}) := 6\sum_{l=1}^{d} \sum_{|h_l| \le b_l} \left\{ \prod_{l=1}^{d} k\left(\frac{h_l}{b_l}\right) \right\} \widehat{C}^3(\boldsymbol{h}), \tag{13}$$

$$\hat{\phi}_{\mathrm{K}}^2 = 24\sum_{\boldsymbol{h}} \omega_{\boldsymbol{b}}(\boldsymbol{h})\widehat{C}^4(\boldsymbol{h}) := 24\sum_{l=1}^{d} \sum_{|h_l| \le b_l} \left\{ \prod_{l=1}^{d} k\left(\frac{h_l}{b_l}\right) \right\} \widehat{C}^4(\boldsymbol{h}), \tag{14}$$

where $\widehat{C}(\boldsymbol{h})$ is the sample auto-covariance function for the standardised observations with spatial lag $\boldsymbol{h} = (h_1, \ldots, h_d)^\top$; $k(\cdot)$ 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 $\widehat{C}(\boldsymbol{h})$, 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 develop a test for the multivariate case, we adopt the union-intersection testing approach originally proposed by Roy (1957). The union-intersection principle can be formulated as follows. Suppose we have a p-variate spatial dataset $\mathcal{X} = \{\boldsymbol{X}(\boldsymbol{s}_1), \boldsymbol{X}(\boldsymbol{s}_2), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)\}$, where $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n\}$ 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 $\boldsymbol{s}_i$, $i = 1, \ldots, n$. Note that the hypothesis $H_0: F_{\boldsymbol{X}} \in \mathcal{N}_p$ holds true exactly if and only if the projection $\boldsymbol{a}^\top \boldsymbol{X}$ has a univariate normal distribution for all vectors $\boldsymbol{a} \in \mathbb{R}^p$. For each $\boldsymbol{a} \in \mathbb{R}^p$, we construct a test $H_{\boldsymbol{a}}: \boldsymbol{a}^\top \boldsymbol{X}$ is normal against the alternative $H_{\boldsymbol{a}}^c: \boldsymbol{a}^\top \boldsymbol{X}$ is not normal, with acceptance region $\mathcal{A}_{\boldsymbol{a}}$ and rejection region $\mathcal{R}_{\boldsymbol{a}}$. Then the union-intersection test identifies the acceptance region for $H_0: F_{\boldsymbol{X}} \in \mathcal{N}_p$ as $\mathcal{A} = \bigcap_{\boldsymbol{a} \in \mathbb{R}^p} \mathcal{A}_{\boldsymbol{a}}$, and the rejection region as $\mathcal{R} = \mathcal{A}^c = \bigcup_{\boldsymbol{a} \in \mathbb{R}^p} \mathcal{R}_{\boldsymbol{a}}$; that is, the union-intersection test accepts H_0 exactly if $H_{\boldsymbol{a}}$ is accepted for all $\boldsymbol{a} \in \mathbb{R}^p$, and rejects H_0 if $H_{\boldsymbol{a}}$ is rejected for at least one vector $\boldsymbol{a} \in \mathbb{R}^p$.

For a fixed $\boldsymbol{a} \in \mathbb{R}^p$, the projected sample $\mathcal{X}_1 = \{\boldsymbol{a}^\top \boldsymbol{X}(\boldsymbol{s}_1), \boldsymbol{a}^\top \boldsymbol{X}(\boldsymbol{s}_2), \ldots, \boldsymbol{a}^\top \boldsymbol{X}(\boldsymbol{s}_n)\}$ is a univariate spatial dataset, and thus we can apply the method in Horváth *et al.* (2020) to test '$H_{\boldsymbol{a}}: \boldsymbol{a}^\top \boldsymbol{X}$ is normal' based on the new sample, under the following assumption.

Assumption 1. *Assume that under H_0, the observations $\mathcal{X} = \{\boldsymbol{X}(\boldsymbol{s}_1), \boldsymbol{X}(\boldsymbol{s}_2), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)\}$ follow a multivariate Gaussian spatial moving average (or kernel convolution) process:*

$$X_l(\boldsymbol{s}) = \mu_l + \sigma_l \sum_{\boldsymbol{t} \in \mathbb{Z}^d} k_l(\boldsymbol{s} - \boldsymbol{t})\omega(\boldsymbol{t}), \quad \boldsymbol{s} \in \mathbb{Z}^d, \ l = 1, \ldots, p, \tag{15}$$

where μ_l is the unknown mean, σ_l is the unknown standard deviation, $k_l(\cdot)$, $l = 1, \ldots, p$, is a set of p square integrable kernel functions on $\mathbb{Z}^d$ with $k_l(\mathbf{0}) = 1$, and $\omega(\cdot)$ is a zero-mean, unit-variance Gaussian random field on $\mathbb{Z}^d$ with a certain correlation function ρ.

Assumption 1 implies that under H_0, the linear combination $\boldsymbol{a}^\top \boldsymbol{X}$, for each $\boldsymbol{a} \in \mathbb{R}^p$, is from a strictly stationary Gaussian spatial moving average process as defined in Equation (11), so that the test of Horváth *et al.* (2020) can be applied. Under Assumption 1, $\boldsymbol{X}$ is 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

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

The kernel convolution technique (Gelfand & Banerjee, 2010) in Assumption 1 is a well-known approach for creating rich classes of stationary processes (Bernardo *et al.*, 2003). Therefore, our new test for MVN can be applied to spatial datasets with this big class of dependence structures.

Now, denote the JB-type test statistic for each $H_{\boldsymbol{a}}$ as $\mathrm{JB}_{\boldsymbol{a}}$, computed from Equation (12) based on the univariate sample $\mathcal{X}_1 = \{\boldsymbol{a}^\top \boldsymbol{X}(\boldsymbol{s}_1), \boldsymbol{a}^\top \boldsymbol{X}(\boldsymbol{s}_2), \ldots, \boldsymbol{a}^\top \boldsymbol{X}(\boldsymbol{s}_n)\}$. Suppose that the corresponding acceptance region is $\mathcal{A}_{\boldsymbol{a}} = \{\mathcal{X}_1 \colon \mathrm{JB}_{\boldsymbol{a}} \le c\}$ and the rejection region is $\mathcal{R}_{\boldsymbol{a}} = \{\mathcal{X}_1 \colon \mathrm{JB}_{\boldsymbol{a}} > c\}$, where c is a properly chosen constant (critical value) that does not depend on $\boldsymbol{a}$. Then the union-intersection test accepts H_0 exactly if $\max_{\boldsymbol{a} \in \mathbb{R}^p, \boldsymbol{a} \neq \boldsymbol{0}} \mathrm{JB}_{\boldsymbol{a}} \le c$. The critical value c for the test must be determined by the distribution of the statistic $\max_{\boldsymbol{a} \in \mathbb{R}^p, \boldsymbol{a} \neq \boldsymbol{0}} \mathrm{JB}_{\boldsymbol{a}}$, which is difficult to obtain in the current setting. In fact, this union-intersection test consists of infinitely many univariate tests. In practice, we can randomly select a large number of vectors, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K \in \mathbb{R}^p$, and do multiple testing; if at least one test $H_{\boldsymbol{a}}$ is rejected, then H_0 is also rejected; otherwise, if all tests $H_{\boldsymbol{a}_1}, \ldots, H_{\boldsymbol{a}_K}$ are not rejected, then this provides an evidence of not rejecting H_0. The number of tests, K, can be chosen as large as feasible for computation. In order to have a certain significance level α for the original test, the individual univariate tests cannot have the same level (Flury, 2013). Suppose that each test has a level α, then the chance of a false rejection of the null for each test is α, but the chance of at least one false rejection is much higher. In order to control the false discovery rate (FDR), which is the expected proportion of false rejections, the multiple testing procedure can be conducted based on the Benjamini–Hochberg (BH) method (Benjamini & Hochberg, 1995). Specifically, denote the ordered p-values for the K univariate tests as $P_{(1)}, \ldots, P_{(K)}$, and $R = \max\left\{i \colon P_{(i)} < \frac{i\alpha}{K}\right\}$. The BH rejection threshold is defined as $T = P_{(R)}$, and the hypothesis $H_{\boldsymbol{a}_i}$ is rejected if $P_i \le T$. If this procedure is applied, then it can be shown that FDR $\le \alpha$. Because the new test is a JB-type test, it is affine invariant and universally consistent.

6.2 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 with various configurations of 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 spatial applications) from the spatial moving average (kernel convolution) process of Equation (15). Specifically, each variable is generated from the spatial moving average model defined in Haining (1978), located on the points of a rectangular square lattice $\mathbb{Z}^2$:

$$X_l(i,j) = \theta_l\{e(i-1,j)+e(i+1,j)+e(i,j-1)+e(i,j+1)\}+e(i,j), \quad l = 1, \ldots, p, \tag{16}$$

where i and j are integers satisfying $1 \le i \le M$ and $1 \le j \le N$, $e(\cdot, \cdot)$ 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 $1 \le i \le M$ and $1 \le j \le N$. When $|\theta_l| \le 1/4$, 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 (16) is a special case of the spatial kernel convolution process of Equation (15), 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 5, with varying effective ranges.

Based on the aforementioned settings, we consider the bivariate case (i.e. $p = 2$), set $\theta_1 = 1/5$, $\theta_2 = -1/5$, simulate the random field at an $N \times N$ 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. Without loss of generality, suppose that the K vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K \in \mathbb{R}^p$ all have norm 1, and they are chosen as $\boldsymbol{a}_i = (\cos(\theta_i), \sin(\theta_i))^\top$, where θ_i is the coordinate direction angle in the polar coordinate system, randomly drawn from a uniform distribution in $[0, 2\pi]$. The null hypothesis, H_0, is rejected when at least one of the K univariate hypotheses based on the projection data is rejected using the BH method. The kernel function in the univariate test statistics for projection data is chosen as the Bartlett kernel defined as $k(t) = (1 - |t|)I\{|t| \leq 1\}$ with the bandwidth $b = \lfloor 4(N/100)^{2/9} \rfloor$; this selection of kernel and smoothing bandwidth is also used in Horváth *et al.* (2020), and it works well for our purpose. For comparison, we also apply several tests for MVN that do not account for the spatial dependence in the data, that is, Mardia's tests, MS and MK, defined in Equation (1), and the test of Doornik & Hansen (2008), $\mathrm{JB_{DH}}$, defined in Equation (3).

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 & Pewsey, 2009) on the data from Gaussian random fields, that is, the data used earlier for assessing the type I error. The corresponding transformation parameter

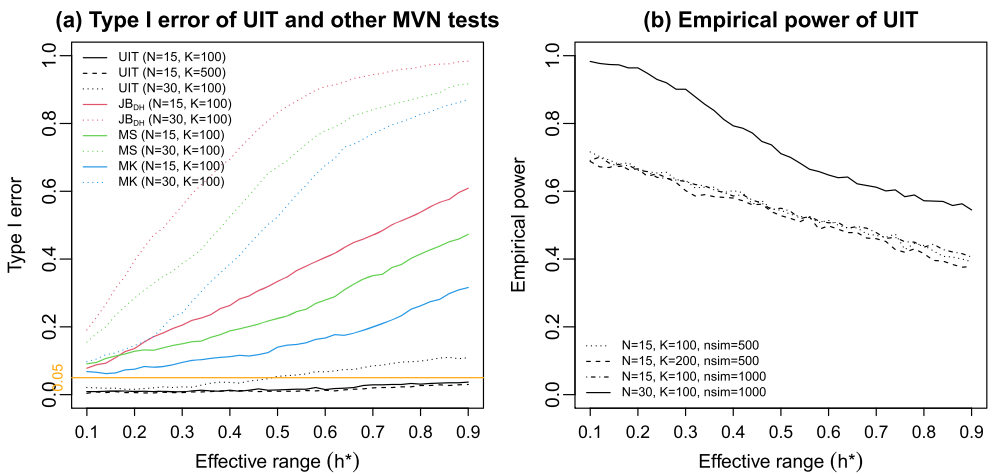

Figure 2. *(a) Type I error, as a function of the spatial dependence indicated by the effective range, of the new test (UIT, the union-intersection test) for MVN under spatial dependence (black curves) and three MVN tests for i.i.d. data (coloured curves) for* $N = 15$ *(solid curves) and* $N = 30$ *(dotted curves), based on 1,000 simulations for the nominal significance level of* $\alpha = 5\%$ *(the orange horizontal line).* $\mathrm{JB_{DH}}$ *(in red) represents the test of Doornik & Hansen (2008), and* MS *(in green) and* MK *(in blue) represent the tests of Mardia (1970). The black solid curve represents the type I error of UIT for* $N = 15$ *and* $K = 100$*, the black dashed curve represents that for* $N = 15$ *and* $K = 500$*, and the black dotted curve represents that for* $N = 30$ *and* $K = 100$*. (b) Empirical power of the new test, UIT, as a function of the spatial dependence indicated by the effective range, for the nominal significance level of* $\alpha = 5\%$ *for different values of* N*,* K *and the number of simulations denoted by 'nsim'*

setting is $(0.5, 0.5)$ for the first variable and $(0.3, 0.5)$ for the second variable, both have positive skewness and heavier tails than the normal distribution. Again, we use the same kernel function as above, and the empirical power is approximated by the proportion of null hypothesis rejection.

Figure 2(a) shows the results of type I errors of our new test (UIT, union-intersection test), compared with three MVN tests for i.i.d. data for different values of N and K based on 1,000 simulations. 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. From Figure 2(a), we can see that when $N = 15$, the type I error of our new test (the black solid curve) is bounded below and not too far from the nominal significance level of $\alpha = 5\%$ for all levels of spatial dependence, while the type I errors of the three MVN tests for i.i.d. data (the solid coloured curves) are all severely inflated and increase as the spatial dependence gets stronger. Note that the black solid curve (with $N = 15$ and $K = 100$) is very close to the black dashed curve (with $N = 15$ and $K = 500$), indicating that $K = 100$ is a large enough number of projections for the UIT test. When $N = 30$, the type I error of our new test (the black dotted curve) increases as the effective range h^* increases, and is slightly inflated when $h^* > 0.5$, that is, under strong spatial dependence; in contrast, all the three MVN tests for i.i.d. data exhibit inflated type I errors, even more severely than the case when $N = 15$ and much higher than the type I error of the UIT test. The slightly inflated type I error of the UIT test for $N = 30$ and $h^* > 0.5$ is probably caused by the strong spatial dependence in the unit square, which cannot be accurately accounted for in the asymptotic variance estimators expressed by Equations (13) and (14). The results from Figure 2(a) indicate that the MVN tests for i.i.d. data cannot be used for spatially correlated data, because they have severely inflated type I errors especially for data with strong dependence, whereas our new test can be used for spatially correlated data, and it only becomes problematic when the spatial dependence is very strong.

Figure 2(b) shows the empirical powers of our new test, UIT, for different values of N, K and number of simulations. When $N = 15$, the empirical power is not much affected by K (the number of projections) and 'nsim' (the number of simulations), because the three non-solid curves are close to each other. When $N = 30$, the empirical power (shown in black solid curve) is much higher than those in the case of smaller sample size, $N = 15$. In addition, all power curves go down as the effective range h^* increases; moreover, when $N = 30$, the power is close to one when h^* is small. The results from Figure 2 suggest that our new test would perform best in terms of type I error and empirical power when the sample size is large and the spatial dependence is not very strong. To give a more comprehensive picture for the power performance of our new testing procedure, more investigations are needed by considering a variety of alternative non-Gaussian distributions.

7 Wind Data Application

In this section, we present a data application using our new multivariate normality test for spatial data. The raw gridded data are daily U (zonal velocity) and V (meridional velocity) wind speed components during 1976–2005 over the Arabian Peninsula from the publicly available MENA CORDEX dataset (Zittis & Hadjinicolaou, 2017). We use the fourth simulated historical run with a spatial resolution of $0.22° \times 0.22°$ (latitude × longitude), which has been identified in Chen *et al.* (2018) as having the highest skill in capturing the spatio-temporal variability of reanalysis data. Following the common practice in the literature (e.g. Chen *et al.*, 2021), we apply the square root transform to the wind components, which stabilises the variance over space and

makes the marginal distributions approximately normal. Furthermore, in order to avoid modelling the complex seasonality in the data, we investigate the monthly average data over transformed U and V wind components in July during 30 years from 1976 to 2005. Finally, to make the spatial data approximately stationary, we deduct the long term averages from the monthly mean winds following Horváth *et al.* (2020), yielding monthly anomalies (residuals) of U and V components. Based on the pre-processed bivariate spatial data, we then test the bivariate normality over six small regions where local stationarity can be assumed, instead of the whole region where different topographies lead to spatial patterns and nonstationarity. The six regions (referred to as R1–R6 in Figure 3) are selected similarly to those in Chen *et al.* (2018). In each region, we have $12 \times 12 = 144$ spatial locations in each of the 30 years of bivariate wind data.

We apply our new MVN test (UIT) designed for gridded stationary spatial data as well as three MVN tests designed for i.i.d. data (i.e. JB_{DH}, MS and MK) at the nominal significance level of $\alpha = 5\%$. Table 2 shows the proportion of rejections on bivariate normality among the 30 years of July anomalies of wind U and V components over the six selected regions. We can see that in most cases, our UIT test has smaller proportions of rejections; that is, it suggests bivariate normality more often than the other three tests. In Regions 4 and 6 in particular, the UIT test does not reject normality for all 30 years, while the JB_{DH} and MS tests reject normality for almost all 30 years. Also in Regions 1 and 2, our UIT test rejects normality in only a small number of years, while the JB_{DH} and MS tests reject normality for all 30 years. These results imply that the MVN tests designed for i.i.d. data are usually too conservative when applied to spatially correlated data; that is, data from a Gaussian random field with spatial dependence tend to be detected as being non-Gaussian. The MK test rejects normality less often than the UIT test in a few cases (i.e. in Regions 1, 2 and 5); this can happen because the MK test is a directed test which only considers departure from multivariate normality revealed by kurtosis, and failure to reject the null hypothesis does not necessarily imply normality, as there might be departures from normality in other ways.

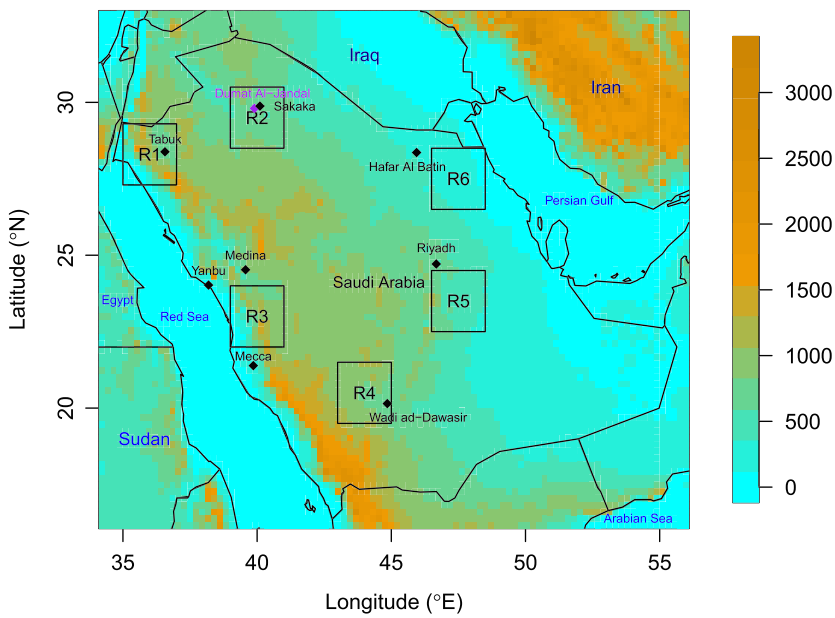

Figure 3. *Six selected regions (denoted as R1–R6) over Saudi Arabia for testing bivariate normality of spatial data. The colour shading indicates terrain elevation (in meters)*

Table 2. *Proportion of rejections of the bivariate normality hypothesis among 30 years during 1976–2005 for July anomalies of wind U and V components over the six selected regions in Saudi Arabia*

Test	*Region 1*	*Region 2*	*Region 3*	*Region 4*	*Region 5*	*Region 6*
UIT	0.3	0.4	0.23	0	0.53	0
JB_{DH}	1	1	0.83	1	1	0.97
MS	1	1	0.87	1	0.83	1
MK	0.2	0.27	0.67	0.37	0.4	0.7

8 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, based on the union-intersection test approach. The new test has a good control of the type I error, and it is inappropriate only when the spatial dependence in the data is very strong. In addition, the new test has a fairly high empirical power at all levels of spatial dependence, especially for large sample sizes.

Our new test is constructed under the stationarity assumption, which should be validated before applying our test. The test for spatial stationarity proposed by, for example, Fuentes (2005) and Jun & Genton (2012), can be used to check if some marginal spatial processes are nonstationary, and graphical tools such as contour plots can be used to identify possible nonstationary patterns in the cross-covariance functions. If the original data are detected as nonstationary, it is a common practice to transform them into stationarity, using the deformation approach proposed by, for example, Schmidt & O'Hagan (2003) and Fouedjio *et al.* (2015). One can also fit a nonstationary regression model (see, e.g. Schabenberger & Gotway, 2005) which captures most of the nonstationary features in the data, so that the residuals to be tested remain stationary.

The new test serves as a simple and useful diagnostic tool: if the null is not rejected, it lends confidence in the applications of various methodologies based on the multivariate normality assumption; if the null is rejected, it provides a caution on the validity of conclusions, and necessary pre-processing procedures may be needed before applying the methodologies, or alternative non-Gaussian methods should be considered, as illustrated in the next paragraph.

The rejection of the null hypothesis only means that the current multivariate data cannot be treated as realisations from a multivariate stationary Gaussian random field. To reveal a clearer picture of the multivariate data, the univariate normality test proposed by Horváth *et al.* (2020) can be applied to each component of the variables, and the new multivariate test for spatial data we proposed can be applied to subsets of the variables, to check if some marginal processes are not normal. If that is the case, then we may need data transformations (such as the log, power or square root transformation) in order to approximately have a Gaussian process. Of course, all marginals being normal does not mean being jointly normal, so these marginal transformations may only partly help; in this case, we should be aware of the effect of conducting the current statistical procedures under the violated Gaussian assumption, and consider to switch to non-Gaussian methods (e.g. Xu & Genton, 2017 and Yan *et al.*, 2020).

Also note that when the sample size is large, the estimation of the auto-covariance function, that is, $\widehat{C}(\boldsymbol{h})$, in Equations (13) and (14) can be computationally prohibitive. One solution is that

we can fit a parametric covariance model (such as the Matérn model) for $C(\boldsymbol{h})$, and obtain $\widehat{C}(\boldsymbol{h})$ by using the software `ExaGeoStat` (Abdulah *et al.*, 2018), which allows for exact maximum likelihood estimation with dense full covariance matrices, using high performance computations. In addition, various approximation methods for large spatial datasets have also been proposed to reduce the computational burden; recent reviews include Sun *et al.* (2012), Heaton *et al.* (2019) and Huang *et al.* (2021).

One limitation of the new test, similarly to the univariate Horváth *et al.* (2020) test, is that it can only be used for 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. Nevertheless, 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, a 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 future work.

ACKNOWLEDGEMENTS

This research was supported by the National Key Research and Development Program of China (2021YFA1000101), Zhejiang Provincial Natural Science Foundation of China (LZJWY22E090009), Natural Science Foundation of Shanghai (22ZR1420500), and the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU and King Abdullah University of Science and Technology (KAUST).

References

Abdulah, S., Ltaief, H., Sun, Y., Genton, M.G. & Keyes, D.E. (2018). Exageostat: A high performance unified software for geostatistics on manycore systems. *IEEE Trans. Parallel Distrib. Syst.*, **29**(12), 2771–2784.

Ahmad, F. & Khan, R.A. (2015). A power comparison of various normality tests. *Pakistan J. Stat. Operat. Res.*, **11**(3), 331–345.

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

Baringhaus, L. & 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. & Zografos, K. (2013). A necessary power divergence type family tests of multivariate normality. *Commun. Stat.-Simul. Comput.*, **42**(10), 2253–2271.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodological)*, **57**(1), 289–300.

Bera, A. & John, S. (1983). Tests for multivariate normality with Pearson alternatives. *Commun. Stat.-Theory Methods*, **12**(1), 103–117.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A. & 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*. Oxford University Press: USA, pp. 181.

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. John Wiley and Sons: New York.
Bowman, K. & 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. & Ferreira, D. (2010). Multivariate extension of chi-squared univariate normality test. *J. Stat. Comput. Simul.*, **80**(5), 513–526.
Chen, W., Castruccio, S., Genton, M.G. & Crippa, P. (2018). Current and future estimates of wind energy potential over Saudi Arabia. *J. Geophys. Res.: Atmosph.*, **123**(12), 6443–6459.
Chen, W., Genton, M.G. & Sun, Y. (2021). Space-time covariance structures and models. *Ann. Rev. Stat. Appl.*, **8**, 191–215.
Cressie, N. & Read, T.R. (1984). Multinomial goodness-of-fit tests. *J. R. Stat. Soc.: Ser. B (Methodolog.)*, **46**(3), 440–464.
Csörgő, S. (1989). Consistency of some tests for multivariate normality. *Metrika*, **36**(1), 107–116.
D'Agostino, R.B. & Lee, A.F. (1977). Robustness of location estimators under changes of population kurtosis. *J. Am. Stat. Assoc.*, **72**(358), 393–396.
D'Agostino, R.B. & Stephens, M.A. (1986). *Goodness-Of-Fit Techniques*, Volume 68 of Statistics: A Series of Textbooks and Monographs. Marcel Dekker: New York.
Doornik, J.A. & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bull. Econ. Stat.*, **70**, 927–939.
Dutta, S. & Genton, M.G. (2014). A non-Gaussian multivariate distribution with all lower-dimensional Gaussians and related families. *J. Multivariate Anal.*, **132**, 82–93.
Eaton, M.L. & Perlman, M.D. (1973). The non-singularity of generalized sample covariance matrices. *Ann. Stat.*, **1**(4), 710–717.
Ebner, B. & Henze, N. (2020). Tests for multivariate normality–a critical review with emphasis on weighted L^2-statistics. *TEST*, **29**, 845–892.
Enomoto, R., Okamoto, N. & Seo, T. (2012). Multivariate normality test using Srivastava's skewness and kurtosis. *SUT J. Math.*, **48**(1), 103–115.
Epps, T. (1999). Limiting behavior of the ICF test for normality under Gram–Charlier alternatives. *Stat. Probab. Lett.*, **42**(2), 175–184.
Epps, T.W. & Pulley, L.B. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, **70**(3), 723–726.
Farrell, P.J. & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. *J. Stat. Comput. Simul.*, **76**(9), 803–816.
Farrell, P.J., Salibian-Barrera, M. & Naczk, K. (2007). On tests for multivariate normality and associated simulation studies. *J. Stat. Comput. Simul.*, **77**(12), 1065–1080.
Flury, B. (2013). *A first course in multivariate statistics*. Springer Science & Business Media.
Fouedjio, F., Desassis, N. & Romary, T. (2015). Estimation of space deformation model for non-stationary random functions. *Spatial Stat.*, **13**, 45–61.
Fuentes, M. (2005). A formal test for nonstationarity of spatial stochastic processes. *J. Multivariate Anal.*, **96**(1), 30–54.
Geary, R.C. (1947). Testing for normality. *Biometrika*, **34**(3/4), 209–242.
Gelfand, A.E. & Banerjee, S. (2010). Multivariate spatial process models. In *Handbook of Spatial Statistics*, Eds. Gelfand, A.E., Diggle, P.J., Fuentes, M. & Guttorp, P., Chapman and Hall–CRC: Boca Raton, pp. 495–515.
Gelfand, A.E. & Schliep, E.M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Stat.*, **18**, 86–104.
Genton, M.G. & Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics (with discussion). *Stat. Sci.*, **30**(2), 147–163.
Gnanadesikan, R. & Kettenring, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, **28**(1), 81–124.
Gneiting, T., Kleiber, W. & Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Am. Stat. Assoc.*, **105**(491), 1167–1177.
Guhaniyogi, R. & 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. *Trans. Inst. British Geograph.*, **3**(2), 202–225.
Hanusz, Z., Enomoto, R., Seo, T. & Koizumi, K. (2018). A Monte Carlo comparison of Jarque–Bera type tests and Henze–Zirkler test of multivariate normality. *Commun. Stat.-Simul. Comput.*, **47**(5), 1439–1452.
Hanusz, Z. & Tarasińska, J. (2012). New tests for multivariate normality based on Small's and Srivastava's graphical methods. *J. Stat. Comput. Simul.*, **82**(12), 1743–1752.

Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D., Sun, F. & Zammit-Mangion, A. (2019). A case study competition among methods for analyzing large spatial data. *J. Agricult. Biolog. Environm. Stat.*, **24**(3), 398–425.
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. *Stat. Probab. Lett.*, **35**(3), 203–213.
Henze, N. (2002). Invariant tests for multivariate normality: A critical review. *Stat. Papers*, **43**(4), 467–506.
Henze, N. & 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. & Meintanis, S.G. (2019). Characterizations of multinormality and corresponding tests of fit, including for GARCH models. *Econometric Theory*, **35**(3), 510–546.
Henze, N. & Wagner, T. (1997). A new approach to the BHEP tests for multivariate normality. *J. Multivar. Anal.*, **62**(1), 1–23.
Henze, N. & Zirkler, B. (1990). A class of invariant and consistent tests for multivariate normality. *Commun. Stat.-Theory Methods*, **19**(10), 3595–3617.
Horswell, R.L. & Looney, S.W. (1992). A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. *J. Stat. Comput. Simul.*, **42**(1-2), 21–38.
Horváth, L., Kokoszka, P. & Wang, S. (2020). Testing normality of data on a multivariate grid. *J. Multivar. Anal.*, **179**, 104640.
Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Stat.*, **2**(3), 360–378.
Huang, H., Abdulah, S., Sun, Y., Ltaief, H., Keyes, D.E. & Genton, M.G. (2021). Competition on spatial statistics for large datasets. *J. Agricult. Biolog. Environm. Stat.*, **26**(4), 580–595.
Irvine, K.M., Gitelman, A.I. & Hoeting, J.A. (2007). Spatial designs and properties of spatial correlation: Effects on covariance estimation. *J. Agricult. Biolog. Environm. Stat.*, **12**(4), 450–469.
Islam, T.U. (2017). Stringency-based ranking of normality tests. *Commun. Stat.–Simul. Comput.*, **46**(1), 655–668.
Isogai, T. (1982). On a measure of multivariate skewness and a test for multivariate normality. *Ann. Inst. Stat. Math.*, **34**(1), 531–541.
Jarque, C.M. & Bera, A.K. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Econ. Lett.*, **7**(4), 313–318.
Joenssen, D.W. & Vogel, J. (2014). A power study of goodness-of-fit tests for multivariate normality implemented in R. *J. Stat. Comput. Simul.*, **84**(5), 1055–1078.
Jones, M.C. & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, **96**(4), 761–780.
Jönsson, K. (2011). A robust test for multivariate normality. *Econ. Lett.*, **113**(2), 199–201.
Jun, M. & Genton, M.G. (2012). A test for stationarity of spatio-temporal random fields on planar and spherical domains. *Stat. Sin.*, **22**, 1737–1764.
Kankainen, A., Taskinen, S. & Oja, H. (2007). Tests of multinormality based on location vectors and scatter matrices. *Stat. Methods Appl.*, **16**(3), 357–379.
Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *J. Am. Stat. Assoc.*, **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. *J. Appl. Sci. Res.*, **2**(5), 296–300.
Kim, N. (2016). A robustified Jarque–Bera test for multivariate normality. *Econ. Lett.*, **140**, 48–52.
Kim, I. & Park, S. (2018). Likelihood ratio tests for multivariate normality. *Commun. Stat.-Theory Methods*, **47**(8), 1923–1934.
Koizumi, K., Hyodo, M. & Pavlenko, T. (2014). Modified Jarque–Bera type tests for multivariate normality in a high-dimensional framework. *J. Stat. Theory Pract.*, **8**(2), 382–399.
Koizumi, K., Okamoto, N. & Seo, T. (2009). On Jarque-Bera tests for assessing multivariate normality. *J. Stat.: Adv. Theory Appl.*, **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.
Loève, M. (1977). *Probability Theory*, 4, Vol. **1**. Springer: New York.
Looney, S.W. (1995). How to use tests for univariate normality to assess multivariate normality. *Am. Statistician*, **49**(1), 64–70.
Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer: New York.
Madukaife, M.S. & Okafor, F.C. (2019). A new large sample goodness of fit test for multivariate normality based on chi-squared probability plots. *Commun. Stat.-Simul. Comput.*, **48**(6), 1651–1664.

Majerski, P. & Szkutnik, Z. (2010). Approximations to most powerful invariant tests for multinormality against some irregular alternatives. *TEST*, **19**(1), 113–130.
Malkovich, J.F. & Afifi, A. (1973). On tests for multivariate normality. *J. Am. Stat. Assoc.*, **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: Indian J. Stat. Ser. B*, **36**, 115–128.
Mardia, K.V. (1980). 9 tests of unvariate and multivariate normality. In *Handbook of Statistics*, Ed. Krishnaiah, P.R., Vol. **1**, Elsevier: New York, pp. 279–320.
Mardia, K.V. & Foster, K. (1983). Omnibus tests of multinormality based on skewness and kurtosis. *Commun. Stat.-Theory Methods*, **12**(2), 207–221.
Mardia, K.V. & Kent, J. (1991). Rao score tests for goodness of fit and independence. *Biometrika*, **78**(2), 355–363.
Mason, R.L. & Young, J.C. (1985). Re-examining two tests for bivariate normality. *Commun. Stat.-Theory Methods*, **14**(7), 1531–1546.
Mecklin, C.J. & Mundfrom, D.J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *J. Stat. Comput. Simul.*, **75**(2), 93–107.
Moore, D.S. (1986). Tests of chi-squared type. In *Goodness-of-Fit Techniques*, Eds. D'Agostino, R.B. & Stephens, M. A., Marcel Dekker: New York, pp. 63–96.
Moore, D.S. & Stubblebine, J.B. (1981). Chi-square tests for multivariate normality with application to common stock prices. *Commun. Stat.-Theory Methods*, **10**(8), 713–738.
Móri, T.F., Rohatgi, V.K. & Székely, G. (1994). On multivariate skewness and kurtosis. *Theory Probab. Its Appl.*, **38**(3), 547–551.
Noughabi, H.A. & Arghami, N.R. (2011). Monte carlo comparison of seven normality tests. *J. Stat. Comput. Simul.*, **81**(8), 965–972.
Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F. & Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comput. Graph. Stat.*, **24**(2), 579–599.
Olea, R.A. & Pawlowsky-Glahn, V. (2009). Kolmogorov–Smirnov test for spatially correlated data. *Stochast. Environment. Res. Risk Assess.*, **23**(6), 749–757.
Öztuna, D., Elhan, A.H. & Tüccar, E. (2006). Investigation of four different normality tests in terms of type I error rate and power under different distributions. *Turkish J. Med. Sci.*, **36**(3), 171–176.
Paciorek, C.J., Lipshitz, B., Zhuo, W., Kaufman, C.G. & Thomas, R.C. (2015). Parallelizing Gaussian process calculations in R. *J. Stat. Softw.*, **63**(i10).
Pardo-Igúzquiza, E. & Dowd, P.A. (2004). Normality tests for spatially correlated data. *Math. Geology*, **36**(6), 659–681.
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. *London Edinburgh Dublin Philosoph. Mag. J. Sci.*, **50**(302), 157–175.
Pearson, E.S., D'Agostino, R.B. & Bowman, K.O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, **64**(2), 231–246.
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. & Vávra, M. (2020). Normality tests for dependent data: Large-sample and bootstrap approaches. *Commun. Stat. Simul. Comput.*, **49**(2), 283–304.
Pudełko, J (2005). On a new affine invariant and consistent test for multivariate normality. *Probab. Math. Stat.*, **25**, 43–54.
Read, T.R. & Cressie, N.A. (2012). *Goodness-Of-Fit Statistics for Discrete Multivariate Data*. Springer: New York.
Romao, X., Delgado, R. & Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *J. Stat. Comput. Simul.*, **80**(5), 545–591.
Romeu, J.L. & Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality. *J. Multivar. Anal.*, **46**(2), 309–334.
Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*. Wiley: New York.
Sánchez-Espigares, J.A., Grima, P. & Marco-Almagro, L. (2019). Graphical comparison of normality tests for unimodal distribution data. *J. Stat. Comput. Simul.*, **89**(1), 145–154.
Schabenberger, O. & Gotway, C.A. (2005). *Statistical methods for spatial data analysis: Texts in statistical science*. Chapman and Hall/CRC.
Schmidt, A.M. & O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)*, **65**(3), 743–758.
Seo, T. & Ariga, M. (2011). On the distribution of sample measure of multivariate kurtosis. *J. Combinatorics Inform. Syst. Sci.*, **36**(1-4), 179.

Shapiro, S.S., Wilk, M.B. & Chen, H.J. (1968). A comparative study of various tests for normality. *J. Am. Stati. Assoc.*, **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. *Stat. Probab. Lett.*, **2**(5), 263–267.
Subrahmaniam, K., Subrahmaniam, K. & Messeri, J. (1975). On the robustness of some tests of significance in sampling from a compound normal population. *J. Am. Stat. Assoc.*, **70**(350), 435–438.
Sun, Y. & Genton, M.G. (2011). Functional boxplots. *J. Comput. Graph. Stat.*, **20**(2), 316–334.
Sun, Y., Li, B. & Genton, M.G. (2012). Geostatistics for large datasets. In *Advances and Challenges in Space-Time Modelling of Natural Events*, Vol. **207**, Springer, pp. 55–77.
Székely, G.J. & Rizzo, M.L. (2005). A new test for multivariate normality. *J. Multivar. Anal.*, **93**(1), 58–80.
Tenreiro, C. (2011). An affine invariant multiple test procedure for assessing multivariate normality. *Comput. Stat. Data Anal.*, **55**(5), 1980–1992.
Tenreiro, C. (2017). A new test for multivariate normality by combining extreme and nonextreme BHEP tests. *Commun. Stat. Simul. Comput.*, **46**(3), 1746–1759.
Thadewald, T. & Büning, H. (2007). Jarque–Bera test and its competitors for testing normality–a power comparison. *J. Appl. Stat.*, **34**(1), 87–105.
Thode, H.C. (2002). *Testing for normality*, Vol. **164**. Chapman and Hall/CRC.
Thulin, M. (2014). Tests for multivariate normality based on canonical correlations. *Stat. Methods Appl.*, **23**(2), 189–208.
Tukey, J.W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, Eds. Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G. & Mann, H.B., Stanford University Press: CA, pp. 448–485.
Villasenor, A., José, A. & Estrada, E.G. (2009). A generalization of Shapiro–Wilk's test for multivariate normality. *Commun. Stat.–Theory Methods*, **38**(11), 1870–1883.
Voinov, V., Pya, N., Makarov, R. & Voinov, Y. (2016). New invariant and consistent chi-squared type goodness-of-fit tests for multivariate normality and a related comparative simulation study. *Commun. Stat.-Theory Methods*, **45**(11), 3249–3263.
Volkmer, H. (2014). A characterization of the normal distribution. *J. Stat. Theory Appl.*, **13**(1), 83–85.
Wilson, E.B. & Hilferty, M.M. (1931). The distribution of chi-square. *Proc. Nat. Acad. Sci. United States Am.*, **17**(12), 684.
Xu, G. & Genton, M.G. (2017). Tukey g-and-h random fields. *J. Am. Stat. Assoc.*, **112**(519), 1236–1249.
Yamada, T., Romer, M.M. & Richards, D.S.P. (2015). Kurtosis tests for multivariate normality with monotone incomplete data. *TEST*, **24**(3), 532–557.
Yan, Y., Jeong, J. & Genton, M.G. (2020). Multivariate transformed Gaussian processes. *Japanese J. Stat. Data Sci.*, **3**(1), 129–152.
Yap, B.W. & Sim, C.H. (2011). Comparisons of various types of normality tests. *J. Stat. Comput. Simul.*, **81**(12), 2141–2155.
Yazici, B. & Yolacan, S. (2007). A comparison of various tests of normality. *J. Stat. Comput. Simul.*, **77**(2), 175–183.
Zheng, W. (2019). Kolmogorov–Smirnov type tests under spatial correlations. Ph.D. thesis, UT School of Public Health Dissertations. (Open Access).
Zhou, M. & Shao, Y. (2014). A powerful test for multivariate normality. *J. Appl. Stat.*, **41**(2), 351–363.
Zimmerman, D. & Stein, M. (2010). Classical geostatistical methods. In *Handbook of Spatial Statistics*, Eds. Gelfand, A.E., Diggle, P.J., Fuentes, M. & Guttorp, P., Chapman and Hall–CRC: Boca Raton, pp. 29–44.
Zittis, G. & Hadjinicolaou, P. (2017). The effect of radiation parameterization schemes on surface temperature in regional climate simulations over the MENA-CORDEX domain. *Int. J. Climatology*, **37**(10), 3847–3862.

[*Received September 2021; accepted May 2022*]

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.