International Statistical Review

International Statistical Review (2020), 88, 3, 677–697 doi:10.1111/insr.12362

Tests of Normality of Functional Data

Tomasz Górecki[1], Lajos Horváth[2] and Piotr Kokoszka[3]

[1]*Faculty of Mathematics and Computer Science, Adam Mickiewicz University, 61-614 Poznań, Poland*

[2]*Department of Mathematics, University of Utah, Salt Lake City, UT 84112-0090, USA*

[3]*Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA*
E-mail: Piotr.Kokoszka@colostate.edu

Summary

The paper is concerned with testing normality in samples of curves and error curves estimated from functional regression models. We propose a general paradigm based on the application of multivariate normality tests to vectors of functional principal components scores. We examine finite sample performance of a number of such tests and select the best performing tests. We apply them to several extensively used functional data sets and determine which can be treated as normal, possibly after a suitable transformation. We also offer practical guidance on software implementations of all tests we study and develop large sample justification for tests based on sample skewness and kurtosis of functional principal component scores.

Key words: functional data; normal distribution; significance tests.

1 Introduction

The objective of this paper is to propose and evaluate in finite samples several tests of normality of functional data. The assumption of the normality of observations or model errors has motivated much of the development of statistics since the origins of the field. Consequently, many tests of normality have been proposed. Perhaps the best known is the Shapiro and Wilk test, Shapiro & Wilk (1965), which has been extended and improved in many directions (Royston, 1982, 1983, 1992). Tests based on the empirical distribution function have also been extensively used (Anderson & Darling, 1954; Stephens, 1974; Scholz & Stephens, 1997). A great number of other approaches have been proposed: Mardia (1970, 1974), D'Agostino *et al.* (1990), Henze & Zirkler (1990), Doornik & Hansen (2008), among many others. The Jarque–Bera test (Jarque & Bera, 1980; 1987) is commonly used to check if model errors are normal. Testing the normality of models residuals have received considerable attention (see, e.g. Pierce & Gray, 1982; Duchesne *et al*., 2016 and references therein.

Normality of functional error curves is assumed quite often (e.g. Crainiceanu *et al*., 2009; Gromenko *et al*., 2017). Some procedures are derived using normal likelihoods (e.g. Constantinou *et al*., 2017; Hörmann *et al*., 2018). In many settings, formulas become much simpler if normality is assumed (e.g. Panaretos *et al*., 2010; Kraus & Panaretos, 2012; Fremdt *et al*., 2013; Aston et al., 2017). A test that verifies that the assumption of normality is reasonable for a sample of curves, either observed or computed as model residuals, will thus bolster confidence in the conclusions of these and many other functional data analysis procedures. Normality is at the core of curve reconstruction techniques introduced in Yao, Müller and Wang (2005a,

2005b) and recently extended to spatially correlated data by Liu *et al.* (2017). A test of normality applied to reconstructed curves would confirm the applicability of these approaches to any specific data set.

We are aware of two approaches to testing normality of functional data: (a) the graphical tools of Chiou *et al.* (2004) [various scatter plots of functional principal components (FPCs) scores] and (b) The Jarque–Bera type tests of Górecki *et al.* (2018). Tools based on scatter plots and QQ plots should form the main ingredient of exploratory analysis to be followed by inferential analysis. Our objective is to propose a broad approach and determine which specific tests perform best. The work of Górecki *et al.* (2018) focused on extensions of the Jarque–Bera test that are robust to temporal dependence. We will see that in the context of independently and identically distributed (iid) curves or regression residuals, other approaches are superior.

A well-known challenge of working with functional data is that to perform computations, data must be reduced to finite-dimensional objects. This is typically done by projecting on deterministic or random systems, with the latter often offering a more effective dimension reduction. In case of a projection on a system estimated from the data, the effects of the estimation and truncation must be taken into account. We explore this issue in the context of testing the normality of functional data.

The contribution of the paper can be summarised as follows:

(a) We review many multivariate normality test, and explain how they can be used to tests normality of functional data.
(b) We provide information on which tests perform best in finite samples in several commonly encountered scenarios.
(c) We provide precise information on the implementation of these tests in `R`.
(d) We develop large sample theory that justifies the application to functional data of multivariate tests based on sample skewness and kurtosis.
(e) We examine many extensively studied functional data sets and inform the reader if they can be considered as coming from a normal distribution in a functions space.

The remainder of the paper is organised as follows. In Section 2, we review multivariate normality tests on which functional normality tests can be based. In Section 3, we explain our testing approach. Section 4 contains the results of our simulation study and our recommendations. In Section 5, we examine many well–known data sets used in functional data analysis research. We specify which of them can be treated as coming from a normal distribution in a function space and, in some cases, suggest suitable transformations to normality. Large sample theory is developed in Section 6.

2 Multivariate Normality Tests

In this section, we describe the multivariate normality tests whose extensions to functional data we study. We do not aim at studying all available tests but rather those that are conveniently implemented in `R`. A comprehensive review is given by Mecklin and Mundfrom (2004).

We denote by p the dimension of the data vectors $\mathbf{X}_i$. We observe a random sample of N such vectors. The null distributions used to obtain critical values are generally finite sample approximations.

There are several equivalent definitions of a normal random vector of dimension p. Normality can be defined in terms of the joint density function, the characteristic function or univariate projections. The formula for the multivariate normal density, the characteristic function and the moment generating function can be found in most statistics textbooks (see, e.g chapter 2 of Seber & Lee, 2003). Less well–known is the fact that a random vector

$\mathbf{X} = [X_1, X_2, \ldots, X_p]^\top$ is multivariate normal if and only if each linear combination, or projection, $\langle \mathbf{a}, \mathbf{X} \rangle = \sum_{i=1}^{p} a_i X_i$ is univariate normal (see, e.g. theorem 2.3 in Seber & Lee, 2003). Using this fact, we see that multivariate normality of $\mathbf{X}$ is equivalent to the multivariate normality of any vector of the form $[\langle \mathbf{a}_1, \mathbf{X} \rangle, \ldots, \langle \mathbf{a}_q, \mathbf{X} \rangle]^\top$. Some tests of multivariate normality are based on choosing suitable vectors $\mathbf{a}_1, \ldots, \mathbf{a}_q$ and generally taking $q = p$. This idea can be extended, with suitable modifications, to the setting of functional data, as explained at the beginning of Section 3.

2.1 Jarque–Bera Test

In addition to normal QQ plots, the Jarque–Bera test is perhaps the most often used normality test because of its simplicity and the fact that its P values are automatically reported in many software packages. Denoted by τ_j, the sample skewness of the j-th component of the vectors $\mathbf{X}_i$, i.e. of $X_{1j}, X_{2j}, \ldots, X_{Nj}$, and by κ_j, their sample kurtosis. Then, assuming the $\mathbf{X}_i$ are independent and normal,

$$\mathrm{JB}_{j,N} = N \left(\frac{\tau_j^2}{6} + \frac{(\kappa_j - 3)^2}{24} \right) \sim \chi^2_2, \tag{2.1}$$

where $\sim$ indicates an approximate equality in distribution. If the collections $X_{1j}, X_{2j}, \ldots, X_{Nj}$ and $X_{1k}, X_{2k}, \ldots, X_{Nk}$ are independent for $k \neq j$, then

$$\mathrm{JB}_N^{(p)} = \sum_{j=1}^{p} \mathrm{JB}_{j,N} \sim \chi^2_{2p}. \tag{2.2}$$

We will explain in Section 3 how Test (2.2) can be justified in the context of functional data. The approximate independence of the components holds by choosing the right projections. However, the independence is only approximate and even the approximation in (2.1) is known to be far from perfect. In the following sections, we therefore describe alternative approaches.

2.2 Mardia Tests

Assuming $N > p$, Mardia (1970) developed multivariate extensions of skewness and kurtosis. The sample statistic for multivariate skewness is

$$\hat{\tau}_{1,p} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} m_{ij}^3,$$

and for multivariate kurtosis is

$$\hat{\kappa}_{2,p} = \frac{1}{N} \sum_{i=1}^{N} m_{ij}^2,$$

where $m_{ij} = (\mathbf{X}_i - \bar{\mathbf{X}})^\top \mathbf{S}^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})$, i.e the squared Mahalanobis distance. The test statistic for skewness,

$$Z_{1,p} = \frac{N}{6} \hat{\tau}_{1,p},$$

is approximately χ^2 distributed with $p(p+1)(p+2)/6$ degrees of freedom. The normality assumption is rejected for large skewness values. The test statistic for kurtosis,

$$Z_{2,p} = \sqrt{\frac{N}{8p(p+2)}}[\hat{\kappa}_{2,p} - p(p+2)],$$

is approximately $N(0,1)$. The normality assumption is rejected for large or small kurtosis values.

Mardia (1974) proposed finite sample adjustments and introduced the following test statistics as improvements on $Z_{1,p}$ and $Z_{2,p}$:

$$Z^*_{1,p} = Z_{1,p}\frac{(p+1)(N+1)(N+3)}{N[(N+1)(p+1)-6]} \sim \chi^2_{p(p+1)(p+2)/6},$$

and

$$Z^*_{2,p} = \frac{\sqrt{(N+3)(N+5)}[(N+1)\hat{\kappa}_{2,p} - p(p+2)(N-1)]}{\sqrt{8p(p+2)(N-3)(N-p-1)(N-p+1)}} \sim N(0,1).$$

Koizumi *et al.* (2009) defined new test statistics using Mardia's measures as follows:

$$\text{MJB}_{\boldsymbol{M}} = Z_{1,p} + Z^2_{2,p}, \quad \text{MJB}^*_{\boldsymbol{M}} = Z^*_{1,p} + Z^{*2}_{2,p}. \tag{2.3}$$

Both are approximately distributed as χ^2 with $p(p+1)(p+2)/6+1$ degrees of freedom.

2.3 Royston Test

Royston (1983) extended the Shapiro–Wilk test to the multivariate case. First, we calculate the univariate Shapiro–Wilk statistic W_j for each variable. Then each W_j is normalised via the following transformation:

$$Z_j = \frac{(1-W_j)^\lambda - \mu}{\sigma},$$

where the constants λ, μ and σ for a given sample size are given in a table in Royston (1982). Next we define the statistic

$$R_j = \left[\Phi^{-1}\left(\frac{\Phi(-Z_j)}{2}\right)\right]^2.$$

Finally, we use the test statistic

$$H = \frac{e\sum_{j=1}^{p} R_j}{p},$$

where $1 \le e \le p$ is a constant equal to

$$e = \frac{p}{1+(p-1)\bar{c}},$$

where $\bar{c}$ is an estimate of the average correlation among the R_j's. The statistic H has an approximate χ^2_e distribution, and multivariate normality is rejected if it exceeds the appropriate critical value. Unfortunately, the statistic H is unable to achieve the nominal significance level. Royston (1992) revised the procedure by introducing a finite sample adjustment to Shapiro–Wilk statistic W_j.

2.4 Henze–Zirkler Test

Henze and Zirkler (1990) test is based on a non-negative functional distance that measures the distance between two distribution functions:

$$D_\beta(P, Q) = \int |\hat{P}(t) - \hat{Q}(t)|^2 \varphi_\beta(t) dt,$$

where $\hat{P}(t)$ is the characteristic function of the multivariate standard normal and $\hat{Q}(t)$ is the empirical characteristic function of the standardised observations. The function also involves a kernel (weighting) function $\varphi_\beta(t)$, where $\beta \in \mathbb{R}$ is a smoothing parameter that needs to be selected. Extensive derivations yielded a closed form for the Henze–Zirkler statistic $T_\beta(p)$. If data are distributed as multivariate normal, the test statistic is approximately log-normally distributed. The kernel function must be chosen. Henze and Zirkel chose the density of $N_p(\mathbf{0}, \beta^2 \mathbf{I}_p)$ as a kernel function. They also provide a formula for determining an optimal choice of β for each N and p (the Henze–Zirkler statistic seems to behave similarly to Mardia's skewness as $\beta \to 0$ and to Mardia's kurtosis as $\beta \to \infty$). An appealing property of this test is that it is a consistent and invariant under linear transformations of the data.

2.5 Energy Test

Székeley and Rizzo (2005) proposed the energy test. It begins by standardising the multivariate observations to $\mathbf{Z}_i = \mathbf{S}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$. The test compares the Euclidean distances between the $\mathbf{Z}_i$ and randomly sampled, uncorrelated multivariate normal random variables $\mathbf{Y}_i$. The energy test can be written as

$$E = 2A - B - C.$$

Here, B is the average Euclidean distance between the $\mathbf{Z}_i$, and C is the average distance between observations made on a simulated multivariate random sample $\mathbf{Y}_i$ with zero means and identity covariance matrix. The A term is the average distance between the $\mathbf{Y}$'s and the $\mathbf{Z}$'s. Because the statistic follows no known distribution under the null hypothesis, the critical values are determined via simulation.

2.6 Doornik–Hansen and Lobato–Velasco Tests

Doornik and Hansen (2008) proposed a simple multivariate normality test based on transformed skewness and kurtosis. The test requires $N > 7 > p$. Wickham (2015) modified this test by including a correction proposed by Lobato and Velasco (2004), which makes the test robust to temporal dependence. To differentiate the modified test from the Doornik and Hansen test, we refer to it as the Lobato–Velasco test. In both cases, the approximate null distribution is χ^2_{2p}. Because these tests perform quite well in our simulations, we provide more details.

The data are vectors $\mathbf{X}_i = [X_{i1}, X_{i2}, \ldots, X_{ip}]^\top$, $1 \leq i \leq N$. Their sample mean and covariance matrix are

$$\bar{\mathbf{X}} = N^{-1} \sum_{i=1}^{N} \mathbf{X}_i, \quad \mathbf{S} = N^{-1} \sum_{i=1}^{N} [\mathbf{X}_i - \bar{\mathbf{X}}][\mathbf{X}_i - \bar{\mathbf{X}}]^\top.$$

Denote by $\mathbf{V}$ the diagonal matrix with the diagonal elements of $\mathbf{S}$ and define the correlation matrix by $\mathbf{C} = \mathbf{V}^{-1/2}\mathbf{S}\mathbf{V}^{-1/2}$. The matrix $\mathbf{C}$ is symmetric and positive definite, so it has positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Set $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p) = \mathbf{U}^\top \mathbf{C} \mathbf{U}$, where

$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$ is an orthonormal matrix and $\mathbf{C}\mathbf{u}_j = \lambda_j \mathbf{u}_j$. Define the orthogonalised vectors by

$$\mathbf{Y}_i = \mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^\top \mathbf{V}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}}).$$

Define the sample skewness and kurtosis of their components by

$$\tau_{Y,j} = \frac{m_{Y,3,j}}{m_{Y,2,j}^{3/2}}, \quad \kappa_{Y,j} = \frac{m_{Y,4,j}}{m_{Y,2,j}^{2}}, \quad m_{Y,k,j} = N^{-1}\sum_{i=1}^{N}(Y_{ij} - \bar{Y}_j)^k. \tag{2.4}$$

Using the Lobato–Velasco correction consists in modifying the denominators of $\tau_{Y,j}$ and $\kappa_{Y,j}$. The modified skewness and kurtosis are given by

$$\tau^{\star}_{Y,j} = \frac{m_{Y,3,j}}{v_{Y,3,j}^{1/2}}, \quad \kappa^{\star}_{Y,j} = \frac{m_{Y,4,j}}{v_{Y,4,j}^{1/2}}, \tag{2.5}$$

where

$$v_{Y,3,j} = \hat{\gamma}^3_{0,j} + 2\sum_{h=1}^{N-1}\hat{\gamma}^3_{h,j} \quad v_{Y,4,j} = \hat{\gamma}^4_{0,j} + 2\sum_{h=1}^{N-1}\hat{\gamma}^4_{h,j},$$

and where $\hat{\gamma}_{h,j}$ is the lag-h sample autocovariance of the $Y_{ij}, 1 \leq i \leq N$.

Next, using Formulas (2) and (3) in Doornik and Hansen (2008), we compute

$$z_{1j} = f_{1,N}(\tau_{Y,j}), \quad z_{2j} = f_{2,N}(\kappa_{Y,j}, \tau_{Y,j}).$$

For ease of reference, the functions $f_{1,N}$ and $f_{2,N}$ are defined in Section 2 of the supporting information. The test statistic is given by

$$E_d = \sum_{j=1}^{p}\left(z_{1j}^2 + z_{2j}^2\right).$$

3 Application of the Tests to Functional Data

We consider functions on a compact set $\mathcal{T} \subset \mathbb{R}$. If X is a random function, we say that it is square integrable if $E\int_{\mathcal{T}} X^2(t)dt < \infty$. Square integrable random functions are almost surely elements of the Hilbert space $L^2(\mathcal{T})$. Just like for multivariate random vectors, there are several equivalent definitions of a normal random element X in a Hilbert space (see, e.g. section 11.3 of Kokoszka & Reimherr, 2017). For any deterministic function $a \in L^2(\mathcal{T})$, we define the projection $\langle X, a\rangle = \int_{\mathcal{T}} X(t)a(t)dt$, which is a random variable. For testing the normality of X, the essential property is that X is a normal random function if and only if all vectors of the form (for any p)

$$[\langle X, a_1\rangle, \langle X, a_2\rangle, \ldots, \langle X, a_p\rangle]^T,$$

are multivariate normal. This property can be taken as the definition of the normality of X. Every square integrable random function, in particular every normal random function, admits the Karhunen–Loéve expansion.

$$X(t) = \mu(t) + \sum_{j=1}^{\infty}\sqrt{\lambda_j}Z_j v_j(t), \quad t \in \mathcal{T}. \tag{3.1}$$

The orthonormal functions v_j are called FPCs, and $\langle X, v_j \rangle = \sqrt{\lambda_j} Z_j$ their scores. The Z_j are uncorrelated, have mean zero, and unit variance. The λ_j are given by

$$\lambda_j = E\left\{\int_{\mathcal{T}} [X(t) - \mu(t)] v_j(t) dt\right\}^2.$$

We test the null hypothesis

$$H_0 : X \text{ has a normal distribution in } L^2.$$

Using (3.1), it is easy to verify that H_0 is equivalent to the assertion that the Z_j are independent standard normal. For this reason, one can expect that the most efficient tests of the normality of X will be based on the projections $\langle X, v_j \rangle = \sqrt{\lambda_j} Z_j$. These projections are however not available because the functions v_j are not observable and must be estimated from the sample $X_1, X_2, \ldots, X_N$. We assume that the X_i have the same distribution as X. Also, only a finite number of projections can be used in practice. The sample analogue of (3.1) is

$$X_i(t) \approx \bar{X}_N(t) + \sum_{j=1}^{p} \hat{\xi}_{ij} \hat{v}_j(t), \tag{3.2}$$

where the $\hat{v}_j$ are estimators of the v_j and $\hat{\xi}_{ij} = \int (X_i(t) - \bar{X}_N(t)) \hat{v}_j(t) dt$. The advantage of using the $\hat{v}_j$ over any other orthonormal system is that for any given p, they give a better approximation in (3.2) than any other system (see, e.g. chapter 3 of Horváth & Kokoszka, 2012). In practice, this means that a relatively small p is sufficient to approximate the functional data well, whereas for other system, a large p would, in general, be required, which would make the multivariate normality tests less efficient.

Under H_0, the sample scores $\hat{\xi}_{ij}$ are *approximately* normal and *approximately* independent across j. If the functions X_i are independent, then the score vectors $\hat{\xi}_i = [\hat{\xi}_{i1}, \hat{\xi}_{i2}, \ldots, \hat{\xi}_{ip}]^\top$ are also independent. The multivariate tests are applied to these scores, both with fixed and data-driven dimension p. There are several ways of selecting a suitable truncation level p. The simplest and most commonly used approach, which is coded in relevant software packages, is to choose p such that the first p-estimated FPCs explain a prespecified percentage of total variance (details are given by Horváth & Kokoszka, 2012, pp. 41–42).

We emphasise that even when the Z_j in (3.1) are standard normal, the $\hat{\xi}_{ij}$ in (3.2) are not exactly normal. This is because $\hat{\xi}_{ij} = \langle X_i, \hat{v}_j \rangle$ and the $\hat{v}_j$ are functions of all $X_i, 1 \le i \le N$. The $\hat{v}_j$ are only asymptotically close to v_j (see, e.g. section 2.5 of Horváth & Kokoszka, 2012). Because of these estimation and selection of p steps, it is not clear which multivariate tests will be most accurate and powerful.

Functional models commonly assume that error functions are normal. Performing a test on residuals introduces another estimation step, which impacts testing the normality of the errors. To illustrate, consider the function-on-function regression

$$Y_i(t) = \mu(t) + \int \psi(t, s) X_i(s) ds + \varepsilon_i(t), \quad 1 \le i \le N, \tag{3.3}$$

with iid mean zero ε_i distributed as ε. The null hypothesis now is

$$H_0 : \varepsilon \text{has a normal distribution in} L^2.$$

A test can be applied only to the residuals

$$\hat{\varepsilon}_i(t) = Y_i(t) - \hat{\mu}_N(t) - \int \hat{\psi}(t,s)X_i(s)ds.$$

Under conditions stated in Section 6, test statistics based on the $\hat{\varepsilon}_i$ will have the same asymptotic null distribution as their counterparts based on the unobservable errors ε_i. However, in finite samples, there is a difference between the unobservable $\langle \varepsilon_i, v_j \rangle$ and the available $\langle \hat{\varepsilon}_i, \hat{v}_j \rangle$. In particular, the latter projections are not independent across i because both $\hat{\varepsilon}_i$ and $\hat{v}_j$ depend on the whole data set. Simulations must thus be used to determine which tests perform best in finite samples.

4 Simulation Results

The section is organised as follows. In Section 4.1, we list `R` packages and functions used to perform the simulations. In Section 4.2, we define the data-generating processes (DGPs). Section 4.3 reports the results of preliminary screening intended to eliminate from further analysis the tests that are not competitive with the best tests whose performance is analysed in greater detail in Section 4.4.

4.1 R Software

Initial stages of computations are performed with the following functions:

(a) Raw data are converted to functional objects with `fda::create.bspline.basis` and `fda::smooth.basisPar` with default parameters.
(b) The scores ξ_{ij} are computed using the function `fda::pca.fd`.
(c) The regression residuals defined in Section 4.2 are computed with `fda::fRegress` and `fda::predict` with default parameters.
(d) Kurtosis and skewness are computed using functions from the `moments` library.

The multivariate normality tests are performed using the following `R` software:

(a) Mardia's multivariate normality tests: `MVN::mvn(mvnTest = 'mardia')`, MJB_M and MJB^*_M tests of Koizumi *et al.* (2009) (own implementation).
(b) Royston's multivariate normality test: `MVN::mvn(mvnTest = 'royston')`.
(c) Henze–Zirkler's multivariate normality test: `MVN::mvn(mvnTest = 'hz')`.
(d) Energy test: `MVN::mvn(mvnTest = 'energy')`.
(e) Doornik–Hansen Test: `normwhn.test:normality.test1`, its Lobato and Velasco's modification: `normwhn.test:normality.test2`.

4.2 Data-Generating Processes

We consider testing normality in a random sample of functions, as well as testing the normality of regression errors.

IID functions We work with two types of random functions. The first one is the random walk on the unit interval:

$$W\left(\frac{k}{K}\right) = \frac{1}{\sqrt{K}} \sum_{i=1}^{k} Z_i. \quad 1 \le k \le K. \tag{4.1}$$

If the Z_i are independent standard normal, then it is a Gaussian random function that approximates the standard Wiener process. We used $K = 75$ because this value was sufficient to produce visually almost continuous trajectories resembling realisations of the Wiener process, the most extensively studied Gaussian process. The second type are functions of the form

$$V(t) = Z_1 t + \frac{1}{2} Z_2 \cos(\pi t) + \frac{1}{4} Z_3 \sin(\pi t), \quad t \in [0, 1], \tag{4.2}$$

which have a different FPC structure. Again, if the Z_i are independent standard normal, $V(\cdot)$ is a Gaussian random function. If the Z_i are iid with a non-normal distribution, (4.1) and (4.2) define non-normal random functions.

Regression errors We consider the fully functional linear regression

$$Y_i(t) = \int \psi(t, s) X_i(s) ds + \varepsilon_i(t), \quad 1 \le i \le N. \tag{4.3}$$

We consider regression (4.3) with

$$X_i(t) = it, \quad \psi(t, s) = ts, \quad 0 \le s, t \le 1.$$

The errors ε_n are either the random walks (4.1) or the functions (4.2). As noted in Section 4.1, the estimator $\hat{\psi}$ is obtained as a linear combination of the products of univariate spline bases. We compute the residual curves

$$\hat{\varepsilon}_i(t) = Y_n(t) - \int \hat{\psi}(t, s) X_i(s) ds.$$

To test H_0: the ε_n are iid normal in L^2, we apply the test procedures to the $\hat{\varepsilon}_i$.

Non-normal distributions There are many non-normal distributions that could be used to evaluate the power of the tests. In the tables in Appendix A, we use the following:

(a) t_4; skewness = 0, kurtosis = ∞.
(b) EXP(1) -1 (centred exponential with unit rate); skewness = 2, kurtosis = 9.
(c) SN($\xi = 0, \omega^2 = 1, \alpha = 10$); skewness ≈ 0.96, kurtosis ≈ 3.82.

The skew-normal distribution, SN(ξ, ω^2, α), is treated in Azzalini (2014), where formulas for its skewness and kurtosis are given.

DGP abbreviations For ease of reference, we use the following abbreviations for the DGPs:

IID RW	iid random walks (4.1)
IID LT	iid linear + trigonometric functions (4.2)
RES RW	residuals of regression (4.3) with RW errors
RES LT	residuals of regression (4.3) with LT errors

Under H_0, the Z_j are standard normal. Under H_A, their distributions are specified previously.

4.3 Preliminary Screening

Because of the large number of normality tests we considered, we first applied them all to the DGPs described in Section 4.2. For sample sizes $N = 50, 150$, and 450, we recorded the counts of rejections in one thousand replications for nominal sizes $\alpha = 0.10, 0.05$, and 0.01. This initial screening was intended to eliminate from further analysis tests that are not competitive with better tests. They might be useful in some settings we did not consider, but in the following, we focus on tests that performed best in our preliminary study. Tables of all rejections rates are presented in the supporting information. In this section, we discuss the results.

As the first criterion, we considered empirical size. As 'acceptable', we considered tests with the count of rejections between 80 and 120 when 100 were expected, between 40 and 60 when 50 were expected and less than 20 if 10 were expected. Generally, all tests improve as N increases, but some are much worse than others for $N = 50$ and $N = 150$ or for some specific α. This screening thus eliminated tests that are not competitive for smaller sample sizes, which are common in functional data analysis. We examined in how many cases of N and α each test was 'unacceptable'. The Jarque–Bera and Royston and $\text{MJB}^*_{\boldsymbol{M}}$ tests were the worst in all four size tables. The $\text{MJB}_{\boldsymbol{M}}$, $\text{MJB}^*_{\boldsymbol{M}}$ and Mardia kurtosis were not competitive with the best tests.

When considering empirical power, in each of the four tables, we selected about five tests that were better than others for most combinations of N and α and considered the remaining tests as not competitive. In all four tables, the Henze–Zirkler and energy tests were not competitive, and the Mardia skewness or kurtosis tests were not competitive in some tables. These tests are designed for specific alternatives, and their modification $\text{MJB}_{\boldsymbol{M}}$ that targets both skewness and kurtosis has competitive power in most cases; $\text{MJB}^*_{\boldsymbol{M}}$ has higher power, but its size is often unacceptably higher than the nominal size.

In the next section, we therefore focus on the Doornik–Hansen test and its Lobato and Velasco modification and on the $\text{MJB}_{\boldsymbol{M}}$ test. We also include the Jarque–Bera test because it is often used.

4.4 Performance of the Best Tests

Appendix A contains the most informative tables obtained in the course of our simulation study aimed at comparing the four tests that emerged as the best tests after the preliminary screening described in Section 4.3. Many other tables are included in the supporting information. All tables except those for preliminary screening are based on 5 000 replications. Like with all simulation studies, subjectivity enters through the choice of the DGPs. We tried to select a representative range of the DGPs. We now list the main conclusion that follows from the tables in Appendix A and those presented in the supporting information.

(a) The Doornik–Hansen and the Lobato–Velasco tests generally have empirical sizes closest to the nominal sizes, and this advantage is most pronounced for sample sizes $N = 50$ and $N = 150$.
(b) Mardia's $\text{MJB}_{\boldsymbol{M}}$ test is generally the most powerful. The Lobato–Velasco test is often the second most powerful test, competing with the Jarque–Bera test for the second place.
(c) Mardia's $\text{MJB}_{\boldsymbol{M}}$ test has too conservative size for $N = 50$. For $N = 150$ and $N = 450$, its size is almost as accurate as that of the Doornik–Hansen and the Lobato–Velasco tests.
(d) The sizes of the four tests considered in this section are not impacted much by the percentage of variance explained used to determine p. In almost all cases, the difference in empirical sizes is less than 0.01; in most cases, it is less than 0.05, which is about or less (depending on the nominal size α) than two standard errors. There is no obvious pattern that would apply even to a specific test.

(e) The same is generally true for power. In the case of the IID random walk process, the power generally increases with the percentage of variance explained, but this does not hold for other DGPs.

We can state the following recommendations:

(a) If $N \geq 150$, Mardia's $\text{MJB}_{\boldsymbol{M}}$ test has accurate size and is most powerful.
(b) For $N \approx 50$, the Doornik–Hansen and the Lobato–Velasco tests have accurate size and power comparable with Mardia's $\text{MJB}_{\boldsymbol{M}}$ test.
(c) Using the established 85% rule is therefore safe.

Tables A7 and A8 in Appendix A report the counts (out of 5 000) of replications with a specific p selected by the 85% of explained variance criterion. We see that as N increases, these counts concentrate on a single value of p. Denoting by $\hat{p}_N$ the selected p, we may thus empirically assert that for some p_0, which depends on the DGP,

$$\lim_{N\to\infty} P(\hat{p}_N = p_0) = 1.$$

The value p_0 is equal to the value that explains 85% of variance for the population model.

5 Normality of Well-Known Functional Data Sets

In this section, we apply the tests we studied in Section 4.4 to several well-known and extensively studied functional data sets. Our objective is to determine which of them can be assumed to follow a normal distribution in the space L^2. We study the following data.

(a) Growth curves of $N = 54$ girls and $N = 39$ boys. The data are part of the object `growth` in the `fda` package (Ramsay *et al*., 2018).
(b) Spectral curves for $N = 60$ gasoline samples. These data are available as `gasoline` in the `refund` package (Goldsmith *et al*., 2018). We report the results for the raw data and for their reciprocals.
(c) Spectral curves for $N = 215$ meat samples. These data are available as `tecator` in the `fda.usc` package.
(d) Fractional anisotropy tract profiles of $N = 376$ patients. These data are part of the `DTI` data set in the `refund` package. These data contain two groups: MS patients and controls.
(e) CD4 cell counts for $N = 366$ human immunodeficiency virus-infected individuals. These data are available as data set `CD4` in the `refund` package.

The first four data sets are used and described in some detail in Kokoszka and Reimherr (2017) (and dozens of papers and a few books). The data set of CD4 counts is described and used in Goldsmith *et al.* (2013). It is a sparse functional data set, with 1 to 11 observations per curve, with the median 5. To apply the normality tests, the curves were reconstructed using principal analysis by conditional estimation methodology using the `R` library `fdapace`.

Table 1 displays the results. There is no strong evidence that the growth curves are not normal, but relatively small samples sizes must be kept in mind. The Jarque–Bera test indicates non-normality of the gasoline data, with the other tests providing marginal evidence against normality. These curves have intervals of values close to 0 with a few spikes (see, e.g. figure 4.2 in Kokoszka & Reimherr, 2017). Thus normality could indeed be questionable. The left panel of Figure 1 displays 10 randomly selected and centred curves, i.e. $X_r(t) - \bar{X}_N(t)$, $r = 1, 2, \ldots 10$. The right panel shows the curves $Y_r(t) - \bar{Y}_N(t)$, where $Y_i(t) = (X_i(t) + 1)^{-1}$. Table 1 shows

Table 1. *P values for real data sets and the selected truncation level* $\boldsymbol{p}$.

	Jarque–Bera	MJB_M	Doornik–Hansen	Lobato–Velasco	$\hat{p}$
Growth, female	0.417	0.556	0.049	0.065	2
Growth, male	0.385	0.458	0.154	0.111	2
Gasoline, raw	0.015	0.074	0.050	0.107	2
Gasoline, transformed	0.249	0.238	0.114	0.136	2
tecator	0.000	0.000	0.000	0.000	2
DTI, control	0.757	0.570	0.761	0.766	4
DTI, MS	0.004	0.000	0.002	0.001	4
CD4	0.000	0.000	0.000	0.000	2

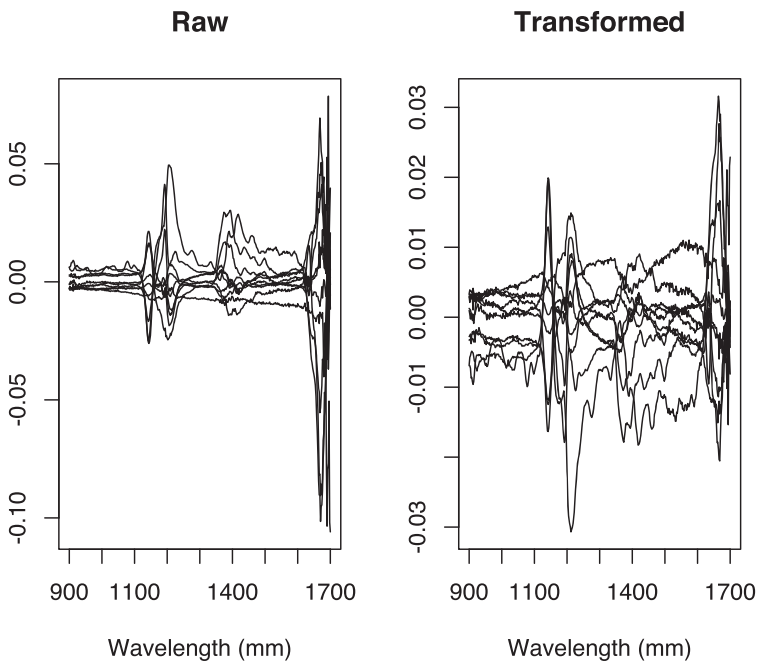

Figure 1. *Left panel: 10 randomly selected centred gasoline spectral curves. Right panel: the same curves transformed and then centred. According to our tests, there is some evidence that the curves in the left panel do not come from a normal distribution, while there is no evidence against the normality of the curves in the right panel.*

that the gasoline curves so transformed can be considered normal. The extensively studied tecator data, made popular after the publication of the monograph of Ferraty and Vieu (2006), are definitely not normal. There are several curves that show visibly greater variability than the remaining curves.

The last three data sets produce very interesting results. We see a clear-cut difference between the DTI curves of MS patients and healthy controls. The DTI curves of healthy people can be assumed to come from a normal distribution: those of MS patients form a highly non-normal distribution. Examples of centred curves from both groups are shown in Figure 2. It is difficult to say by visual examination that the curves in the left panel come from a normal distribution in L^2, while those in the right panel from a non-normal distribution. Figure 3 shows all raw DTI curves in both groups (without the mean function subtracted). One can see that there are a few curves in the MS group that might be considered outliers. If these curves are removed

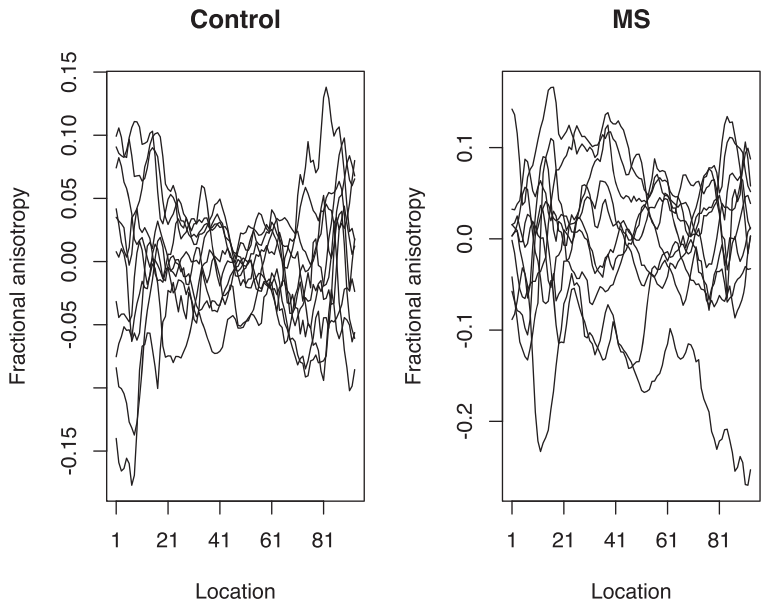

Figure 2. *Left panel: 10 randomly selected centred fractional anisotropy tract profiles from the control group. Right panel: 10 randomly selected centred fractional anisotropy tract from the MS group. According to our tests, the curves in the left panel come from a normal distribution, while those in the right panel from a non-normal distribution.*

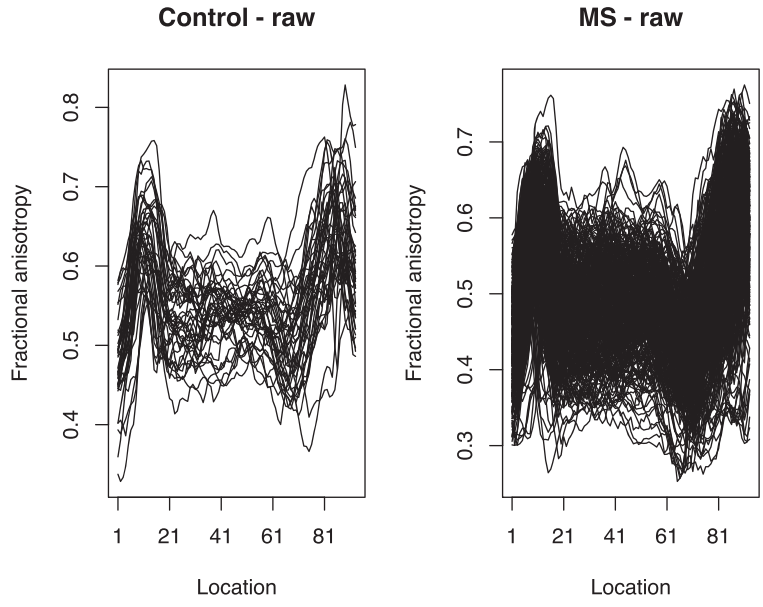

Figure 3. *DTI curves in the control and MS groups.*

from the MS group, the evidence against normality becomes weaker; the P values become 0.375, 0.004, 0.435, and 0.032. It is however not clear that these curves should be considered as outliers, as similarly 'outlying curves' are present in the control group. An observation should be considered an outlier if there is some evidence that it comes from a distribution different than other observations, not if it just a little bit different than the bulk of the observations. Without access to more detailed clinical information, this issue cannot be resolved. The DTI data are described in greater detail in Goldsmith *et al.* (2011) and Goldsmith *et al.* (2012) and introduced briefly in section 1.5 of Kokoszka and Reimherr (2017).

The reconstructed CD4 curves are highly non-normal. This poses an interesting question. The principal analysis by conditional estimation methodology assumes that the scores and the errors with which the curves are sparsely observed are jointly Gaussian. The reconstruction is based on conditional expectations computed from a multivariate Gaussian distribution. If the reconstructed curves are not Gaussian, it means that these assumptions are not valid for the sparsely observed CD4 data. This example thus motivates a need to develop reconstruction methods robust to the assumption of normality.

6 Large Sample Theory

We have considered several normality tests in this paper. They follow the same general paradigm: a multivariate test is applied to projections of the data or suitable residuals on estimated FPCs $\hat{v}_j$. A natural way of showing the asymptotic validity of such tests is to prove that they have the same limit distributions as the multivariate tests based on projections on the unobservable population FPCs v_j. The latter projections form multivariate iid samples, for which the limit distributions were found in the papers cited in Section 2. We focus on Mardia's tests and regression (3.3). Similar theory can be developed for other tests, but the specific technical arguments would be different. The theory is much simpler if $\psi \equiv 0$, i.e. if the tests are based on observations rather than residuals.

We assume that all functions are elements of the Hilbert space $L^2(\mathcal{T})$ with the inner product $\langle f, g\rangle = \int f(t)g(t)dt$ and the norm $\|\cdot\|$ it induces. We will also refer to the space $\mathcal{S}$ of Hilbert–Schmidt operators on L^2, whose norm is denoted $\|\cdot\|_{\mathcal{S}}$. Details are presented, e.g. in chapter 2 of Horváth and Kokoszka (2012).

In Model (3.3), we assume that

$$\text{the sequences } \{\varepsilon_i\} \text{ and } \{X_j\} \text{ are independent,} \tag{6.1}$$

$$\text{the } \varepsilon_i \text{ are } iid \text{ in } L^2 \text{ with } E\varepsilon_0(t) = 0 \text{ and } E\|\varepsilon_0\|^6 < \infty \tag{6.2}$$

and

$$\text{the } X_i \text{ are } iid \text{ in } L^2 \text{ with } EX_0(t) = 0 \text{ and } E\|X_0\|^2 < \infty. \tag{6.3}$$

We assume that the kernel ψ can be consistently estimated by $\hat{\psi}_N$, i.e.

$$\left\|\hat{\psi}_N - \psi\right\|_{\mathcal{S}}^2 := \iint \left\{\hat{\psi}_N(t,s) - \psi(t,s)\right\}^2 dsdt \xrightarrow{P} 0. \tag{6.4}$$

An estimator based on FPCs satisfies (6.4) (see, e.g. Horváth & Kokoszka (2012, p. 134).

Recall that the FPCs v_ℓ are the eigenfunctions of the covariance kernel $c(t,s) = E[\varepsilon_0(t)\varepsilon_0(s)]$. Assumption (6.2) implies that

$$\left\| \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i(t)\varepsilon_i(s) - c(t,s) \right\|_{\mathcal{S}} = o_P(1), \tag{6.5}$$

and

$$\|\hat{v}_\ell - v_\ell\| = o_P(1). \tag{6.6}$$

For these relations to hold, $\|\varepsilon_0\|^4 < \infty$ is enough (see, e.g. chapter 2 of Horváth & Kokoszka (2012), but we must assume finite sixth moment to establish the equivalence for Mardia's statistics because they are based on the third moment of cross covariances. Because the FPCs $\hat{v}_\ell$ and v_ℓ are determined only up to a sign, to be precise, the left-hand side of (6.6) should involve $\hat{v}_\ell - c_{\ell,N} v_\ell$ with $c_{\ell,N} = sign\langle \hat{v}_\ell, v_\ell\rangle$. The formulas that follow are invariant to $c_{\ell,N}$, so to reduce their size, we assume that $c_{\ell,N} = 1$.

Our tests are based on the centred residuals

$$\hat{\varepsilon}_i(t) - \hat{\varepsilon}^\star(t), \quad \hat{\varepsilon}^\star_N(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i(t),$$

where

$$\hat{\varepsilon}_i(t) = Y_i(t) - \int \hat{\psi}_N(t,s) X_i(s) ds.$$

To ease the notation, we set

$$\hat{u}_i(t) = \hat{\varepsilon}_i(t) - \hat{\varepsilon}^\star_N(t), \quad u^\star_i(t) = \varepsilon_i(t) - \varepsilon^\star_N(t),$$

where

$$\varepsilon^\star_N(t) = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i(t).$$

We consider the projections

$$\hat{u}_{i,\ell} = \langle \hat{u}_i, \hat{v}_\ell\rangle, \;\; u^\star_{i,\ell} = \langle u^\star_i, \hat{v}_\ell\rangle, \;\; a_{i,\ell} = \langle u^\star_i, v_\ell\rangle, \quad 1 \le \ell \le d, \; 1 \le i \le N,$$

which we combine to d-dimensional vectors $\mathbf{u}_i$, $\mathbf{u}^\star_i$ and $\mathbf{a}_i$. For example,

$$\mathbf{a}_i = [a_{i,1}, a_{i,2}, \ldots, a_{i,d}]^\top, \;\; 1 \le i \le N.$$

The difference between the vectors $\mathbf{a}_i$ and $\mathbf{u}^\star_i$ is that the former involves projections on the population FPCs v_ℓ, whereas the $\mathbf{u}^\star_i$ involves projections of the estimated FPCs $\hat{v}_\ell$.

We normalize the projected residuals with the corresponding $d \times d$ covariance matrices

$$\widehat{\mathbf{S}}_N = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top, \quad \mathbf{S}^\star_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}^\star_i \mathbf{u}^{\star\top}_i \quad and \quad \mathbf{S}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_i \mathbf{a}_i^\top.$$

Relations (6.5), (6.6) and the law of large numbers imply that

$$\widehat{\mathbf{S}}_N \xrightarrow{P} \mathbf{S}, \quad \mathbf{S}^\star_N \xrightarrow{P} \mathbf{S}, and \quad \mathbf{S}_N \xrightarrow{P} \mathbf{S}, \tag{6.7}$$

where

$$\mathbf{S} = \left[\iint c(t,s) v_\ell(t) v_k(s) dt ds,\ 1 \leq \ell, k \leq d \right].$$

We assume that

$$\mathbf{S}^{-1} \text{ exists.} \tag{6.8}$$

Note that Condition (6.8) means that the d largest eigenvalues of the kernel $c(\cdot,\cdot)$ are positive, i.e.

$$\lambda_1 > \lambda_2 \cdots > \lambda_d > \lambda_{d+1} > 0. \tag{6.9}$$

Set

$$\hat{m}_{i,j} = \hat{\mathbf{u}}_i^\top \widehat{\mathbf{S}}_N^{-1} \hat{\mathbf{u}}_j, \quad m^\star_{i,j} = \mathbf{u}_i^{\star\top} \mathbf{S}_N^{\star\,-1} \mathbf{u}_j^\star \quad \text{and} \quad m_{i,j} = \mathbf{a}_i^\top \mathbf{S}_N^{-1} \mathbf{a}_j.$$

Theorem 6.1. *If assumptions* (6.1)*,* (6.2)*,* (6.3)*,* (6.4) *and* (6.9) *(equivalently* (6.8)*) hold, then*

$$\sum_{i,j=1}^{N} \left(\hat{m}_{i,j}^3 - m_{i,j}^3 \right) = o_P(N), \tag{6.10}$$

and

$$\sum_{i,j=1}^{N} \left(\hat{m}_{i,i}^2 - m_{i,i}^2 \right) = o_P(N^{1/2}). \tag{6.11}$$

The proof of Theorem 6.1 is presented in the supporting information. It involves a number of complex bounds that show that the terms quantifying the transition from Mardia's statistics based on the $m_{i,j}$ to the functional statistics based on the $\hat{m}_{i,j}$ are asymptotically negligible. Recall that $N^{-2} \sum_{i,j=1}^{N} m_{i,j}^3$ and $N^{-1} \sum_{i,j=1}^{N} m_{i,i}^2$ are Mardia's statistics, used to construct the tests described in Section 2.2, computed from iid random vectors

$$[\langle \varepsilon_i, v_1 \rangle, \langle \varepsilon_i, v_2 \rangle, \ldots \langle \varepsilon_i, v_d \rangle]^\top. \tag{6.12}$$

Theorem 6.1 thus shows that any form of Mardia's statistic computed from the vectors

$$[\langle \hat{\varepsilon}_i - \hat{\varepsilon}_N^\star, \hat{v}_1 \rangle, \langle \hat{\varepsilon}_i - \hat{\varepsilon}_N^\star, \hat{v}_2 \rangle, \ldots \langle \hat{\varepsilon}_i - \hat{\varepsilon}_N^\star, \hat{v}_d \rangle]^\top, \tag{6.13}$$

has the same asymptotic distribution as the corresponding test statistic computed from the unobservable vectors (6.12). Vectors (6.13) are used to construct the tests considered in this paper.

Remark 6.1. *We have worked with the sample and population FPCs. Examination of the proof of Theorem 6.1 shows that it remains valid if the $\hat{v}_\ell$ and v_ℓ form any, respectively, data-driven and deterministic orthonormal system for which* (6.6) *holds.*

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.

References

Anderson, T. W. & Darling, D. A. (1954). A test of goodness-of-fit. *J. Am. Stat. Assoc.*, **49**, 765–769.

Aston, J. A. D., Pigoli, D. & Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *Ann. Stat.*, **45**, 1431–1461.

Azzalini, A. (2014). *The Skew–Normal and Related Families.* Shaker Heights, OH: IMS.

Chiou, J.-M., Müller, H.-G. & Wang, J.-L. (2004). Functional response models. *Stat. Sin.*, **14**, 675–693.

Constantinou, P., Kokoszka, P. & Reimherr, M. (2017). Testing separability of space-time functional processes. *Biometrika*, **104**, 425–437.

Crainiceanu, C. M., Staicu, A.-M. & Di, C.-Z. (2009). Generalized multilevel functional regression. *J. Am. Stat. Assoc.*, **104**, 1550–1561.

D'Agostino, R. B., Belanger, A. & D'Agostino Jr., R. B. (1990). A suggestion for using powerful and informative tests of normality. *Am. Stat.*, **44**, 316–321.

Doornik, J. A. & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxf. Bull. Econ. Stat.*, **70**, 927–939.

Duchesne, P., Lafaye de Micheaux, P. & Tatsinkou, J. (2016). Estimating the mean and its effects on Neyman smooth tests of normality for ARMA models. *Can. J. Stat.*, **44**, 241–270.

Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* New York, NY: Springer.

Fremdt, S., Horváth, L., Kokoszka, P. & Steinebach, J. (2013). Testing the equality of covariance operators in functional samples. *Scand. J. Stat.*, **40**, 138–152.

Górecki, T., Hörmann, S, Horváth, L. & Kokoszka, P. (2018). Testing normality of functional time series. *J. Time Ser. Anal.*, **39**, 471–487.

Goldsmith, J., Bobb, J., Crainiceanu, C., Caffo, B. & Reich, D. (2011). Penalized functional regression. *J. Comput. Graph. Stat.*, **20**, 830–851.

Goldsmith, J., Crainiceanu, C., Caffo, B. & Reich, D. (2012). Longitudinal penalized functional regression for cognitive outocomes on neuronal tract measurements. *J. R. Stat. Soc. (C)*, **61**, 453–469.

Goldsmith, J., Greven, S. & Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biom.*, **69**, 41–51.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C. & Reiss, P. T. (2018). *Refund: Regression with functional data*. https://CRAN.R-project.org/package=refund, R package version 0.1-17.

Gromenko, O., Kokoszka, P. & Sojka, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Biol. Stat.*, **11**, 898–918.

Hörmann, S., Kokoszka, P. & Nisol, G. (2018). Testing for periodicity in functional time series. *Ann. Stat.*, **46**, 2960–2984.

Henze, N. & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Commun. Stat. - Theory Methods*, **19**(10), 3595–3617.

Horváth, L. & Kokoszka, P. (2012). *Inference for Functional Data with Applications*: Springer.

Jarque, C. M. & Bera, A. K. (1980). Efficient tests for normality, homoskedasticity and serial independence of regression residuals. *Econ. Lett.*, **6**, 255–259.

Jarque, C. M. & Bera, A. K. (1987). A test of normality of observations and regression residual. *Int. Stat. Rev.*, **55**, 163–172.

Koizumi, K., Okamoto, N. & Seo, T. (2009). On Jarque–Bera tests for assessing multivariate normality. *J. Stat. Adv theory appl.*, **1**, 207–220.

Kokoszka, P. & Reimherr, M. (2017). *Introduction to Functional Data Analysis.* Boca Raton, FL: CRC Press.

Kraus, D. & Panaretos, V. M. (2012). Dispersion operators and resistant second-order analysis of functional data. *Biometrika*, **99**, 813–832.

Liu, C., Ray, S. & Hooker, G. (2017). Functional principal component analysis of spatially correlated data. *Stat. Comput.*, **27**, 1639–1654.

Lobato, I. & Velasco, C. (2004). A simple test of normality for time series. *Econ Theory*, **20**, 671–689.

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

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya (B)*, **36**, 115–128.

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

Panaretos, V. M., Kraus, D. & Maddocks, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Am. Stat. Assoc.*, **105**, 670–682.

Pierce, D. A. & Gray, R. J. (1982). Testing normality of errors in regression models. *Biometrika*, **69**, 233–236.

Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. (2018). *FDA: Functional data analysis*. https://CRAN.R-project.org/package=fda, R package version 2.4.8.
Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *J. R. Stat. Soc. Series C Appl. Stat.*, **31**(2), 115–124.
Royston, J. P. (1983). Some techniques for assessing multivarate normality based on the Shapiro–Wilk W. *J. Roy. Stat. Soc. (C)*, **32**(2), 121–133.
Royston, J. P. (1992). Approximating the Shapiro–Wilk W-test for non-normality. *Stat. Comput.*, **2**(3), 117–119.
Scholz, F. W. & Stephens, M. A. (1997). K-sample Anderson–Darling tests. *J. Am. Stat. Assoc.*, **82**, 918–924.
Seber, G. A. F. & Lee, A. J. (2003). *Linear Regression Analysis* New York: Wiley.
Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.
Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.*, **69**, 730–737.
Székely, G. J. & Rizzo, M. L. (2005). A new test for multivariate normality. *J. Multivar. Anal.*, **93**, 58–80.
Wickham, P. (2015). *Package 'normwhn.test'*. R reference manual.
Yao, F., Müller, H.-G. & Wang, J.-L. (2005a). Functional linear regression analysis for longitudinal data. *Ann. Stat.*, **33**, 2873–2903.
Yao, F., Müller, H.-G. & Wang, J.-L. (2005b). Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.*, **100**, 577–590.

[*Received November 2018, accepted December 2019*]

Appendix A Tables

Empirical sizes are reported in Tables A1 and A2; power in Tables A3, A4, A5 and A6; and selected p in Tables A7 and A8.

Table A1. *Size of the tests for independently and identically distributed functions; standard errors are about 0.004 and 0.003. 0.002 for* $\alpha = 0.10, 0.05$ *and .01. Top panel RW, bottom LT (85%).*

	$N = 50$			$N = 150$			$N = 450$		
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
Jarque–Bera	0.068	0.048	0.025	0.077	0.047	0.020	0.089	0.048	0.018
MJB_M	0.069	0.042	0.016	0.080	0.042	0.014	0.091	0.049	0.013
Doornik–Hansen	0.099	0.053	0.012	0.093	0.046	0.012	0.090	0.046	0.012
Lobato–Velasco	0.103	0.057	0.011	0.096	0.048	0.012	0.091	0.047	0.011
Jarque–Bera	0.081	0.059	0.029	0.096	0.063	0.030	0.098	0.062	0.023
MJB_M	0.065	0.037	0.012	0.089	0.048	0.012	0.099	0.055	0.014
Doornik–Hansen	0.093	0.052	0.012	0.099	0.055	0.013	0.101	0.056	0.012
Lobato–Velasco	0.102	0.055	0.014	0.102	0.054	0.013	0.102	0.057	0.012

Table A2. *Size of the tests for RESidual functions; standard errors are about 0.004 and 0.003. 0.002 for* $\alpha = 0.10, 0.05$ *.01. Top panel RW, bottom LT (85%).*

	$N = 50$			$N = 150$			$N = 450$		
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
Jarque–Bera	0.059	0.041	0.024	0.083	0.056	0.029	0.086	0.050	0.017
MJB_M	0.064	0.039	0.015	0.085	0.048	0.019	0.085	0.046	0.010
Doornik–Hansen	0.090	0.047	0.012	0.100	0.057	0.017	0.095	0.052	0.010
Lobato–Velasco	0.100	0.053	0.015	0.101	0.057	0.018	0.095	0.051	0.010
Jarque-Bera	0.069	0.045	0.022	0.083	0.056	0.022	0.096	0.056	0.019
MJB_M	0.062	0.040	0.016	0.087	0.048	0.017	0.095	0.055	0.014
Doornik–Hansen	0.099	0.055	0.015	0.102	0.051	0.013	0.100	0.053	0.016
Lobato–Velasco	0.109	0.061	0.016	0.107	0.054	0.014	0.099	0.052	0.016

Table A3. *Power and independently and identically distributed RW functions. All standard errors are smaller than 0.007 (85%).*

	$N = 50$			$N = 150$			$N = 450$		
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
				t_4					
Jarque–Bera	0.132	0.103	0.067	0.232	0.186	0.126	0.369	0.310	0.218
MJB_M	0.160	0.116	0.076	0.272	0.218	0.149	0.460	0.384	0.274
Doornik–Hansen	0.166	0.108	0.046	0.238	0.170	0.096	0.371	0.295	0.187
Lobato–Velasco	0.170	0.105	0.044	0.233	0.165	0.093	0.365	0.292	0.185
				$\text{Exp}(1) - 1$					
Jarque–Bera	0.121	0.086	0.047	0.266	0.200	0.115	0.610	0.500	0.313
MJB_M	0.187	0.128	0.066	0.491	0.387	0.220	0.939	0.892	0.751
Doornik–Hansen	0.154	0.090	0.027	0.284	0.193	0.079	0.633	0.518	0.294
Lobato–Velasco	0.169	0.099	0.032	0.291	0.197	0.082	0.635	0.521	0.297
				$\text{SN}(\xi = 0, \omega = 1, \alpha = 10)$					
Jarque–Bera	0.069	0.051	0.024	0.124	0.083	0.042	0.211	0.137	0.058
MJB_M	0.081	0.052	0.022	0.171	0.113	0.047	0.392	0.280	0.125
Doornik–Hansen	0.100	0.055	0.015	0.138	0.080	0.028	0.216	0.138	0.045
Lobato–Velasco	0.111	0.063	0.015	0.142	0.083	0.029	0.219	0.140	0.046

Table A4. *Power and independently and identically distributed LT functions. All standard errors are smaller than 0.007 (85%).*

	$N = 50$			$N = 150$			$N = 450$		
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
				t_4					
Jarque–Bera	0.389	0.343	0.279	0.761	0.718	0.639	0.987	0.982	0.965
MJB_M	0.417	0.355	0.256	0.751	0.702	0.602	0.982	0.973	0.949
Doornik–Hansen	0.412	0.325	0.205	0.766	0.706	0.588	0.989	0.980	0.959
Lobato–Velasco	0.404	0.321	0.201	0.758	0.695	0.578	0.988	0.978	0.958
				$\text{Exp}(1) - 1$					
Jarque–Bera	0.512	0.447	0.345	0.975	0.958	0.907	1	1	1
MJB_M	0.640	0.550	0.394	0.998	0.997	0.989	1	1	1
Doornik–Hansen	0.560	0.459	0.282	0.983	0.967	0.911	1	1	1
Lobato–Velasco	0.595	0.489	0.311	0.984	0.970	0.916	1	1	1
				$\text{SN}(\xi = 0, \omega = 1, \alpha = 10)$					
Jarque–Bera	0.166	0.126	0.074	0.505	0.404	0.260	0.978	0.962	0.895
MJB_M	0.163	0.104	0.042	0.603	0.481	0.285	0.998	0.993	0.970
Doornik–Hansen	0.182	0.115	0.036	0.540	0.423	0.222	0.982	0.968	0.911
Lobato–Velasco	0.207	0.132	0.043	0.553	0.440	0.234	0.984	0.968	0.916

Table A5. *Power and RW RESiduals. All standard errors are smaller than 0.007 (85%).*

	$N = 50$			$N = 150$			$N = 450$		
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
				t_4					
Jarque–Bera	0.140	0.109	0.069	0.252	0.204	0.145	0.430	0.371	0.275
MJB_M	0.142	0.102	0.059	0.256	0.198	0.127	0.443	0.369	0.267
Doornik–Hansen	0.169	0.110	0.050	0.261	0.192	0.112	0.433	0.357	0.240
Lobato–Velasco	0.175	0.112	0.050	0.256	0.189	0.109	0.426	0.355	0.237
				$\text{Exp}(1) - 1$					
Jarque–Bera	0.159	0.120	0.069	0.399	0.306	0.182	0.875	0.797	0.622
MJB_M	0.168	0.117	0.058	0.460	0.352	0.191	0.926	0.879	0.725
Doornik–Hansen	0.185	0.115	0.040	0.427	0.314	0.137	0.887	0.820	0.629
Lobato–Velasco	0.206	0.134	0.050	0.440	0.323	0.147	0.891	0.823	0.634
				$\text{SN}(\xi = 0, \omega = 1, \alpha = 10)$					
Jarque–Bera	0.080	0.053	0.030	0.152	0.103	0.050	0.328	0.230	0.104
MJB_M	0.076	0.050	0.022	0.164	0.105	0.044	0.379	0.268	0.120
Doornik–Hansen	0.108	0.060	0.015	0.164	0.094	0.032	0.343	0.233	0.088
Lobato–Velasco	0.122	0.067	0.018	0.168	0.101	0.033	0.348	0.238	0.089

Table A6. *Power and LT RESiduals. All standard errors are smaller than 0.007 (85%).*

	$N = 50$			$N = 150$			$N = 450$		
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
				t_4					
Jarque–Bera	0.225	0.183	0.135	0.450	0.392	0.305	0.778	0.726	0.631
MJB_M	0.219	0.173	0.119	0.422	0.356	0.269	0.736	0.675	0.571
Doornik–Hansen	0.258	0.187	0.099	0.460	0.382	0.266	0.782	0.721	0.602
Lobato–Velasco	0.257	0.187	0.098	0.455	0.374	0.260	0.777	0.715	0.596
				$\text{Exp}(1) - 1$					
Jarque–Bera	0.290	0.235	0.160	0.792	0.706	0.542	1	0.997	0.988
MJB_M	0.280	0.212	0.124	0.789	0.695	0.514	1	0.999	0.987
Doornik–Hansen	0.347	0.243	0.109	0.827	0.743	0.535	1	0.999	0.991
Lobato–Velasco	0.390	0.277	0.129	0.838	0.758	0.553	1	1.000	0.992
				$\text{SN}(\xi = 0, \omega = 1, \alpha = 10)$					
Jarque–Bera	0.115	0.080	0.045	0.273	0.199	0.109	0.722	0.606	0.384
MJB_M	0.106	0.074	0.033	0.277	0.194	0.091	0.700	0.582	0.355
Doornik–Hansen	0.137	0.081	0.025	0.311	0.205	0.080	0.750	0.641	0.406
Lobato–Velasco	0.157	0.095	0.031	0.326	0.218	0.086	0.755	0.647	0.414

Table A7. *Number of functional principal components selected according to the 85% of explained variance criterion; independently and identically distributed functions.*

Size	$N = 50$					$N = 150$					$N = 450$				
p	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
						$N(0,1)$									
RW	4999	1	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	0	0	796	4204	0	0	0	2	4988	0	0	0	0	5000	0
						t_4									
RW	4999	1	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	1	4	1086	3907	2	0	1	79	4919	1	0	0	8	4992	0
						$\text{Exp}(1) - 1$									
RW	5000	0	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	0	0	1054	3946	0	0	0	29	4971	0	0	0	0	5000	0
						$\text{SN}(\xi = 0, \omega = 1, \alpha = 10)$									
RW	5000	0	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	0	0	877	4123	0	0	0	2	4998	0	0	0	0	5000	0

Table A8. *Number of functional principal components selected according to the 85% of explained variance criterion; RESidual functions.*

Size	$N = 50$					$N = 150$					$N = 450$				
p	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
						$N(0,1)$									
RW	5000	0	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	0	0	812	4188	0	0	0	6	4994	0	0	0	0	5000	0
						t_4									
RW	4997	3	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	1	1	1008	3890	0	0	0	75	4925	0	0	0	7	4992	1
						$\text{Exp}(1) - 1$									
RW	4999	1	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	0	0	1087	3913	0	0	0	28	4972	0	0	0	0	5000	0
						$\text{SN}(\xi = 0, \omega = 1, \alpha = 10)$									
RW	5000	0	0	0	0	5000	0	0	0	0	5000	0	0	0	0
LT	0	0	829	4171	0	0	0	3	4997	0	0	0	0	5000	0