JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **39**: 471–487 (2018)
Published online 29 December 2017 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12281

ORIGINAL ARTICLE

TESTING NORMALITY OF FUNCTIONAL TIME SERIES

TOMASZ GÓRECKI,[a] SIEGFRIED HÖRMANN,[b] LAJOS HORVÁTH[c*] AND PIOTR KOKOSZKA[d]

[a] *Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland*
[b] *Institute for Statistics, Graz University of Technology, Graz, Austria*
[c] *Department of Mathematics, University of Utah, Salt Lake City, UT, USA*
[d] *Department of Statistics, Colorado State University, Fort Collins, CO, USA*

We develop tests of normality for time series of functions. The tests are related to the commonly used Jarque–Bera test. The assumption of normality has played an important role in many methodological and theoretical developments in the field of functional data analysis. Yet, no inferential procedures to verify it have been proposed so far, even for i.i.d. functions. We propose several approaches which handle two paramount challenges: (i) the unknown temporal dependence structure and (ii) the estimation of the optimal finite-dimensional projection space. We evaluate the tests via simulations and establish their large sample validity under general conditions. We obtain useful insights by applying them to pollution and intraday price curves. While the pollution curves can be treated as normal, the normality of high-frequency price curves is rejected.

Received 25 August 2017; Accepted 21 November 2017

Keywords: Functional data; Jarque–Bera test; normal distribution; time series

JEL classification: C12; C32; C15

MOS subject classification: 62M10; 62M15; 62M07.

1. INTRODUCTION

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–Wilk test (Shapiro and 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 and Darling, 1954; Stephens, 1974; Scholz and Stephens, 1987). A number of other approaches have also been proposed (Mardia, 1970, 1974; D'Agostino *et al.*, 1990; Henze and Zirkler, 1990; Doornik and Hansen, 2008; among many others).

In time series analysis and econometrics, the Jarque–Bera test of normality (Jarque–Bera, 1980, 1987) has been most extensively used because of its simplicity and satisfactory performance. It is included as output of most numerical implementations of various time series estimation and goodness-of-fit procedures, and is explained in many textbooks (e.g. Shumway and Stoffer, 2011; Ruppert, 2011). The test is typically applied to model residuals to verify the assumption of i.i.d. normal errors, which is generally needed to compute prediction intervals. This is useful within the Autoregressive Integrated Moving Average (ARIMA) framework, which provides efficient algorithms for the computation of the residuals. In such a linear framework, normality of errors implies normality of the observations. This is no longer the case for nonlinear models, even though testing for normality of errors can still be useful (e.g. Kulperger and Yu, 2005). If temporally dependent observations follow a nonlinear model, or if the dependence can only be quantified nonparametrically (e.g. by mixing or cumulant conditions), then the

* Correspondence to: Lajos Horváth, Department of Mathematics, University of Utah, Salt Lake City, UT 84112-0090, USA. E-mail: horvath@math.utah.edu

usual Jarque–Bera test cannot be applied, as it assumes i.i.d. data (like the other tests listed above). For this reason, several extensions to time series data have been proposed. We review them in Section 2.1.

This work is motivated by recent advances in the emerging field of the analysis of functional time series. A unit observation is a curve, say $X_i(t),\ t \in \mathcal{T}$. For example, $X_i(t)$ can be the pollution level, at some location, at time t of day i; numerous examples are given in the books of Horváth and Kokoszka (2012) and Kokoszka and Reimherr (2017). Many procedures for functional data require that, or are simplified if, the curves can be assumed to be realizations of a Gaussian random process. Yet, at present, no tests of normality have been developed in the context of functional data. The purpose of this work is to fill this gap by developing extensions of the Jarque–Bera testing paradigm to functional data. While our chief motivation comes from applications to functional time series, the tests we propose can also be applied to random samples. We give a few examples of recent research in functional data analysis (FDA) where the normality assumption has been used, without any claim that our list is representative. 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, formulae become much simpler if normality is assumed (e.g. Panaretos *et al.*, 2010; Kraus and 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 will bolster confidence in the conclusions of these and many other functional data analysis (FDA) procedures.

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 reduction. In either case, extending the notions of symmetry and kurtosis to vectors so obtained requires thought. In case of a projection on a system estimated from the data, the effect of the estimation must be taken into account. In the context of this article, test statistics are computed from such projections. We establish conditions under which the effect of estimation of the projection subspace is asymptotically negligible. Moreover, it turns out that, in finite samples, the usual functional principal components (FPCs) do not lead to the best tests, and a different data-driven system leads to tests with a better balance of size and power.

It is hoped that this work will stimulate further research on procedures for testing for or transformation to normality in the context of functional data. The Jarque–Bera paradigm is just one of the possible approaches. It is well known that it has many disadvantages. Like most moment-based statistics, the test statistic has breakdown value 0. Outliers for functional data are generally more difficult to detect than for scalar or vector observations (see Arribas-Gil and Romo, 2014 and references therein), so robust approaches can be of value; the ideas of Brys *et al.* (2004) could be potentially extended to the functional context or other robust tests developed. We note that it is always useful to complement significance tests by exploratory tools such as normal QQ plots of the scores or other tools described, for example, in Kosiorowski and Zawadzki (2017).

This article is organized as follows. In Section 2, we provide background on Jarque–Bera tests and heuristically derive our tests for functional data. In Section 3, we evaluate them by means of a simulation study and applications to real functional time series. Large-sample justification of our tests is presented in Section 4. The proofs involve several delicate steps and require extensive theoretical background, so they are moved to supplemental material, which exceeds in length the expository part of the article. The sections of the Supporting Information are numbered with roman capitals to help identify them.

2. DERIVATION OF THE TESTS

We begin by reviewing in Section 2.1 relevant work on Jarque–Bera type tests for scalar time series. In Section 2.2, we derive several approaches to testing normality of functional time series. Our tests, like all tests in this family, are based on the comparison of sample third and fourth moments to those implied by the normal distribution. They will therefore generally not be consistent if the data are not normal but have these two moments as the corresponding Gaussian data structure. Additionally, in the functional setting, the tests involve projections on finite-dimensional subspaces. While these subspaces are constructed to capture most of the variability of the data, the tests will not detect a non-Gaussian component present in a 'small' subspace orthogonal to the projection subspace.

  J. Time Ser. Anal. **39**: 471–487 (2018)
DOI: 10.1111/jtsa.12281

2.1. Jarque–Bera Type Tests for Scalar Time Series

Suppose we observe a realization $X_1, X_2, \dots, X_N$ of a strictly stationary *scalar* time series $\{X_i\}$. We use the index i to denote time because starting with Section 2.2, we will use t as the argument of the functions forming a functional time series. Set $\mu = EX_i$ and $\mu_r = E[(X_i - \mu)^r]$, $r = 2, 3, \dots$. Recall that the skewness and kurtosis are respectively defined by

$$\tau = \frac{\mu_3}{\sigma^3} \quad \text{and} \quad \kappa = \frac{\mu_4}{\sigma^4}. \tag{2.1}$$

The sample skewness and kurtosis are defined by replacing in (2.1) the population moments by the corresponding sample moments. If the X_i are i.i.d. normal, then the convergence

$$\mathrm{JB}_N := N\left(\frac{\hat{\tau}^2}{6} + \frac{(\hat{\kappa} - 3)^2}{24}\right) \xrightarrow{D} \chi^2(2)$$

leads to the standard Jarque–Bera test. It no longer holds if the assumption of independence is violated. Several extensions have been proposed.

We first focus on the test of Bai and Ng (2005). Introduce random vectors

$$\mathbf{Z}_i = \begin{bmatrix} (X_i - \mu)^3 \\ X_i - \mu \end{bmatrix}$$

and their long-run covariance matrix

$$\mathbf{\Gamma}_Z = \lim_{N\to\infty} N^{-1} \sum_{n,m=1}^{N} E[\mathbf{Z}_n \mathbf{Z}_m^\top].$$

Next, introduce the vector $\boldsymbol{\alpha} = [1,\ -3\sigma^2]$. If $\tau = 0$, then, under appropriate assumptions

$$\sqrt{N}\hat{\tau} \xrightarrow{D} N(0,\ \sigma^{-6}\boldsymbol{\alpha}\mathbf{\Gamma}_Z\boldsymbol{\alpha}^\top). \tag{2.2}$$

To specify the limiting null distribution of the estimated kurtosis, introduce the vectors

$$\mathbf{W}_i = \begin{bmatrix} (X_i - \mu)^4 - \mu_4 \\ X_i - \mu \\ (X_i - \mu)^2 - \sigma^2 \end{bmatrix}$$

and their long-run covariance matrix

$$\mathbf{\Gamma}_W = \lim_{N\to\infty} N^{-1} \sum_{n,m=1}^{N} E[\mathbf{W}_n \mathbf{W}_m^\top].$$

Next, introduce the vector $\boldsymbol{\beta} = [1,\ -4\mu_3,\ -6\sigma^2]$. Then, if $\kappa = 3$

$$\sqrt{N}(\hat{\kappa} - 3) \xrightarrow{D} N(0,\ \sigma^{-8}\boldsymbol{\beta}\mathbf{\Gamma}_W\boldsymbol{\beta}^\top). \tag{2.3}$$

Bai and Ng (2005) claim that the relation (2.3) holds for scalar sequences stationary up to eighth order, but do not provide a complete proof. Some weak dependence assumption or normality is likely needed. Bai and Ng (2005)

DOI: 10.1111/jtsa.12281

argue, citing Lomnicki (1961), that under suitable assumptions, the limits in (2.2) and (2.3) are independent, and conclude that

$$\mathrm{JBD}_N := N\left\{\frac{\hat{\sigma}^6\hat{\tau}^2}{\hat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\Gamma}_Z}\hat{\boldsymbol{\alpha}}^\top} + \frac{\hat{\sigma}^8(\hat{\kappa}-3)^2}{\hat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Gamma}_W}\hat{\boldsymbol{\beta}}^\top}\right\} \xrightarrow{D} \chi^2(2).$$

The estimates of the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are constructed using the sample variance and setting $\mu_3 = 0$. To compute the sample counterpart of $\mathbf{W}_i$, σ^2 is replaced by the sample variance. There are two options for μ_4: $\hat{\mu}_4$ or $3\hat{\sigma}^2$. We obtained similar empirical rejection rates for either option when working with this test. The matrices $\boldsymbol{\Gamma}_Z$ and $\boldsymbol{\Gamma}_W$ are estimated by one of the consistent estimators of the long-run covariance matrix, denoted $\widehat{\boldsymbol{\Gamma}}_Z$ and $\widehat{\boldsymbol{\Gamma}}_W$ in the following.

We now turn to the approach of Lobato and Velasco (2004). It is based on the convergence of empirical moments noted by Lomnicki (1961) and Gasser (1975)

$$\sqrt{N}\begin{bmatrix}\hat{\mu}_3\\ \hat{\mu}_4 - 3\hat{\mu}_2\end{bmatrix} \xrightarrow{D} N\begin{bmatrix}6F^{(3)} & 0\\ 0 & 24F^{(4)}\end{bmatrix},$$

where

$$F^{(k)} = \{\gamma(0)\}^k + 2\sum_{h=1}^{\infty}\{\gamma(h)\}^k,$$

and where $\gamma(i) = \mathrm{Cov}(X_0, X_i)$. Lobato and Velasco (2004) propose to estimate the quantities $F^{(k)}$ by

$$\widehat{F}^{(k)} = \{\hat{\gamma}(0)\}^k + 2\sum_{h=1}^{N-1}\{\hat{\gamma}(h)\}^k$$

and shown their consistency. They conclude that under normality

$$\mathrm{LV}_N = \frac{N\hat{\mu}_3^2}{6\widehat{F}^{(3)}} + \frac{N(\hat{\mu}_4 - 3\hat{\mu}_2^2)^2}{24\widehat{F}^{(4)}} \xrightarrow{D} \chi^2(2). \tag{2.4}$$

We conclude this section by explaining how the above two approaches are actually closely related. Note that JBD_N can be rewritten as

$$\mathrm{JBD}_N = \frac{N\hat{\mu}_3^2}{\hat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\Gamma}}_Z\hat{\boldsymbol{\alpha}}^\top} + \frac{N(\hat{\mu}_4 - 3\hat{\mu}_2^2)^2}{\hat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Gamma}}_W\hat{\boldsymbol{\beta}}^\top}.$$

Using Isserlis' theorem (see Section A of the Supporting Information), it can be shown that under normality

$$\boldsymbol{\alpha}\boldsymbol{\Gamma}_Z\boldsymbol{\alpha}' = [1, -3\sigma^2]\begin{bmatrix}9\sigma^4F^{(1)} + 6F^{(3)} & 3\sigma^2F^{(1)}\\ 3\sigma^2F^{(1)} & F^{(1)}\end{bmatrix}\begin{bmatrix}1\\ -3\sigma^2\end{bmatrix} = 6F^{(3)}.$$

Similarly, it holds that $\boldsymbol{\beta}\boldsymbol{\Gamma}_W\boldsymbol{\beta}' = 24F^{(4)}$. Hence, the two test statistics differ only by different estimators of the variance term. In finite samples, this may, however, lead to substantially different performance. This is shown in Section 3 in the context of functional data.

The formulae stated by Lomnicki (1961) and Gasser (1975) are crucial for the derivation of both tests, of Bai and Ng (2005) and of Lobato and Velasco (2004). They are, however, not rigorously proven in those articles. As a by-product of our work, we provide rigorous proofs.

  J. Time Ser. Anal. **39**: 471–487 (2018)
DOI: 10.1111/jtsa.12281

2.2. Tests for Functional Time Series

We derive several tests of normality for time series of functions. The presentation is informal and is geared toward explaining which approaches can be expected to work and which cannot, and then defining the test algorithms. A precise asymptotic framework is formulated in Section 4.

In practice, functional data are observed on a discrete set of time points and then transformed in a preprocessing step into functional objects. Depending on the data and problem at hand, the transformation into curves can be done in many different ways, see, for example, Ramsay and Silverman (2005) for densely observed curves, Yao *et al.* (2005) for sparsely observed curves, and Kraus (2015) for curves with missing segments. In the latter two cases, the assumption of normality plays an important role in curve reconstruction, so an application of a normality test following such a reconstruction would be of particular value. *Our tests proposed below are designed for a functional sample obtained after some preprocessing/reconstruction step.* We thus work with a sequence of fully observed functions $X_1, X_2, \ldots, X_N$ defined on the same domain, which, for simplicity of notation, is assumed to be the unit interval $[0,1]$. These functions are assumed to be a realization of a strictly stationary time series in $L^2 = L^2([0,1])$. The null hypothesis we want to test is

$$\mathcal{H}_0: \text{ the process } \{X_i\} \text{ is Gaussian.}$$

If $\langle \cdot, \cdot \rangle$ denotes the inner product on L^2, that is, $\langle u, v\rangle = \int u(t)v(t)\mathrm{d}t$, then $\mathcal{H}_0$ may be equivalently stated as

$$\mathcal{H}_0: (\langle X_{i_1}, \phi_1\rangle, \ldots, \langle X_{i_n}, \phi_n\rangle)^{\mathrm{T}} \text{ is multi-variate normal,}$$
$$\text{for all } \{i_1, \ldots, i_n\} \subset \mathbb{Z},\, n \geq 1, \text{ and } \phi_i \in L^2.$$

The null hypothesis is thus very complex because of the infinite-dimensional nature of the data. An efficient dimension reduction is required to render the problem feasible. We consider tests based on two dimension reduction techniques.

2.2.1. *Functional Principal Component Analysis-Based Tests*

We recall the well-known population and sample Karhunen–Loéve expansions (e.g. Chapter 2 of Horváth and Kokoszka, 2012). Denote by X a random function which has the same distribution in L^2 as each function X_i. If $\mathcal{H}_0$ is true, then X admits the decomposition

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \sqrt{\lambda_j} Z_j v_j(t), \quad t \in [0,1], \tag{2.5}$$

where the v_j are the FPCs, Z_j are independent standard normal, and $\lambda_j = E\langle X-\mu, v_j\rangle^2$. The sample analog of (2.5) is

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

with $\hat{\xi}_{ij} = \langle X_i - \overline{X}_N, \hat{v}_j\rangle$ being the empirical FPC scores. If the functions X_i are independent, then under $\mathcal{H}_0$, the scores $\hat{\xi}_{ij}$ are approximately normal and independent across i and across j. Denote by $\hat{\tau}_j$ the estimated skewness of the pseudo-observations $\hat{\xi}_{1j}, \hat{\xi}_{2j}, \ldots, \hat{\xi}_{Nj}$, and by $\hat{\kappa}_j$ their estimated kurtosis. Then

$$\mathrm{JB}_N^{(j)} := N\left(\frac{\hat{\tau}_j^2}{6} + \frac{(\hat{\kappa}_j - 3)^2}{24}\right) \xrightarrow{d} \chi^2(2).$$

DOI: 10.1111/jtsa.12281

Since the $\mathrm{JB}_N^{(j)}$ are approximately independent, it follows that

$$\hat{T}_{1,p} := N \sum_{j=1}^{p} \left(\frac{\hat{\tau}_j^2}{6} + \frac{(\hat{\kappa}_j - 3)^2}{24} \right) \xrightarrow{D} \chi^2(2p). \tag{2.6}$$

We thus reject $\mathcal{H}_0$ at a nominal significance level α if $\hat{T}_{1,p} > \chi^2_\alpha(2p)$, where $\chi^2_\alpha(2p)$ is the $(1-\alpha)$th quantile of the $\chi^2(2p)$ distribution.

For functional time series, the (population) scores $\xi_{ij} = \langle X_i - \mu, v_j \rangle,\ i = 1, 2, \ldots$, are correlated, so tests based on convergence (2.6) will, in general, not be usable. Under $\mathcal{H}_0$, for each i, the random variables $Z_{ij} = \lambda_j^{-1/2} \xi_{ij}$ are independent standard normal. Introduce the random vector of length $2p$ defined as

$$\mathbf{U}_i = [Z_{i1}^3, Z_{i1}^4, Z_{i2}^3, Z_{i2}^4, \ldots, Z_{ip}^3, Z_{ip}^4]^\top. \tag{2.7}$$

The stationarity of the X_i implies that the distribution of $\mathbf{U}_i$ does not depend on i. Under $\mathcal{H}_0$

$$\boldsymbol{\mu}_U = E\mathbf{U}_i = [0, 3, 0, 3, \ldots, 0, 3]^\top. \tag{2.8}$$

To construct a test statistic, set $\hat{Z}_{ij} = \hat{\lambda}_j^{-1/2} \hat{\xi}_{ij}$, where the $\hat{\lambda}_j$ are the sample eigenvalues and the $\hat{\xi}_{ij}$ are the sample scores. This allows us to define the estimated vectors $\hat{\mathbf{U}}_i$ analogous to the vectors $\mathbf{U}_i$. Denote by $\hat{\mathfrak{D}}$ a consistent estimator of the long-run covariance matrix of the vectors $\hat{\mathbf{U}}_i$. The test statistic is then

$$\hat{T}_{2,p} = \frac{1}{N} \left[\sum_{i=1}^{N} (\hat{\mathbf{U}}_i - \boldsymbol{\mu}_U) \right]^\top \left[\hat{\mathfrak{D}} \right]^{-1} \left[\sum_{i=1}^{N} (\hat{\mathbf{U}}_i - \boldsymbol{\mu}_U) \right]. \tag{2.9}$$

We reject $\mathcal{H}_0$ if $\hat{T}_{2,p} > \chi^2_\alpha(2p)$.

2.2.2. *Dynamic Functional Principal Component Analysis-Based Tests*

The above approach has dealt with the correlation of the scores of the usual FPCs by constructing a suitable quadratic form which has a chi-square limit distribution. An alternative approach is to define score-like objects which are uncorrelated. A suitable set of scores has recently been derived by Hörmann *et al.* (2015). Their application will allow us to extend the approaches of Bai and Ng (2005) and Lobato and Velasco (2004). The definition of the *dynamic scores* of Hörmann *et al.* (2015) requires defining several objects used in the spectral theory of functional time series; we present their details in Section 4.1. At this point, we merely note that the jth dynamic score of function X_i is defined by

$$Y_{ij} = \sum_{\ell \in \mathbb{Z}} \langle X_{i-\ell} - \mu, \phi_{j\ell} \rangle, \tag{2.10}$$

where the $\phi_{j\ell}$ are some deterministic functions. Therefore, if the functions X_i are jointly normal, so are their dynamic scores. Their fundamental property is

$$E[Y_{ij} Y_{i'j'}] = 0 \quad \forall i, i' \quad \text{if } j \neq j'. \tag{2.11}$$

Note that (2.11) is not true for the usual FPC scores – there it only holds for $i = i'$. It follows that under $\mathcal{H}_0$, the time series $\{Y_{ij}\},\ i = 1, 2, \ldots$, are independent. This property allows us to construct tests analogous to the approaches reviewed in Section 2.1. Their validity can be justified if the effect of the approximation by the sample analogs of the Y_{ij} is asymptotically negligible. The details and the asymptotic framework are complex, so they

DOI: 10.1111/jtsa.12281

are postponed to Sections 4.1 and B (Supporting Information). Below we define the test statistics and state their asymptotic distribution.

Suppose $\hat{Y}_{ij}$ approximate the unobservable Y_{ij} sufficiently well; a specific choice is to set $\hat{Y}_{ij} = V_{ij}$, with the V_{ij} defined by (4.1). Recall the definition of the statistic JBD_N introduced in Section 2.1. Use the subscript j to indicate that the quantities defining it are computed from the $\hat{Y}_{ij}$, $1 \le i \le N$, in place of scalar observations. With this convention, we define the test statistic

$$\widehat{T}_{3,p} = N \sum_{j=1}^{p} \left\{ \frac{\hat{\sigma}_j^6 \hat{\tau}_j^2}{\hat{\boldsymbol{\alpha}}_j \widehat{\boldsymbol{\Gamma}}_{Z,j} \hat{\boldsymbol{\alpha}}_j^\top} + \frac{\hat{\sigma}_j^8 (\hat{\kappa}_j - 3)^2}{\hat{\boldsymbol{\beta}}_j \widehat{\boldsymbol{\Gamma}}_{W,j} \hat{\boldsymbol{\beta}}_j^\top} \right\}. \tag{2.12}$$

Under general assumptions, if $\mathcal{H}_0$ holds, then $\widehat{T}_{3,p}$ converges to a chi-square distribution with $2p$ degrees of freedom. The intuitive justification is that the independence of the Y_{ij} in j is sufficiently well inherited by the $\hat{Y}_{ij}$, so the p terms defining $\widehat{T}_{3,p}$ are asymptotically independent. We thus reject $\mathcal{H}_0$ if $\widehat{T}_{3,p} > \chi^2_\alpha(2p)$.

An analogous argument utilizes the statistic LV_N. Define the test statistic as

$$\widehat{T}_{4,p} = \sum_{j=1}^{p} \mathrm{LV}_{j,N}, \tag{2.13}$$

where $\mathrm{LV}_{j,N}$ is the statistic LV_N given by (2.4) computed from the jth series $\hat{Y}_{ij}, i = 1, 2, \ldots, N$. We reject $\mathcal{H}_0$ if $\widehat{T}_{4,p} > \chi^2_\alpha(2p)$.

For ease of reference, we call the four tests introduced above T_1, T_2, T_3, and T_4 and list them in Table I. For all tests, the critical values are those of the chi-square distribution with $2p$ degrees of freedom. In our simulation study and applications, we choose p as the smallest number that explains at least 85% of the variance of the functional time series. This approach is well known when using the usual FPC scores (static scores), for example, Ramsay and Silverman (2005) or Horváth and Kokoszka (2012). It can also be defined for the dynamic scores (see Hörmann *et al.*, 2015). In Section 3, we evaluate the finite sample performance of these tests and make practical recommendations for their application.

3. FINITE SAMPLE PERFORMANCE AND APPLICATIONS

In Section 3.1, we assess the performance of the tests proposed in Section 2 by means of a simulation study. We apply them to two types of functional time series in Section 3.2.

3.1. A Simulation Study

Functional time series form a very broad class of stochastic models, and any simulation study will cover only a small part of this class. We use data generating processes (DGPs) that reflect the aspects relevant to the contribution of this article: (i) no temporal dependence versus temporal dependence and (ii) normal distribution versus non-normal distribution, with the second aspect being the focus of our study.

Table I. Summary of the four functional tests

Test	Data	PCA
T_1 (2.6)	Independent data	Static PCA
T_2 (2.9)	Dependent data	Static PCA
T_3 (2.12)	Dependent data	Dynamic PCA
T_4 (2.13)	Dependent data	Dynamic PCA

We use error functions

$$\varepsilon_n(t) = Z_{n1}\sin(\pi t) + \frac{1}{2}Z_{n2}\cos(2\pi t), \quad t \in [0,1].$$

Z_{nk} are independent in $n \geq 1$ and $k = 1, 2$. Under $\mathcal{H}_0$, they are standard normal. Under the alternative, they follow Johnson's S_U distribution (Johnson, 1949), which is often used to generate non-normal distributions. It is a four-parameter family in which skewness and kurtosis can be readily specified.

We consider two types of DGPs:

(i) IID : $X_n(t) = \varepsilon_n(t)$;
(ii) FAR : $X_n(t) = \int \varphi(t,s)X_{n-1}(s)\mathrm{d}s + \varepsilon_n(t)$.

We use the kernel $\varphi(t,s) = \alpha ts$ with $\alpha = 9/4$. Functions are generated at 75 points in [0, 1]; 15 splines of order 3 are used to convert them to functional objects.

Tables II and III report empirical size, and Tables IV and V report empirical power. Rates closest to nominal level and the highest power are highlighted by bold characters. All entries are based on 1000 replications, so the standard error for size is about 1(%).

As expected, T_1 performs very well in the IID case, but over-rejects in the FAR case. Generally, in the IID case, the three tests designed to take temporal dependence into account tend to over-reject for sample sizes $N = 150$ and $N = 450$. For sample size $N = 900$, test T_4 has the best empirical size in IID and FAR settings. Tests T_2 and T_4 perform better if temporal dependence is actually present. Based on the empirical size, T_4 emerges as the overall best choice. The empirical sizes of our tests, while not perfect, are comparable to those reported for the scalar Jarque–Bera type tests discussed in Section 2.1.

We now turn to the empirical power. The magnitude of the departure from normality was selected so that the differences in the power of the tests are visible. For larger departures, for example, heavier tails, the tests have generally higher power. For the selected distributions, for most tests, sample sizes exceeding $N = 450$ are needed to detect non-normality reliably. In the IID case, T_1 is most powerful. In the FAR case, T_4 is most powerful. The test T_1 is not far behind, but it cannot be used in the presence of temporal dependence because of it inflated size. Test T_3 is also powerful, but it may again be due to inflated size. Test T_2 test has low power for $N \leq 450$, especially if the skewness $\tau = 0$.

Based on the balance of performance with respect to size and power and temporal dependence or lack thereof, test T_4 emerges as our recommendation.

Table II. Empirical size (%) for IID functions

	$N = 150$			$N = 450$			$N = 900$		
α (%)	10	5	1	10	5	1	10	5	1
T_1	**9.3**	**6.6**	3.9	**9.6**	**5.9**	1.8	10.2	5.7	1.8
T_2	14.5	10.5	5.1	10.9	7.7	3.6	6.9	3.6	1.5
T_3	19.0	10.3	**2.4**	17.8	12.1	2.9	14.8	8.9	1.8
T_4	12.7	7.3	2.8	10.7	6.7	**1.4**	**9.9**	**5.4**	**1.5**

Table III. Empirical size (%) for the FAR series

	$N = 150$			$N = 450$			$N = 900$		
α (%)	10	5	1	10	5	1	10	5	1
T_1	14.1	9.0	4.6	17.1	11.4	5.2	18.1	11.4	4.7
T_2	6.4	4.4	2.1	7.4	3.9	2.3	7.6	4.2	2.4
T_3	16.9	8.2	**1.2**	14.4	7.7	**1.7**	14.1	8.2	2.1
T_4	**7.8**	**5.1**	2.0	**9.1**	**5.6**	2.2	**9.9**	**4.9**	**1.4**

DOI: 10.1111/jtsa.12281

Table IV. Empirical power (%) for IID functions

	$N = 150$			$N = 450$			$N = 900$		
α (%)	10	5	1	10	5	1	10	5	1
SU ($\mu = 0, \sigma = 1, \tau = 0, \kappa = 6$)									
T_1	**33.6**	**27.5**	**18.5**	**69.7**	**61.4**	**47.4**	**92.7**	**90.2**	**80.2**
T_2	9.6	7.3	3.7	31.2	19.9	18.5	53.7	33.2	27.6
T_3	25.1	16.0	5.2	48.7	31.5	24.3	63.9	48.6	32.6
T_4	14.3	9.4	4.9	43.5	29.2	23.7	65.1	50.5	34.1
SU ($\mu = 0, \sigma = 1, \tau = 1, \kappa = 6$)									
T_1	**50.9**	**43.3**	**29.9**	**96.7**	**94.8**	**88.7**	**100.0**	**100.0**	**99.9**
T_2	33.5	17.8	10.3	56.0	49.5	29.9	80.9	71.9	49.7
T_3	25.6	15.9	5.2	60.3	41.9	23.7	77.7	60.5	43.1
T_4	23.8	9.9	4.2	54.1	38.1	23.5	77.1	61.3	45.1

Table V. Empirical power (%) for FAR time series

	$N = 150$			$N = 450$			$N = 900$		
α (%)	10	5	1	10	5	1	10	5	1
SU ($\mu = 0, \sigma = 1, \tau = 0, \kappa = 6$)									
T_1	56.5	49.9	41.5	90.5	87.9	81.0	98.5	97.9	96.0
T_2	14.9	6.8	4.7	31.2	22.5	9.8	74.2	51.6	33.6
T_3	22.7	12.3	2.8	56.6	36.2	12.1	82.3	68.7	38.2
T_4	**65.2**	**59.9**	**50.4**	**98.4**	**97.7**	**94.8**	**100**	**100**	**99.9**
SU ($\mu = 0, \sigma = 1, \tau = 1, \kappa = 6$)									
T_1	75.9	70.0	57.8	98.7	98.2	96.9	99.9	99.9	99.8
T_2	15.2	9.8	7.4	52.0	31.7	24.3	83.2	67.4	45.2
T_3	58.0	41.3	14.6	96.1	92.2	75.1	99.3	98.4	95.5
T_4	**84.3**	**78.9**	**69.9**	**100.0**	**99.9**	**99.6**	**100**	**100**	**100**

The long-run covariance matrix $\widehat{\mathfrak{D}}$ in (2.9) was estimated using the function `lrvar` in the `R` package `sandwich` without prewhitening but otherwise default settings. Prewhitening does not affect size but results in slightly smaller power. (If $\mathbf{\Gamma}^{(R)}$ is the estimated long-run covariance matrix obtained as output of `lrvar`, then $\widehat{\mathbf{\Gamma}} = N\mathbf{\Gamma}^{(R)}$, where $\widehat{\mathbf{\Gamma}}$ estimates the long-run covariance matrix as defined in this article.) The dynamic scores were computed using functions `dprcomp` and `dpca.scores` in the `R` package `freqdom` (also with default settings). For the test T_4, following the recommendation of Lobato and Velasco (2004), in place of the $\hat{Y}_{ij}$ we used their standardized versions $(\hat{Y}_{ij} - \bar{Y}_j)/S_j$, where $S_j^2 = (N-1)^{-1} \sum_{i=1}^{N} (\hat{Y}_{ij} - \bar{Y}_j)^2,\ j = 1, \ldots, p$. Such a standardization improves the performance of T_4, but seemed to have no impact on the performance of T_3.

3.2. Application to Pollution and Stock Market Data

The purpose of this section is to see what conclusions our tests lead to when applied to two types of functional time series of importance. These applications also allow us to confirm some insights into their finite sample performance, which we gained in Section 3.1.

The first group of series are air quality data. The Environmental Protection Agency collects massive amounts of air quality data which are available through its website. The records consist of data for six common pollutants, collected by outdoor monitors in hundreds of locations across the United States. The number and frequency of the observations varies greatly by location, but some locations have as many as three decades worth of daily measurements. We focus on nitrogen dioxide, a common pollutant emitted by combustion engines and power stations. We consider nine locations along the east coast that have relatively complete records since 2000: Allentown, Baltimore, Boston, Harrisburg, Lancaster, New York City, Philadelphia, Pittsburgh, and Washington, DC. We use the data for the years 2000–2012. Each functional observation $X_{i,s}(t)$ consists of the daily maximum 1-h nitrogen dioxide concentration measured in parts per billion for day t, month i ($N = 156$), and at location s. We thus have a

DOI: 10.1111/jtsa.12281

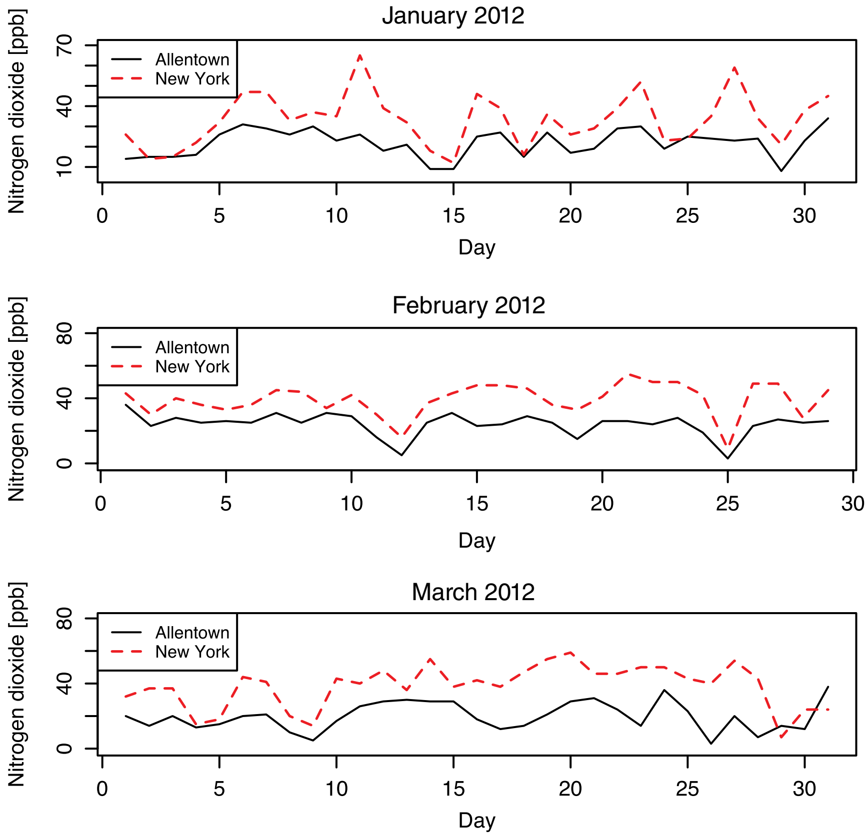

Figure 1. Maximum 1-h nitrogen dioxide curves for three consecutive months of 2012 at two locations [Color figure can be viewed at wileyonlinelibrary.com]

panel of $S = 9$ functional time series (one at every location), $X_{i,s}(t), s = 1, 2, \ldots, 9,\ i = 1, 2, \ldots, 156 = N$. Figure 1 shows the data for two locations during the first 3 months of 2012. Before the application of the test, the curves were deseasonalized by removing the monthly mean from each curve.

Table VI shows that all tests except T_1 accept the null hypothesis of normality. The test T_1 cannot be trusted in this context because the pollution curves may exhibit some form of temporal dependence, and we have seen in Section 3.1 that T_1 over-rejects if dependence is present, cf. Table III. In the presence of dependence, for the sample size $N = 156$, T_4 has reasonable power (cf. Table V), so it should detect a departure from normality. The application of our tests thus indicates that it is reasonable to assume that these curve are normal. This partially

Table VI. *p*-Values for the pollution data

	T_1	T_2	T_3	T_4
Allentown	0.001	0.190	0.403	0.564
Baltimore	0.000	0.436	0.458	0.673
Boston	0.000	0.915	0.270	0.310
Harrisburg	0.000	0.816	0.107	0.312
Lancaster	0.191	0.985	0.194	0.925
New York	0.049	0.814	0.667	0.827
Philadelphia	0.000	0.824	0.361	0.779
Pittsburgh	0.000	0.774	0.612	0.126
Washington, DC	0.017	0.245	0.864	0.848

DOI: 10.1111/jtsa.12281

justifies the tests of Constantinou *et al.* (2017), which were aimed at detecting a nonseparable spatiotemporal covariance structure, and were derived using Gaussian likelihood. On the other hand, the test of Constantinou *et al.* (2017) was derived assuming independent curves. The discussion above indicates that an extension to a setting that accounts for temporal dependence would be useful.

We now turn to an application to a stock portfolio. Cumulative intraday returns have recently been studied in several articles, including Kokoszka and Reimherr (2013b), Lucca and Moench (2015), and Kokoszka *et al.* (2015). If $P_i(t)$ is the price of a stock at minute t of the trading day i, then the cumulative intraday return curve on day i is defined by

$$R_i(t) = \log(P_i(t)) - \log(P_i(0)),$$

where time 0 corresponds to the opening of the market (9:30 EST for the New York Stock Exchange). Horváth *et al.* (2014) showed that such time series of functions are stationary. Figure 2 shows the curves R_i on three consecutive days in 2001 for two corporations whose returns we study. Specifically, we consider stock values, recorded every minute, from October 10, 2001 to April 2, 2007 (1378 trading days) for the following 10 companies: Bank of America (BOA), Citi Bank (CITI), Coca Cola (COCA), Chevron Corporation (CVX), Walt Disney Company (DIS), International Business Machines (IBM), McDonald's Corporation (MCD), Microsoft Corporation (MSFT), Walmart Stores (WMT), and Exxon Mobil Corporation Common (XOM). On each trading day, there are 390 discrete observations. There is an outlier on August 26, 2004 for BOA, which is due to a stock split. That day is discarded from further analysis, so the sample size is $N = 1377$.

Table VII. p-Values for stock data

	T_1	T_2	T_3	T_4
BOA	0.000	0.076	0.006	0.000
CITI	0.000	0.135	0.000	0.000
COCA	0.000	0.124	0.007	0.000
CVX	0.000	0.227	0.392	0.000
DIS	0.000	0.258	0.000	0.000
IBM	0.000	0.264	0.000	0.000
MCD	0.000	0.124	0.000	0.000
MSFT	0.000	0.160	0.000	0.000
WMT	0.000	0.571	0.010	0.000
XOM	0.000	0.312	0.011	0.000

Table VII shows that, except T_2, all tests reject the null hypothesis of normality (T_3 does not reject for CVX). The test T_2 has low power if skewness is close to 0, and return data are generally characterized by excess kurtosis with only small negative skewness. We, however, emphasize that these well-known properties of returns on speculative assets refer to point-to-point returns, not the curves of cumulative intraday returns, whose distributional properties have not been extensively studied. Our tests show that these curves cannot be treated as Gaussian elements in the L^2 space. We remark that it is difficult to glean from graphs like those in Figures 1 and 2 which curves might be normal and which might not. The curves in Figure 1 exhibit spiky swings, and the curves in Figure 2 resemble trajectories of the Brownian motion. However, they can exhibit large swings, mostly at the opening of the trading day, which are apparently incompatible with the normality of the Brownian motion. The latter, after a transformation to an exponential martingale, is a suitable model for annual returns. Our analysis shows that an extension of this model to continuous time is questionable, as at small time scales the exponential martingale is approximately equal to the Brownian motion. This justifies models with a Poisson component and other non-Gaussian approaches.

4. ASYMPTOTIC THEORY

The chief challenge in the development of any theory that justifies the validity of the tests introduced in Section 2 is to guarantee that the replacement of random population basis systems (FPCs and dynamic FPCs (DFPCs)) by

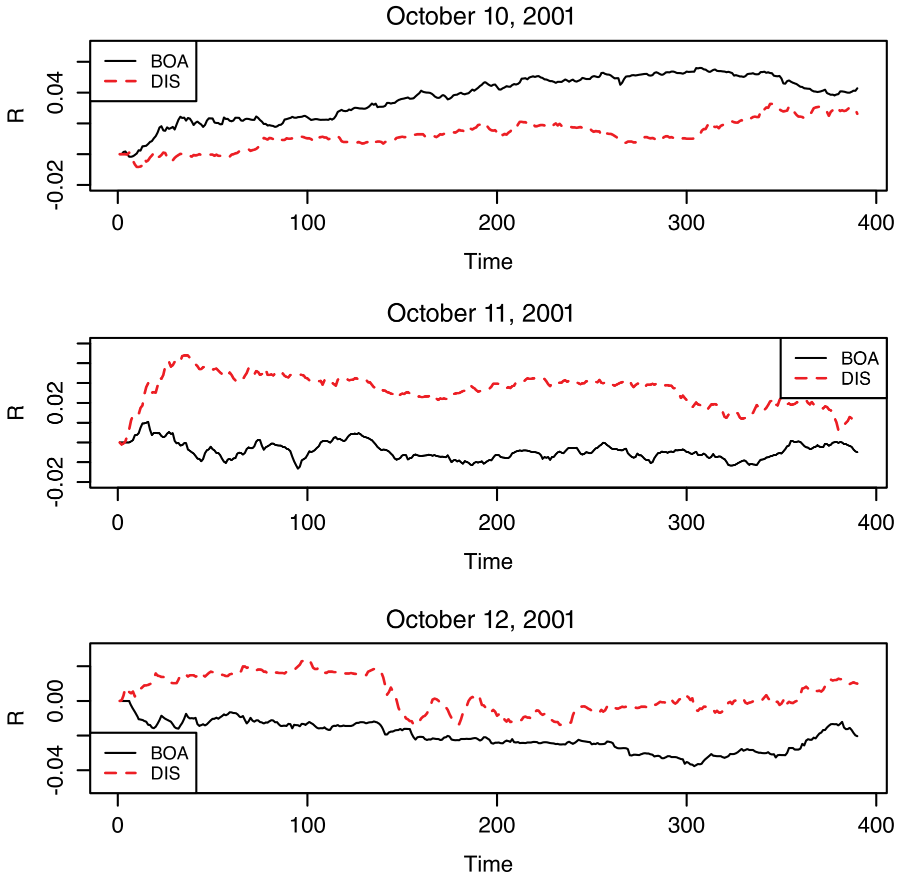

Figure 2. Cumulative intraday return curves for two companies on three consecutive days in October 2001 [Color figure can be viewed at wileyonlinelibrary.com]

their estimates does not affect the limit distributions. This is a particularly delicate task for DFPC-based tests (T_3 and T_4), where each population score is computed from a double-infinite sequence of basis functions. Justification of T_1 and T_2 is easier because the dimension reduction through the FPCs has been extensively studied, and several exiting results can be used. We will see, however, in Section D of the Supporting Information that the estimation of the mean function μ in (2.5) has a non-negligible impact on the asymptotic distribution of the sample third moments. We begin with the theory for the test T_4, which is our recommended procedure. A justification of test T_3 is similar, but in view of the weaker empirical performance being less relevant, it is not presented. Justification of the tests based on the FPCs T_1 and T_2 is presented because classical principal component analysis (PCA)-based approaches are still most popular among practitioners.

As noted in Section 2.2, the observed functions are assumed to be elements of the Hilbert space $L^2 = L^2([0, 1])$ equipped with the usual inner product and norm, denoted $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ (see, e.g. Chapter 2 of Horváth and Kokoszka, 2012).

4.1. Large-Sample Justification of Test T_4

Before proceeding with the asymptotic justifications of test T_4 defined by (2.13), we must explain in greater detail how this statistic is computed. The definitions below assume that the series $\{X_i\}$ is strictly stationary; more specific

DOI: 10.1111/jtsa.12281

conditions are stated in Assumption 4.1. The filter function ϕ_{jk} in (2.10) is defined as

$$\phi_{jk} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \varphi_j(\theta)e^{-ik\theta}\mathrm{d}\theta,$$

where $\varphi_j(\theta)$ is the jth eigenfunction of the spectral density operator

$$\mathcal{F}_\theta = \sum_h C_h e^{-ih\theta}.$$

The autocovariance operator $C_h = E[(X_h - \mu) \otimes (X_0 - \mu)]$ is defined by its action on a function $x \in L^2$: $C_h(x) = E[(X_h - \mu)\langle x, (X_0 - \mu)\rangle]$. The spectral density operator $\mathcal{F}_\theta$ can be estimated by

$$\widehat{\mathcal{F}}_\theta = \sum_{|h|\le q}\left(1 - \frac{|h|}{q}\right)\widehat{C}_h e^{-ih\theta},$$

where $\widehat{C}_h$ is the usual sample autocovariance operator at lag h, and q is a bandwidth parameter. (See Hörmann *et al.*, 2015.) This yields the estimators

$$\hat{\phi}_{jk} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{\varphi}_j(\theta)e^{-ik\theta}\mathrm{d}\theta,$$

where the $\hat{\varphi}_j(\theta)$ are the eigenfunctions of $\widehat{\mathcal{F}}_\theta$. Next, the infinite sum in (2.10) must be truncated, so we approximate the dynamic scores Y_{ij} by their sample versions

$$V_{ij} = \sum_{|\ell|<K}\langle X_{i-\ell} - \bar{X}, \hat{\phi}_{j\ell}\rangle. \tag{4.1}$$

Recommendations on how to choose the truncation level K in practice are given in Hörmann *et al.* (2015). The statistic $\widehat{T}_{4,p}$ is therefore computed using V_{ij}. Specifically, denoting by $\hat{\mu}^V_{k,j}$ the kth central sample moment of $V_{1j}, V_{2j}, \dots, V_{Nj}$, we have

$$\widehat{T}_{4,p} \approx \widehat{T}^V_{4,p} := N\sum_{j=1}^{p}\left\{\frac{[\hat{\mu}^V_{3,j}]^2}{6\widehat{F}^{(3)}_j} + \frac{(\hat{\mu}^V_{4,j} - 3[\hat{\mu}^V_{2,j}]^2)^2}{24\widehat{F}^{(4)}_j}\right\}, \tag{4.2}$$

where $\widehat{F}^{(k)}_j$ estimate $F^{(k)}_j = \sum_{h\in\mathbb{Z}}(\gamma^Y_j(h))^k$, with $\gamma^Y_j(h)$ being the lag h autocovariances of the process defined in (2.10). In principle, any consistent estimator can be used. We will discuss two possible choices:

(i) Let $\hat{\gamma}^Y_j(h)$ be the empirical ACF (autocorrelation function) obtained from $(Y_{ij} : 1 \le i \le N)$. If we are able to directly observe the Y_{ij}, then we could set

$$\widehat{F}^{(k)}_j = \{\hat{\gamma}^Y_j(0)\}^k + 2\sum_{h=1}^{N-1}\{\hat{\gamma}^Y_j(h)\}^k, \quad k = 3, 4.$$

By Lemma 1 in Lobato and Velasco (2004), these form consistent estimates of $F^{(k)}_j$. In practice, we have to replace $\hat{\gamma}^Y_j(h)$ by $\hat{\gamma}^V_j(h)$ – the empirical ACF of the process $(V_{ij} : 1 \le i \le N)$. It is tedious to show that

DOI: 10.1111/jtsa.12281

this replacement leads to a consistent estimator; lengthy arguments similar to those used in Section B of the Supporting Information would need to be developed.

(ii) An alternative estimator is given by

$$\widehat{F}_j^{(k)} = \sum_{h\in\mathbb{Z}} \{\tilde{\gamma}_j(h)\}^k, \quad k = 3, 4, \tag{4.3}$$

with $\tilde{\gamma}_j(h) = (1/2\pi)\int_{-\pi}^{\pi} \hat{\lambda}_j(\omega)e^{ih\omega}\mathrm{d}\omega$ and $\hat{\lambda}_j(\omega)$ the jth largest eigenvalue of $\widehat{\mathcal{F}}_\omega$. For this estimator, consistency is established in Section C.2 of the Supporting Information.

We now proceed with the statements of assumptions. The first assumption pertains to the dependence of the functional time series $\{X_i\}$.

Assumption 4.1. The sequence $\{X_i\}$ is a stationary Gaussian sequence satisfying

$$\sum_{h\in\mathbb{Z}} \|C_h\|_1 < \infty. \tag{4.4}$$

Here, $\|\cdot\|_1$ is the Schatten-1 norm. This assumption implies that the spectral density operator is trace-class (see Panaretos and Tavakoli, 2013).

Next, we specify conditions on the filter functions ϕ_{jk}, their estimators $\hat{\phi}_{jk}$, and the truncation level K.

Assumption 4.2. (a) The filter functions ϕ_{jk} are absolutely summable, that is, for each j, $\sum_k \|\phi_{jk}\| < \infty$. (b) Their estimators $\hat{\phi}_{jk}$ and the truncation level $K = K_N \to \infty$ satisfy

$$K \sup_{|k|\le K} \|\phi_{jk} - \hat{\phi}_{jk}\| = o_P(1), \quad 1 \le j \le p. \tag{4.5}$$

In Section C of the Supporting Information, we state a spectral domain assumption, Assumption C.1, which, together with Assumption 4.1, implies $\sum_k \|\phi_{jk}\| < \infty$. It follows from Hörmann *et al.* (2015) (see the proof of their Theorem 3) that equation (4.5) holds if Assumption C.1 holds and if the $\hat{\varphi}_j(\theta)$ are obtained from a mean square consistent estimator of the spectral density operator. However, to emphasize what is actually needed, we prefer to state (4.5) directly.

Theorem 4.1. Suppose Assumptions 4.1 and 4.2 hold. Let $\widehat{F}_j^{(k)}$ be a consistent estimator of $F_j^{(k)}$, $k = 1, 2$. Then, the statistic $\widehat{T}_{4,p}^V$ defined in (4.2) converges to the chi-square distribution with $2p$ degrees of freedom.

Theorem 4.1 is proven in Section B of the Supporting Information. In Section C of the Supporting Information, we provide further discussion concerning the consistency of the proposed test and the assumptions.

4.2. Asymptotic Justification of Tests T_1 and T_2

We work with the Karhunen–Loéve expansion

$$X_i(t) = \mu(t) + \sum_{\ell=1}^{\infty} \lambda_\ell^{1/2} Z_{i\ell} v_\ell(t), \tag{4.6}$$

  J. Time Ser. Anal. **39**: 471–487 (2018)
DOI: 10.1111/jtsa.12281

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues and $v_1, v_2, \ldots$ are the corresponding eigenfunctions of the operator defined by the covariance kernel $c(t,s) = \text{cov}(X_0(t), X_0(s))$, that is

$$\lambda_\ell v_\ell(t) = \int c(t,s) v_\ell(s) \mathrm{d}s, \quad \ell = 1, 2, \ldots .$$

The following assumption is commonly made to ensure that the largest p eigenvalues and the corresponding eigenfunctions can be consistently estimated.

Assumption 4.3. $\lambda_1 > \lambda_2 > \cdots \lambda_p > \lambda_{p+1} \geq \lambda_{p+2} \cdots$.

λ_ℓ and the v_ℓ are estimated from the sample. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2$ and $\hat{v}_1, \hat{v}_2, \ldots$ be the eigenvalues and the corresponding eigenfunctions of the empirical covariance operator

$$\hat{c}_N(t,s) = \frac{1}{N}\sum_{\ell=1}^{N}(X_\ell(t) - \bar{X}_N(t))(X_\ell(s) - \bar{X}_N(s)) \quad \text{with} \quad \bar{X}_N(t) = \frac{1}{N}\sum_{\ell=1}^{N} X_\ell(t),$$

i.e.

$$\hat{\lambda}_\ell \hat{v}_\ell(t) = \int \hat{c}_N(t,s)\hat{v}_\ell(s)\mathrm{d}s, \quad \ell = 1, 2, \ldots .$$

Define the sample versions of the $Z_{i\ell}$ by

$$\hat{Z}_{i\ell} = \hat{\lambda}_\ell^{-1/2}\langle X_i - \bar{X}_N, \hat{v}_\ell\rangle, \quad 1 \leq i \leq N,\ 1 \leq \ell \leq p.$$

Now recall the definitions of the vectors $\mathbf{U}_i \in R^{2p}$ given by (2.7). We define analogously

$$\widehat{\mathbf{U}}_i = [\hat{Z}_{i1}^3, \hat{Z}_{i1}^4, \hat{Z}_{i2}^3, \hat{Z}_{i2}^4, \ldots, \hat{Z}_{ip}^3, \hat{Z}_{ip}^4]^\top .$$

Consider the kernel estimator

$$\widehat{\mathfrak{D}}_N = \sum_{\ell=-(N-1)}^{N-1} \mathcal{K}\left(\frac{\ell}{h}\right)\mathfrak{R}_{\ell,N}, \tag{4.7}$$

where

$$\mathfrak{R}_{\ell,N} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{j=1}^{N-\ell}(\widehat{\mathbf{U}}_{j+\ell} - \bar{\mathbf{U}})(\widehat{\mathbf{U}}_j - \bar{\mathbf{U}}), & \ell \geq 0, \\ \dfrac{1}{N}\displaystyle\sum_{j=1-\ell}^{N}(\widehat{\mathbf{U}}_{j+\ell} - \bar{\mathbf{U}})(\widehat{\mathbf{U}}_j - \bar{\mathbf{U}}), & \ell < 0, \end{cases}$$

and where $\bar{\mathbf{U}} = N^{-1}\sum_{i=1}^{N}\widehat{\mathbf{U}}_i$. The kernel function $\mathcal{K}$ is assumed to satisfy

(i) $\mathcal{K}(0) = 1$.
(ii) $\mathcal{K}$ is symmetric around 0, $\mathcal{K}(u) = 0$, if $u > c$ with some $c > 0$.
(iii) $\mathcal{K}$ is Lipschitz continuous on $[-c, c]$, where c is given in (ii).

DOI: 10.1111/jtsa.12281

(iv) $h = h(N), h \to \infty$ and $h/N \to 0$.

We refer to Taniguchi and Kakizawa (2000) for examples of kernel functions.

The following theorem provides a large sample justification of test T_2.

Theorem 4.2. Suppose Assumption 4.1 holds, and the matrix $\mathfrak{D}$ defined by (D.16 in Supporting Information) is nonsingular. Then, under H_0

$$\frac{1}{N}\left(\sum_{i=1}^{N}(\widehat{\mathbf{U}}_i - \boldsymbol{\mu}_U)\right)\widehat{\mathfrak{D}}_N^{-1}\left(\sum_{i=1}^{N}(\widehat{\mathbf{U}}_i - \boldsymbol{\mu}_U)\right)^{\top} \xrightarrow{\mathcal{D}} \chi^2(2p),$$

where $\chi^2(2p)$ stands for a χ^2 random variable with $2p$ degrees of freedom.

Theorem 4.2 is proved in Section D of the Supporting Information.

If the observed functions are i.i.d., there is no need for the normalization in the quadratic form. The matrix $\mathfrak{D}$ is then diagonal with 6 and 24 as alternating diagonal elements. The proof of Theorem 4.2 thus implies the following corollary.

Corollary 4.1. If $X_1, X_2, \ldots$ are i.i.d. Gaussian random curves, then $\widehat{T}_{1,p} \xrightarrow{D} \chi^2(2p)$.

ACKNOWLEDGMENTS

This work was partially supported by United States NSF Grant DMS-1462067, Communauté francaise de Belgique, Actions de Recherche Concertées, Projects Consolidation 2016–2021, and Interuniversity Attraction Poles Programme (IAP-network P7/06) of the Belgian Science Policy Office. The data used in Section 3.2 were kindly shared with us by Panayiotis Constantinou.

SUPPORTING INFORMATION

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

REFERENCES

Anderson TW, Darling DA. 1954. A test of goodness-of-fit. *Journal of the American Statistical Association* **49**: 765–769.

Arribas-Gil A, Romo J. 2014. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics* **15**: 603–619.

Aston JAD, Pigoli D, Tavakoli S. 2017. Tests for separability in nonparametric covariance operators of random surfaces. *Annals of Statistics* **45**: 1431–1461.

Bai J, Ng S. 2005. Tests for skewness, kurtosis and normality for time series data. *Journal of Business & Economic Statistics* **23**: 49–60.

Brys G, Hubert M, Struyf A. 2004. A robustification of the Jarque–Bera test of normality. In: *COMPSTAT'2004 Symposium*. Physica-Verlag/Springer; 753–760.

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

Crainiceanu CM, Staicu A-M, Di C-Z. 2009. Generalized multilevel functional regression. *Journal of the American Statistical Association* **104**: 1550–1561.

D'Agostino RB, Belanger A, D'Agostino Jr. RB. 1990. A suggestion for using powerful and informative tests of normality. *The American Statistician* **44**: 316–321.

Doornik JA, Hansen H. 2008. An omnibus test for univariate and multi-variate normality. *Oxford Bulletin of Economics and Statistics* **70**: 927–939.

Fremdt S, Horváth L, Kokoszka P, Steinebach J. 2013. Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics* **40**: 138–152.

DOI: 10.1111/jtsa.12281

Gasser T. 1975. Goodness-of-fit tests for correlated data. *Biometrika* **62**: 563–570.
Gromenko O, Kokoszka P, Sojka J. 2017. Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *The Annals of Applied Statistics* **11**: 898–918.
Henze N, Zirkler B. 1990. A class of invariant consistent tests for multi-variate normality. *Communications in Statistics – Theory and Methods* **19**: 3595–3617. number 10.
Hörmann S, Kidzi ski L, Hallin M. 2015. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B* **77**: 319–348.
Hörmann S, Kokoszka P, Nisol G. 2018. Testing for periodicity in functional time series. *The Annals of Statistics*. Forthcoming.
Horváth L, Kokoszka P. 2012. *Inference for Functional Data with Applications*. New York: Springer.
Horváth L, Kokoszka P, Rice G. 2014. Testing stationarity of functional time series. *Journal of Econometrics* **179**: 66–82.
Jarque CM, Bera AK. 1980. Efficient tests for normality, homoskedasticity and serial independence of regression residuals. *Economic Letters* **6**: 255–259.
Jarque CM, Bera AK. 1987. A test of normality of observations and regression residual. *International Statistical Review* **55**: 163–172.
Johnson NL. 1949. Systems of frequency curves generated by methods of translation. *Biometrika* **36**: 149–176.
Kokoszka P, Reimherr M. 2013b. Predictability of shapes of intraday price curves. *The Econometrics Journal* **16**: 285–308.
Kokoszka P, Reimherr M. 2017. *Introduction to Functional Data Analysis*. Boca Raton: CRC Press.
Kokoszka P, Miao H, Zhang X. 2015. Functional dynamic factor model for intraday price curves. *Journal of Financial Econometrics* **13**: 456–477.
Kosiorowski D, Zawadzki Z. 2017. *Depthproc: an R package for robust exploration of multidimensional economic phenomena*. Kraków, Poland: Cracow University of Economics. arXiv.
Kraus D. 2015. Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B* **77**: 777–801.
Kraus D, Panaretos VM. 2012. Dispersion operators and resistant second-order analysis of functional data. *Biometrika* **99**: 813–832.
Kulperger R, Yu H. 2005. High moment partial sum processes of residuals in GARCH models and their applications. *The Annals of Statistics* **33**: 2395–2422.
Lobato I, Velasco C. 2004. A simple test of normality for time series. *Econometric Theory* **20**: 671–689.
Lomnicki Z. 1961. Tests for departure from normality in the case of linear stochastic processes. *Metrika* **4**: 37–62.
Lucca DO, Moench E. 2015. The pre-FOMC announcement drift. *The Journal of Finance* **70**: 329–371.
Mardia KV. 1970. Measures of multi-variate skewness and kurtosis with applications. *Biometrika* **57**: 519–530.
Mardia KV. 1974. Applications of some measures of multi-variate skewness and kurtosis in testing normality and robustness studies. *Sankhya (B)* **36**: 115–128.
Panaretos VM, Kraus D, Maddocks JH. 2010. Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association* **105**: 670–682.
Panaretos VM, Tavakoli S. 2013. Fourier analysis of stationary time series in function space. *The Annals of Statistics* **41**: 568–603.
Ramsay JO, Silverman BW. 2005. *Functional Data Analysis*. New York: Springer.
Royston JP. 1982. An extension of Shapiro and Wilk's w test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **31**: 115–124. number 2.
Royston JP. 1983. Some techniques for assessing multivariate normality based on the Shapiro–Wilk W. *Journal of the Royal Statistical Society: Series C* **32**: 121–133. number 2.
Royston JP. 1992. Approximating the Shapiro–Wilk W-test for non-normality. *Statistics and Computing* **2**: 117–119. number 3.
Ruppert D. 2011. *Statistics and Data Analysis for Financial Engineering*. New York: Springer.
Scholz FW, Stephens MA. 1987. K-sample Anderson–Darling tests. *Journal of the American Statistical Association* **82**: 918–924.
Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika* **52**: 591–611.
Shumway RH, Stoffer DS. 2011. *Time Series Analysis and Its Applications with R Examples*. New York: Springer.
Stephens MA. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* **69**: 730–737.
Taniguchi M, Kakizawa Y. 2000. *Asymptotic Theory of Statistical Inference for Time Series*. New York: Springer.
Yao F, Müller H-G, Wang J-L. 2005. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**: 577–590.

DOI: 10.1111/jtsa.12281