Modeling Functional Data. A Test Procedure

Enea G. Bongiorno, Aldo Goia and Philippe Vieu

November 7, 2019

Abstract The paper deals with a test procedure able to state the compatibility of observed data with a reference model, by using an estimate of the volumetric part in the small-ball probability factorization which plays the role of a real complexity index. As a preliminary by-product we state some asymptotics for a new estimator of the complexity index. A suitable test statistic is derived and, referring to the U–statistics theory, its asymptotic null distribution is obtained. A study of level and power of the test for finite sample sizes and a comparison with a competitor are carried out by Monte Carlo simulations. The test procedure is performed over a financial time series.

Keywords U-Statistics, Small–Ball Probability, Multiple test, Holm-Bonferroni procedure

1 Introduction

During the last twenty years, various theoretical results and statistical techniques have been developed to describe, analyze and modelize statistical problems involving functional data, that is, (discretized) trajectories of a random element valued in suitable functional space (for a review, see for instance Bosq, 2000; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012; Ramsay and Silverman, 2005 and, for recent developments, Goia and Vieu, 2016; Aneiros et al., 2017).

Consider a sample of observations drawn from a random process $X(t)$ defined over a compact interval that, without loss of generality, we take equal to $[0,1]$.

Enea G. Bongiorno
Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale,
E-mail: enea.bongiorno@uniupo.it

Aldo Goia
Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale,
E-mail: aldo.goia@uniupo.it

Philippe Vieu
Institut de Mathématiques de Toulouse, Université Paul Sabatier,
E-mail: philippe.vieu@math.univ-toulouse.fr

One problem in dealing with this kind of data is to model the underlying process $X(t)$ because of descriptive (e.g. dimensionality reduction) and forecasting needs (e.g. model misspecification could lead to drastically wrong predictions). A notable example often occurs in financial settings, where stock prices, interest rates, currency exchange rates, or commodity prices are modelled by means of Gaussian diffusion processes (see Fusai and Roncoroni, 2007).

In practice, starting from the sample, the goal is to identify a compatible model for $X(t)$. This problem has been faced up in the literature, in a goodness-of-fit perspective, for instance in Cuesta-Albertos et al. (2007). In Bongiorno et al. (2017) a new approach to explore the modelling nature of functional data has been introduced and discussed. It is based on the information that the small–ball probability (SmBP) $\varphi_\chi(\varepsilon)$ of X (that is the probability that X belongs to a ball of center χ and radius ε with the latter tending to zero) brings on the nature of the process. In particular, assuming that $\varphi_\chi(\varepsilon)$ can be factorized in two terms depending on the center of the ball and the radius of it and denoted by $f(\chi)$ and $\phi(\varepsilon)$ respectively, it is shown how the volumetric term $\phi(\varepsilon)$ may reveal some latent features of the process so that it could be interpreted as *complexity index* and then used in detecting a model for X.

In this paper, we develop a goodness-of-fir test procedure able to state the compatibility of observed functional data with a reference model, by using as test statistic a simple estimate of the complexity index $\phi(\varepsilon)$, which takes the form of a second order U-statistic, conveniently standardized. In practice, we implements a multiple test procedure where one compares an estimate of $\phi(\varepsilon)$ with a benchmark $\phi_0(\varepsilon)$ for some selected ε. Referring to the theory of U-statistics, the asymptotic null distribution is derived and a studentized version is introduced. To asses the performance of the test procedure for finite sample sizes, a study of level and power of the test is carried out by Monte Carlo simulations under various experimental conditions and for different processes (finite and infinite dimensional). Moreover, the empirical level and power of the proposed test are compared, by Monte Carlo experiments, with the ones obtained by the random projection test introduced in Cuesta-Albertos et al. (2007). Finally, as a by-product, it is shown how to extend directly the one-sample test procedure to the context of comparison of two samples.

The outline of the paper is the following: in Section 2 the main notations, definitions and the main theoretical results are introduced, in Section 3 the hypothesis are formalized, the test statistic is defined and its null distribution derived. Moreover, the practical implementation of the test is illustrated and two extensions are proposed: the case of two-population and the situation in which nuisance parameters appear. Section 4 collects results of numerical experiments, Section 5 provides a comparison with the competitor, whereas the last Section 6 illustrates an application to financial datasets. Mathematical aspects are collected in the Appendix.

2 Statistical background

Consider a random curve X valued in a Hilbert space $\mathcal{F}$ endowed with the L^2 norm $\|\cdot\|$. The Small Ball Probability (SmBP) $\varphi_\chi(\varepsilon)$ is the probability that X belongs to the ball of center $\chi \in \mathcal{F}$ and radius $\varepsilon > 0$ when ε is small:

$$\varphi_\chi(\varepsilon) = \mathbb{P}(\|X - \chi\| \leq \varepsilon).$$

Operatively, it is useful to assume that the SmBP satisfies the following factorization

$$\varphi_{\chi}(\varepsilon) \sim f(\chi)\,\phi(\varepsilon) \qquad \text{as } \varepsilon \to 0 \tag{1}$$

with the constraint

$$\mathbb{E}[f(X)] = 1 \tag{2}$$

to ensure the identifiability of the decomposition. The factorization (1) isolates the manner in which the SmBP depends upon χ and ε. In Bongiorno and Goia (2017) some theoretical conditions for which the factorization (1) holds are provided in a Hilbert setting and a kernel–type estimator only for the spatial factor $f(X)$ is defined: its asymptotic properties are studied and its performances in the the finite–sample case are considered. That term is used to effectively implement some density–based classification procedures in Bongiorno and Goia (2016).

To define inferential procedures (like as tests) based on the complexity index $\phi(\varepsilon)$, one firstly needs asymptotic properties of a related estimate. The literature on this topic is not so much developed and this is why, in a first step, we state a result in this sense, namely Proposition 1 that gives asymptotic normality of a nonparametric complexity index estimate. This result will serve as a crucial preliminary tool for getting asymptotic distributions of our testing procedures (see Section 3), but it should be noted that these asymptotic results could of course be used in the future for other purposes.

Given a sample $\{X_1, \ldots, X_n\}$ of i.i.d. random functions as X, an estimate of $\phi(\varepsilon)$ is given by

$$\widehat{\phi}_n(\varepsilon) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}_{\{\|X_i - X_j\| \leq \varepsilon\}} \tag{3}$$

which is a U-statistic of order two with kernel $g(x,y) = \mathbb{I}_{\{\|x-y\| \leq \varepsilon\}}$. From a practical point of view, since the functional data X_i are observed over a grid of points in $[0,1]$, the computation of the norm must be approximated by summation. It is worth to noticing that $\widehat{\phi}_n$ is invariant by deterministic translations applied to the process X.

In order to provide some properties of such an estimator, we need some technical hypothesis. In particular, as done in Ferraty et al. (2012), we assume that:

(H1) for any $\varepsilon > 0$, $\varphi_{\chi}(\varepsilon) > 0$

(H2) $\sup_{\chi \in \mathcal{F}} \left| \dfrac{\varphi_{\chi}(\varepsilon)}{\phi(\varepsilon)} - f(\chi) \right| = o(1)$

(H3) ϕ is increasing on a neighbourhood of zero, strictly positive and tends to zero as $\varepsilon \to 0$

(H4) f is bounded and $f(\chi) > 0$.

Given these, estimator (3) satisfies the following proposition.

Proposition 1 *Under (H1)–(H4) and when $\varepsilon \to 0$ the estimator $\widehat{\phi}_n(\varepsilon)$ is asymptotically unbiased with variance*

$$\operatorname{Var}\left(\widehat{\phi}_n(\varepsilon)\right) = \frac{4(n-2)}{n(n-1)}\sigma_1^2(\varepsilon) + \frac{2}{n(n-2)}\sigma_2^2(\varepsilon) \tag{4}$$

where $\sigma_1^2(\varepsilon) = \mathrm{Var}(\varphi_{X_2}(\varepsilon))$ *and* $\sigma_2^2(\varepsilon) = \mathrm{Var}\left(\mathbb{I}_{\{\|X_2 - X_1\| \leq \varepsilon\}}\right)$ *are positive and finite. Moreover, its standardized version converges in law to a standard Gaussian distribution as* $n \to \infty$.

Even if, the U-statistic (3) is built from a sample of random curves, its kernel takes its values in $\{0, 1\}$ and then $\widehat{\phi}_n(\varepsilon)$ is a real random variable for each ε. Hence, to obtain results in Proposition 1, one can evoke standard results on U–statistics (see, e.g. Lee, 1990; Lehmann, 1999): as shown in the technical details collected in the Appendix, the functional nature of data appear only along the calculations of bias and variance.

3 A general test procedure

In order to identify a compatible model for a sample drawn from $X(t)$ up to a deterministic translation, we exploit the information carried by the complexity index $\phi(\varepsilon)$. In this view, we compare the latter with a benchmark $\phi_0(\varepsilon)$. Thus, to test compatibility of the model ϕ_0 at the significance level $\alpha \in (0, 1)$ consider the following hypothesis:

$$\begin{aligned} &H_0 : \phi(\varepsilon) = \phi_0(\varepsilon) \qquad \text{for any } \varepsilon \in E \setminus \{0\} \\ &H_1 : \exists \varepsilon : \phi(\varepsilon) \neq \phi_0(\varepsilon) \end{aligned}$$

where E is a suitable right neighbourhood of zero.

From a practical point of view, it is not possible to explore the whole E and one has to fix a finite set of possible values for the radius. This leads to design a multiple test with the following hypothesis:

$$\begin{aligned} &H_0 : \phi(h) = \phi_0(h) \qquad \text{for any } h \in \mathcal{H} \\ &H_1 : \exists h \in \mathcal{H} : \phi(h) \neq \phi_0(h) \end{aligned}$$

where

$$\mathcal{H} = \{h_1, \dots, h_m\} \subset E, \qquad m \in \mathbb{N}^\star$$

being m the number of tests to be run.

The remaining part of this section is divided in four subsections: in Subsection 3.1 the test statistic is introduced and its limit distribution derived; Subsection 3.2 addresses some issues concerning the test statistic and the benchmark complexity index whereas Subsection 3.3 provides some practical details about the implementation of the algorithm; finally, a generalization to the two-sample case is presented in Subsection 3.4.

3.1 The test statistic and its limit distribution

Denote by H_0^k the k–th marginal null hypothesis, that is $H_0^k : \phi(h_k) = \phi_0(h_k)$. We define the m–dimensional set of test statistics $\{D_k^2, k = 1, \dots, m\}$, where

$$D_k^2 = \frac{\left(\widehat{\phi}_n(h_k) - \phi_0(h_k)\right)^2}{\mathrm{Var}\left(\widehat{\phi}_n(h_k)\right)}, \qquad k = 1, \dots, m. \tag{5}$$

The convergence in law result in Proposition 1 allows to derive the asymptotic null distribution of each test statistic:

Proposition 2 *For any $k = 1, \ldots, m$, under H_0^k the test statistic D_k^2 is asymptotically distributed as a chi-square distribution with one degree of freedom.*

It is worth noting that the null distribution does not depend neither from k nor from the complexity index ϕ_0 (and then, from the conjectured model). Thanks to Proposition 2, the asymptotic p-value p_k associated to H_0^k can be calculated directly as follows:

$$p_k = 1 - \mathcal{C}_1^2 \left(d_k^2 \right)$$

where $\mathcal{C}_1^2$ is the pdf of the r.v. $\chi^2 (1)$ and d_k^2 is an estimate over a sample.

Since we deal with a multiple testing procedure, the decision rule can be based on different strategies. The simplest is the well-known Bonferroni correction consisting in rejecting the null hypothesis if at least one of the p-values p_k is less than α/m. It is known that this approach has some drawbacks, in particular it conduces to a too conservative test procedure. Thus, we implement the Holm-Bonferroni method (see Holm, 1979) which is less conservative and controls the familywise error rate. According to the latter approach, the decision rule is the following: order p-values $p_{(1)} \leq \cdots \leq p_{(m)}$ and reject H_0 if $p_{(k)} \leq \alpha/(m+1-k)$ for at least one k.

To conclude this section, we prove that our test is consistent with respect to some special alternatives.

Proposition 3 *Consider the following multiple hypothesis:*

$$\begin{aligned} &H_0 : \phi(h) = \phi_0(h) && \textit{for any } h \in \mathcal{H} \\ &H_1 : \phi(h) = \phi_1(h) && \textit{with } \phi_1(h) \neq \phi_0(h) \ \textit{ for any } h \in \mathcal{H}. \end{aligned}$$

Then the test based on the statistics $D_1^2, \ldots, D_m^2$ is consistent with respect to the marginal alternatives $H_1^1, \ldots, H_1^m$.

The proof is a direct consequence of the consistency of the marginal alternatives, more details are given in Appendix.

3.2 Operationalize the test procedure

The introduced test procedure needs some adjustments in order to allow its practical applicability. The first problem concerns the nuisance parameters $\sigma_1^2(h_k)$ and $\sigma_1^2(h_k)$ contained in the variance $\text{Var}\left(\widehat{\phi}_n(h_k)\right)$ that appears in (4). Since a direct calculation is hard to obtain, it is convenient to estimate the variance following a different strategy. We propose to use the jackknife variance estimator:

$$V_n = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\phi}_n^{[-i]}(h_k) - \widehat{\phi}_n(h_k) \right)^2 \tag{6}$$

where $\widehat{\phi}_n^{[-i]}$ is the estimates of ϕ leaving out the i-th observation X_i, whose theoretical properties are widely studied in the literature (see e.g. Maesono, 1998)

The studentized version of (5), obtained by replacing $\text{Var}\left(\widehat{\phi}_n(h_k)\right)$ with its jackknife estimator V_n, satisfies the following asymptotic result.

Proposition 4 *For any $k = 1, \ldots, m$, under H_0^k the studentized version of D_k^2 is asymptotically distributed as a chi-square distribution with one degree of freedom $\chi^2(1)$.*

The latter proposition allows to compute the asymptotic p-value for each test and then to apply the Holm-Bonferroni procedure.

The second issue concerns the exact expression of $\phi_0(h)$. It is rarely available and, in such cases, often some unknown constants appear making it unusable for practical purpose. To overcome this shortcoming, ϕ_0 can be estimated from an artificial sample, with a suitable sample size n_0 and generated according to the benchmark model which is supposed to be true. As a consequence, ϕ_0 is replaced in D_k^2 with the random object $\widehat{\phi}_0$ and the variance of the test statistics is given by the sum of variances of $\widehat{\phi}_n$ and $\widehat{\phi}_0$. The new test statistics write as follows:

$$\widetilde{D}_k^2 = \frac{\left(\widehat{\phi}_n(h_k) - \widehat{\phi}_0(h_k)\right)^2}{\operatorname{Var}\left(\widehat{\phi}_n(h_k)\right) + \operatorname{Var}\left(\widehat{\phi}_0(h_k)\right)} \qquad k = 1, \ldots, m.$$

Also in this case the variance of $\widehat{\phi}_0(h_k)$ can be estimated by using the jackknife estimator (6) leading to the following asymptotic result for the studentized test statistic, obtained by replacing the variances with their jackknife estimators, and allowing to compute the asymptotic p-value for each test and then to apply the Holm-Bonferroni procedure.

Proposition 5 *For any $k = 1, \ldots, m$, under H_0^k the studentized version of the test statistic $\widetilde{D}_k^2$ is asymptotically distributed as a chi-square distribution with one degree of freedom, $\chi^2(1)$, under each marginal null hypothesis.*

3.3 Algorithm details

The testing procedure depends on many parameters: the number of marginal tests m, the values of $h_j \in \mathcal{H}$, and the size n_0 of the artificial sample used to generate the target $\phi_0(h)$.

About the choice of m, one has to balance different conflicting aspects. On the one hand, if one chooses m large, a better exploration of the range E would be possible with potential beneficial effects on the power, but it could conduce to a too conservative test (effective level is less than the nominal one) for practical purposes as known from multiple test literature, with a possible time consuming procedure. On the other hand, take $m = 1$ produces a robust test (effective level equals the nominal one) and a fast procedure, but it could reduce drastically the power. Moreover, it is reasonable to take m depending on n, since when n is small, one could obtain the same value of (3) and hence of the test statistic for different h_k. The rule of thumb we adopt is to choose m as the integer part of $\log n$ that seems to offset the different needs.

Once m is fixed, one has to select a range of values where to pick h_k. Taking into account the fact that the factorization of the SmBP holds when h tends to zero, one should take h as smaller as possible. The problem is that a too small value of h nullifies estimator (3). To avoid this, one can refer to distances $\delta_{ij} = \|X_i - X_j\|$, with $j > i$, and select h_1 and h_m as low quantiles of the set of numbers

$\varDelta = \{\delta_{ij}, i = 1, \ldots, n, j > i\}$. A solution which provides reasonably good results in simulation is to take h_1 and h_m as the quantiles of order 5% and 25% respectively. Quantiles of lower order can be taken when n is large enough. It is worth to noticing that the quality of the test results could be related to the approximation of the integral in computing the distances δ_{ij}: if the trajectories of the process are discretized over a sparse mesh, one could experience a deterioration of the test abilities. In this view it is preferable to have a process observed over a relative dense grid (our simulations suggest that a grid of 100 points on $[0, 1]$ is enough).

For what concerns n_0, that is the size of the artificial sample used to generate the target index $\phi_0(h)$, simulations suggest that $n_0 = 200$ is enough to obtain satisfying results.

3.4 A two-sample generalization

Consider two independent samples of random curves $\{X_1, \ldots, X_{n_1}\}$ and $\{Y_1, \ldots, Y_{n_2}\}$ drawn from the random processes X and Y. One may wonder if X and Y can be modelled in the same way. To do this, consider $\phi_1(\varepsilon)$ and $\phi_2(\varepsilon)$ the complexity indexes associated to X and Y respectively. The hypotheses write:

$$\begin{aligned} &H_0 : \phi_1(\varepsilon) = \phi_2(\varepsilon) \qquad \text{for any } \varepsilon \in E \setminus \{0\} \\ &H_1 : \exists \varepsilon : \phi_1(\varepsilon) \neq \phi_2(\varepsilon). \end{aligned}$$

Following similar arguments as above in this section, and using similar notations, one writes the test hypothesis:

$$\begin{aligned} &H_0 : \phi_1(h) = \phi_2(h) \qquad \text{for any } h \in \mathcal{H} \\ &H_1 : \exists h \in \mathcal{H} : \phi_1(h) \neq \phi_2(h), \end{aligned}$$

whose test statistic can be adapted from $\widetilde{D}_k^2$ to the two sample problem as follows:

$$\widetilde{D}^2_{[2],k} = \frac{\left(\widehat{\phi}_{1,n_1}(h_k) - \widehat{\phi}_{1,n_2}(h_k)\right)^2}{\mathrm{Var}\left(\widehat{\phi}_{1,n_1}(h_k)\right) + \mathrm{Var}\left(\widehat{\phi}_{1,n_2}(h_k)\right)} \qquad k = 1, \ldots, m$$

where $\widehat{\phi}_{j,n_j}(h_k)$ is the estimate (3) obtained for the j-th sample. Moreover, the following result applies when n_1 and n_2 are large enough.

Proposition 6 *For any $k = 1, \ldots, m$, under H_0^k the studentized version of $\widetilde{D}^2_{[2],k}$, obtained by replacing the variances with their jackknife estimators, is asymptotically distributed as a chi-square distribution with one degree of freedom.*

The latter result allows to compute the marginal p-values and then to apply the Holm-Bonferroni procedure.

3.5 The case of nuisance parameters

So far we have analyzed tests with simple null hypothesis. This is enough if one works with some specific random processes $X(t)$ which do not depend on real parameters, as in the Wiener case, the Brownian Bridge and the most finite dimensional processes. In this section we complete the framework by dealing with the case of composite null hypothesis.

Consider the random process $Z(t) = H(X(t), \rho)$ where H is an invertible known real function and $\rho \in \mathbb{R}^p$ $(p \geq 1)$ is a vector of parameters. An example is the Geometric Brownian Motion (GBM), which is a transformation of a Wiener process characterized by $\rho = (\mu, \sigma)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are called drift term and volatility rate respectively.

Given a sample $\{Z_1, \dots, Z_n\}$ drawn from Z, whenever ρ is known, it is sufficient to invert H and to perform the test procedure illustrated in the previous sections by using the sample $\left\{X_i = H^{-1}(Z_i, \rho), i = 1, \dots, n\right\}$.

If ρ is unknown, one has to test the composite null hypothesis $\phi(h) = \phi_0(h, \rho)$, $h \in \mathcal{H}$. In this case, as usually done in the goodness-of-fit framework, the test statistic is computed on the sample $\left\{X_i^\star = H^{-1}(Z_i, \widehat{\rho}), i = 1, \dots, n\right\}$. For instance, in the case of GBM, μ and σ^2 can be estimated by using the maximum likelihood estimators. More in general, one can suppose that each sampled curve Z_i depends on a specific parameter ρ_i so that $Z_i(t) = H(X_i(t), \rho_i)$. In such case each $\widehat{\rho}_i$ is estimated from the discretization points of Z_i.

Denote $\widehat{\phi}_n^\star(\varepsilon)$ the estimator computed over the sample $\{X_i^\star, i = 1, \dots, n\}$ and $\widehat{\phi}_n(\varepsilon)$ the corresponding one computed if ρ was known. If $\left|\widehat{\phi}_n^\star(\varepsilon) - \widehat{\phi}_n(\varepsilon)\right| = o\left(\sqrt{Var\left(\widehat{\phi}_n(\varepsilon)\right)}\right)$ for any ε, direct computations give

$$\mathbb{E}\left[\widehat{\phi}_n^\star(\varepsilon)\right] = \mathbb{E}\left[\widehat{\phi}_n(\varepsilon)\right] + o\left(\sqrt{Var\left(\widehat{\phi}_n(\varepsilon)\right)}\right)$$

and

$$Var\left(\widehat{\phi}_n^\star(\varepsilon)\right) = Var\left(\widehat{\phi}_n(\varepsilon)\right) + o\left(Var\left(\widehat{\phi}_n(\varepsilon)\right)\right).$$

Consequently, the test statistics computed using $\widehat{\phi}_n^\star$ are asymptotically equivalent to those obtained with $\widehat{\phi}_n$ and hence the statements of the propositions in the previous subsections still hold when the test statistic is computed on $\{X_i^\star, i = 1, \dots, n\}$.

4 Simulation study

The aim of this section is to explore finite sample properties of the test by evaluating the empirical level and the power for different families of processes (finite and infinite dimensional) and sample sizes.

All the experiments are conducted using the software R (R Core Team, 2013) under the following general experimental conditions:

- sample sizes $n = 25, 50, 75, 100, 150, 200$;
- number of marginal tests $m = \lfloor \log n \rfloor$ (where $\lfloor a \rfloor$ is the integer part of a);

- $\{h_1, \ldots, h_m\}$ are m equispaced points from $\max\{q_{0.05}(\varDelta), q_{0.05}(\varDelta_0)\}$ and $\min\{q_{0.25}(\varDelta), q_{0.25}(\varDelta_0)\}$ (where q_v denotes the v-quantile, $\varDelta$ and $\varDelta_0$ are the sets of distances δ_{ij}, with $j > i$, computed for the observed data and the target ones respectively);
- the targets $\phi_0(h_j)$ are computed from an artificial sample of $n_0 = 200$ curves generated according to a process whose volumetric term is the one in the null hypothesis;
- all the curves are discretized over a mesh of 100 equispaced points on $[0, 1]$.

For each of the treated family of process, the power is estimated as the proportion of times that the null hypothesis H_0 is rejected at the nominal level $\alpha = 5\%$ over 1000 Monte Carlo replications. The p-values are computed using the asymptotic null distribution.

The family of processes we deal with are the following:

Experiment 1 - Finite-dimensional processes - In the first experiment we consider finite dimensional processes X generated according to the model:

$$X(t) = \sum_{j=1}^{d} \xi_j \sqrt{\lambda_j} v_j(t) \qquad t \in [0, 1]$$

where $\xi_j \sim \mathcal{N}(0, 1)$ i.i.d., $\{v_j\}$ is the Fourier basis, and $\lambda_j = \beta^{-j}$. The aim is to evaluate the power of our test procedure when one wants to discriminate the null hypothesis $d = 3$ against the alternatives $d = 2, 4, 5$ taking $\beta = 2, 3$.

Experiment 2 - Wiener against finite-dimensional processes - Consider the process X generated according to

$$X(t) = \sqrt{2} \sum_{j=1}^{d} \xi_j \frac{2}{(2j-1)} \sin\left(\frac{1}{2}(2j-1)\pi t\right) \qquad t \in [0, 1]$$

where $\xi_j \sim \mathcal{N}(0, 1)$ i.i.d.. If $d = \infty$ this is a Wiener process. The aim is to test the null hypothesis that the process is Wiener against the alternatives $d = 3, 5, 10, 20$. The trajectories of the Wiener process are generated according to cumulative summation of independent standard Gaussian r.v.s.

Experiment 3 - Wiener against a generalized Wiener processes - Consider the Ornstein-Uhlenbeck process, a process of the family of generalized Wiener, and generated according to the following stochastic differential equation:

$$dX(t) = \theta(\mu - X(t))\, dt + \sigma dW(t) \qquad t \in [0, 1]$$

where $W(t)$ is a Wiener process and $\theta, \mu \geq 0$ and $\sigma > 0$. Set $X(0) = 0$, and $\mu = 5$, $\sigma = 1$, one tests the null hypothesis $\theta = 0$ (that is, X is Wiener) against $\theta = 0.5, 1, 3$. The Ornstein-Uhlenbeck trajectories (for $\theta > 0$) are generated by using the classical Euler-Maruyama simulation scheme (see e.g. Fusai and Roncoroni, 2007).

In figures 1 and 2 the results of above illustrated experiments are reproduced. The plots depict the behaviour of the estimated level and power varying the sample size for each considered family of processes. In all the plots, the full horizontal line indicates the nominal level of the test (that is 5%).

As expected, the test is rather conservative, and its performances improve as n increases. More in details, we note that the performances are relatively good for all the defined families of processes. In the finite dimensional case (see Figure 1), the test provides better results, also for small sample size, and with the smallest eigenvalue decay rate: if the weight λ_3 is too small, it is more difficult to discriminate the case $d = 3$ from $d = 4$. When one has to discriminate a Wiener process from finite dimensional ones (see the left panel in Figure 2), the sample size plays a central role: in fact, if one wants to distinguish an infinite dimensional process from a relatively high dimensional (but finite) one, the sample size must be large, otherwise the test often fails. Finally, for what concerns the test of Wiener against an Ornstein-Uhlenbeck process, the results (see the right panel in Figure 2) show how the power behaves varying θ: if θ is relatively small, a large sample size is necessary to discriminate the Ornstein-Uhlenbeck process from the Wiener one, whereas if θ is large the test performs well also for small sample sizes.

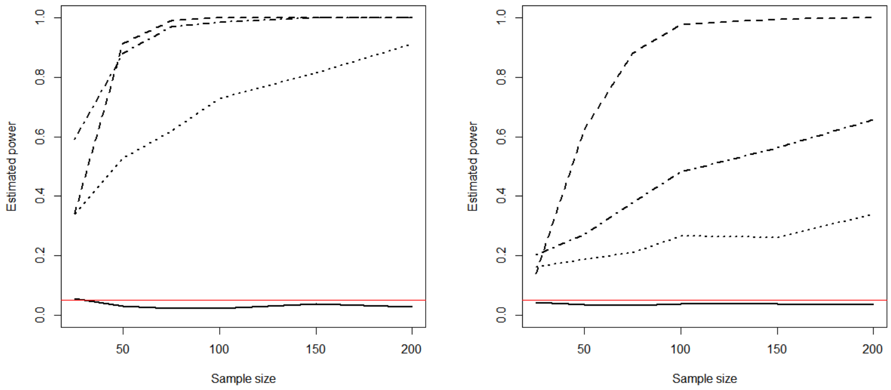

Fig. 1 Estimated power for experiment 1 with $\beta = 2$ (left panel) and $\beta = 3$ (right panel). The lines represent: $d = 3$ (level): solid line, $d = 2$: dashed line, $d = 4$: dotted line, and $d = 5$: dot-dashed lines.

5 Comparison with the random project test

Through Monte Carlo simulations, we compare the empirical level and power of our test with the random projections test proposed in Cuesta-Albertos et al. (2007). Curves are generated according to the following family of models:

$$X(t) = (1 + a_1 t^2 + a_2 \sin(2\pi t) + a_3 e^t)W(t)$$

where $W(t)$ is a Wiener Process and a_j are constants. If a_1, a_2, a_3 are null, one tests the null hypothesis that X is a Wiener Process, otherwise, one considers local alternatives.

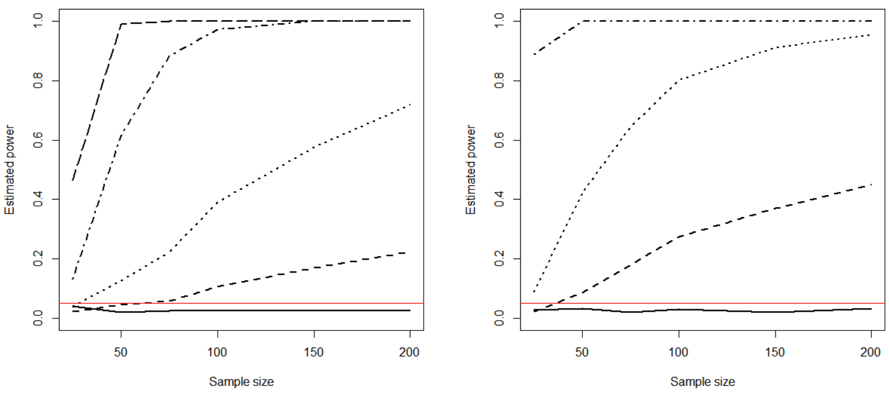

Fig. 2 Left panel - Estimated power for experiment 2. The lines represent: $d = \infty$ (level): solid line, $d = 3$: long-dashed line, $d = 5$: dot-dashed line, $d = 10$: dotted line, and $d = 20$: dashed. Right panel - Estimated power for experiment 3. The lines represent: $\theta = 0$ (level): solid line, $\theta = 0.5$: dashed line, $\theta = 1$: dotted line, and $\theta = 3$: dot-dashed lines.

The experimental setting is the following: the sample size is $n = 50, 100$, the power is estimated through 1000 Monte Carlo replications, $m = 1, 2, 3, 4$ and the h_j and $\phi_0(h_j)$ are computed as in Section 4. The nominal level is $\alpha = 5\%$.

The results are reported in Table 1 together with the ones obtained by the random projections test using k-dimensional projections with $k = 3, 5$ and $B = 200$ bootstrap iterations to compute the critical values.

In general, the test based on the complexity index performs well if compared with the random projection one. As already shown in Section 4 the first one is rather conservative, in particular for $n = 50$ and when m increases; this can be also explained by the fact that we used the asymptotic null distribution. For a lot of parameter constellations, our test is equivalent or outperforms the competitor. One exception is the case $a_1 = 0, a_2 = 1, a_3 = 0$ where the complexity index is not able to discriminate the generated process with a Wiener one: in order to obtain better results it is necessary to increase a_2. In fact for $a_1 = 0, a_2 = 1, a_3 = 0$ the estimated powers are comparable with those obtained through the random projections.

6 Application to real data

A crucial problem in finance is the modelization of stock prices time series with the aim of building models to evaluate derivatives and other contracts, which have these prices as underlying. Along the years, various approaches have been proposed: a common assumption is that the prices follow a Geometric Brownian Motion (GBM), with drift and volatility which evolve during the time (see e.g. Campbell et al., 2012 or Fusai and Roncoroni, 2007).

The verification of such hypothesis is still an open problem and only indirect empirical evidences have been provided to support it (for instance, by testing

Parameters				Our Test				Projection Test	
a_1	a_2	a_3	n	$m=1$	$m=2$	$m=3$	$m=4$	$k=3$	$k=5$
0	0	0	50	0.044	0.039	0.029	0.022	0.043	0.058
			100	0.046	0.04	0.031	0.031	0.041	0.043
1	0	0	50	1.000	1.000	1.000	1.000	0.136	0.154
			100	1.000	1.000	1.000	1.000	0.251	0.271
0	1	0	50	0.085	0.064	0.05	0.044	0.968	0.991
			100	0.079	0.063	0.042	0.036	0.993	0.999
0	2	0	50	1.000	1.000	1.000	1.000	0.991	1.000
			100	1.000	1.000	1.000	1.000	0.998	1.000
0	0	1	50	1.000	1.000	1.000	1.000	0.104	0.102
			100	1.000	1.000	1.000	1.000	0.178	0.142
0.5	1	0	50	0.966	0.938	0.928	0.918	0.823	0.86
			100	0.998	0.995	0.993	0.992	0.957	0.979
1	0	1	50	1.000	1.000	1.000	1.000	0.161	0.148
			100	1.000	1.000	1.000	1.000	0.282	0.31
0	2	1	50	1.000	1.000	1.000	1.000	0.541	0.588
			100	1.000	1.000	1.000	1.000	0.776	0.834
1	2	0.5	50	1.000	1.000	1.000	1.000	0.692	0.749
			100	1.000	1.000	1.000	1.000	0.912	0.944

Table 1 Estimated power for our test and for the random projections test under different local alternatives, with sample sizes $n = 50, 100$.

marginal Gaussianity, serial correlation of increments, and so on; see among many others, Marathe and Ryan, 2005 and Yen and Yen, 1999).

In this section we apply the procedure illustrated above in order to test if financial data are compatible with the GBM assumption. In particular, we handle the S&P500 index recorded with 1 minute frequency during the period 14th October 2016, 6th May 2017 (the dataset consists in 54810 observations).[1] The trajectory of the whole time series is depicted in the left panel of Figure 3.

To implement the test procedure, a preliminary step is to cut the time series in order to obtain a sample of functional data. Let $S(t)$ be the S&P500 observed at time t, from the time series $\{S(t_j), j = 1, \ldots, N\}$ one builds a sample of n discretized functional data X_i by dividing the interval $\mathcal{T} = [t_1, t_N]$ in n disjoint intervals $\mathcal{T}_i$ with constant width τ (a positive integer, so that $N = n\tau$) and then cutting the whole trajectory as follows:

$$X_i(t_j) = S((i-1)\tau + t_j) \qquad t_j \in [0, \tau), \ i = 1, \ldots, n.$$

If one assumes that the underlying continuous process $X(t)$ which generates data follows the GBM model, then

$$X_i(t) = X_i(0) \exp\left\{ \left(\mu_i - \frac{1}{2}\sigma_i^2 \right) t + \sigma_i W_i(t) \right\} \qquad t \in [0, \tau)$$

where μ_i and σ_i are the specific drift term and the specific volatility rate associated to $\mathcal{T}_i$ and $W(t)$ is a Wiener process. In such a way:

$$W_i(t) = \left[\log\left(X_i(t) / X_i(0) \right) - \left(\mu_i - \sigma_i^2/2 \right) t \right] / \sigma_i.$$

[1] Data have been weekly downloaded by using the link:
https://www.google.com/finance/getprices?i=60&p=200d&f=d,o,h,l,c,v&df=cpct&q=.INX

In our analysis we decided to divide the whole intervals in subintervals of 145 minutes each one, in order to have a discretization mesh dense enough. This leads to a relatively large sample of size $n = 378$. The parameters μ_i and σ_i are estimated from each curve by using the classical maximum likelihood approach. The sample of curves $W_i(t)$ obtained after these transformations is plotted in the right panel of Figure 3.

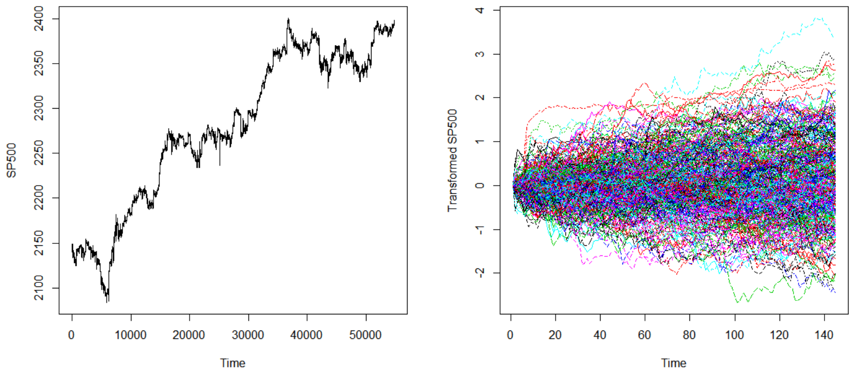

Fig. 3 Trajectory of S&P500 value from 14th October 2016 to 6th May 2017 with 1 minute frequency (left panel) and the functional sample after the transformations (right panel).

At this stage, one wants to test the null hypothesis that the underlying process $W(t)$ is Wiener. The target ϕ_0 is built by using an artificial sample of size $n_0 = 500$ (we initialize the random generator of R with the seed 1234567890), the number of marginal test m is selected according to the rule of thumb $\lfloor \log n \rfloor$ and the range of $\mathcal{H}$ is based on the quantile of order 1% and 10%.

The first line in Table 2 collects the sorted p-values: the decision (the test is conducted at the level 5%) is to accept the null hypothesis. We test also the null hypothesis that the process $W(t)$ is finite dimensional with $d = 5, 10, 15, 20$. Results in Table 2 tell us that the process can not be modelled with a finite dimensional process with dimension smaller or equal to 20. From the modelling and simulation point of view, higher dimensions are not considered since they are redundant.

	Ordered p-values					
H_0	$p_{(1)}$	$p_{(2)}$	$p_{(3)}$	$p_{(4)}$	$p_{(5)}$	Decision
Wiener	0.5499	0.5597	0.5673	0.5804	0.6133	Accept
$d = 5$	$0.0000^{(*)}$	$0.0000^{(*)}$	$0.0000^{(*)}$	$0.0000^{(*)}$	$0.0000^{(*)}$	Reject
$d = 10$	$0.0000^{(*)}$	$0.0000^{(*)}$	$0.0000^{(*)}$	$0.0001^{(*)}$	$0.0008^{(*)}$	Reject
$d = 15$	$0.0000^{(*)}$	$0.0000^{(*)}$	$0.0000^{(*)}$	$0.0003^{(*)}$	$0.0013^{(*)}$	Reject
$d = 20$	$0.0001^{(*)}$	$0.0030^{(*)}$	$0.0119^{(*)}$	0.0394	0.0867	Reject

Table 2 Estimated p-values and decision under various target model. The notation $^{(*)}$ means that $p_{(k)} \leq \alpha / (m - k - 1)$.

In order to evaluate the stability of results with respect to the way in which we built the sample of functional data, we repeated the test by using different cutting criteria: choosing 105 and 203 minutes for each subinterval (to which correspond samples of size $n = 522$ and $n = 270$ respectively). All of the explored cases present very similar result in terms of p-values (that therefore are omitted), leading to the same conclusion: the GBM assumption for the underlying process is compatible with data.

A – Appendix: Theoretical results

A.1 Proof of Proposition 1 and some further details

The proof is based on similar arguments as in Corollary 5.1 in Ferraty et al. (2012).

Bias

Compute the mean of the estimator:

$$\begin{aligned}\mathbb{E}\left[\widehat{\phi}_n(h)\right] &= \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{I}_{\{\|X_i-X_j\|\leq h\}}\right]\\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right]\\ &= \mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right].\end{aligned}$$

Using the law of total expectation, one has

$$\mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right] = \mathbb{E}\left[\ \mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}|X_2\right]\ \right] = \mathbb{E}\left[\varphi_{X_2}(h)\right].$$

Thanks to (H2) and the constraint (2) it follows

$$\mathbb{E}\left[\varphi_{X_2}(h)\right] = \left(\mathbb{E}\left[f(X_2)\right] + o(1)\right)\phi(h) = \phi(h) + o(\phi(h)). \tag{7}$$

Combining the results one gets

$$\mathbb{E}\left[\widehat{\phi}_n(h)\right] = \phi(h) + o(\phi(h)) \tag{8}$$

that allows to conclude that the estimator is unbiased when $h \to 0$ and $n \to \infty$.

Variance

By using classical results on U-statistics (see e.g. Lehmann, 1999, Theorem 6.1.1) it follows

$$\mathrm{Var}\left(\widehat{\phi}_n(h)\right) = \frac{4(n-2)}{n(n-1)}\sigma_1^2(h) + \frac{2}{n(n-2)}\sigma_2^2(h)$$

where

$\sigma_1^2(h) = \mathrm{Var}\left(\mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}|X_2\right]\right) = \mathrm{Var}\left(\varphi_{X_2}(h)\right),$
$\sigma_2^2(h) = \mathrm{Var}\left(\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right).$

Consider the second term $\sigma_2^2(h)$, one has

$$\mathrm{Var}\left(\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right) = \mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right]\left(1-\mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right]\right).$$

Since (see equation (7))

$$\mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}|X_2\right]\right] = \mathbb{E}\left[\varphi_{X_2}(h)\right] = \phi(h) + o(\phi(h))$$

it follows

$$\sigma_2^2(h) = (\phi(h)+o(\phi(h)))(1-\phi(h)-o(\phi(h))) = \phi(h)+o(\phi(h)).$$

About the first term $\sigma_1^2(h)$, one has

$$\mathrm{Var}\left(\varphi_{X_2}(h)\right) = \mathbb{E}\left[\varphi_{X_2}^2(h)\right] - \mathbb{E}^2\left[\varphi_{X_2}(h)\right]$$

where

$$\mathbb{E}^2\left[\varphi_{X_2}(h)\right] = \phi^2(h) + o\left(\phi^2(h)\right).$$

Since

$$\begin{aligned}\varphi_{X_2}^2(h) &= \mathbb{E}^2\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}|X_2\right] \leq \mathbb{E}\left[\mathbb{I}^2_{\{\|X_2-X_1\|\leq h\}}|X_2\right]\\ &= \mathbb{E}\left[\mathbb{I}_{\{\|X_2-X_1\|\leq h\}}|X_2\right] = \varphi_{X_2}(h)\end{aligned}$$

then

$$\mathbb{E}\left[\varphi_{X_2}^2(h)\right] \leq \mathbb{E}\left[\varphi_{X_2}(h)\right] = \phi(h)+o(\phi(h)).$$

Concluding, there exist two finite positive constants c_1 and c_2 depending on h, such that

$$\mathrm{Var}\left(\widehat{\phi}_n(h)\right) \leq \frac{4(n-2)}{n(n-1)}c_1(h) + \frac{2}{n(n-2)}c_2(h)$$

and then

$$\mathrm{Var}\left(\widehat{\phi}_n(h)\right) = O\left(\frac{1}{n}\right). \tag{9}$$

Asymptotic distribution

Using classical asymptotic results on U-statistics (see e.g. Lehmann, 1999, theorems 3.3.1 and 6.1.2) since $0<\sigma_1^2(h)<\infty$ and $0<\sigma_2^2(h)<\infty$ (thanks to (H3) and results above), one gets, for $h\to 0$, and $n\to\infty$,

$$\frac{\widehat{\phi}_n(h) - \mathbb{E}\left[\widehat{\phi}_n(h)\right]}{\sqrt{\mathrm{Var}\left(\widehat{\phi}_n(h)\right)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{10}$$

It is worth to noting that combining (10) with (8) we get, as $n\to\infty$,

$$\frac{\widehat{\phi}_n(h) - \phi(h)}{\sqrt{\mathrm{Var}\left(\widehat{\phi}_n(h)\right)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{11}$$

A.2 Proof of propositions 2, 4, 5 and 6

For what concerns Proposition 2, the result is a consequence of asymptotic normality of the estimator $\widehat{\phi}_n$ and its unbiasness (8). In particular, for any $k=1,\ldots,m$, under the marginal null hypothesis H_0^k, when $h\to 0$ from (11) it follows:

$$\frac{\left(\widehat{\phi}_n(h) - \phi_0(h)\right)^2}{\mathrm{Var}\left(\widehat{\phi}_n(h)\right)} \xrightarrow{d} \chi^2(1) \qquad \text{as } n\to\infty.$$

About the statements in propositions 4, 5 and 6, the estimators consistency, their asymptotic unbiasness and normality, together with consistency of the jackknife variance estimators allow to invoke Slustky's Theorem and lead to the results (see e.g. Maesono (1995)).

Proof of Proposition 3

Recalling that a multiple test is consistent w.r.t. the marginal alternatives $H_1^1, \dots, H_1^m$ if each marginal test is consistent (see e.g. Alt, 2005), then we have to prove that, under each H_1^k one has

$$D_k^2 \longrightarrow +\infty \qquad \text{in probability as } n \to \infty.$$

Observe that for any k,

$$D_k^2 = \left(\frac{\widehat{\phi}_n(h) - \phi_1(h)}{\operatorname{Var}\left(\widehat{\phi}_n(h)\right)} + \frac{\phi_1(h) - \phi_0(h)}{\operatorname{Var}\left(\widehat{\phi}_n(h)\right)} \right)^2 = (A_n + B_n)^2 .$$

On the one hand, under each H_1^k, thanks to (11), the sequence of random variables A_n is bounded in probability. On the other hand, under each H_1^k, thanks to (9), the deterministic sequence B_n diverges with n. The conclusion follows immediately.

Acknowledgements

The Authors thank an Associate Editor and two anonymous referees for their valuable remarks and comments that allowed to improve the presentation of the paper. Special thanks go to Professor J. A. Cuesta Albertos, Universidad de Cantabria, who provided MatLab code for simulations in Section 5. E. Bongiorno and A. Goia are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The financial support of CRoNoS - COST Action IC1408 is acknowledged by the first author.

References

Alt, R., 2005. Multiple hypothesis testing in linear regression model with applications to economics and finance. Curvillier, Verlag.

Aneiros, G., Bongiorno, E. G., Cao, R., Vieu, P., 2017. Functional Statistics and Related Fields. Springer.

Bongiorno, E. G., Goia, A., 2016. Classification methods for hilbert data based on surrogate density. Comput. Statist. Data Anal. 99, 204 – 222.

Bongiorno, E. G., Goia, A., 2017. Some insights about the small ball probability factorization for Hilbert random elements. Statist. Sinica 27, 1949–1965.

Bongiorno, E. G., Goia, A., Vieu, P., 2017. Evaluating the complexity of functional data. Preprint.

Bosq, D., 2000. Linear processes in function spaces. Vol. 149 of Lecture Notes in Statistics. Springer-Verlag, New York.

Campbell, J. Y., Lo, A. W.-C., MacKinlay, A. C., 2012. The econometrics of financial markets. princeton University press.

Cuesta-Albertos, J. A., del Barrio, E., Fraiman, R., Matrán, C., 2007. The random projection method in goodness of fit for functional data. Comput. Statist. Data Anal. 51 (10), 4814–4831.

Ferraty, F., Kudraszow, N., Vieu, P., 2012. Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. J. Nonparametr. Stat. 24 (2), 447–464.

Ferraty, F., Vieu, P., 2006. Nonparametric functional data analysis. Springer Series in Statistics. Springer, New York.

Fusai, G., Roncoroni, A., 2007. Implementing models in quantitative finance: methods and cases. Springer Science & Business Media.

Goia, A., Vieu, P., 2016. An introduction to recent advances in high/infinite dimensional statistics [Editorial]. J. Multivariate Anal. 146, 1–6.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6 (2), 65–70.

Horváth, L., Kokoszka, P., 2012. Inference for functional data with applications. Springer Series in Statistics. Springer, New York.

Lee, J., 1990. U-statistics: Theory and Practice. Citeseer.

Lehmann, E. L., 1999. Elements of large-sample theory. Springer Science & Business Media.

Maesono, Y., 1995. On the normal approximations of Studentized U-statistic. J. Japan Statist. Soc. 25 (1), 19–33.

Maesono, Y., 1998. Asymptotic mean square errors of variance estimators for U-statistics and their Edgeworth expansions. J. Japan Statist. Soc. 28 (1), 1–19.

Marathe, R. R., Ryan, S. M., 2005. On the validity of the geometric Brownian motion assumption. The Engineering Economist 50 (2), 159–192.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O., Silverman, B. W., 2005. Functional data analysis, 2nd Edition. Springer Series in Statistics. Springer, New York.

Yen, G., Yen, E. C., 1999. On the validity of the Wiener process assumption in option pricing models: Contradictory evidence from Taiwan. Review of Quantitative Finance and Accounting 12 (4), 327–340.