SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 9–40
doi:https://doi.org/10.37830/SJS.2022.1.02

INVITED ARTICLE

A review on specification tests for models with functional data

Wenceslao González-Manteiga
Centre for Mathematical Research and Technology Transfer of Galicia (CITMAga).
Department of Statistics, Mathematical Analysis and Optimization.
Universidad de Santiago de Compostela, Santiago de Compostela,
wenceslao.gonzalez@usc.es

Received: Novermber 30, 2022. Accepted: December 14, 2022

Abstract: Nowadays, due to the progress in technological advances, massive amounts of data are generated. As a result, new statistical methodology is needed to properly manage this information. The functional data are an example of special importance. These are mainly obtained by means of high-frequency measurements (spectrometric curves, stock prices recording, etc.). Since the beginning of this century, this type of data has achieved great popularity. This fact has generated new distribution or regression models, among others, appropriate to the functional context. In the last 10 years, novel specification tests are proposed for those models. These are generalizations of methodologies developed for the vectorial framework over the last century. Besides, innovative procedures based on distance correlation ideas have been proposed as well. This article reviews the most notable developments in this context, providing some illustrations from real data sets.

Keywords: distance correlation, functional data, goodness-of-fit, regression models

MSC: 62R10, 62G10

1 Introduction

The invention of computers meant a real change in statistical methodology in the last century. The scientific developments, derived mainly from the first half of the 20th century, were headed to understand existing real data sets information. These were of medium size, obtained wit a great effort in many cases. Other developments in Statistics during the first part of the 20th century were leaded to the design of algorithms for the estimation and testing of different models parameters. In all of these, the computational burden was considerable for the available and quite limited calculus capacity of that moment. Real parameters were estimated, but not curves due to poor graphics resources. The behavior of the statistics distributions were analyzed, under some parametric hypoth-

esis, because of the obvious impossibility of working with large sample sizes in a nonparametric way.

In the 80s, the versatility provided by advances in computer calculus generated new statistical procedures. These are based on simulating artificial data, as is the case of "Bootstrap". Nevertheless, it is not until the next decades, motivated by Internet use, new technologies information, development of distributed as well as parallelized computing, and computational costs reduction for storage and data processing, when the beginning of the "Big Data" age can be established. This phenomenon has a great impact on the development of modern technology in Statistics as well as on all its applications.

Currently, many companies already have continuous and real-time monitoring systems: stock quotes can be measured as high-frequency data, information generated by web pages, social media data or just the credit cards transactions are some examples of massive information generation sources. Other example can be found in the electric market, where high-frequency measures about energy consumption or demand are available as well. In all these cases it is quite relevant to be able to correctly process and control the information.

It is, precisely, in this context of massive, high-frequency or related data, where functional data arise. This kind of data gains an immense popularity with Ramsay and Silverman (2005), Ferraty and Vieu (2006) or more recently with Horváth and Kokoszka (2012), Hsing and Eubank (2015) and Kokoszka and Reimherr (2017), among others. Functional data allows to summary a great amount of information through a curve, surface or, in general, using a "statistical object". This last is typically modeled in a functional space, such as Hilbert spaces.

The management of functional data guide us, naturally, to the consideration of models based on these (distribution models, regression models, etc.), employed for prediction purposes, using interpretation of the results in diverse applications. Thus, the necessity of mechanism for specification testing devoted to models with functional data appears. In this article, the diverse procedures that have emerged during this period are reviewed. These have been mainly developed in the last 10 years, generalizing classic procedures introduced in the first half of the 20th century, essentially based on the empirical distribution, and the more recent advances in the last part of the 20th century making use of nonparametric estimations of the density or regression function.

Although the Goodness-of-Fit (GoF) term is due to Pearson, at the beginning of the 20th century, it is not until the 70s with Durbin (1973) and Bickel and Rosenblatt (1973) where modern specification tests start. These are based on distances between nonparametric estimators of the distribution or density function with respect to hypothetical estimations under the null hypothesis of the model.

In this way, formally, assume that $\{X_1, \ldots, X_n\}$ is an identically and independent distributed (iid) sample of a random variable X with (unknown) distribution F (or density f, if that is the case). If the target function is the distribution F, then the GoF testing problem can be formulated as testing $H_0 : F \in \mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ vs. $H_1 : F \notin \mathcal{F}_\Theta$, where $\mathcal{F}_\Theta$ stands for a parametric family of distributions indexed in some finite-dimensional set Θ. A general test statistic for this problem can be written as $T_n = T(F_n, F_{\hat{\theta}})$, with the functional T denoting, here and henceforth, some kind of distance between a nonparametric estimate, given in this case by the mentioned empirical cumulative distribution function $F_n(x) = n^{-1} \sum_{i=1}^n \mathbb{I}(X_i \leq x)$, and an estimate obtained under the null hypothesis H_0, $F_{\hat{\theta}}$ in this case. Similarly, for the case of a parametric density model,

the testing problem is formulated as $H_0 : f \in f_\Theta = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ vs. $H_1 : f \notin f_\Theta$ and can be approached with the general test statistic $T_n = T(f_{nh}, f_{\hat{\theta}})$. In this setting, $f_{\hat{\theta}}$ is the density estimate under H_0 and f_{nh} denotes a general nonparametric density estimate, as for example, the kernel density estimator $f_{nh}(x) = n^{-1}\sum_{i=1}^n K_h(x - X_i)$ introduced by Parzen (1962) and Rosenblatt (1956) where $K_h(\cdot) = K(\cdot/h)/h$, K is the kernel function $\left(K(x) \geq 0 \text{ and } \int K(x)dx = 1\right)$, and h is the smoothing bandwidth.

More recent procedures were generalized to the context of regression models in the 1990s. Consider a nonparametric, random design, regression model such that $Y = m(X) + \varepsilon$, with $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, $m(x) = \mathbb{E}[Y|X = x]$ and $\mathbb{E}[\varepsilon|X = x] = 0$. Denote by $\{(X_i, Y_i)\}_{i=1}^n$ an iid sample of (X, Y) satisfying such a model. In this context, the GoF goal is to test $H_0 : m \in \mathcal{M}_\Theta = \{m_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ vs. $H_1 : m \notin \mathcal{M}_\Theta$, where $\mathcal{M}_\Theta$ represents a parametric family of regression functions indexed in Θ. Following Durbin (1973) and Bickel and Rosenblatt (1973) ideas, the seminal works of Stute (1997) and Härdle and Mammen (1993), respectively, introduced two types of GoF tests for regression models:

a) Tests based on empirical regression processes, considering distances between estimates of the integrated regression function $I(x) = \int_{-\infty}^x m(t)\,\mathrm{d}F(t)$ (F being the marginal distribution of X under H_0 and H_1). Specifically, the test statistics are constructed as $T_n = T(I_n, I_{\hat{\theta}})$, with $I_n(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x)Y_i$ and $I_{\hat{\theta}}(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x)m_{\hat{\theta}}(X_i)$.
b) Smoothing-based tests, using distances between estimated regression functions, $T_n = T(m_{nh}, m_{\hat{\theta}})$, with m_{nh} a smooth regression estimator. As a particular case, $m_{nh}(x) = \sum_{i=1}^n W_{nh,i}(x)Y_i$, with $W_{nh,i}(x)$ some weights depending on a smoothing parameter h. Such an estimator can be obtained, for example, with Nadaraya–Watson or local linear weights (see, e.g., Wand and Jones (1995)).

A complete review of these methodologies, related to specification tests, can be consulted in González-Manteiga and Crujeiras (2013). This is an invited article with discussion for the TEST journal. In this reference, different contributions on this topic since 1990 are reviewed. Resulting statistics for diverse specification tests as well as its distribution calibration, by means of asymptotic techniques or resampling procedures like the Bootstrap, are studied. This review analyzes more than 20 years of developments, being very scarce or almost non-existent the procedures designed for functional data. Very recently, in a chapter of the book González-Manteiga et al. (2022), we perform a review of the existing methodologies for specification tests in the functional data context. These procedures are mainly based on extensions of the methodology introduced in the 90s for specification tests in the vectorial framework to the functional case.

In this paper, an update of the chapter corresponding to the book González-Manteiga et al. (2022) is provided in the next section. Later, in Section 3, the "fundamental case" of the manuscript is presented. This is covered with a detailed review of specification tests based on "distance correlation" ideas and their novel extension to the functional data context. In Section 4 some applications to real data sets for specification testing in the functional framework, applying techniques introduced in previous sections, are displayed. Finally, some conclusions arise in Section 5 and the document finishes with an exhaustive revision of relevant references.

2 Testing specification models for functional data using smoothing or empirical processes

In this section we review the most notable results for specification tests in terms of the distribution function or regression models in the functional data context. For the development of these procedures it is necessary to include functional data in complex structures associated to general spaces (metric or topological ones), as Hilbert spaces. These represent a natural and quite employed way for adequate model description in the functional context.

2.1 GoF for distribution models for functional data

Let $\mathcal{H}$ denote a Hilbert space over $\mathbb{R}$, the norm of which is given by its scalar product as $\|x\| = \sqrt{\langle x, x\rangle}$. Consider $\{X_1, \ldots, X_n\}$ iid copies of the random variable $X : (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$, with $(\Omega, \mathcal{A}, \mathbb{P})$ the probability space where the random sample is defined and $\mathcal{B}(\mathcal{H})$ the Borel σ-field on $\mathcal{H}$. The general GoF problem for the distribution of X consists on testing $H_0 : \mathrm{P}_X \in \mathcal{P}_\Theta = \{\mathrm{P}_\theta : \theta \in \Theta\}$ vs. $H_0 : \mathrm{P}_X \notin \mathcal{P}_\Theta$, where $\mathcal{P}_\Theta$ is a class of probability measures on $\mathcal{H}$ indexed in a parameter set Θ, now possibly infinite-dimensional, and P_X is the (unknown) probability distribution of X induced over $\mathcal{H}$.

When the goal is to test the simple null hypothesis $H_0 : \mathrm{P}_X \in \{\mathrm{P}_0\}$, a general feasible approach that enables the construction of different test statistics is based on projections $\pi : \mathcal{H} \to \mathbb{R}$, in such a way that the test statistics are defined from the projected sample $\{\pi(X_1), \ldots, \pi(X_n)\}$. Such an approach can be taken on the projected distribution function: $T_{n,\pi} = T(F_{n,\pi}, F_{0,\pi})$ with $F_{n,\pi}(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(\pi(X_i) \leq x)$ and $F_{0,\pi}(x) = \mathbb{P}_{H_0}(\pi(X) \leq x)$. Some specific examples are given by the adaptation to this context of the Kolmogorov–Smirnov, Cramer–von Mises, or Anderson–Darling type tests. As an alternative, based on smoothing techniques tests presented in Section 1, a test statistic can also be built as $T_{n,\pi} = T(f_{nh,\pi}, \mathbb{E}_{H_0}[f_{nh,\pi}])$ with $f_{nh,\pi}(x) = n^{-1}\sum_{i=1}^n K_h(x - \pi(X_i))$ the density estimate of $\pi(x)$. It should be also noted that, when embracing the projection approach, the test statistic may take into account 'all' the projections within a certain space, e.g. by considering $T_n = \int T_{n,\pi}\, \mathrm{d}W(\pi)$ for W a probability measure on the space of the different projections, or take just $T_n = T_{n,\hat{\pi}}$ with $\hat{\pi}$ being a randomly-sampled projection from a certain non-degenerate probability measure W.

Now, when the goal is to test the composite null hypothesis $H_0 : \mathrm{P}_X \in \mathcal{P}_\Theta$, the previous generic approaches are still valid if replacing $\mathrm{P}_{0,\pi}(x)$ with $\mathrm{P}_{\hat{\theta},\pi}(x) = \mathbb{P}_{\mathrm{P}_{\hat{\theta}}}(\pi(X) \leq x)$. Cuesta-Albertos et al. (2006) and Cuesta-Albertos et al. (2007) provide a characterization of the composite null hypothesis by means of random projections, and provide a bootstrap procedure for calibration, see also Bugni et al. (2009) and Ditzhaus and Gaigall (2018). In the space of real square-integrable functions $\mathcal{H} = L^2[0,1]$, one may take $\pi_h(x) = \langle x, h\rangle$, with $h \in \mathcal{H}$. The previous references provide also some approaches for the calibration of the tests under the null hypothesis of the rejection region $\{T_n > c_\alpha\}$, where $\mathbb{P}(T_n > c_\alpha) \leq \alpha$.

A very relevant alternative to the procedures based on projections is the use of the so-called "energy statistics" Székely and Rizzo (2017). Working with $\mathcal{H}$ a general Hilbert separable space (as it can be seen in Lyons (2013)) if $X \sim \mathrm{P}_X$ and $Y \sim \mathrm{P}_Y = \mathrm{P}_0$ (P_0 being the distribution under the null)

then

$$E = E(X,Y) = 2\mathbb{E}[\|X - Y\|] - \mathbb{E}[\|X - X'\|] - \mathbb{E}[\|Y - Y'\|] \geq 0, \tag{1}$$

with $\{X, X'\}$ and $\{Y, Y'\}$ iid copies of the variables with distributions P_X and P_Y, respectively. Importantly, (1) equals 0 if and only if $\mathrm{P}_X = \mathrm{P}_Y$, a characterization that serves as basis for a GoF test. See the nice review of Székely and Rizzo (2017), where a motivation is given for the duality between the expression displayed in (1) and the well-known energy formula of Einstein.

The energy statistic in (1) can be empirically estimated from a sample $\{X_1, \ldots, X_n\}$ as

$$\hat{E}^* = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_i - Y_j^*\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_i - X_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|Y_i^* - Y_j^*\|,$$

The distribution of $\hat{E}^*$, $\mathbb{P}^*\left\{\hat{E}^* \leq x\right\}$ can be approximated by simulation of the artificial variable $Y^* \sim P_Y$, resulting in $\left\{Y_1^{*b}, \ldots, Y_n^{*b}\right\}$ with $b = 1, \ldots, B$. The critic point for a given α level can be obtained in a natural way from the quantiles of the sorted sample: $\hat{E}^{*(1)} \leq \cdots \leq \hat{E}^{*(B)}$ as a result of the Monte Carlo replicates of the artificial samples.

The most studied case is the Gaussian one, where P_Y follows the distribution of a Gaussian process. This is analyzed in recent literature as in Kellner and Celisse (2019), Kolkiewicz et al. (2021), Górecki and Łukasz (2019), Henze and Jiménez-Gamero (2021) and Bongiorno et al. (2019). In these works, diverse alternatives to the mentioned procedures are provided and reviewed to specification of a functional model of Gaussian distribution or related.

Finally, in the context of tests for distributions, it is worth it to mention the related two-sample problem, a common offspring of the simple-hypothesis one-sample GoF problem. This topic has been extensively studied for scalar random data in the last decades. However, the situation involving functional random data has attracted less attention until now. Three main related approaches have been considered in this setting recently, namely,

a) Comparison of functional means using, e.g., principal component approaches (Horváth and Rice (2015), Ghale-Joogh and E. Hosseini-Nasab (2018)) or adapting the ideas of the F-test to the functional context (Cuevas et al. (2004), González-Rodrígez et al. (2012), Górecki and Łukasz (2019), Lee et al. (2015), Zhang and Liang (2014), Qiu et al. (2021)).

b) Comparison of covariance structures (Boente et al. (2018), Fremdt et al. (2013), Guo et al. (2018), Guo et al. (2019), Guo et al. (2019)).

c) Comparison of the distribution structure in various ways. Tests based on smoothing discrete observed data in potential functional data are developed in Bárcenas et al. (2017) and, similarly, in Estévez-Pérez and Vilar (2013) or Pomann et al. (2016). Empirical processes have been used in Bárcenas et al. (2017). An L^2-type criterion based on empirical distribution functions is used in Jiang et al. (2019). Some Cramér-von Mises-type statistics adapted to the functional case are employed in Bugni and Horowitz (2021).

2.2 GoF for regression models with functional data based on smoothing or empirical processes

We assume in the following, for easier presentation of the different methods, that both the predictor X and response Y are centered, so that the intercepts of the linear functional regression models are null.

A particular case of a regression model with functional predictor and scalar response is the so-called functional linear model. For $\mathcal{H}_X = L^2[0,1]$, this parametric model is given by

$$Y = m_\beta(X) + \varepsilon, \quad m_\beta(x) = \langle x, \beta \rangle = \int_0^1 x(t)\beta(t)\,\mathrm{d}t, \tag{2}$$

for some unknown $\beta \in \mathcal{H}_X$ indexing the functional form of the model and $\mathbb{E}\left[\varepsilon|_X\right] = 0$. This model is the natural generalization of the classical and popular linear (Euclidean) regression models.

In general, there have been two approaches for the inference on (2): (*i*) testing the significance of the trend within the linear model, i.e., testing $H_0 : m \in \{m_{\beta_0}\}$ vs. $H_1 : m \in \{m_\beta : \beta \in \mathcal{H}_X, \beta \neq \beta_0\}$, usually with $\beta_0 = 0$; (*ii*) testing the linearity of m, i.e., testing $H_0 : m \in \mathcal{L} = \{m_\beta : \beta \in \mathcal{H}_X\}$ vs. $H_1 : m \notin \mathcal{L}$.

For the GoF testing problem presented in (*ii*), given an iid sample $\{(X_i, Y_i)\}_{i=1}^n$ and following Härdle and Mammen (1993) ideas in the vectorial case, a test statistic structure can be given by $T_n = T(m_{nh}, m_{\hat{\beta}})$, where $\hat{\beta}$ is a suitable estimator for β and

$$m_{nh}(x) = \sum_{i=1}^n W_{ni}(x) Y_i = \sum_{i=1}^n \frac{K_h(\|x - X_i\|)}{\sum_{j=1}^n K_h(\|x - X_j\|)} Y_i \tag{3}$$

is the Nadaraya–Watson estimator with a functional predictor. In Delsol et al. (2011), a L_2 distance is offered,

$$T_n = \int \left(m_{nh}(x) - m_{nh,\hat{\beta}}(x)\right)^2 \omega(x)\,\mathrm{dP}_X(x),$$

where $m_{nh,\hat{\beta}}$ is a smoothed version of the parametric estimator that follows by replacing Y_i with $m_{\hat{\beta}}(X_i)$ in (3). A crucial problem is the computation of the critical region $\{T_n > c_\alpha\}$, which depends on the selection of h when a class of estimators for β is used under the null. This class of smoothed-based tests, or related, were deeply studied in the Euclidean setting (see González-Manteiga and Crujeiras (2013)). Nevertheless, this is not the case in the functional context, except for this mentioned contribution and others more recent by Maistre and Patilea (2020) and Patilea and Sánchez-Sellero (2020).

As in the vectorial case, it is possible to avoid the bandwidth selection problem using tests based on empirical regression processes. For this purpose, a key element is the empirical counterpart of the integrated regression function $I_n(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x) Y_i$, where $X_i \leq x$ means that $X_i(t) \leq x(t)$, for all $t \in [0,1]$. In this scenario, the test statistic can be formulated as $T_n(I_n, I_{\hat{\beta}})$, where $I_{\hat{\beta}}(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x)\hat{Y}_i$, where $\hat{Y}_i = \langle X_i, \hat{\beta}\rangle$. Deriving the theoretical behavior of an empirical regression process indexed by $x \in \mathcal{H}_X$, namely $R_n(x) = \sqrt{n}(I_n(x) - I_{\hat{\beta}}(x))$ is, still today, a challenging task. Yet, as previously presented, the useful projection approach over $\mathcal{H}_X$ can be

considered. The null hypothesis $H_0 : m \in \mathcal{L}$ can be formulated by means of

$$H_0 : \mathbb{E}[(Y - \langle X, \beta\rangle)\mathbb{I}(\langle X, \gamma\rangle \leq u)] = 0, \text{ for a } \beta \in \mathcal{H}_X \text{ and for all } \gamma \in \mathcal{H}_X,$$

which in turn is equivalent to replacing 'for all $\gamma \in \mathcal{H}_X$' with 'for all $\gamma \in \mathcal{S}_{\mathcal{H}_X}$' or 'for all $\gamma \in \mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^{\infty}}$, for all $p \geq 1$', where

$$\mathcal{S}_{\mathcal{H}_X} = \{\rho \in \mathcal{H}_X : \|\rho\| = 1\}, \quad \mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^{\infty}} = \left\{\rho = \sum_{j=1}^{p} r_j\psi_j : \|\rho\| = 1\right\}$$

are infinite- and finite-dimensional spheres on $\mathcal{H}_X$, $\{\psi_j\}_{j=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}_X$, and $\{r_j\}_{j=1}^{p} \subset \mathbb{R}$. As follows from García-Portugués et al. (2014) a general test statistic can be built aggregating all the projections within a certain subspace: $T_n = \int T_{n,\pi}\,\mathrm{d}W(\pi)$ with $T_{n,\pi} = T(I_{n,\pi}, I_{\hat{\beta},\pi})$ based on

$$I_{n,\pi}(u) = n^{-1}\sum_{i=1}^{n}\mathbb{I}(\pi(X_i) \leq u)Y_i \text{ and } I_{\hat{\beta},\pi}(u) = n^{-1}\sum_{i=1}^{n}\mathbb{I}(\pi(X_i) \leq u)\hat{Y}_i, \tag{4}$$

for $\pi(x) = \langle x, \gamma\rangle$. In this case, W is a probability measure defined in $\mathcal{S}_{\mathcal{H}_X}$ or $\mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^{\infty}}$, for a certain $p \geq 1$. Alternatively, the test statistic can be based on only one random projection: $T_n = T_{n,\hat{\pi}}$. More generally, T_n may consider the aggregation of a finite number of random projections, as advocated in the test statistic of Cuesta-Albertos et al. (2019). Both types of tests, all-projections and finite-random-projections, may feature several distances for T, such as Kolmogorov–Smirnov or Cramár–von Mises types.

In the last years, more general procedures for model (2) focus on model specification with scalar response and functional covariate are defined. McLean et al. (2015) consider the functional generalized additive model

$$Y = m_{\mathcal{F}} + \varepsilon = \eta + \int_0^1 \mathcal{F}(X(t), t)dt, \tag{5}$$

being (2) a particular case of (5) taking $\mathcal{F}(x, t) = x\beta(t)$ and $\eta = 0$, whereas Horváth and Reeder (2013) take under consideration the functional quadratic regression model

$$Y = \int_0^1 \beta(t)X(t)dt + \int_0^1\int_0^1 \gamma(s, t)X(t)X(s)dtds + \varepsilon \tag{6}$$

where (2) corresponds with taking $\gamma = 0$ in (6).

Besides, we can highlight some recent alternatives: generalizing the well-known F-test for specification testing or, more generally, the likelihood ratio test. See McLean et al. (2015) or Kong et al. (2016). New ones which establish alternative tests with easy to calibrate distribution as Shi et al. (2022) or devoted to speed computational tasks as in Zhao et al. (2022).

It is also worth mentioning literature comparing the above mentioned procedures. See Tekbudak et al. (2019) for an extensive comparative between procedures based on smoothing techniques, empirical processes and adapted statistics from the likelihood ratio test.

When both the predictor and the response, X and Y, are functional random variables evaluated in $\mathcal{H}_X = L^2[a,b]$ and $\mathcal{H}_Y = L^2[c,d]$, the regression model $Y = m(X) + \varepsilon$ is related with the operator $m : \mathcal{H}_X \to \mathcal{H}_Y$. Perhaps the most popular operator specification is a (linear) Hilbert–Schmidt integral operator, expressible as

$$m_\beta(x)(t) = \langle x, \beta(\cdot,t)\rangle = \int_a^b \beta(s,t)x(s)\,\mathrm{d}s, \quad t \in [c,d], \tag{7}$$

for $\beta \in \mathcal{H}_X \otimes \mathcal{H}_Y$, which is simply referred to as the functional linear model with functional response. The kernel β can be represented as $\beta = \sum_{j=1}^\infty \sum_{k=1}^\infty b_{jk}(\psi_j \otimes \phi_k)$, with $\{\psi_j\}_{j=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ being orthonormal bases of $\mathcal{H}_X$ and $\mathcal{H}_Y$, respectively.

Similarly to the case with scalar response, performing inference on (7) have attracted the analogous two mainstream approaches: (*i*) testing $H_0 : m \in \{m_{\beta_0}\}$ vs. $H_1 : m \in \{m_\beta : \beta \in \mathcal{H}_X \otimes \mathcal{H}_Y, \beta \neq \beta_0\}$, usually with $\beta_0 = 0$; (*ii*) testing $H_0 : m \in \mathcal{L} = \{m_\beta : \beta \in \mathcal{H}_X \otimes \mathcal{H}_Y\}$ vs. $H_1 : m \notin \mathcal{L}$. The GoF problem given in (*ii*) can be approached by considering a double-projection mechanism based on $\pi_X : \mathcal{H}_X \to \mathbb{R}$ and $\pi_Y : \mathcal{H}_Y \to \mathbb{R}$. Given an iid sample $\{(X_i, Y_i)\}_{i=1}^n$, a general test statistic follows (see García-Portugués et al. (2021)) as $T_n = \int T_{n,\pi_X,\pi_Y}\,\mathrm{d}W(\pi_X \times \pi_Y)$ with $T_{n,\pi_X,\pi_Y} = T(I_{n,\pi_X,\pi_Y}, I_{\hat\beta,\pi_X,\pi_Y})$, where I_{n,π_1,π_2} and $I_{\hat\beta,\pi_1,\pi_2}$ follows from (4) by replacing π with π_X, and Y_i and $\hat Y_i$ with $\pi_Y(Y_i)$ and $\pi_Y(\hat Y_i)$, respectively. In this case, W is a probability measure is defined in $\mathcal{S}_{\mathcal{H}_X} \times \mathcal{S}_{\mathcal{H}_Y}$ or $\mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^\infty} \times \mathcal{S}^{q-1}_{\mathcal{H}_Y,\{\phi_k\}_{k=1}^\infty}$, for certain $p, q \geq 1$. The projection approach is immediately adaptable to the GoF of (7) with $\mathcal{H}_X = \mathbb{R}$, and allows graphical tools for that can help detecting the deviations from the null, see García-Portugués et al. (2020). An alternative route considering projections just for X is presented by Chen et al. (2020).

The above generalization to the case of functional response is certainly more difficult for the class of tests based on the likelihood ratios. Regarding the smoothing-based tests, Patilea et al. (2016) introduced a kernel-based significance test consistent for nonlinear alternative. Moreover, Smaga (2022) extends the F-test to the context of functional response making use of projections.

3 A new generation of procedures for testing in regression models based on distance correlation

Since the article of Székely et al. (2007), with the first correlation distance methodology development, there has been a huge variety of works using its ideas for independence tests. Some of them focused on the specification testing field. Very recently, in the last five years, new procedures for specification testing have been derived extending correlation distance ideas. These have resulted in novel covariates selection or GoF approaches. In case of covariates selection, this translates in testing if all considered $X_1, \ldots, X_p$ covariates are relevant to explain a variable Y or some can be excluded from the model. For this aim, the covariates selection problem is rewritten as an independence test and distance correlation methodology is used to construct proper statistics. For GoF the model is estimated under the null hypothesis assumptions and then, the independence between the estimation of the model error and covariates is tested. As a result, specification tests result in independence ones which can be performed using distance correlation ideas.

In this section, a first timeline review of classical methods for independence or significance testing in regression models, being special cases of specification tests, is carried out in Section 3.1. We

highlight the most notable procedures and expose their drawbacks. Then, the benefits of the distance correlation based tests, specially in the high-dimensional context of $p > n$, are motivated. Next, a review of the distance correlation and derivatives methodology is introduced in Sections 3.2, 3.3 and 3.4. The distance correlation, the martingale difference divergence and the conditional distance correlation coefficients, as well as their associated independence tests, are described for the vectorial framework in these sections, respectively. Eventually, specific advances for statistics based on distances in the functional data context are detailed in Section 3.5.

3.1 Previous considerations of correlation measures based on distances

During the last decades, covariates selection procedures have received special attention. This study has been specially focused on the big data context, in which the number of covariates (p) is high, even larger than the sample size (n), $p > n$. As a result, several covariates selection techniques have been developed for this framework.

From the beginning, one of the first and well-known dependence measures for random vectors is the correlation coefficient. See, for example, Pearson (1920). This allows to perform covariates selection taking under consideration only covariates with the greatest correlation value with the response. However, this is only able to correctly detect linear relations. As a result, we can only select covariates if we can assume a linear structure in the regression model. With the aim of identifying other types of dependence, other coefficients measures based on ranks were proposed. These are the Spearman's coefficient (Wissler (1905)) or the Kendall's τ (Kendall (1938)). These measures are robust to outliers and detect any type of monotone dependence pattern. Nevertheless, it is not possible to identify non-monotone structures, being unsuitable for some regression models. These techniques only measure the grade of dependence for each covariate separately and do not pay attention to the information provided by the rest of them in the process. Moreover, the computational cost increases in terms of the p size.

If certain structure of the regression model can be assumed, this information can be employed to perform significance tests for covariates selection. For example, under the linearity assumption with Gaussian errors, we can resort to the well-known F-test. Nonetheless, these methodologies are not available in the $p > n$ case and other approaches are needed. In this framework, the most important covariates selection methods are those based on regularizations. These have been specifically proposed for the covariates selection problem in the big data context of $p > n$ to face the problem of the curse of dimensionality. In this way a sparse parameter vector associated with a linear regression model is estimated and those covariates with negligible associated coefficient are excluded. Some examples are the LASSO (Tibshirani (1996)), the SCAD (Fan and Li (2001)), the adaptive LASSO (Zou (2006)) or the Dantzig selector (Candes and Tao (2007)) to name a few. See the review of Freijeiro-González et al. (2022) for in-depth details. However, these procedures and their extensions have some restrictions in practice: it is necessary to assume certain structure in the regression model, which can not always be a reliable assumption, and their behavior is worse when p increases faster than n. Furthermore, some of these techniques require of high computational time and resources for a large number of covariates.

Motivated by these previous limitations, Székely et al. (2007) introduced the concept of distance correlation (DC). This coefficient detects all types of possible dependence relations and, as a result, solve the main drawbacks of the previous correlation coefficients. Besides, no structure assumption

is needed in comparison with regularization techniques. Hence, a covariates selection approach can be performed using the DC coefficient no matter the regression model structure. Consequently, innovative techniques for covariates selection were proposed using DC ideas of Székely et al. (2007). Some examples are the procedure of Székely et al. (2007), the DC-SIS (distance covariance sure independence screening) procedure of Li et al. (2012), using the SIS (sure independence screening) algorithm for linear models of Fan and Lv (2008), or the partial distance correlation methodology introduced in Székely and Rizzo (2014). First and third approaches apply independence tests considering an adequate statistic based on DC ideas. In contrast, the DC-SIS sorts out covariates using the distance correlation values and then applies some cutoff or threshold to consider only the most important ones in model explanation terms, which corresponds with the greatest DC values between covariates and response.

In the last years, two new measures of dependence related with the DC were introduced. The martingale difference divergence (MDD) of Shao and Zhang (2014) and the conditional distance correlation (CDC) of Wang et al. (2015). The MDD is used to test the causality of a vector $Y \in \mathbb{R}^q$ conditioned to a scalar random variable $X \in \mathbb{R}$, whereas the CDC tests the conditional dependence of two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ conditioned to a third one, $Z \in \mathbb{R}^r$. Both coefficients can be employed to derive specification tests and to implement covariates selection procedures. See, for example, the work of Shao and Zhang (2014) and Zhang et al. (2018) for the MDD case and all the details of the procedure proposed in Wang et al. (2015) for the CDC performance.

The necessity of covariates selection and specification testing procedures for the functional data context has motivated the recently development of new procedures for this framework. Here, classic methodologies are not available and thus, new ones are needed. Works as the one developed by Gretton et al. (2005) or Febrero-Bande et al. (2019) in the machine learning context, are examples of novel screening tools and bring out the complexity of the functional data case. In Section 3.5 a review of novel specification tests using DC ideas for the functional context is introduced.

In the following, more details about DC, MDD and CDC are given for a deeper understanding of these three kinds of dependence measures for random vectors in Sections 3.2, 3.3 and 3.4, respectively. Next, recent advances in the functional data context using these ideas are described in Section 3.5.

3.2 Distance correlation

The DC is a measure of dependence which detects all types of relations between two random vectors of different dimensions. This coefficient is introduced for the first time by Székely et al. (2007). The main DC interest is to test if two random vectors, $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with $p, q \geq 1$, are independent. This results in testing

$$H_0 : X \perp Y \quad \text{vs.} \quad H_1 : X \not\perp Y, \tag{8}$$

where $X \perp Y$ denotes independence between X and Y.

Two random vectors are said to be independent if they verify $F_{X,Y} = F_X F_Y$, being F_X, F_Y the distribution functions of X and Y, respectively, and $F_{X,Y}$ their joint distribution. This condition can be rewritten in terms of the characteristic functions and the independence test can be formulated as

$$H_0 : \varphi_{X,Y} = \varphi_X \varphi_Y \quad \text{vs.} \quad H_1 : \varphi_{X,Y} \neq \varphi_X \varphi_Y \tag{9}$$

being $\varphi_{X,Y}$ the joint characteristic function and φ_X, φ_Y the marginal characteristic functions of X, Y.

So, for the testing of the null hypothesis (9) it is needed a statistic measuring if the difference $\varphi_{X,Y} - \varphi_X\varphi_Y$ is significant. This is the main motivation for the introduction of the DC coefficient (Székely et al. (2007), Székely and Rizzo (2017)).

In order to measure the difference between $\varphi_{X,Y}$ and $\varphi_X\varphi_Y$ a weighted L_2 norm ($\|\cdot\|_w^2$) in the $\mathbb{R}^p \times \mathbb{R}^q$ space of complex functions is applied. This is defined as

$$\|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|_w^2 = \int_{\mathbb{R}^p\times\mathbb{R}^q} |\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2 w(t,s)\, dt\, ds \tag{10}$$

where $w(\cdot,\cdot)$ is a weight function properly selected to guarantee the existence of the above integral and $|f| = f\bar{f}$ for $f(\cdot)$, a complex value function with conjugate $\bar{f}(\cdot)$.

Then, once the weight function $w(\cdot,\cdot)$ has been selected, we can take as a measure of dependence $\mathcal{V}^2(X,Y;w) = \|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|_w^2$ satisfying that $\mathcal{V}^2(X,Y;w) = 0$ if and only if X and Y are independent. Particularly, dividing $\mathcal{V}^2(X,Y;w)$ by $\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}$, where

$$\mathcal{V}^2(X;w) = \int_{\mathbb{R}^{2p}} |\varphi_{X,X}(t,s) - \varphi_X(t)\varphi_X(s)|^2 w(t,s)\, dt\, ds \tag{11}$$

we obtain a type of unsigned correlation $\mathcal{R}_w$.

Following these guidelines, in Székely et al. (2007) it is taken

$$w(t,s) = (c_p c_q |t|_p^{1+p} |t|_q^{1+q})^{-1} dt\, ds \quad \text{for} \quad c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)} \ \text{ and } \ c_q = \frac{\pi^{(q+1)/2}}{\Gamma((q+1)/2)}, \tag{12}$$

denoting by $\|\cdot\|_p$ and $\|\cdot\|_q$ the euclidean norms in $\mathbb{R}^p$ and $\mathbb{R}^q$ and $\Gamma(\cdot)$ the gamma function.

For simplicity, we write $\|\cdot\|^2$ henceforth, instead of $\|\cdot\|_\omega^2$, as the L_2 norm using this weight function. Thus, for finiteness of $\|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|^2$, it is sufficient that $\mathbb{E}[\|X\|_p] < \infty$ and $\mathbb{E}[\|Y\|_q] < \infty$. With this notation, the DC between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{V}^2(X,Y)$ defined by expression (13)

$$\mathcal{V}^2(X,Y) = \|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|^2 = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2}{\|t\|_p^{p+1}\|s\|_q^{q+1}} dt\, ds \tag{13}$$

Similarly, distance variance is given as the square root of

$$\mathcal{V}^2(X) = \mathcal{V}^2(X,X) = \|\varphi_{X,X}(t,s) - \varphi_X(t)\varphi_X(s)\|^2. \tag{14}$$

The DC coefficient between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{R}(X,Y)$ defined by

$$\mathcal{R}(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0, \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \tag{15}$$

It is verified that $0 \le \mathcal{R}(X,Y) \le 1$, and $\mathcal{R}(X,Y) = 0$ if and only if X and Y are independent.

Alternative expressions for (13) are

$$\begin{aligned}\mathcal{V}^2(X,Y) =& \mathbb{E}\left[\|X'-X''\|_p\|Y'-Y''\|_q\right] \\ &+\mathbb{E}\left[\|X'-X''\|_p\right]\mathbb{E}\left[\|Y'-Y''\|_q\right]-2\mathbb{E}\left[\|X'-X''\|_p\|Y'-Y'''\|_q\right]\end{aligned} \tag{16}$$

and

$$\begin{aligned}\mathcal{V}^2(X,Y) =& \mathbb{E}_{X'Y'}\left[\mathbb{E}_{X''Y''}\left[\|X'-X''\|_p\|Y'-Y''\|_q\right]\right] \\ &+\mathbb{E}_{X'X''}\left[\|X'-X''\|_p\right]\mathbb{E}_{Y'Y''}\left[\|Y'-Y''\|_q\right] \\ &-2\mathbb{E}_{X'Y'}\left[\mathbb{E}_{X''}\left[\|X'-X''\|_p\right]\mathbb{E}_{Y''}\left[\|Y'-Y''\|_q\right]\right]\end{aligned} \tag{17}$$

being (X',Y'), (X'',Y'') and (X''',Y''') iid copies of (X,Y). See Székely et al. (2007) for more details.

Given $(\mathbf{X_n},\mathbf{Y_n}) = \{(X_i,Y_i), i=1,\ldots,n\}$ an iid sample from the joint distribution function of $(X,Y)\in\mathbb{R}^p\times\mathbb{R}^q$, the empirical sample versions of the estimator of $\mathcal{V}^2(\cdot,\cdot)$ can be obtained as follows. Defining $A_{il} = a_{il}-\bar{a}_{i\cdot}-\bar{a}_{\cdot l}+\bar{a}_{\cdot\cdot}$ by means of quantities

$$a_{il}=\|X_i-X_l\|_p,\quad \bar{a}_{i\cdot}=\frac{1}{n}\sum_{l=1}^n a_{il},\quad \bar{a}_{\cdot l}=\frac{1}{n}\sum_{i=1}^n a_{il}\quad\text{and}\quad \bar{a}_{\cdot\cdot}=\frac{1}{n^2}\sum_{i,l=1}^n a_{il}, \tag{18}$$

similarly $B_{il}=b_{il}-\bar{b}_{i\cdot}-\bar{b}_{\cdot l}+\bar{b}_{\cdot\cdot}$ with $b_{il}=\|Y_i-Y_l\|_q$. The empirical distance covariance $\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})$, based on the empirical estimator of (13), is the nonnegative number given by

$$\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})=\frac{1}{n^2}\sum_{i,l=1}^n A_{il}B_{il}. \tag{19}$$

Respectively, $\mathcal{V}_n^2(\mathbf{X_n})$ is the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X_n})=\mathcal{V}_n^2(\mathbf{X_n},\mathbf{X_n})=\frac{1}{n^2}\sum_{i,l=1}^n A_{il}^2. \tag{20}$$

In summary, the estimation given in (19) is an easier way of obtaining an estimator of $\mathcal{V}^2(X,Y)$ just centering two times the data.

Furthermore, the empirical DC, $\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})$, is the square root of

$$\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})=\begin{cases}\dfrac{\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})}{\sqrt{\mathcal{V}_n^2(\mathbf{X_n})\mathcal{V}_n^2(\mathbf{Y_n})}}, & \mathcal{V}_n^2(\mathbf{X_n})\mathcal{V}_n^2(\mathbf{Y_n})>0,\\ 0, & \mathcal{V}_n^2(\mathbf{X_n})\mathcal{V}_n^2(\mathbf{Y_n})=0.\end{cases} \tag{21}$$

This coefficient takes values $0\le\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})\le 1$, and verifies that, if $\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})=1$, then there exist a vector a, a nonzero real number b and an orthogonal matrix C such that $\mathbf{Y_n}=a+b\mathbf{X_n}C$. Moreover, it is verified almost surely that $\lim_{n\to\infty}\mathcal{R}_n^2(\mathbf{X_n},\mathbf{Y_n})=\mathcal{R}^2(X,Y)$. For more properties about $\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})$, $\mathcal{V}_n(\mathbf{X_n})$ and $\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})$ we refer to Székely et al. (2007).

Under the null hypothesis of independence, it is verified that $n\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})/S_2$ converges in distribution to a quadratic form $Q=\sum_{m=1}^\infty c_m G_m^2$, where S_2 is a normalizing factor defined in Székely et al. (2007), $\{G_m\}_{m=1}^\infty$ are independent standard normal random variables and $\{c_m\}_{m=1}^\infty$

nonnegative constants that depend on the distribution of (X, Y). Moreover, when this hypothesis is violated, $n\mathcal{V}_n^2(\mathbf{X_n}, \mathbf{Y_n}) \to \infty$ in probability as $n \to \infty$. Thus, a test which rejects H_0 for large values of $n\mathcal{V}_n^2(\mathbf{X_n}, \mathbf{Y_n})$ is consistent in an omnibus way against dependence alternatives. In practice, the limiting distribution can be approximated by resampling techniques, as for example using permutation tests.

DC can be also used to perform proper GoF tests. In Xu and He (2021) a procedure based on DC is used to test the null hypothesis $H_0 : X \perp \varepsilon$ and $m \in \mathcal{M}_\beta$ in the regression model $Y = m(X) + \varepsilon$ with $m \in \mathcal{M}_\beta = \{g(x)^\top \beta : \beta \in \mathbb{R}^p\}$ for a given known function $g(\cdot)$. In this context, $\mathbf{Y_n}$ is built with the residuals of the fitted model.

Despite all discussed desirable qualities of the empirical distance covariance coefficient, this is a biased estimator of (13) and its bias increases with dimension of X and Y, i.e. when $p, q \to \infty$. Besides, the DC statistic introduced in (21), and based on these coefficients, exhibits some drawbacks as well. As it is explained in Székely and Rizzo (2013), although distance correlation characterizes independence, interpretation of the size of $\mathcal{R}_n(\mathbf{X_n}, \mathbf{Y_n})$ without a formal test is difficult in high dimensions. An explanation for this is that $\mathcal{R}_n^2(\mathbf{X_n}, \mathbf{Y_n}) \longrightarrow 1$ as $p, q \to \infty$, even though X and Y are independent. Székely and Rizzo (2013) proposed a new unbiased sample estimator for distance covariance and a modified distance correlation statistic based on plug-in these unbiased version in numerator and denominator of expression (21) and verifying that, under the null hypothesis of independence, this converges to a Student t distribution. This new approach solves the inconsistency problem in high dimensions.

An additional problem is the computational cost of the construction of the distance matrices. Some recent works such as Huo and Székely (2016) or Chaudhuri and Hu (2019) propose alternatives to reduce this. However, only the univariate random variables case is considered. New solutions applying for the vectorial framework need to be considered in the future.

Finally, it is remarkable the natural relation between DC and the Hilbert-Schmidt Independence Criterion (HSIC) of Gretton et al. (2005). The HSIC makes use of the cross-covariance operator between two reproducing kernel Hilbert spaces (RKHSs) to measure if there exists some type of dependence between two random vectors defined in two different RKHSs with universal kernel. These vectors will be independent when the HSIC operator will take the null value. The DC is a particular case of HSIC operator where general kernel distances are replaced by Euclidean versions. In some sense, there was a parallel evolution between the HSIC criteria in the machine learning world, related to RHKSs, and the DC ideas in literature. There are really interesting papers published in the last decade giving a unifying framework that links both fields. See Sejdinovic et al. (2013), Hua and Ghosh (2015), Zhu et al. (2020) or Edelmann and Goeman (2022) for examples of this connection under different perspectives. As a result, the HSIC measure can be used to perform independence tests, an example is the work of Song et al. (2012), as well as specification tests, see Sen and Sen (2014) for simultaneous GoF and error-predictor independence tests in linear models.

3.3 Martingale difference divergence

The MDD is a new dependence coefficient introduced by Shao and Zhang (2014). This metric measures the departure from the conditional mean independence hypothesis. This is based on testing if the conditional mean of $Y \in \mathbb{R}$, given $X \in \mathbb{R}^p$, is independent of X. The testing problem is now

given by

$$H_0 : \mathbb{E}[Y|_X] = \mathbb{E}[Y] \text{ almost surely} \quad \text{vs.} \quad H_1 : \mathbb{E}[Y|_X] \neq \mathbb{E}[Y] \text{ almost surely.} \tag{22}$$

Its name comes from the interpretation of martingale difference concept in probability. This means that if H_0 in (22) is verified, then $Y - \mathbb{E}[Y]$ is a martingale difference with respect to X.

As a result, the MDD coefficient is designed to measure the difference between the conditional mean and the unconditional one to perform (22). The MDD of Y given X is the nonnegative number $MDD^2(Y|_X)$ defined by

$$MDD^2(Y|_X) = \frac{1}{c_p} \int_{\mathbb{R}^p} \frac{|\psi_{Y,X}(t) - \psi_Y \psi_X(t)|^2}{\|t\|_p^{p+1}} dt \tag{23}$$

where $\psi_{Y,X}(t) = \mathbb{E}[Ye^{i<t,X>}]$, $\psi_Y = \mathbb{E}[Y]$ and $\psi_X(t) = \varphi_X(t)$.

The MDD coefficient defined in (23) verifies that $MDD^2(Y|_X) \geq 0$ and takes the null value if and only if the null hypothesis (22) holds. This is called divergence and not distance because $MDD^2(Y|_X) \neq MDD^2(X|_Y)$.

Similar to DC, a scale invariant coefficient can be defined. This gives place to the martingale difference correlation (MDC) given by the square root of

$$MDC^2(Y|_X) = \begin{cases} \frac{MDD^2(Y|_X)}{\sqrt{\text{var}^2(Y)\mathcal{V}^2(X)}}, & \text{var}^2(Y)\mathcal{V}^2(X) > 0, \\ 0, & \text{var}^2(Y)\mathcal{V}^2(X) = 0. \end{cases} \tag{24}$$

where $\mathcal{V}^2(X)$ is the distance variance of X defined in (14). It is verified that $0 \leq MDC^2(Y|_X) \leq 1$. Similar properties as DC for $MDD^2(Y|_X)$ and $MDC^2(Y|_X)$ are collected in Shao and Zhang (2014).

For a sample of $i = 1, \ldots, n$ iid observations $(\mathbf{X_n}, \mathbf{Y_n}) = \{(X_i, Y_i), i = 1, \ldots, n\}$ from the joint distribution of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, it is defined A_{il} as in (18) and $B_{il} = b_{il} - \bar{b}_{i\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$; being now $b_{il} = |Y_i - Y_l|^2/2$, $\bar{b}_{i\cdot} = \frac{1}{n}\sum_{l=1}^n b_{il}$, $\bar{b}_{\cdot l} = \frac{1}{n}\sum_{i=1}^n b_{il}$ and $\bar{b}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,l=1}^n b_{il}$ for $i, l = 1, \ldots, n$. The empirical estimation of $MDD^2(Y|_X)$, i.e. the sample martingale difference divergence, can be defined as the nonnegative number

$$MDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) = \frac{1}{n^2} \sum_{i,l=1}^n A_{il} B_{il} \tag{25}$$

and its associated sample martingale difference correlation coefficient is given by

$$MDC_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) = \begin{cases} \frac{MDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})}{\sqrt{\text{var}_n^2(\mathbf{Y_n})\mathcal{V}_n^2(\mathbf{X_n})}}, & \text{var}_n^2(\mathbf{Y_n})\mathcal{V}_n^2(\mathbf{X_n}) > 0, \\ 0, & \text{var}_n^2(\mathbf{Y_n})\mathcal{V}_n^2(\mathbf{X_n}) = 0. \end{cases} \tag{26}$$

where $\text{var}_n(\mathbf{Y_n}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$, for $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, and $\mathcal{V}_n^2(\mathbf{X_n})$ is defined in (20).

See the paper of Park et al. (2015) for a nice connection between MDD and DC coefficients.

If $\mathbb{E}\left[\|X\|_p + |Y|^2\right] < \infty$, both estimators, $MDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})$ and $MDC_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})$, converge to their population versions displayed in (23) and (24) almost surely. A prove of this result can be found in Shao and Zhang (2014). Moreover, under the null hypothesis of independence in mean, it is guaranteed that $nMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) \longrightarrow \|\Gamma(t)\|^2$ in distribution when $n \to \infty$, being $\Gamma(\cdot)$ a Gaussian process. In addition, if $\mathbb{E}[Y^2|_X] = \mathbb{E}[Y^2]$ is also guaranteed, $nMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})/S_n \longrightarrow Q$ in distribution when $n \to \infty$, being Q a nonnegative quadratic form of centered Gaussian random variable with $\mathbb{E}[Q] = 1$ and $S_n = \frac{1}{n^2}\sum_i \sum_l \|X_i - X_l\|_p \frac{1}{n}\sum_i (Y_i - \bar{Y}_n)^2$. Finally, if the null hypothesis is not verified, we have that $nMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})/S_n \longrightarrow \infty$ in probability when $n \to \infty$. We refer to Shao and Zhang (2014) for more details. Although the asymptotic distribution under both, H_0 and H_1 hypothesis is known, resampling procedures can be applied in practice to calibrate the distribution of the test statistic, especially for small sample sizes.

Thus, using the estimators of the MDD or MDC, it is possible to perform covariates selection in regression models, specifying which covariates are the relevant ones. Shao and Zhang (2014) propose a screening procedure sorting out the covariates relevance in terms of the regressor function explanation, i.e. based on $\mathbb{E}[Y|_X]$ explanation, and then they establish a proper threshold to detect the important significance covariates. Authors make use of the MDC criteria to measure covariates relevance. A different approach for covariates selection in terms of causality is introduced in Zhang et al. (2018). They propose a statistic based on the MDD ideas to test the null hypothesis of $H_0 : \mathbb{E}\left[Y|_{X_j}\right] = \mathbb{E}[Y]$ almost surely for all $j = 1 \ldots, p$. A wild bootstrap scheme is proposed to approximate the statistics distribution.

All these ideas can be transferred to GoF testing. An example is the work of Su and Zheng (2017). They test the null hypothesis of $H_0 : \mathbb{P}\left(\mathbb{E}[Y|_X] = g(X, \beta)\right) = 1$ for some $\beta \in \mathcal{B}$, being $\mathcal{B}$ the parameter space and assuming $Y = g(X, \beta) + \varepsilon$, with $g(\cdot)$ a known function. The MDD is applied using the residuals calculated under the null hypothesis, and covariates. Calibration of the test is again done by means of wild bootstrap. A similar, but broader approach, using HSIC is also provided by Teran Hidalgo et al. (2018).

3.4 Conditional distance correlation

The CDC was introduced in Wang et al. (2015) to measure the dependence of two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ conditioned to a third one, $Z \in \mathbb{R}^r$. For this purpose, conditional characteristic functions are employed and ideas of the distance correlation introduced in Section 3.1 are adapted to the conditional framework. The problem to be tested now is

$$H_0 : X \perp_{|Z} Y \text{ almost surely} \quad \text{vs.} \quad H_1 : \mathbb{P}\left(X \not\perp_{|Z} Y\right) > 0 \tag{27}$$

where $X \perp_{|Z} Y$ denotes independence of X and Y conditioned to Z.

Using similar DC arguments, it is possible to rewrite (27) in terms of characteristic functions. The new test is given by

$$H_0 : \varphi_{X,Y|Z} = \varphi_{X|Z}\varphi_{Y|Z} \quad \text{vs.} \quad H_1 : \varphi_{X,Y|Z} \neq \varphi_{X|Z}\varphi_{Y|Z} \tag{28}$$

where $\varphi_{X,Y|Z}$, $\varphi_{X|Z}$ and $\varphi_{Y|Z}$ are the joint and marginal conditional characteristic functions.

Then, the CDC with finite first moments given Z ($\mathbb{E}[|X|_p + |Y|_q|_Z] < \infty$), is defined as the square root of

$$\begin{aligned} CDC^2(X,Y|_Z) =& \|\varphi_{X,Y|_Z}(t,s) - \varphi_{X|_Z}(t)\varphi_{Y|_Z}(s)\|^2 \\ =& \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y|_Z}(t,s) - \varphi_{X|_Z}(t)\varphi_{Y|_Z}(s)|^2}{\|t\|_p^{p+1}\|s\|_q^{q+1}} dt\, ds \end{aligned} \tag{29}$$

where c_p and c_q are the ones defined in (12) and conditional distance variance is the square root of

$$CDC^2(X|_Z) = CDC^2(X,X|_Z) = \|\varphi_{X,X|_Z}(t,s) - \varphi_{X|_Z}(t)\varphi_{X|_Z}(s)\|^2,$$

being $\|\cdot\|$ the weighted norm defined in Section 3.2.

The CDC coefficient defined in (29) has analogues properties to the unconditional version of (13). Particularly, it is verified that $CDC(X,Y|_Z) \geq 0$ if and only if X and Y are conditionally independent given Z.

The conditional distance correlation (CDCor) is the square root of

$$CDCor(X,Y|_Z) = \begin{cases} \frac{CDC^2(X,Y|_Z)}{\sqrt{CDC^2(X|_Z)CDC^2(Y|_Z)}}, & CDC^2(X|_Z)CDC^2(Y|_Z) > 0, \\ 0, & CDC^2(X|_Z)CDC^2(Y|_Z) = 0. \end{cases} \tag{30}$$

and this verifies that $0 \leq CDCor(X,Y|_Z) \leq 1$ and $CDCor(X,Y|_Z) = 0$ if and only if X and Y are conditionally independent given Z.

To construct an estimator of $CDC^2(X,Y|_Z)$ the empirical characteristic functions conditioned to Z are plugged in (29). Note that, for the estimation of conditional characteristic functions, it is needed to resort to some kind of smoothing techniques as for example kernel-type estimators. We refer to Wang et al. (2015) for more details. Denote by $W_i = (X_i, Y_i, Z_i),\ i = 1,\ldots,n$ a sample iid from a random vector $W = (X,Y,Z) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r$, $\mathbf{X_n} = \{X_1,\ldots,X_n\}$, $\mathbf{Y_n} = \{Y_1,\ldots,Y_n\}$, $\mathbf{Z_n} = \{Z_1,\ldots,Z_n\}$, and $\mathbf{W_n} = (\mathbf{X_n}, \mathbf{Y_n}, \mathbf{Z_n})$. As a result, the sample conditional distance covariance $CDC_n(\mathbf{X_n}, \mathbf{Y_n}|_{\mathbf{Z_n}})$ is the positive quantity defined by

$$\widetilde{CDC}_n^2(\mathbf{X_n}, \mathbf{Y_n}|_{\mathbf{Z_n}}) = \|\varphi^n_{X,Y|_Z}(t,s) - \varphi^n_{X|_Z}(t)\varphi^n_{Y|_Z}(s)\|^2. \tag{31}$$

being $\varphi^n_{X,Y|_Z}$, $\varphi^n_{X|_Z}$ and $\varphi^n_{Y|_Z}$ the corresponding empirical conditional characteristic functions.

Following Wang et al. (2015), letting $d_{ijkl} = \left(a^X_{ij} + a^X_{kl} - a^X_{ik} - a^X_{jl}\right)\left(b^Y_{ij} + b^Y_{kl} - b^Y_{ik} - b^Y_{jl}\right)$ and $d^S_{ijkl} = d_{ijkl} + d_{ijlk} + d_{ilkj}$ for $i,j,k,l = 1,\ldots,n$, where a_{ij} and b_{ij} are defined in (18), and Z_1, Z_2, Z_3 and Z_4 are iid copies of Z, it is verified that

$$CDC^2(X,Y|_{Z=z}) = \frac{1}{12}\mathbb{E}[d^S_{1234}|_{Z_1=z,Z_2=z,Z_3=z,Z_4=z}]$$

As a result, the conditional dependence measures can be estimated by applying kernel regression smoothing ideas to the above expectation estimation. This results in a V-process. The sample conditional distance covariance is defined as the square root of

$$CDC^2_n(\mathbf{W_n}|_Z) = CDC^2_n(\mathbf{X_n}, \mathbf{Y_n}, \mathbf{Z_n}|_Z) = \frac{1}{n^4}\sum_{ijkl} \Psi_n(W_i, W_j, W_k, W_l; Z) \tag{32}$$

where Ψ_n is the symmetric random kernel of degree 4 defined in Schick (1997):

$$\Psi_n(W_i, W_j, W_k, W_l; Z) = \frac{n^4 \Phi_i(Z)\Phi_j(Z)\Phi_k(Z)\Phi_l(Z)}{12\Phi^4(Z)} d^S_{ijkl}$$

for $\Phi_i(Z) = K_H(Z - Z_i)$ and $\Phi(Z) = \sum_{i=1}^n \Phi_i(Z)$, being K a kernel function and H a bandwidth matrix r-dim.

Let $\mathbf{W_{X_n}} = (\mathbf{X_n}, \mathbf{X_n}, \mathbf{Z_n})$ and $\mathbf{W_{Y_n}} = (\mathbf{Y_n}, \mathbf{Y_n}, \mathbf{Z_n})$. Analogously, the sample conditional distance correlation can be defined as the square root of

$$CDCor_n(\mathbf{W_n}|_Z) = \begin{cases} \frac{CDC_n^2(\mathbf{W_n}|_Z)}{\sqrt{CDC_n^2(\mathbf{W_{X_n}}|_Z) CDC_n^2(\mathbf{W_{Y_n}}|_Z)}}, & CDC_n^2(\mathbf{W_{X_n}}|_Z) CDC_n^2(\mathbf{W_{Y_n}}|_Z) > 0, \\ 0, & CDC_n^2(\mathbf{W_{X_n}}|_Z) CDC_n^2(\mathbf{W_{Y_n}}|_Z) = 0. \end{cases}$$

It is verified that $\widetilde{CDC}_n^2(\mathbf{W_n}|_Z) = CDC_n^2(\mathbf{W_n}|_Z)$ given $\mathbf{W_n} = \{W_1, \ldots, W_n\}$ a sample from the joint distribution of (X, Y, Z). Furthermore, if $\mathbb{E}\left[\|X\|_p + \|Y\|_q|_Z\right] < \infty$ and $\Phi(Z)/n$ is a consistent density function estimator of Z, then $CDC_n^2(\mathbf{W_n}|_Z) \longrightarrow CDC^2(X, Y|_Z)$ in probability for each value of Z as $n \to \infty$. See Wang et al. (2015) for more details and properties of $CDC_n^2(\mathbf{W_n}|_Z)$. Analogously, an unbiased version of (32) can be defined with similar properties. For this purpose, U-processes theory is applied.

Wang et al. (2015) make use of these ideas to perform the conditional independence test displayed in (27), applying conditioned covariates selection. In particular, they define a statistic based on the CDC coefficient and implement a test calibrated by means of a local bootstrap. Other procedures related with screening techniques in terms of conditional dependence are the recent works of Song et al. (2020) and Lu and Lin (2020). The first one adapt the ideas of Liu et al. (2014) using the CDCor to specify significant covariates for general varying-coefficient models in regression. Covariates are sorting out based on their CDCor value and then a cutoff is applied. In contrast, Lu and Lin (2020) select an initial set of covariates and measures the importance of remaining ones conditioned to this subset. For this purpose, they use the CDCor, resulting in the CDC-SIS (conditional distance correlation sure independence screening) algorithm.

3.5 A new generation of procedures for testing in regression models based on distance correlation with functional data

In this Section, we assume that both, the explanatory covariate X as well as the output Y of the regression model $Y = m(X) + \varepsilon$, are functions. Here, similar to Section 2, it is assumed that $X \in \mathcal{H}_X$ and $Y \in \mathcal{H}_Y$, being $\mathcal{H}_X$ and $\mathcal{H}_Y$ Hilbert spaces. As it was mentioned in previous sections, results for specification testing in regression models with functional data appear in the last 10 years. However, it is not until very recently when the methodology of the DC is employed, extending procedures of Section 2 to the functional framework.

A first paper is Lee et al. (2020), where it is tested the null hypothesis $H_0 : \mathbb{E}\left[Y|_X\right] = \mathbb{E}[Y]$ as in (22) but now for the functional case. Then, to carry out the test, an statistic based on a generalized version of the MDD coefficient described in Section 3.3 is proposed.

In particular, the vectorial MDD term can be written as (see Shao and Zhang (2014))

$$MDD^2\left(Y|_X\right) = -\mathbb{E}\left[\left(Y - \mathbb{E}[Y]\right)\left(Y' - \mathbb{E}[Y]\right)\|X - X'\|_{\mathcal{H}_X}\right]$$

and this idea is extended to the functional context considering

$$FMDD^2\left(Y|_X\right) = -\mathbb{E}\left[\langle Y-\mathbb{E}[Y], Y'-\mathbb{E}[Y]\rangle_{\mathcal{H}_Y}\|X-X'\|_{\mathcal{H}_X}\right]$$

being, in both cases (vectorial and functional), (X', Y') iid copies of (X, Y).

Hence, based on an iid sample $(\mathbf{X_n}, \mathbf{Y_n}) = \{(X_i, Y_i), i = 1, \ldots, n\}$ of (X, Y), an unbiased estimator of $FMDD^2$ is obtained with the empirical version

$$FMDD_n^2\left(\mathbf{Y_n}|_{\mathbf{X_n}}\right) = \frac{1}{n(n-3)}\sum_{i\neq j}\tilde{A}_{ij}\tilde{B}_{ij} \tag{33}$$

where $\tilde{A}_{ij}$ and $\tilde{B}_{ij}$ are now the corresponding U-centered versions of (18), being the $(i,j)^{th}$ elements of the matrices defined as

$$\tilde{A}_{ij} = \begin{cases} a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, & i \neq j \\ 0, & i = j \end{cases}$$

$$\tilde{B}_{ij} = \begin{cases} b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}, & i \neq j \\ 0, & i = j \end{cases}$$

with $a_{ij} = \|X_i - X_j\| = \sqrt{\langle X_i - X_j, X_i - X_j\rangle_{\mathcal{H}_X}}$, $\bar{a}_{i\cdot} = \frac{\sum_l a_{il}}{(n-2)}$, $\bar{a}_{\cdot j} = \frac{\sum_k a_{kj}}{(n-2)}$, $\bar{a}_{\cdot\cdot} = \frac{\sum a_{kl}}{(n-1)(n-2)}$, $b_{ij} = \|Y_i - Y_j\|^2_{\mathcal{H}_Y}/2$ and $\bar{b}_{i\cdot}$, $\bar{b}_{\cdot j}$ and $\bar{b}_{\cdot\cdot}$ defined in a similar way.

That is, the modified and adapted empirical unbiased version of the estimation given in (25), but now, for the functional context.

Nice results are obtained in Lee et al. (2020) under the assumptions of $\mathbb{E}\left[\|X\|^2_{\mathcal{H}_X} + \|Y\|^2_{\mathcal{H}_Y}\right] < \infty$, $\mathbb{E}\left[\|X-\mathbb{E}[X]\|^2_{\mathcal{H}_X} + \|Y-\mathbb{E}[Y]\|^2_{\mathcal{H}_Y}\right] < \infty$ and the null hypothesis is true: $nFMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) \longrightarrow \sum_{k=1}^{\infty}\lambda_k(G_k^2-1)$ in distribution, being $\{\lambda_k\}_{k=1}^{\infty}$ the eigenvalues corresponding to the eigenfunctions $\{\Psi_k(\cdot)\}_{k=1}^{\infty}$ such that $J(z,z') = \sum_{k=1}^{\infty}\lambda_k\Psi_k(z)\Psi_k(z')$ with $z = (x,y)$ and $J(z,z') = U(x,x')V(y,y')$, where $U(x,x') = \|x-x'\|_{\mathcal{H}_X} + \mathbb{E}\left[\|X-X'\|_{\mathcal{H}_X}\right] - \mathbb{E}\left[\|x-X'\|_{\mathcal{H}_X}\right] - \mathbb{E}\left[\|X-x'\|_{\mathcal{H}_X}\right]$ and $V(y,y') = -\langle y-\mathbb{E}[Y], y'-\mathbb{E}[Y]\rangle_{\mathcal{H}_Y}$. Here $\{\Psi_k\}$ is an orthogonal sequence in the sense that $\mathbb{E}\left[\Psi_j(z)\Psi_k(z)\right] = \mathbb{I}(j=k)$ and $\{G_k\}_{k=1}^{\infty}$ is a sequence of iid $N(0,1)$ random variables.

This represent the limit distribution of a degenerate U-statistic with kernel $h(\cdot)$, being

$$FMDD_n^2\left(\mathbf{Y_n}|_{\mathbf{X_n}}\right) = \frac{1}{\binom{n}{4}}\sum_{i<j<k<l}h(Z_i, Z_j, Z_k, Z_l)$$

with $h(Z_i, Z_j, Z_k, Z_l) = \frac{1}{4!}\sum_{(s,t,u,v)}^{(i,j,k,l)}(a_{st}b_{uv} + a_{st}b_{st} - a_{st}b_{su} - a_{st}b_{tv})$ the sum over the 24 possible permutations of the indexes (i,j,k,l).

In Lee et al. (2020) it is proposed to reject the null hypothesis of conditional mean independence if and only if $T_n = nFMDD_n^2\left(\mathbf{Y_n}|_{\mathbf{X_n}}\right) > C$, where C is a constant taken based on the α significance level. The power of the test is studied and demonstrated to be consistent under both, local and fixed alternatives, using a consistent wild bootstrap calibration.

A second recent contribution is the paper of Lai et al. (2020), devoted to test a modified null hypothesis: $\widetilde{H}_0$: "X is independent of ε and m satisfies the linear model given by (7) in Section 2". Using the recent results related with the distance covariance (see Székely et al. (2007), Lyons (2013) and Sejdinovic et al. (2013)). Consider now $(\mathcal{X}, \rho_{\widetilde{X}})$ and $(\mathcal{Y}, \rho_{\widetilde{Y}})$ two semimetric spaces of negative type, where $\rho_{\widetilde{X}}$ and $\rho_{\widetilde{Y}}$ are the corresponding semimetrics. Denote by $(\widetilde{X}, \widetilde{Y})$ a random element with joint distribution $\mathrm{P}_{\widetilde{X}\widetilde{Y}}$ and marginals $\mathrm{P}_{\widetilde{X}}$ and $\mathrm{P}_{\widetilde{Y}}$, respectively, and take $(\widetilde{X}', \widetilde{Y}')$ an iid copy of $(\widetilde{X}, \widetilde{Y})$. The generalized distance covariance $(\widetilde{X}, \widetilde{Y})$ is given by

$$\begin{aligned}\theta(\widetilde{X}, \widetilde{Y}) = \mathbb{E}\big[\rho_{\widetilde{X}}(\widetilde{X}, \widetilde{X}')\rho_{\widetilde{Y}}(\widetilde{Y}, \widetilde{Y}')\big] \\ + \mathbb{E}\big[\rho_{\widetilde{X}}(\widetilde{X}, \widetilde{X}')\big]\mathbb{E}\big[\rho_{\widetilde{Y}}(\widetilde{Y}, \widetilde{Y}')\big] \\ - 2\mathbb{E}_{(\widetilde{X},\widetilde{Y})}\big[\mathbb{E}_{\widetilde{X}'}\big[\rho_{\widetilde{X}}(\widetilde{X}, \widetilde{X}')\big]\mathbb{E}_{\widetilde{Y}'}\big[\rho_{\widetilde{Y}}(\widetilde{Y}, \widetilde{Y}')\big]\big].\end{aligned}$$

This corresponds with expression (16) and (17) for the vectorial case.

As noted by Lai et al. (2020) the generalized distance covariance can be alternatively written as

$$\theta(\widetilde{X}, \widetilde{Y}) = \int \rho_{\widetilde{X}}(\widetilde{x}, \widetilde{x}')\rho_{\widetilde{Y}}(\widetilde{y}, \widetilde{y}')\, \mathrm{d}[(\mathrm{P}_{\widetilde{X}\widetilde{Y}} - \mathrm{P}_{\widetilde{X}}\mathrm{P}_{\widetilde{Y}}) \times (\mathrm{P}_{\widetilde{X}\widetilde{Y}} - \mathrm{P}_{\widetilde{X}}\mathrm{P}_{\widetilde{Y}})].$$

where $d[\cdot]$ denotes the differential term of the integral.

Note that $\theta(\widetilde{X}, \widetilde{Y}) = 0$ if and only if $\widetilde{X}$ and $\widetilde{Y}$ are independent. Given an iid sample $\{(\widetilde{X}_i, \widetilde{Y}_i)\}_{i=1}^n$ of $(\widetilde{X}, \widetilde{Y})$, an empirical estimator of θ is given by

$$\theta_n(\widetilde{X}, \widetilde{Y}) = \frac{1}{n^2}\sum_{i,j} k_{ij}\ell_{ij} + \frac{1}{n^4}\sum_{i,j,q,\tau} k_{ij}\ell_{q\tau} - \frac{2}{n^3}\sum_{i,j,q} k_{ij}\ell_{iq}$$

with $k_{ij} = \rho_{\widetilde{X}}(\widetilde{X}_i, \widetilde{X}_j)$ and $\ell_{ij} = \rho_{\widetilde{Y}}(\widetilde{Y}_i, \widetilde{Y}_j)$. Taking $\widetilde{X} = X$ and $\widetilde{Y} = \varepsilon = Y - \langle X, \beta\rangle_{\mathcal{H}_X}$, $\rho_{\widetilde{Y}}$ is the absolute value and $\rho_{\widetilde{X}}$ is the distance associated to the Hilbert space $\mathcal{H}_X$. The test statistic is $T_n = \theta_n(\hat{\varepsilon}, X)$ and is based on $\{(X_i, Y_i - \langle X_i, \hat{\beta}\rangle_{\mathcal{H}_X})\}_{i=1}^n$.

In other recent papers Hu et al. (2020) and Zhao et al. (2022), the null hypothesis about the linearity given in (7) is tested using related approximations based on the MDD adapted to the functional context.

All the tests described in this section have challenging limit distributions and need to be calibrated with resampling techniques.

The references mentioned above are for the extension of DC and MDD coefficients to specification tests in the functional data context. Specification tests, in general, for independence testing between two functional variables X and Y, conditioned to a third one Z, are a really though problem. Some very relevant and recent papers in this topic are the ones of Shah and Peters (2020) or Lundborg et al. (2022). A deep study of the CDC in the functional framework is still an open problem of interest for future research.

4 Applications

In this last section, we illustrate some of the recently developed new methodologies for specification tests in the functional framework introduced along the document. Three real datasets examples with functional nature are employed.

The first application is an illustration of the test of equality of distribution functions. This is devoted to the Medflies data (Carey et al. (1998)). In this example, the Mediterranean fruit flies' lifetime distributions are compared with respect to their fertility (number of eggs). The distinction is done in terms of short-lived or long-lived individuals. As a result, a test of equality of distribution for functional data is performed (Section 4.1).

Secondly, other well-known data set in the functional framework is employed. This is the Tecator database (see Ferraty and Vieu (2006)). In this application, it is wanted to determine if the spectrometric functional variable (absorbance), as well as its first and second derivatives, support relevant information to explain the fat content in a regression model. For this purpose, significance tests are applied over the considered functional covariates (Section 4.2).

Finally, a GoF test based on CD is applied to check if a Ornstein-Uhlenbeck diffusion process explains the evolution of high-frequency financial data. In particular, Johnson & Johnson stock prices from August 2018 to August 2019 are analyzed (Section 4.3).

4.1 Testing equality of distributions in the Medflies data set

Medflies is a functional dataset usually used for classification purposes. See Carey et al. (1998) for more details. This is available in the `ddalpha` (Pokotylo et al. (2019)) package of `R` (R Core Team (2022)). This contains the medflies trajectories for number of eggs laid differentiation between short or long-lived individuals. The goal is to classify a group of Mediterranean flies as short-lived or long-lived (alive after day 50), X and Y populations, respectively, given their fertility up to day 35. The dataset contains 278 trajectories of long-lived and 256 for the short-lived group. This is considered a hard classification problem and the best overall ratio is around 60%. As a result, it makes sense to wonder if it is possible to correctly discriminate between both groups. For this purpose, a test for comparison of populations in the function context is applied.

First, data on flies that have not laid any egg are removed to avoid outliers. We found this type of individuals for both classes: short-lived as well as long-lived flies. This results in a total of 266 long-lived trajectories and 246 of the short-lived ones for the new dataset. As the number of eggs laid is a discrete variable we considered the raw data as well as a logaritmic transformation to avoid heterogeneity problems. Next, functional data is smoothed using nonparametric kernel estimation. This is done using the `optim.np` function of `fda.usc` library (Febrero-Bande and Oviedo de la Fuente (2012)). We have employed a bandwidth parameter value of $h = 1$, other values could be considered as well. However, we have appreciated a suitable smoothing taking this quantity. Results for first 30^{th} resulting samples are displayed in Figure 1 for both groups.

Thus, after smoothing both functional variables X and Y, corresponding to short-lived and long-lived populations, we are interested in determining if there exits significance differences between both groups in terms of their distributions. For this purpose, a test for comparison of the

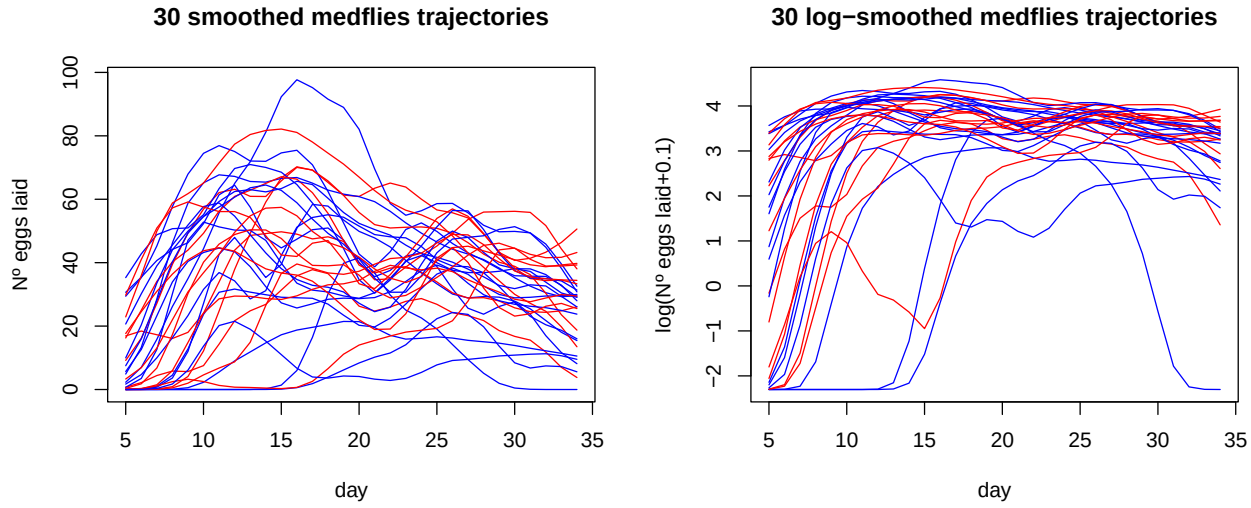

Figure 1: First 30^{th} smoothed medflies trajectories for number of eggs laid (left) and $\log$(number of eggs laid +0.1) (right) for short-lived individuals (—) and long-lived ones (—).

distribution of the populations in the functional framework is needed. This results in testing the null hypothesis of $H_0 : X \sim Y$.

We resort to random projections in functional data (Cuesta-Albertos et al. (2007)) to construct a proper statistic for the test. Once our data is projected, scalar procedures for comparison of populations can be employed. In an illustrative way, we decided to use a total of 10 random projections and then apply Kolmogorov-Smirnov (KS10) and Anderson-Darling (AD10) techniques. For this aim, function `XYRP.test` of the `fda.usc` library (Febrero-Bande and Oviedo de la Fuente (2012)) is applied. Obtained results are collected in Table 1.

	KS10	AD10
smfl	1.3×10^{-4}	6.5×10^{-4}
$\log(\text{smfl} + 0.1)$	0.00804	0.01547

Table 1: Resulting p-values for Kolmogorov-Smirnov (KS10) and Anderson-Darling (AD10) tests using 10 random projections for smoothed medflies trajectories (sfml) and its logarithmic version ($\log(\text{smfl} + 0.1)$).

In view of the results, all p-values< 0.0155, we have evidences to reject the null hypothesis that the number of eggs laid for fruit flies are equally distributed for short and long lived individuals. Thus, we can conclude that there exists difference between the fertility of a fly with long life expectancy compared to one with a lower rate. Therefore, new classification methodologies for the functional data context are needed to correctly discriminate between both groups.

4.2 Significance tests with functional covariates for the Tecator database

The Tecator data set records the content of water, fat and protein percentages jointly with absorbances spectrometric curves, measured in a 100-channel spectrum, of a total of $n = 215$ meat samples. This is available in the `fda.usc` (Febrero-Bande and Oviedo de la Fuente (2012)) package of `R` (R Core Team (2022)). This database is a well-studied real data example in the functional framework. Some examples where this data set is considered are the works of Ferraty and Vieu (2006), García-Portugués et al. (2014), Lee et al. (2020) and Shi et al. (2022) among others. Following previous studies guidelines, we are interested on model the percentage of fat (Y) using the spectrometric curves information (X). In particular, we consider the absorbance (ab) and its first and second derivatives (ab1 and ab2). Representation of the considered covariates is collected in Figure 2.

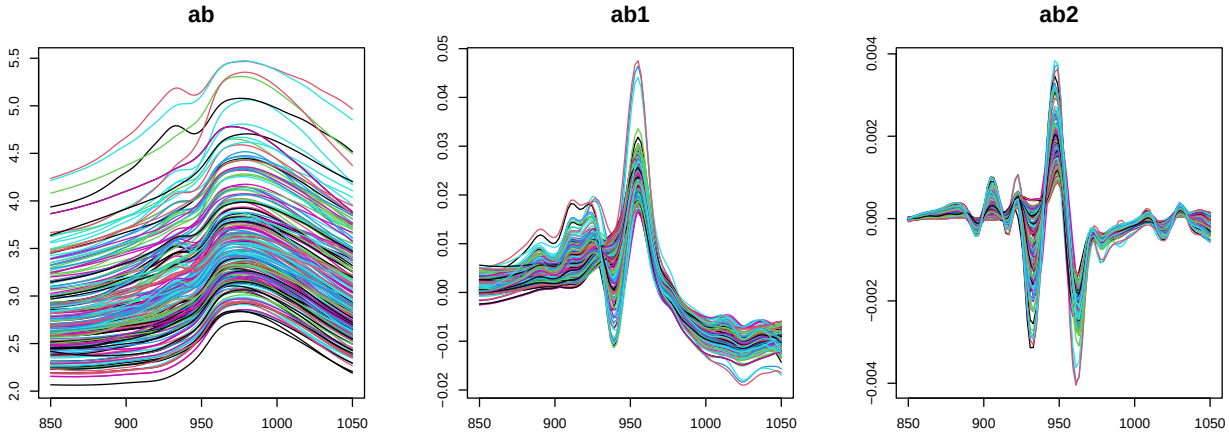

Figure 2: Left: absorbance curves. Middle: first derivative of absorbance curves. Right: second derivative of absorbance curves.

To verify if all considered covariates are relevant in the fat percentage explanation or if some do not support enough information, we can resort to significance tests. In particular, we want to test

$$H_0 : \mathbb{E}\left[Y|_X\right] = \mathbb{E}[Y] \text{ almost surely} \quad \text{vs.} \quad H_1 : \mathbb{P}\left(\mathbb{E}\left[Y|_X\right] \neq \mathbb{E}[Y]\right) > 0,$$

where X can be the absorbance information as well as its first or second derivative. This corresponds with the test displayed in expression (22) of Section 3.3 for the vectorial case.

Following similar ideas to Lee et al. (2020) we can implement the test using the FMDD coefficient introduced previously in Section 3.5. Using a $B = 1000$ resampling wild bootstrap calibration procedure, we obtain null p-values for raw absorbance (ab), first (ab1) and second derivative (ab2). As a result, we have evidences to reject the null hypothesis of conditional mean independence and to claim that these three covariates provide relevant information in the fat percentage explanation. It is interesting to note that no model assumption or structure is needed, as all types of possible dependence in mean are collected in the considered H_0.

Moreover, we can go a step further to detect which covariates are the most and least relevant ones. For this aim, we can define a scale invariant functional martingale difference correlation coefficient (FMDC). This is an extension of the $MDC^2(Y|_X)$ term introduced in (24) for the vectorial context

	ab	ab1	ab2
DC	0.2	0.78	0.91
FMDC	0.45	0.88	0.94

Table 2: Results of distance correlation (DC) and functional martingale difference correlation (FMDC) coefficients for Absorbances (ab), first Absorbances' derivative (ab1) and second one (ab2).

just applying the considered metrics for the functional martingale calculation. Now, we build our functional scale invariant coefficient using the unbiased $FMDD_n^2\left(\mathbf{Y_n}|\mathbf{X_n}\right)$ estimator formula introduced in (33). Results are displayed in Table 2. We see as the ab2 term is the one with the greatest explanation capability, following for ab1 and ab. This highlights the fact that employing the second derivative instead of ab increases the explanatory power of the regression model. Besides, we calculate the DC coefficient and similar results are obtained.

4.3 GoF test for a high-frequency dynamic model example: Johnson & Johnson company stock prices

To illustrate a real-data application for dynamic models, we apply the ideas of DC to test a GoF of the Ornstein-Uhlenbeck process as an autorregresive Hilbertian (ARH) process. We refer the reader to Bosq (2000) for more details about ARH processes.

Let $\{X_t\}_{t\in\mathbb{R}^+}$ be a continuous-time zero-mean stochastic process. Following the ideas in Álvarez-Liébana et al. (2022), we split the path, corresponding to the observed domain of the $t \in \mathbb{R}^+$ term of the stochastic process, as $\mathcal{X}_n(t) = X_{nh+t}$, with $t \in [0,h]$, and $\mathcal{X}_n \in \mathcal{H} = L^2([0,h])$, for each $n \in \mathbb{Z}^+$, constituting an infinite-dimensional discrete-time process. The zero-mean autoregressive Hilbertian process of order one $\mathcal{X} = \{\mathcal{X}_n\}_{n\in\mathbb{Z}^+}$, denoted as ARH(1), satisfies the state equation

$$\mathcal{X}_n(t) = \Lambda\left(\mathcal{X}_{n-1}\right)(t) + \mathcal{E}_n(t), \qquad n \in \mathbb{Z}^+, \quad t \in [0,h],$$

with $\mathcal{X}_n,\ \mathcal{E}_n \in \mathcal{H} = L^2([0,h])$, Λ the linear autocorrelation operator, and $\{\mathcal{E}_n\}_{n\in\mathbb{Z}}$ an independent sequence of Gaussian processes with null mean (strong-white noise) with iid components (see Assumptions considered in Álvarez-Liébana et al., 2022). The Ornstein-Uhlenbeck process,

$$X_t = X_0 e^{-\theta t} + \mu\left(1 - e^{-\theta t}\right) + \sigma\int_0^t e^{-\theta(t-s)}\,\mathrm{d}W_s, \qquad t,s \in \mathbb{R}^+,$$

with X_0 the initial condition at $t_0 = 0$, can be characterized as an ARH(1) process. Let $\mathcal{H}$ be a separable Hilbert space given by $\mathcal{H} = L^2\left([0,h], \mathcal{B}_{[0,h]}, \lambda + \delta_{(h)}\right)$, with $\mathcal{B}_{[0,h]}$ the σ-algebra generated by the subintervals $[0,h]$, λ the Lebesgue measure and $\delta_{(h)}(s) = \delta(s-h)$ the Dirac measure at h. Given a centered process $\{X_t\}_{t\in\mathbb{R}^+}$, the Ornstein-Uhlenbeck process can be characterized as a zero-mean stationary ARH(1) model $\{\mathcal{X}_n(t) := X_{nh+t},\ t \in [0,h]\}_{n\in\mathbb{Z}^+}$, given by

$$\mathcal{X}_n(t) = e^{-\theta t}\mathcal{X}_{n-1}(h) + \sigma\int_{nh}^{nh+t} e^{-\theta(nh+t-s)}\,\mathrm{d}W_s = \Lambda_\theta\left(\mathcal{X}_{n-1}\right)(t) + \mathcal{E}_n(t),$$

with $n \in \mathbb{Z}^+$ and where $\left\{\mathcal{E}_n(t) := \sigma\int_{nh}^{nh+t} e^{-\theta(nh+t-s)}\,\mathrm{d}W_s\right\}_{n\in\mathbb{Z}^+}$ constitutes a $\mathcal{H}$-valued strong white noise and Γ_θ is a bounded linear operator, for each $\theta > 0$.

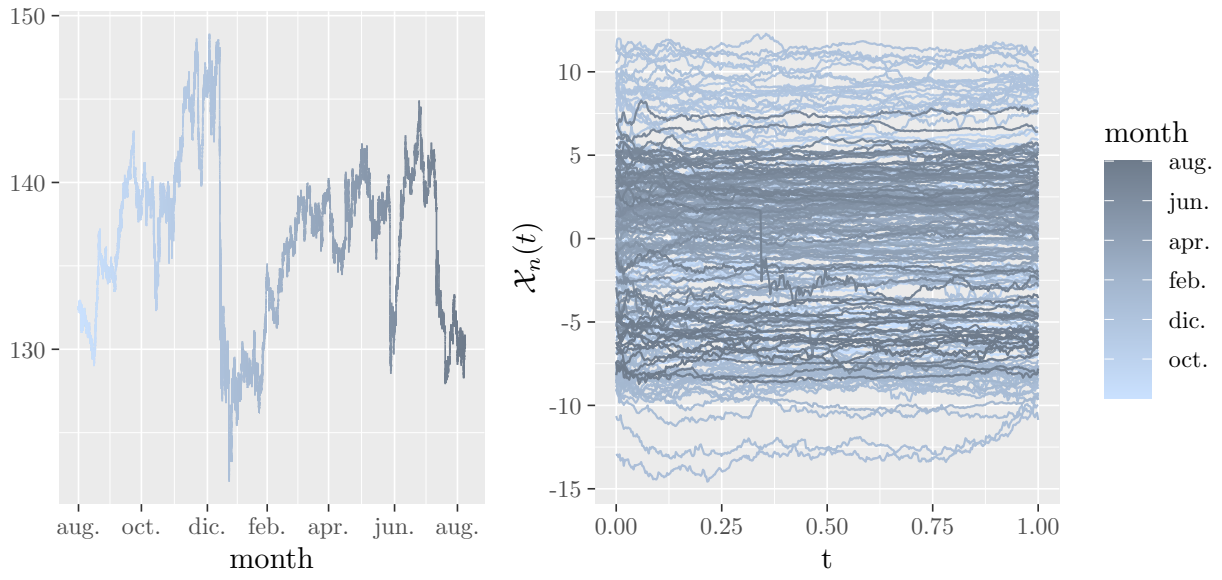

Figure 3: Johnson & Johnson stock prices recorded every minute from August 2018 to August 2019. Observed path (left) and centered daily price curves (right).

The dataset considered consist on Johnson & Johnson stock prices from August 1, 2018 to August 7, 2019, recorded every minute. Figure 3 shows the price path (left) with $98\,280$ observations and the daily curves $\{\mathcal{X}_i(t)\}_{i=1}^n$ with $n = 252$ curves (right) discretized in 390 equispaced grid points, that is, 1-minute data. The daily price curves are evaluated in $\mathcal{H} = L^2\left([0,1], \mathcal{B}_{[0,1]}, \lambda + \delta_{(1)}\right)$, where the $[0,1]$ interval corresponds to a 1-day observation window. We test the parametric form of the Ornstein-Uhlenbeck process, that is, that the daily curves $\{\mathcal{X}_i(t)\}_{i=1}^n$ constitute an ARH(1) process $\mathcal{X}_n(t) = \Lambda(\mathcal{X}_{n-1})(t) + \mathcal{E}_n(t)$ with $\Lambda(\mathcal{X})(t) := \Lambda_\theta(\mathcal{X})(t) = e^{-\theta t}\mathcal{X}(h)$. As in López-Pérez, Febrero-Bande, and González-Manteiga (López-Pérez et al.), to test the specification of the process using the independence test, we have the test

$$H_0 : \mathcal{E}_n(t) \perp \mathcal{X}_n(t) \quad \text{vs.} \quad H_1 : \mathcal{E}_n(t) \not\perp \mathcal{X}_n(t)$$

which is equivalent to test the Ornstein-Uhlenbeck specification. The p-value obtained is 0.0011, therefore the null hypothesis is rejected, as significant evidence is found against the Ornstein-Uhlenbeck as a ARH(1) process for sensible significance levels. Explaining the dynamic of the stock price may require a more intricate model, or coupling the model with jumps, as there was a decline in December due to allegations against the company.

5 Conclusions

In this article, existing procedures about specification tests in the presence of functional data are reviewed. These can be fundamentally differentiated into two types:

a) Extensions of classic procedures developed for the vectorial framework. These are based on distances between a nonparametric universally consistent pilot estimator and another one estimated under the null hypothesis assumptions.

b) Using correlation generalized coefficients. These are employed to measure independence, conditional mean independence and conditional independence. These correspond with the analyzed DC, MDD and CDC coefficients, respectively.

The development of specification tests for the functional context is not an easy task. In fact, most of the references date from the last decade. This field has attracted great interest, resulting in a very fast evolution in the recent years. These novel procedures face some important limitations as the curse of dimensionality in the big data context involving functional data. For this reason, it is currently an interesting line of research for the big data processing.

All the manuscript review is performed for specification tests in static functional models. Nevertheless, an example of specification testing for a functional continuous-time process is given in Section 4.3 to illustrate their possible adaptations.

An open line for future research is the development of specification tests for functional time series. Articles such as the ones of Edelmann et al. (2019), Davis et al. (2018), Dehling et al. (2020), Lee and Shao (2018) or Meintanis et al. (2022) could be a good starting point for construction of new specification tests in dynamic models.

There are several practical problems where functional data are of potential interest. Specially, in the medical context, where a continuous monitoring of patients features can be desirable. An example is the glucose monitoring in diabetes disease. The case of cure models, from the Survival Analysis, is specially relevant. Works as the ones of Zhang et al. (2021) or Edelmann et al. (2022) in the vectorial context based on DC ideas could bridge a gap for specification tests in cure models with functional data.

Eventually, it is important to remark that all the exposition was developed for functional data in Hilbert spaces. There are papers, like Castro-Prado and González-Manteiga (2020) or the excellent review of Jansen (2021), which extend the results to broader spaces. In these last references, a unified version of dependence measures in general metric spaces, being the Hilbertian ones a particular case, is performed. Specification tests for not only the Hilbertian case, but also for general metric spaces, is another open problem for future research.

Acknowledgments

This paper is a consequence of the invitation from Editor of this journal and the President of the Instituto Nacional de Estadística (INE) to produce an article as the second recipient of the Premio Nacional de Estadística. I am really grateful to both of them. The creation of the Spanish National Prize in Statistics was an excellent initiative of professor Juan Manuel Rodríguez Poo (the before mentioned president). This represents a fantastic link between the INE and the developments of Statistics in Spain. This work would not have been possible without the collaboration of all my co-authors: Rosa María Crujeiras Casais, Manuel Febrero Bande, Laura Freijeiro González, Eduardo García Portugués y Alejandra María. López Pérez. My acknowledgment also extends to the rest of my co-authors and all my past students in my academic life. I am an researcher interested in Statistics as a consequence of all I have learned from them.

The research of Wenceslao González-Manteiga is supported by Project PID2020-116587GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" and the Competitive Reference Groups 2021 – 2024 (ED431C 2021/24) from the Xunta de Galicia through the ERDF. Besides, I acknowledge the computational resources from the Supercomputing Centre of Galicia (CESGA).

References

Álvarez-Liébana, J., A. López-Pérez, M. Febrero-Bande, and W. González-Manteiga (2022). A goodness-of-fit test for functional time series with applications to Ornstein-Uhlenbeck processes. arXiv Preprint, https://arxiv.org/abs/2206.12821.

Bárcenas, R., J. Ortega, and A. J. Quiroz (2017). Quadratic forms of the empirical processes for the two-sample problem for functional data. *TEST 26*(3), 503–526.

Bickel, P. J. and M. Rosenblatt (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics 1*(6), 1071–1095.

Boente, G., D. Rodríguez, and M. Sued (2018). Testing equality between several populations covariance operators. *Annals of the Institute of Statistical Mathematics 70*(4), 919–950.

Bongiorno, E. G., A. Goia, and P. Vieu (2019). Modeling functional data: a test procedure. *Computational Statistics 34*(2), 451–468.

Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, Volume 149. Springer Science & Business Media.

Bugni, F. A., P. Hall, J. L. Horowitz, and G. R. Neumann (2009). Goodness-of-fit tests for functional data. *The Econometrics Journal 12*(S1), S1–S18.

Bugni, F. A. and J. L. Horowitz (2021). Permutation tests for equality of distributions of functional data. *Journal of Applied Econometrics 36*(7), 861–877.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313 – 2351.

Carey, J. R., P. Liedo, H. G. Müller, J. L. Wang, and J. M. Chiou (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of mediterranean fruit fly females. *The Journals of Grontology. Series A, Biological Sciences and Medical Sciences 53 4*, B245–51.

Castro-Prado, Fernando and Wenceslao González-Manteiga (2020). Nonparametric independence tests in metric spaces: What is known and what is not. https://arxiv.org/abs/2009.14150.

Chaudhuri, A. and W. Hu (2019). A fast algorithm for computing distance correlation. *Computational Statistics and Data Analysis 135*, 15–24.

Chen, F., Q. Jiang, Z. Feng, and L. Zhu (2020). Model checks for functional linear regression models based on projected empirical processes. *Computational Statistics & Data Analysis 144*, 106897.

Cuesta-Albertos, J. A., E. del Barrio, R. Fraiman, and C. Matrán (2007). The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis 51*(10), 4814–4831.

Cuesta-Albertos, J. A., R. Fraiman, and T. Ransford (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society 37*(4), 477–501.

Cuesta-Albertos, J. A., E. García-Portugués, M. Febrero-Bande, and W. González-Manteiga (2019). Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics 47*(1), 439–467.

Cuevas, A., M. Febrero, and R. Fraiman (2004). An anova test for functional data. *Computational Statistics & Data Analysis 47*(1), 111–122.

Davis, R. A., M. Matsui, T. Mikosch, and P. Wan (2018). Applications of distance correlation to time series. *Bernoulli 24*(4A), 3087 – 3116.

Dehling, H., M. Matsui, T. Mikosch, G. Samorodnitsky, and L. Tafakori (2020). Distance covariance for discretized stochastic processes. *Bernoulli 26*(4), 2758 – 2789.

Delsol, L., F. Ferraty, and P. Vieu (2011). Structural test in regression on functional variables. *Journal of Multivariate Analysis 102*(3), 422–447.

Ditzhaus, M. and D. Gaigall (2018). A consistent goodness-of-fit test for huge dimensional and functional data. *Journal of Nonparametric Statistics 30*(4), 834–859.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics 1*(2), 279–290.

Edelmann, D., K. Fokianos, and M. Pitsillou (2019). An Updated Literature Review of Distance Correlation and Its Applications to Time Series. *International Statistical Review 87*(2), 237–262.

Edelmann, D. and J. Goeman (2022). A Regression Perspective on Generalized Distance Covariance and the Hilbert-Schmidt Independence Criterion. *Statistical Science 37*(4), 562 – 579.

Edelmann, D., T. Welchowski, and A. Benner (2022). A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. *Biometrics 78*(3), 867–879.

Estévez-Pérez, G. and José A. Vilar (2013). Functional anova starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics 20*(3), 495–517.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Febrero-Bande, M., W. González-Manteiga, and M. Oviedo de la Fuente (2019). Variable selection in functional additive regression models. *Computational Statistics 34*(2), 469–487.

Febrero-Bande, M. and M. Oviedo de la Fuente (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software 51*(4), 1–28.

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

Freijeiro-González, L., M. Febrero-Bande, and W. González-Manteiga (2022). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review 90*(1), 118–145.

Fremdt, S., J. G. Steinebach, L. Horváth, and P. Kokoszka (2013). Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics 40*(1), 138–152.

García-Portugués, E., J. Álvarez-Liébana, G. Álvarez-Pérez, and W. González-Manteiga (2020). Goodness-of-fit tests for functional linear models based on integrated projections. In Germán Aneiros, Ivana Horová, Marie Hušková, and Philippe Vieu (Eds.), *Functional and High-Dimensional Statistics and Related Fields*, Cham, pp. 107–114. Springer International Publishing.

García-Portugués, E., J. Álvarez-Liébana, G. Álvarez-Pérez, and W. González-Manteiga (2021). A goodness-of-fit test for the functional linear model with functional response. *Scandinavian Journal of Statistics 48*(2), 502–528.

García-Portugués, E., W. González-Manteiga, and M. Febrero-Bande (2014). A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics 23*(3), 761–778.

Ghale-Joogh, H. S. and S. M. E. Hosseini-Nasab (2018). A two-sample test for mean functions with increasing number of projections. *Statistics 52*(4), 852–873.

González-Manteiga, W. and R. M. Crujeiras (2013). An updated review of goodness-of-fit tests for regression models. *Test 22*(3), 361–411.

González-Manteiga, W., R. M. Crujeiras, and E. García-Portugués (2022). *Trends in Mathematical, Information and Data Sciences*, Volume 445 of *Studies in Systems, Decision and Control*, Chapter A Review of Goodness-of-Fit Tests forÂăModels Involving Functional Data, pp. 349–358. Cham: Springer International Publishing.

González-Rodrígez, G., A. Colubi, and M. A. Gil (2012). Fuzzy data treated as functional data: A one-way anova test approach. *Computational Statistics & Data Analysis 56*(4), 943–955.

Górecki, T. and S. Łukasz (2019). fdanova: an r software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics 34*(2), 571–597.

Gretton, Arthur, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf (2005). Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita (Eds.), *Algorithmic Learning Theory*, Berlin, Heidelberg, pp. 63–77. Springer Berlin Heidelberg.

Guo, J., B. Zhou, J. Chen, and J.-T. Zhang (2019). An L^2-norm-based test for equality of several covariance functions: a further study. *TEST 28*(4), 1092–1112.

Guo, J., B. Zhou, and J.-T. Zhang (2018). Testing the equality of several covariance functions for functional data: A supremum-norm based test. *Computational Statistics & Data Analysis 124*, 15–26.

Guo, J., B. Zhou, and J.-T. Zhang (2019). New tests for equality of several covariance functions for functional data. *Journal of the American Statistical Association 114*(527), 1251–1263.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 21*(4), 1926–1947.

Henze, N. and M. D. Jiménez-Gamero (2021). A test for Gaussianity in Hilbert spaces via the empirical characteristic functional. *Scandinavian Journal of Statistics 48*(2), 406–428.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. New York: Springer.

Horváth, L. and R. Reeder (2013). A test of significance in functional quadratic regression. *Bernoulli 19*(5A), 2130–2151.

Horváth, L. and G. Rice (2015). An introduction to functional data analysis and a principal component approach for testing the equality of mean curves. *Revista Matemática Complutense 28*(3), 505–548.

Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, Volume 997. John Wiley & Sons.

Hu, Wenjuan, Nan Lin, and Baoxue Zhang (2020). Nonparametric testing of lack of dependence in functional linear models. *PLOS ONE 15*(6), 1–24.

Hua, W.-Y. and D. Ghosh (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics 71*(3), 812–820.

Huo, X. and G. J. Székely (2016). Fast computing for distance covariance. *Technometrics 58*(4), 435–447.

Jansen, S. (2021). On distance covariance in metric and Hilbert spaces. *ALEA 18*, 1353–1393.

Jiang, Q., M. Hušková, S. G. Meintanis, and L. Zhu (2019). Asymptotics, finite-sample comparisons and applications for two-sample tests with functional data. *Journal of Multivariate Analysis 170*, 202–220.

Kellner, J. and A. Celisse (2019). A one-sample test for normality with kernel methods. *Bernoulli 25*(3), 1816–1837.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*(1/2), 81–93.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. Texts in Statistical Science Series. CRC Press, Chapman & Hall.

Kolkiewicz, A., G. Rice, and Y. Xie (2021). Projection pursuit based tests of normality with functional data. *Journal of Statistical Planning and Inference 211*, 326–339.

Kong, D., A. M. Staicu, and A. Maity (2016). Classical testing in functional linear models. *Journal of Nonparametric Statistics 28*(4), 813–830.

Lai, T., Z. Zhang, and Y. Wang (2020). Testing independence and goodness-of-fit jointly for functional linear models. *Journal of the Korean Statistical Society 50*.

Lee, C. E. and X. Shao (2018). Martingale Difference Divergence Matrix and Its Application to Dimension Reduction for Stationary Multivariate Time Series. *Journal of the American Statistical Association 113*(521), 216–229.

Lee, C. E., X. Zhang, and X. Shao (2020). Testing conditional mean independence for functional data. *Biometrika 107*(2), 331–346.

Lee, J. S., D. D. Cox, and M. Follen (2015). A two sample test for functional data. *Communications for Statistical Applications and Methods 22*(2), 121–135.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association 109*(505), 266–274.

López-Pérez, A., M. Febrero-Bande, and W. González-Manteiga. A comparative review of specification tests for diffusion models. arXiv Preprint, https://arxiv.org/abs/2208.08420.

Lu, J. and L. Lin (2020). Model-free conditional screening via conditional distance correlation. *Statistical Papers 61*, 225 – 244.

Lundborg, A. R., R. D. Shah, and J. Peters (2022). Conditional independence testing in hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability 41*(5), 3284–3305.

Maistre, S. and V. Patilea (2020). Testing for the significance of functional covariates. *Journal of Multivariate Analysis 179*, 104648.

McLean, M. W., G. Hooker, and D. Ruppert (2015). Restricted likelihood ratio tests for linearity in scalar-on-function regression. *Statistics and Computing 25*(5), 997–1008.

Meintanis, S. G., M. Hušková, and Z. Hlávka (2022). Fourier-type tests of mutual independence between functional time series. *Journal of Multivariate Analysis 189*, 104873.

Park, T., X. Shao, and S. Yao (2015). Partial martingale difference correlation. *Electronic Journal of Statistics 9*(1), 1492 – 1517.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*(3), 1065–1076.

Patilea, V. and C. Sánchez-Sellero (2020). Testing for lack-of-fit in functional regression models against general alternatives. *Journal of Statistical Planning and Inference 209*, 229–251.

Patilea, V., C. Sánchez-Sellero, and M. Saumard (2016). Testing the predictor effect on a functional response. *Journal of the American Statistical Association 111*(516), 1684–1695.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika 13*(1), 25–45.

Pokotylo, O., P. Mozharovskyi, and R. Dyckerhoff (2019). Depth and depth-based classification with R package ddalpha. *Journal of Statistical Software 91*(5), 1–46.

Pomann, G.-M., A.-M. Staicu, and S. Ghosh (2016). A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 65*(3), 395–414.

Qiu, Z., J. Chen, and J.-T. Zhang (2021). Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis 157*, 107160.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics 27*(3), 832–837.

Schick, A. (1997). On U-statistics with random kernels. *Statistics & Probability Letters 34*(3), 275–283.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics 41*(5), 2263 – 2291.

Sen, A. and B. Sen (2014). Testing independence and goodness-of-fit in linear models. *Biometrika 101*(4), 927–942.

Shah, R. D. and J. Peters (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics 48*(3), 1514 – 1538.

Shao, X. and J. Zhang (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association 109*(507), 1302–1318.

Shi, E., Y. Liu, K. Sun, L. Li, and L. Kong (2022). An adaptive model checking test for functional linear model. https://arxiv.org/abs/2204.01831.

Smaga, L. (2022). Projection tests for linear hypothesis in the functional response model. *Communications in Statistics - Theory and Methods 0*(0), 1–18.

Song, F., Y. Chen, and P. Lai (2020). Conditional distance correlation screening for sparse ultrahigh-dimensional models. *Applied Mathematical Modelling 81*, 232–252.

Song, L., A. Smola, A. Gretton, J. Bedo, and K. Borgwardt (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research 13*(1), 1393–1434.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics 25*(2), 613–641.

Su, L. and X. Zheng (2017). A martingale-difference-divergence-based test for specification. *Economics Letters 156*, 162–167.

Székely, G. J. and M. L. Rizzo (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis 117*, 193–213.

Székely, G. J. and M. L. Rizzo (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics 42*(6), 2382 – 2412.

Székely, G. J. and M. L. Rizzo (2017). The energy of data. *Annual Review of Statistics and Its Application 4*, 447–479.

Székely, G. J., M. L Rizzo, N. K. Bakirov, et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

Tekbudak, M. Y., M. Alfaro-Córdoba, A. Maity, and A. M. Staicu (2019). A comparison of testing methods in scalar-on-function regression. *AStA. Advances in Statistical Analysis 103*(3), 411–436.

Teran Hidalgo, S., M. Wu, S. Engel, and M. Kosorok (2018). Goodness-Of-Fit Test for Nonparametric Regression Models: Smoothing Spline ANOVA Models as Example. *Computational Statistics & Data Analysis 122*.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*, Volume 60 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Wang, X., W. Pan, W. Hu, Y. Tian, and H. Zhang (2015). Conditional distance correlation. *Journal of the American Statistical Association 110*(512), 1726–1734.

Wissler, C. (1905). The spearman correlation formula. *Science 22*(558), 309–311.

Xu, K. and D. He (2021). Omnibus model checks of linear assumptions through distance covariance. *Statistica Sinica 31*, 1055–1079.

Zhang, J., Y. Liu, and H. Cui (2021). Model-free feature screening via distance correlation for ultrahigh dimensional survival data. *Statistical Papers 62*(6), 2711–2738.

Zhang, J.-T. and X. Liang (2014). One-Way ANOVA for Functional Data via Globalizing the Pointwise F-test. *Scandinavian Journal of Statistics 41*(1), 51–71.

Zhang, X., S. Yao, and X. Shao (2018). Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics 46*(1), 219 – 246.

Zhao, F., N. Lin, W. Hu, and B. Zhang (2022). A faster U-statistic for testing independence in the functional linear models. *Journal of Statistical Planning and Inference 217*, 188–203.

Zhu, C., X. Zhang, S. Yao, and X. Shao (2020). Distance-based and RKHS-based dependence metrics in high dimension. *The Annals of Statistics 48*(6), 3366 – 3394.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.