Advances in Data Analysis and Classification (2019) 13:1083–1103
https://doi.org/10.1007/s11634-019-00360-z

REGULAR ARTICLE

A Kendall correlation coefficient between functional data

Dalia Valencia[1] · **Rosa E. Lillo**[2] · **Juan Romo**[1]

Received: 26 May 2018 / Revised: 12 May 2019 / Accepted: 17 May 2019 /
Published online: 25 May 2019

Abstract
Measuring dependence is a very important tool to analyze pairs of functional data. The coefficients currently available to quantify association between two sets of curves show a non robust behavior under the presence of outliers. We propose a new robust numerical measure of association for bivariate functional data. We extend in this paper Kendall coefficient for finite dimensional observations to the functional setting. We also study its statistical properties. An extensive simulation study shows the good behavior of this new measure for different types of functional data. Moreover, we apply it to establish association for real data, including microarrays time series in genetics.

Keywords Concordance · Dependence · Functional data · Kendall's tau

Mathematics Subject Classification 62-07 · 62G35 · 62G09

1 Introduction

Many processes currently used in different fields of science and research lead to random observations that can be analyzed as curves. We can also find a large amount of data for which it would be more appropriate to use some interpolation techniques and consider them as functional data. This approach turns out to be essential when data have been observed at different time intervals. Examples of functional data are

✉ Dalia Valencia
djvalega@gmail.com

Rosa E. Lillo
rosaelvira.lillo@uc3m.es

Juan Romo
juan.romo@uc3m.es

[1] Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid 126, Madrid, Spain

[2] Department of Statistics, UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Madrid, Spain

found in areas such as meteorology, where, for example, the ozone level measured during a day is a curve; finance, where, for example, an asset price takes values at very close time instants, and medicine, where the observed gene expressions over time can also be considered as realizations of random curves.

Several multivariate methods have been extended to functional data. Multivariate techniques such as regression (Cardot et al. 1999; He et al. 2000), analysis of variance (Cuevas et al. 2004; Delicado 2007), principal components (Pezulli and Silverman 1993), generalized linear model (Escabias et al. 2004) and depth (López-Pintado and Romo 2007, 2009) have already been extended to a functional context. Other useful methodologies can be found in Ramsay and Silverman (2005). However, there are still some concepts that have not been fully explored for functional data: measures of association and dependence structures between curves, for example.

Leurgans et al. (1993) considered the canonical correlation between two sets of curves. This technique provides a pair of functions called canonical variates and the sample correlation among these variates leads to the canonical correlation between the two sets of curves. He et al. (2000) propose an alternative way of finding the canonical correlation through the extension of multivariate analysis ideas. Opgen-Rhein and Strimmer (2006) proposed an estimator of the dynamic correlation that provides a measure of similarity between pairs of functional observations. It is based on the concept of dynamical correlation introduced by Dubin and Müller (2005) to analyze a nonparametric method to quantify the covariation of components of multivariate longitudinal observations.

In this paper, we extend the Kendall τ correlation coefficient to the functional framework. Kendall's τ allows dependence to be measured in the bivariate case through the definition of concordance, which is based on the idea of order. Since there exists no total order among functions, we use preorders that allow us to sort the functional observations and count the concordant and discordant pairs of a bivariate sample of curves. Once a preorder is introduced, the functional τ coefficient can be defined in a way similar to the bivariate τ coefficient. We show that it fulfils natural properties for a dependence measure and also establish the consistency of the sample version. Finally, we illustrate with simulated and real data the performance of this new dependence measure as well as its robustness, which is a basic characteristic of the Kendall τ in its bivariate version. We analyze a data set corresponding to a microarray time series from a human T-cell experiment with 58 genes, 10 time points and 44 replications. We obtain the functional τ for each pair of genes and construct the partial correlation matrix to compare the gene network resulting from functional τ with those from dynamical correlation.

This paper is organized as follows. In Sect. 2, the functional τ is defined extending the concept of concordance for bivariate random variables. Section 3 is devoted to some properties of this correlation coefficient and to studying convergence results. A summary of the classic techniques, simulation results and sensitivity analysis are given in Sect. 4. Section 5 contains a study of dependence using the genes data set. In Sect. 6, we present a robustness empirical study. Finally, in Sect. 7, we outline the main conclusions of this paper.

2 Functional Kendall correlation coefficient

Kendall (1938) introduced his correlation coefficient τ as a measure for dependence between two random variables X, Y, based on the ranks of sampled observations of X and Y. It makes use of the idea of concordance: two random variables are concordant if large (small) values of one are related to large (small) values of the other. When large (small) values of one are related to small (large) values of the other, the random variables are discordant. More formally, let (x_1, y_1) and (x_2, y_2) be two observations of a random vector (X, Y). We say that (x_1, y_1) and (x_2, y_2) are concordant if $(x_1 - x_2)(y_1 - y_2) > 0$ and discordant if $(x_1 - x_2)(y_1 - y_2) < 0$. This means that they are concordant if either $x_1 < x_2$ and $y_1 < y_2$ or $x_2 < x_1$ and $y_2 < y_1$; in other cases with strict inequality, the observations are discordant.
Kendall's correlation coefficient is defined as the difference between the probabilities of concordance and discordance in two different realizations (X_1, Y_1), (X_2, Y_2) of (X, Y),

$$\tau = P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\}.$$

The above expression can be also written as

$$\tau = 2[P\{X_1 < X_2\ ,\ Y_1 < Y_2\} + P\{X_2 < X_1\ ,\ Y_2 < Y_1\}] - 1.$$

If $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ is a sample from (X, Y), the sample coefficient is

$$\hat{\tau} = \frac{S}{\binom{n}{2}},$$

where $S = cp - dp$ is the difference between the number of concordant pairs (cp) and the number of discordant pairs (dp).

The aim of this paper is to present a functional version of this correlation coefficient where (X, Y) are two random processes $X = \{X(t)|t \in I\}$ and $Y = \{Y(t)|t \in I\}$ that are observed continuously or at disrete time points. Let f and g be in the space $C(I)$ of real continuous functions on the compact interval I. First, we need to introduce relationships allowing the comparison between curves. A natural choice is the usual order, i. e., $f \preceq g \Leftrightarrow f(t) \leq g(t)$, for all $t \in I$. It fulfills the partial order conditions; however, most functions are not comparable with this order. To avoid this difficulty, we waive the antisymmetry condition and use preorders instead of orders.

Definition 1 Let f and g be in $C(I)$. Then, we consider two alternatives.

$$f \preceq_m g \equiv \max_{t \in I} f(t) \leq \max_{t \in I} g(t). \tag{1}$$

$$f \preceq_i g \equiv \int_a^b (g(t) - f(t))dt \geq 0. \tag{2}$$

It follows easily that, for constant functions defined in the same compact interval I, both preorders are equivalent to the usual ordering on the real line. Given any preorder definition among functions, we may define the concordance concept between functions.

Definition 2 (*Functional concordance*) Let $\preceq$ be a preorder between functions, and let $\prec$ address the case without considering ties. Two pairs of functions (f_1, g_1) and (f_2, g_2) are called concordant if either $f_1 \prec f_2$ and $g_1 \prec g_2$ or $f_2 \prec f_1$ and $g_2 \prec g_1$; otherwise they are called discordant.

Definition 2 allows us to extend Kendall's correlation coefficient to the functional case, as described in the next Definition.

Definition 3 (*Functional* τ) If (X_1, Y_1), (X_2, Y_2) are copies of a bivariate stochastic process $\{(X(t), Y(t)) : t \in I\}$, the population version of this dependence measure is

$$\tau = 2[P\{X_1 \prec X_2\ ,\ Y_1 \prec Y_2\} + P\{X_2 \prec X_1\ ,\ Y_2 \prec Y_1\}] - 1. \tag{3}$$

Consider a sample $(x_1, y_1), \ldots, (x_n, y_n)$ of a two-dimensional random process $(X, Y) = \{(X(t), Y(t)) : t \in I\}$ within the compact interval I, with $X, Y \in C(I)$. Then Kendall's extended correlation coefficient is estimated by the empirical version:

$$\hat{\tau} = \binom{n}{2}^{-1} \sum_{i<j}^{n} \left[2I(x_i \prec x_j \quad and \quad y_i \prec y_j) + 2I(x_j \prec x_i \quad and \quad y_j \prec y_i)\right] - 1. \tag{4}$$

Some of the asymptotic properties of the traditional Kendall τ coefficient arise from the fact that it can be expressed as a U-statistic. To obtain an asymptotic result in the functional field, which will be stated in Theorem 2, we need the definition of UB-statistics, U-statistics taking values in a Banach space. We also need some convergence results for this kind of statistics. These concepts can be defined as follows:

Definition 4 (UB-*Statistics*. Borovskikh 1996, page 5) Let B be a real separable Banach space with a norm $\|\cdot\|$ and let B^* be the dual space of real-valued linear functions on B. Denote by $x^*(x)$ the value of functional $x^* \in B^*$ at $x \in B$. Let $X_1, \ldots, X_n$ be independent random variables taking values in the measurable space $(X, \mathfrak{X})$, where $\mathfrak{X}$ is a σ-algebra, and all with identical distribution P. Consider a Bochner integrable symmetric function (kernel) $\Phi : X^m \to B$ of m variables given on X^m and taking values in B . Then, a U-statistic is

$$U_n = \binom{n}{m}^{-1} \sum_{1 \le i_1 < \cdots < i_m \le n} \Phi\{(X_{i1}, \ldots X_{im})\}. \tag{5}$$

It is clear that $U_n \in B$. Hence, the U-statistic (5) with a B-valued kernel Φ is called a UB-statistic. In particular, if $B = R$, it is called a UR-statistic and if $B = H$, where H is a real separable Hilbert space, it is called a UH-statistic.

Let $\mathcal{P} = \{P\}$ be a class a probability distribution on $(X, \mathfrak{X})$. By $\theta : P \to \theta(P)$, we denote a functional given on $\mathcal{P}$ and taking values in B where

$$\theta(P) = \int \dots \int \Phi(x_1, \dots_x m) P(dx_1) \dots P(dx_m)$$

If $E\|\Phi\| < \infty$, then U_n is an unbiased estimate of the B-valued element $\theta(P) = E\Phi(X_1, \dots X_m)$

The following theorem provides an asymptotic result, which will be very useful in what follows.

Theorem 1 (Borovskikh 1996, page 73) *Assume that the B-valued kernel Φ is such that $E\|\Phi\| < \infty$. Then,*

$$U_n \to \theta \quad a.s \quad n \to \infty,$$

and

$$E||U_n - \theta|| \to 0.$$

Now, consider $(X_1, Y_1), \dots, (X_n, Y_n)$ to be independent copies of the bivariate stochastic process (X, Y) with identical distribution P and whose realizations or paths are pairs of functions that take values in the measurable space $(C[a, b] \times C[a, b], \mathfrak{X})$. Then, the functional $\hat{\tau}$ given in (4) can be expressed as a UB-statistic,

$$U_n = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} \Phi\{(X_{i_1}, Y_{i_1}), (X_{i_2}, Y_{i_2})\}, \tag{6}$$

where $\Phi : C^2[a, b] \times C^2[a, b] \to \mathbb{R}$ is a Bochner integrable symmetric function according to Definition 1.3.11 in Schwabik and Guoju (2005), and given by

$$\Phi[(x_i, y_i), (x_j, y_j)] = 2I(x_i \prec x_j \ , \ y_i \prec y_j) + 2I(x_j \prec x_i, y_j \prec y_i) - 1,$$

where I denotes the indicator function.

3 Properties of functional τ

We analyze in this Section some desirable properties of τ as a dependence measure. Scarsini (1984) studies the measures of concordance in terms of copulas and proposes a set of axioms that a concordance measure for ordered pairs of continuous random variables should fulfill. The extension of these axioms to the multivariate case was studied in Taylor (2007, 2008). The following list gives the properties of the functional τ. Some of them come from the axioms proposed by Scarsini (1984). Other properties

are a natural extension of the well known properties of the bivariate τ itself (Kendall 1938).

Let $(X(t), Y(t))$ be a bivariate stochastic process. Then,

1. $\tau(X(t), Y(t)) = \tau(Y(t), X(t))$.
2. $-1 \leq \tau(X(t), Y(t)) \leq 1$.
3. $\tau(-X(t), Y(t)) = -\tau(X(t), Y(t))$.
4. $\tau(X(t), g(X(t))) = 1$, for any monotone increasing function g.
5. $\tau(X(t), g(X(t))) = -1$, for any monotone decreasing g.
6. If $X(t)$ and $Y(t)$ are stochastically independent, then $\tau(X(t), Y(t)) = 0$.
7. The correlation coefficient functional is invariant under strictly increasing and continuous transformations of the functional variables,

$$\tau[\alpha(X(t)), \beta(Y(t))] = \tau(X(t), Y(t)),$$

where α and β are strictly increasing functions.

Note that τ with the preorder (1) verifies 1, 2, 4, 6 and 7, and τ with the preorder (2) verifies 1, 2, 3, 6 but 4, 5 and 7 just for affine transformations. The proof of the properties are given in the Appendix of the Supplementary Material.

The consistency of functional $\hat{\tau}$ is established in the next theorem.

Theorem 2 *Let* $(x_1, y_1), \ldots, (x_n, y_n)$ *be a sample of independent and identical functional observations from* (X, Y). *Then,*

$$\hat{\tau}_n \to \tau \quad a.s. \quad as \quad n \to \infty$$

for the two preorders considered in Definition 1.

It is easy to check that the function

$$\Phi[(x_i, y_i), (x_j, y_j)] = 2I(x_i \prec x_j \ , \ y_i \prec y_j) + 2I(x_j \prec x_i, y_j \prec y_i) - 1,$$

which represents the kernel for the $\hat{\tau}_n$ belongs to the interval $[-1, 3]$. Then, the functional $\hat{\tau}$, given in (4) and expressed as the UB-statistic (6), has associated a kernel Φ such that $E\|\Phi\|$ is finite. Therefore, from Theorem 1, we have that, if Φ is such that $E\|\Phi\| < \infty$, then the UB-statistic will converge almost surely to the parameter τ.

Observe that Theorem 2 is valid in general for any well-defined preorder ($\preceq$).

To illustrate how the functional $\hat{\tau}$ works in simulated functional samples with different kinds of dependence, we provide some examples. From now on, $\hat{\tau}_1$, $\hat{\tau}_2$ denote the functional $\hat{\tau}$ when (1) and (2) preorders are considered, respectively. Consider five joint realizations of the processes $X(t) = t^2 + Z_1$ and $Y(t) = -(t + Z_2)^2 - 8t + Z_2$, where (Z_1, Z_2) follow a bivariate standard normal distribution with correlation σ_{12} representing the random part of the processes. The bivariate functional sample shown in Fig. 1 was generated with a high value of σ_{12} close to 1. Each pair of curves is represented by the same color. In this first case, the ordering for the preorder (1) in the first group is (red > cyan > green > blue > magenta), and for the second group

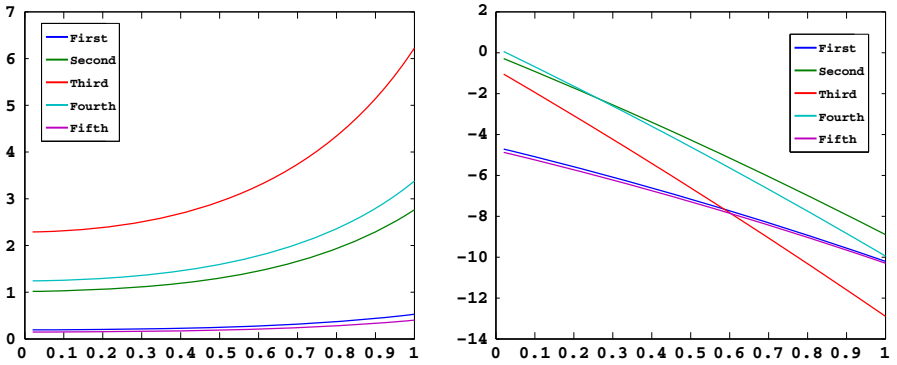

Fig. 1 $\hat{\tau}_1 = 0.6\ \hat{\tau}_2 = 0.4$

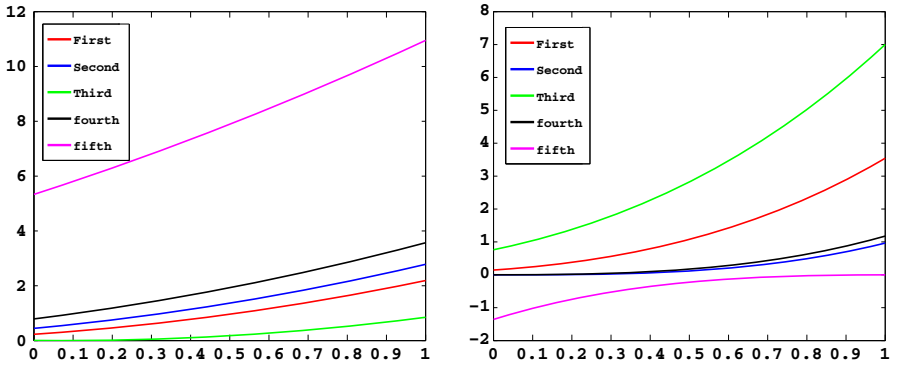

Fig. 2 $\hat{\tau}_1 = -0.8\ \hat{\tau}_2 = -0.8$

it is (cyan > green > red > blue > magenta). In both panels, the cyan and green curves are in the same relative position with respect to the other curves. The blue and magenta curves are also in the same position in both groups. In this case $\hat{\tau}_1 = 0.6$. For the ordering with preorder (2), in the first group it is (red > cyan > green > blue > magenta), and for the second group it is (green > cyan > red > blue > magenta). In both panels, blue and magenta curves are in the same position in the two groups. At the same time, the remainder of the curves are almost completely ordered in the opposite way and $\hat{\tau}_2 = 0.4$. Figure 2 shows another example where both coefficients take the same value.

4 Empirical results and comparisons

In this Section, we illustrate the performance of the functional τ with both preorders given in Eqs. 1 and 2, as well as its behavior with respect to other dependence measures already introduced in the literature. We briefly describe two of them (dynamical correlation and canonical correlation) and, in order to compare our results with these dependence measures, we carry out a simulation study.

A commonly used technique to find the correlation between two groups of functions is the dynamical correlation, which is a measure of similarity between two curves. Dubin and Müller (2005) introduced the dynamical correlation as the following informal idea: "if both trajectories tend to be mostly on the same side of their time average (a constant) then the dynamical correlation is positive; if the opposite occurs, then dynamical correlation is negative". Opgen-Rhein and Strimmer (2006) proposed an estimator for the dynamical correlation considering functional data instead of longitudinal data. We will use the estimator of the dynamical correlation proposed in Opgen-Rhein and Strimmer (2006), which is a slightly revised version of the dynamical correlation introduced in Dubin and Müller (2005):

$$\hat{\rho}_d = \frac{1}{n-1} \sum_{i=1}^{n} \left\langle x_i^s(t), y_i^s(t) \right\rangle,$$

where

$$x^s(t) = \frac{x^c(t)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \langle x_i^c(t), x_i^c(t) \rangle}},$$

and $x^c(t)$ are functions centered in space and time simultaneously, i.e.,

$$x^c(t) = x(t) - \langle \bar{x}(t), 1 \rangle, \quad where \quad \bar{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t),$$

and $\langle \cdot \rangle$ is the usual inner product for functions $\langle x(t), y(t) \rangle = \int_I x(t) y(t) dt$. As we can see, $\hat{\rho}_d$ is an estimator of the population dynamical correlation

$$\rho_d = E \left\langle X^S(t), Y^S(t) \right\rangle,$$

which can been seen as an average of individual correlations.

Another well-known technique to measure functional dependence is the canonical correlation, which was defined in Leurgans et al. (1993). This procedure seeks to investigate which modes of variability in the two sets of curves are most associated with one another. This analysis provides a pair of functions called *canonical variates*

$$(\xi(s), \eta(s))$$

such that $\int \xi X_i$ and $\int \eta Y_i$ are well correlated with one another and the sample correlation between these variables will be what in Leurgans et al. (1993) was called the canonical correlation between the two variables or groups. In a formal way, consider n observed pairs of curves $(x_i(t), y_i(t))$ with t in the same finite interval I and all integrals are taken over I (see Ramsay and Silverman 2005, page 204). Given canonical variates ξ and η, the canonical correlation was defined by Leurgans et al. (1993) as the sample squared correlation of $\int \xi x_i$ and $\int \eta y_i$, i.e.,

Table 1 Steps for calculating $\hat{\tau}_{1,2}$

1. Input: F, G matrices that contain the two sets of functions
2. Compare the functions in each matrix using the maximum preorder or integral preorder
3. Obtain the number of concordant and discordant pairs
4. Calculate $\hat{\tau}_{1,2} = \frac{2*(concordat\ \ pairs - discordant\ \ pairs)}{n*(n-1)}$ (n = total functions)
5. Output: $\hat{\tau}_{1,2}$

$$\hat{\rho}_c(\xi,\eta) = \frac{\left\{cov\left(\int \xi x_i, \int \eta y_i\right)\right\}^2}{\left\{var\left(\int \xi x_i\right) + \lambda||D^2\xi||^2\right\}\left\{var\left(\int \eta y_i\right) + \lambda||D^2\eta||^2\right\}},$$

where λ is a positive smoothing parameter and $||D^2 f||^2 = \int \left(D^2 f\right)^2$, that is, the integrated squared curvature of f that quantifies its roughness. Having a pair of canonical variables with fairly smooth weight functions and correlations that are not excessively low is necessarily a good choice for the smoothing parameter. This parameter can be chosen subjectively, but can also be selected through a cross-validation score if an automatic procedure is required.

As usual in FDA, calculations related to functional data are performed by observing the data in certain points of the functions. Hence, it is necessary before introducing the simulations that are going to be shown in the paper to define the finite dimensional version of these coefficients and the preorders given in Definition (1). Denote by $t_1, \ldots, t_d$ the time points from I where the n sampled functions $x_i = \{x_i(t)|t \in I\}, i = 1, \ldots, n$ have been observed. Then,

- $x_1(t) \preceq_m x_2(t) \Leftrightarrow \max(x_1(t_1), \ldots, x_1(t_d)) \leq \max(x_2(t_1), \ldots, x_2(t_d))$.
- $x_1(t) \preceq_i x_2(t) \Leftrightarrow \frac{t_d - t_1}{2d}[x_1(t_1) + x_1(t_d) + 2\sum_{i=2}^{n-1} x_1(t_i)] \leq \frac{t_d - t_1}{2d}[x_2(t_1) + x_2(t_d) + 2\sum_{i=2}^{n-1} x_2(t_i)]$.

In the same way we can order the functions $y_i(t)$. The last expression corresponds to the composite trapezoidal rule of numerical integration, which has been used for calculating the values of the integrals. Then, the coefficients are calculated as

$$\hat{\tau} = \binom{n}{2}^{-1} \sum_{i<j}^{n} \left[2I(x_i \prec x_j \ \ and \ \ y_i \prec y_j) + 2I(x_j \prec x_i \ \ and \ \ y_j \prec y_i)\right] - 1.$$

Since $\mathbf{R^d}$ is a Banach space, the convergence of the $\hat{\tau}_{n,d} \to \tau_d$ a.s. as $n \to \infty$, follows directly by from a proof similar to Theorem 2. From now on, subscript d is deleted from the notation.

The algorithms to calculate the coefficients were implemented in MATLAB using the preorders defined for the finite dimensional case. Table 1 summarizes the steps necessary to obtain them.

Once we have defined the dependence measures that will be used to compare the performance of our coefficient, we show through simulation exercises the behavior of the measure introduced in this paper and those chosen to compare it.

The data are simulated in the following way. Consider the bivariate stochastic process $(X(t), Y(t)) = [f_1(t, Z_1), f_2(t, Z_2)]$, where (Z_1, Z_2) represents the random part of the process, a bivariate standard normal distribution with correlation σ_{12}. We consider a different structure for the functions $f_i, i = 1, 2$ as well as different values for σ_{12}. In each case, 50 realizations of the process $(X(t), Y(t))$ are generated, where the paths are discretized taking $d = 50$ points over the interval [0, 1] and calculating the measures of dependence previously mentioned. This procedure is carried out 100 times and the results reported refer to the average and deviation over the 100 setups.

Table 2 presents the average of the measures $\hat{\tau}_1$ and $\hat{\tau}_2$ as well as $\hat{\rho}_c$ and $\hat{\rho}_d$, which denote the canonical correlation and dynamical correlation, respectively. The value in brackets reports the standard deviation of the measures considered. We also include, in each case, the value of the correlation σ_{12}. We can see that the coefficients $\hat{\tau}_1$ and $\hat{\tau}_2$ in some cases take different values, which is a consequence of the preorders not sorting the data in the same way. In the case of processes in which one of them is an increasing transformation of the other, both coefficients take value 1, which confirms the perfect dependence between the processes considered. However, this fact does not occur in the measures used for comparison, see for example rows 3 and 4 in Table 2. Indeed, the value of $\hat{\rho}_d$ in row 4 does not reflect the true dependence between those processes, which is positive and perfect. Observe that a similar conclusion can be drawn when the dependence is perfect but negative as may be seen in row 5. There, only our coefficients were able to capture the negative perfect dependence. Note also that in the independent case (row 11), our coefficients reflect this fact better than the other measures. Finally, the standard deviation of $\hat{\tau}_2$ in most cases is the smallest among the other measures.

We can see that the canonical correlation $\hat{\rho}_c$ is always positive, which means that it does not capture the direction of the dependence. This is because it seeks variability in the two sets of curves that maximize the sample correlation between the pairs of canonical variates. Dynamical correlation $\hat{\rho}_d$ reflects the average of individual similarities rather than considering the set of curves as a sample of the same population. This makes the dynamical correlation capture changes at an individual performance level, while Kendall's coefficient detects changes at a more general level, since it is inspired in the bivariate coefficient, which is based on evaluating the relationship of the two sets of data.

Thus, the functional $\hat{\tau}$ is appropriate to indicate how related two functional variables are, regardless of the shape of their realizations. This coefficient measures the joint tendency of the variables to have increasing or decreasing behavior.

In the bivariate case, the decision of which coefficient to use depends on several factors, such as the type of measurement scale in which each variable is expressed, the nature of the distributions (continuous or discrete) and if the dependence sought is linear or nonlinear. Also, each coefficient is designed so that the dependence that it is able to capture measures different aspects such as concordance and discordance in the data or the ordering relation of a group of data with respect to another, etc. Having several coefficients to measure dependency from different points of view is therefore natural in the bivariate case, as happens in the functional case. The final user of these measures will make her decisions based on the joint information offered by the pull of coefficients available to him. In the case of functional data, as in the multivariate

Table 2 Dependence measures in simulated data

	$X(t) = f_1(t, Z_1)$	$Y(t) = f_2(t, Z_2)$	σ_{12}	$\bar{\hat{\tau}}_1$	$\bar{\hat{\tau}}_2$	$\bar{\hat{\rho}}_c$	$\bar{\hat{\rho}}_d$
1	$(t+Z_1)^3+(t+Z_1)^2+3(t+Z_1)$	$(t+Z_2)^2+\frac{7}{8}(t+Z_2)-10$	0.8	0.4861	0.4874	0.7448	0.7098
				(0.0657)	(0.0711)	(0.0898)	(0.1139)
2	$\sin(t+Z_1)$	$\cos(t+Z_2)$	-0.7	0.3084	0.2774	0.5367	0.3605
				(0.0923)	(0.0835)	(0.1004)	(0.11)
3	$(t+Z_1)^2$	$(t+Z_1)^4$	1	1	1	0.9566	0.922
				(0)	(0)	(0.0118)	(0.0125)
4	$(t+Z_1)^2+7(t+Z_1)+2$	$((t+Z_2)^2+7(t+Z_2)+2)^3$	1	1	1	0.9989	0.7779
				(0)	(0)	(0)	(0.0347)
5	$(t+Z_1)^2+7(t+Z_1)+2$	$1-((t+Z_2)^2+7(t+Z_2)+2)^3$	1	-1	-1	0.999	-0.78
				(0)	(0)	(0.0009)	(0.0275)
6	$\exp(t+Z_1)$	$(t+Z_2)^3+(t+Z_2)^2+3(t+Z_2)$	0.6	0.4047	0.4138	0.5098	0.5682
				(0.0811)	(0.0751)	(0.1431)	(0.1301)
7	$\exp(t+Z_1)^2$	$\cos(t+Z_2)$	-0.8	0.3097	0.2982	0.3101	0.0408
				(0.0922)	(0.1035)	(0.07)	(0.1458)
8	$\sin(t+Z_1)$	$(t+Z_2)^2$	0.4	0.1080	0.1059	0.3382	0.1647
				(0.1035)	(0.1021)	(0.1132)	(0.0916)
9	$(t+Z_1)^2+9(t+Z_1)-5$	$\cos(3t+Z_2)$	1	-0.7198	-0.9476	0.9334	-0.7244
				(0.0853)	(0.0358)	(0.0458)	(0.0562)
10	$\exp(t^2+Z_1)$	$(t+Z_2)^2-8t+Z_2$	0.9	0.3621	0.5991	0.8544	0.4620
				(0.1078)	(0.0706)	(0.0485)	(0.1215)
11	$\exp(t+Z_1)$	$\sin(t+Z_2)$	0	-0.0076	0.0087	0.1438	0.0560
				(0.1004)	(0.0883)	(0.0861)	(0.1275)

Table 3 Sensitivity to sample size

Sample size	Model 1 $\bar{\hat{\tau}}_1$	Model 1 $\bar{\hat{\tau}}_2$	Model 2 $\bar{\hat{\tau}}_1$	Model 2 $\bar{\hat{\tau}}_2$
25	0.4035	0.4017	0.2809	0.3014
	(0.1285)	(0.1129)	(0.1475)	(0.1429)
50	0.4044	0.4190	0.3084	0.2774
	(0.0719)	(0.0724)	(0.0923)	(0.0835)
100	0.4130	0.4047	0.2882	0.2945
	(0.0575)	(0.0495)	(0.0600)	(0.0636)
150	0.4093	0.4094	0.2999	0.2880
	(0.0394)	(0.0485)	(0.0517)	(0.0489)
1000	0.4077	0.4096	0.2903	0.2945
	(0.0162)	(0.0185)	(0.0219)	(0.0196)

case, this diversity of coefficients increases because there is no total order of the data, but partial orders that preclude a canonical definition of the dependency measures. This is exactly what happens for example in the definition of depth measurements for functional or multivariate data that, depending on each "idea" of order, originates a different measure.

As we can see, $\hat{\tau}$ depends on the sample size n and on the number of points d used to discretize the functions. In order to assess the stability of the functional $\hat{\tau}$ with respect to (n, d), we perform two sensitivity analysis, using the following two pairs of stochastic processes:

- Model 1: $X(t) = \exp(t + Z_1)$ and $Y(t) = (t + Z_2)^3 + (t + Z_2)^2 + 3(t + Z_2)$ with $\sigma_{12} = 0.6$.
- Model 2: $X(t) = \sin(t + Z_1)$ and $Y(t) = cos(t + Z_2)$ with $\sigma_{12} = -0.7$.

The first analysis has as its objective to evaluate the sensibility with respect to the sample size n. In this case, we move $n = 25, 50, 100, 150$ and 1000 with $d = 50$ (the number of points to discretize the functions) fixed. This procedure is repeated 100 times and we report their average and standard deviation. Table 3 shows that the changes in $\bar{\hat{\tau}}_1$, $\bar{\hat{\tau}}_2$ are negligible and they are quite stable with respect to the sample size.

Now, the same scheme is applied to d. Fix $n = 50$, and consider $d = 25, 50, 100, 150$ and 1000 points. This procedure is repeated 100 times. Table 4 illustrates the sensitivity with respect to d. It is noteworthy that the coefficients present good stability with respect to the number of points taken to discretize the functions. We also carry out the sensitivity analysis for other models, but we do not report them in this work, since we obtain the same conclusions.

We have also smoothed the curves of Table 2 with cubic spline and take the smoothing parameter $\lambda = 0.1, 0.5$, and 1. The results are given in Table 5. In general, the coefficient $\hat{\tau}_2$ is more stable than $\hat{\tau}_1$ when we vary the smoothing parameter as also happened when the data are not smoothed.

Table 4 Sensitivity to the number of points in the discretization

Number of points	Model 1 $\bar{\hat{\tau}}_1$	Model 1 $\bar{\hat{\tau}}_2$	Model 2 $\bar{\hat{\tau}}_1$	Model 2 $\bar{\hat{\tau}}_2$
25	0.4056	0.4163	0.3017	0.2822
	(0.0761)	(0.0770)	(0.0902)	(0.0843)
50	0.4108	0.4071	0.2834	0.3052
	(0.0788)	(0.0769)	(0.0929)	(0.0917)
100	0.4071	0.3911	0.2761	0.2786
	(0.0769)	(0.0772)	(0.0837)	(0.0829)
150	0.4163	0.4109	0.2983	0.2801
	(0.0770)	(0.0804)	(0.0919)	(0.0825)
1000	0.4108	0.4092	0.3005	0.2978
	(0.0788)	(0.0822)	(0.0839)	(0.0834)

Table 5 Sensitivity to smoothing parameter (λ)

λ	$\hat{\tau}_1$			$\hat{\tau}_2$		
	0.1	0.5	1	0.1	0.5	1
1	0.4759	0.4759	0.4759	0.4694	0.4694	0.4694
2	0.1527	0.1559	0.1869	0.2180	0.2180	0.2180
3	1	1	1	1	1	1
4	0.9788	0.9804	1	1	1	1
5	−0.1086	−0.1363	−1	−1	−1	−1
6	0.2735	0.2735	0.2735	0.2735	0.2735	0.2735
7	0.2457	0.2490	0.2180	0.3045	0.3045	0.3045
8	0.2816	0.280	0.2963	0.2816	0.3061	0.3061
9	−0.5135	−0.4857	−0.7502	−0.9478	−0.9478	−0.9478
10	0.5282	0.5282	0.5282	0.6473	0.6473	0.6473
11	0.073	0.0057	0.0743	0.0792	0.0792	0.0792

As already mentioned, the functional Kendall's τ coefficient depends on the preorder used. We have observed that when the sets of curves do not cross each other, either of the preorders would work properly and they would be very similar. However, when the function bundle presents crosses and different shapes, it seems more advisable to use the preorder of the integral (coefficient $\hat{\tau}_2$).

To finish the simulation study, we have obtained the bootstrap intervals to capture the uncertainty related to the estimation of coefficient $\hat{\tau}_2$. For this procedure we have used the standard nonparametric bootstrap interval. Table 6 shows five pairs of processes that have been re-sampled 500 times and provides the estimated coefficient $\hat{\tau}_2$ and the 95% confidence intervals. As can be seen, the intervals contain values very close to the one obtained with the coefficient.

Table 6 Confidence intervals

Stochastic process 1	Stochastic process 2	$\widehat{\tau_2}$	Confidence interval
$(t+Z_1)^2+7(t+Z_1)+2$	$((t+Z_2)^2+7(t+Z_2)+2)^3$	1	(0.9698, 0.9869)
$\exp(t+Z_1)$	$\sin(t+Z_2)$	0.0087	(−0.1755, 0.2278)
$\sin(t+Z_1)$	$(t+Z_2)^2$	0.1059	(0.0433, 0.3673)
$\sin(t+Z_1)$	$\cos(t+Z_2)$	0.2774	(0.0980, 0.4735)
$\exp(t+Z_1)$	$(t+Z_2)^3+(t+Z_2)^2+3(t+Z_2)$	0.4138	(0.2302, 0.5771)

5 Gene data

Existing relationships among genes contain broad information on the structure and functioning of living beings. Accordingly, the interaction between genes allows us to understand many life phenomena. These interactions give rise to the construction of genetic networks. By studying the structural properties of such networks, much more information may be extracted in order to understand the complex functioning of living organisms. Different statistical methodologies have been used to estimate genetic networks, such as graphical models which represent stochastic conditional dependence between the investigated variables. Graphical Gaussian models and the Bayesian network are examples of simple graphical models (see, e.g., Whittaker 1990), but their drawback is that these methods are based on the assumption of identically and independently distributed variables. Opgen-Rhein and Strimmer (2006) studied the graphical Gaussian models from the perspective of functional data, where these two assumptions are not necessary.

Opgen-Rhein and Strimmer (2006) considered the gene expression as a functional observation, rather than describing the individual time points separately. They built the networks in the following way: the network nodes are the genes and the correlations are the connectivity strengths assigned to the edges of the network. They consider the dynamical correlation introduced in Sect. 4. However, they do not use the dynamical correlation itself because it represents only marginal dependencies, including indirect interactions between two variables, since it contains information on the relationships of each variable with the rest. They use the concept of partial correlation, which describes the correlation between any two variables i and j, conditioned on all the other variables, which is the correlation between two variables when the effect of the other is eliminated. Therefore, if the variables are linearly and conditionally associated, the partial correlation coefficient is different from zero.

The partial correlation matrix is constructed as follows: Let $\mathbf{P} = (\rho_{kl})$ be the correlation coefficients, and let Ω be the inverse relationships

$$\Omega = \mathbf{P}^{-1} = (w_{ij});$$

then the partial correlations are given by

$$\tilde{\rho}_{kl} = \frac{-w_{kl}}{\sqrt{w_{kk}w_{ll}}} \quad \text{and} \quad \tilde{\mathbf{P}} = (\tilde{\rho}_{kl}).$$

To test the significance of these correlations and decide which are significant edges, they employ a large-scale simultaneous hypothesis testing, the "local fdr", which is an empirical Bayes estimator of the false discovery rate proposed by Efron (2004) and (2005). This method computes the posterior probability for an edge to be present or absent in the gene network. An important question in the use of this method is whether we can identify a small percentage of interesting cases that deserve further investigation. In this study, these cases will be the edges present in the network.

We propose a new form of finding connectivity strengths (edges) using the functional $\hat{\tau}_2$ and applying the "local fdr" to investigate valid relations.

In order to compare the results, we use the same pre-processed data as in the paper by Opgen-Rhein and Strimmer (2006). The data set characterizes the response of a human T-cell line (Jirkat) to a treatment with PMA and ioconomin. After preprocessing the time course data, 58 genes measured across 10 time points with 44 replications are provided, (see Rangel et al. 2004). Data were smoothed with linear spline, taking four basis functions and a smoothing parameter $\lambda = 0.00001$. Table 7 shows the correlation coefficients including the canonical correlation $\hat{\rho}_c$ and dynamical correlation $\hat{\rho}_d$ for some pairs of genes. Note how the correlations vary depending on the coefficient used, which was considered when we analyzed simulated data in Sect. 4.

In order to compare our results with those obtained by Opgen-Rhein and Strimmer (2006), we calculate the partial correlation matrix from the correlation matrix found with the functional $\hat{\tau}_2$ and we use the "local fdr" algorithm in GeneNet packages, available in library R-software, to find whether significant edges are present or absent in our network, with the same cut-off $= 0.2$ used for calculating the network with dynamical correlation.

Figures 3 and 4 show the network proposed by Opgen-Rhein and Strimmer (2006) and our proposed network, respectively. The network calculated with partial dynamical correlation contains 15 nodes and 9 edges, whereas the network calculated with partial functional $\hat{\tau}_2$ contains 22 nodes and 12 edges. The common nodes in both networks are CASP8, SOD1, MAPK9, CDC2 and CCNA.

The advantage of using functional $\hat{\tau}_2$ instead of the dynamical correlation studied in Opgen-Rhein and Strimmer (2006) is that our coefficient identifies relationships between the variables based on the relative ordering among realizations in each group. And it is not only based on the shape of individual realizations; our coefficient also takes into account the temporal evolution of each gene, so it is able to identify additional and different relationships than those given by the dynamical correlation.

Tables 8 and 9 show the results of partial correlation with dynamical correlation and partial correlation with functional $\hat{\tau}_2$, respectively, which were found through the "local fdr" algorithm. In addition, we can see the $p\ value$ for each of the correlations as well as the nodes included in the networks.

Finally, to explore the relationship between the dynamical correlation and the functional $\hat{\tau}_2$, we make a regression analysis between the partial dynamical correlation and partial functional $\hat{\tau}_2$ for T-cell data. We obtain a $R^2 = 0.0634$, which is low and indicates a weak relationship.

Table 7 Gene data

GEN 1	GEN 2	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\rho}_c$	$\hat{\rho}_d$
RB1	CCNG1	− 0.3425	− 0.3996	0.8296	− 0.3266
TRAF5	CLU	− 0.3975	− 0.3383	0.7322	− 0.2461
MAPK9	SIVA	0.3298	0.3890	0.9031	0.4665
EDG9	ZNFN1A1	− 0.1839	− 0.3858	0.9081	− 0.011
IL4R	MAP2K4	0.2656	0.2706	0.9063	0.4193
JUND	LCK	− 0.2146	− 0.2114	0.9311	− 0.4443
SCYA2	PPSGKA1	− 0.1522	− 0.2622	0.6055	− 0.1518
ITGAM	CTNNB1	0.0962	0.0317	0.8491	0.2373
SMN1	CASP8	− 0.0338	− 0.1755	0.9311	− 0.7743
E2F4	PCNA	0.3869	0.4989	0.9394	0.6312
CCNC	PDE4B	− 0.3087	− 0.5687	0.8562	− 0.5738
IL16	APC	− 0.2474	− 0.3192	0.7916	− 0.1763
ID3	SLA	− 0.4027	− 0.4334	0.8905	− 0.7363
CDK4	EGR1	0.1734	− 0.2421	0.9605	0.2091
TCF12	MCL1	0.3467	0.2960	0.9610	0.8361
CDC2	SOD1	0.0486	0.4080	0.9749	0.4871
CCNA2	PIG3	− 0.4017	− 0.4820	0.9361	− 0.3394
IRAK1	SKIIP	− 0.0560	− 0.1871	0.5658	0.1197
MYD88	CASP4	0.4778	0.4376	0.9266	0.2225
TCF8	API2	− 0.0063	− 0.1966	0.9292	0.5261
GATA3	RBL2	0.3467	0.4038	0.9352	0.5604
C3X1	IFNAR1	0.2653	0.3805	0.8923	0.6694
FYB	IL2R6	− 0.0782	0.5254	0.9301	0.3324
CSF2RA	MPO	− 0.4588	− 0.4778	0.9048	0.0831
API1	CYP19	− 0.3245	0.1036	0.9116	0.1227
CIR	CASP7	− 0.2220	− 0.3827	0.8003	− 0.2234
MAP3K8	JUNB	− 0.3044	− 0.4630	0.8913	− 0.6764
IL3RA	NFKBIA	− 0.4165	− 0.3848	0.7861	− 0.1457
LAT	AKT1	− 0.3404	− 0.1649	0.8210	− 0.0764
RB1	MAPK9	0.5328	0.6964	0.9767	0.7740
RB1	CASP4	− 0.4567	− 0.4207	0.9672	− 0.4748
TRAF5	LCK	0.3647	0.5856	0.8970	0.4583
TRAF5	ITGAM	− 0.4820	− 0.5941	0.9494	− 0.6519
TRAF5	CTNNB1	0.4397	0.5920	0.8145	0.2573
TRAF5	CSF2RA	− 0.5116	− 0.6342	0.9318	− 0.6458
EDG9	C3X1	0.5370	0.7030	0.9626	0.6056
ZNFN1A1	CASP8	− 0.2611	− 0.63	0.9467	− 0.4740
IL4R	ITGAM	0.4926	0.5856	0.9611	0.8036
MAP2K4	IL16	0.1078	0.1015	0.6217	0.0634
JUND	SMN1	− 0.5846	− 0.4419	0.9528	− 0.6019

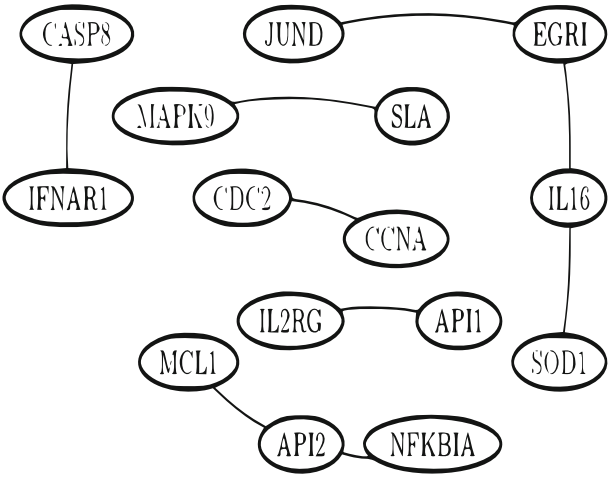

Fig. 3 Gene dependence network using dynamical correlation

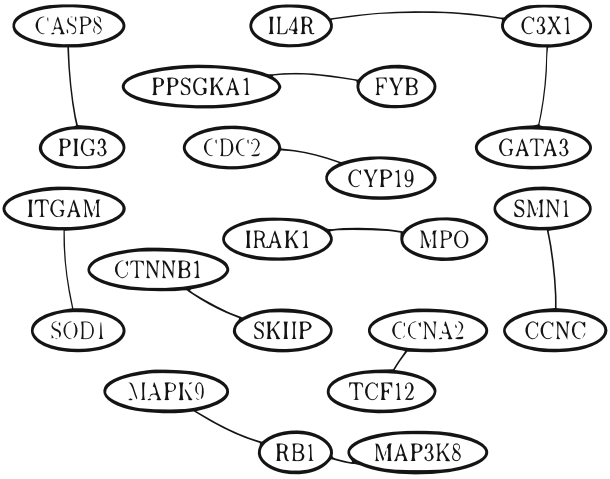

Fig. 4 Gene dependence network using functional $\hat{\tau}_2$

6 Robustness

As commented in the Introduction, we now analyze the robustness of our coefficients $\hat{\tau}_1$ and $\hat{\tau}_2$ and compare them with the results obtained with the dynamical and canonical correlation ($\hat{\rho}_d$ and $\hat{\rho}_c$, respectively). We contaminate the dataset with outliers, defining a functional outlier as in Febrero et al. (2008): a "curve [that] has been generated by a stochastic process with a different distribution than the rest of curves, which are assumed to be identically distributed". Given this definition, we use three types of outliers: shape outliers, magnitude outliers and shape–magnitude outliers.

Table 8 Partial correlation with dynamical correlation

Correlation	node1	node2	pval	prob
0.5196	JUND	EGRI	$4.549e{-}09$	0.9821
0.3971	CDC2	CCNA2	$1.490e{-}05$	0.9821
0.3888	API2	$NFKBIA$	$2.325e{-}05$	0.9821
0.3817	CASP8	IFNAR1	$3.365e{-}05$	0.9778
0.3749	IL16	EGRI	$4.755e{-}05$	0.9317
-0.3543	MAPK9	SLA	$0.2917e{-}04$	0.9317
0.3503	IL16	SOD1	$1.560e{-}04$	0.9317
0.3477	IL2RG	API1	$1.759e{-}04$	0.9079
0.3414	MCL1	API2	$2.337e{-}04$	0.8790

Table 9 Partial correlation with functional $\hat{\tau}_2$

Correlation	node1	node2	pval	prob
-0.3235	PPS6KA1	FYB	$2.286e{-}05$	0.9599
0.3029	IRAK1	MPO	$7.744e{-}05$	0.9599
0.3019	SMN1	$CCNC$	$8.202e{-}05$	0.9599
0.2990	RB1	MAP3K8	$9.678e{-}05$	0.9400
0.2932	RB1	MAPK9	$1.336e{-}04$	0.9287
-0.2842	ITGAM	SOD1	$2.184e{-}04$	0.9287
-0.2839	CDC2	CYP19	$2.211e{-}04$	0.8543
-0.2687	IL4R	C3X1	$4.880e{-}04$	0.8543
-0.2680	GATA3	C3X1	$5.059e{-}04$	0.8543
0.2628	CASP8	PIG3	$6.554e{-}04$	0.8543
0.2627	CTNNB1	SKIIP	$6.586e{-}04$	0.8543
0.2600	TCF12	CCNA2	$7.488e{-}04$	0.8543

We generate 50 curves for the previously studied processes. (Recall that σ_{12} is the correlation between the normal random variables Z_1 and Z_2.)

$$X(t) = \exp(t + Z_1), \quad and \quad Y(t) = (t + Z_2)^3 + (t + Z_2)^2 + 3(t + Z_2), \quad \sigma_{12} = 0.6,$$

and the types of outliers to be considered are:

- Shape outliers. Changing the argument, t to $(1 - t)$.
- Magnitude outliers. Adding a constant to the original process, $X(t)$ to $X(t) + k$. In our case we will use $k = 60$.
- Shape–magnitude outliers. Changing the argument and adding a constant to the original function, $X(t)$ to $X(1 - t) + k$.

Table 10 Contamination with shape outliers

Contaminated groups	Type of outliers	No. outl	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\rho}_d$	$\hat{\rho}_c$
None	None	0	0.454	0.454	0.549	0.544
$X(t)$	Shape	1	0	0	**0.0231**	0.0007
$X(t)$	Shape	2	0	0	0.0242	**0.0669**
$X(t)$	Shape	3	0	0	0.0244	**0.1292**
$X(t)$	Shape	4	0	0	0.0245	**0.1284**
$X(t)$, $Y(t)$ same position	Shape	1	0	0	0	**0.2122**
$X(t)$, $Y(t)$ same position	Shape	2	0	0	0	**0.4137**
$X(t)$, $Y(t)$ same position	Shape	3	0	0	0	**0.2707**
$X(t)$, $Y(t)$ same position	Shape	4	0	0	0	**0.27**
$X(t)$, $Y(t)$ different position	Shape	1	0	0	**0.0296**	0
$X(t)$, $Y(t)$ different position	Shape	2	0	0	0.0301	**0.0698**
$X(t)$, $Y(t)$ different position	Shape	3	0	0	0.0303	**0.1446**
$X(t)$, $Y(t)$ different position	Shape	4	0	0	0.0305	**0.1393**

Bold values indicate the greatest variation

Table 11 Contamination with magnitude outliers

Contaminated groups	Type of outliers	No. outl	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\rho}_d$	$\hat{\rho}_c$
None	None	0	0.454	0.454	0.549	0.544
$X(t)$	Magnitude	1	0.0033	0.0033	**0.096**	0.002
$X(t)$	Magnitude	2	0.0016	0	0.009	**0.043**
$X(t)$	Magnitude	3	0.008	0.008	0.17	**0.18**
$X(t)$	Magnitude	4	0.026	0.026	0.095	**0.126**
$X(t)$, $Y(t)$ same position	Magnitude	1	0.008	0.009	0.16	**0.34**
$X(t)$, $Y(t)$ same position	Magnitude	2	0.0131	0.0147	0.2757	**0.4022**
$X(t)$, $Y(t)$ same position	Magnitude	3	0.0163	0.0196	0.3346	**0.4239**
$X(t)$, $Y(t)$ same position	Magnitude	4	0.0343	0.0375	0.3419	**0.4292**
$X(t)$, $Y(t)$ different position	Magnitude	1	0.0196	0.0245	**0.1786**	0.0079
$X(t)$, $Y(t)$ different position	Magnitude	2	0.0212	0.0261	**0.1766**	0.0384
$X(t)$, $Y(t)$ different position	Magnitude	3	0.0131	0.0196	0.1135	**0.1652**
$X(t)$, $Y(t)$ different position	Magnitude	4	0.1192	0.1274	**0.2091**	0.1076

Bold values indicate the greatest variation

We use different ways to contaminate the data:

1. Contaminating a group.
2. Contaminating two groups in the same position, i.e, the pair (x_i, y_i) is replaced by an outlier.
3. Contaminating two groups in different positions, i.e, the pair (x_i, y_j) where $i \neq j$.

Each measure is calculated before contaminating the data (row 1). Once data have been contaminated with outliers from different types, we report the relative variation

Table 12 Contamination with shape–magnitude outliers

Contaminated groups	Type of outliers	No. outl	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\rho}_d$	$\hat{\rho}_c$
None	None	0	0.454	0.454	0.549	0.544
$X(t)$	Shape-magnit	1	0.003	0.004	**0.09**	0.0008
$X(t)$	Shape-magnit	2	0.001	0	0.006	**0.028**
$X(t)$	Shape-magnit	3	0.008	0.008	0.15	**0.18**
$X(t)$	Shape-magnit	4	0.02	0.02	0.079	**0.11**
$X(t)$, $Y(t)$ same position	Shape-magnit	1	0.008	0.009	0.16	**0.41**
$X(t)$, $Y(t)$ same position	Shape-magnit	2	0.013	0.014	0.27	**0.43**
$X(t)$, $Y(t)$ same position	Shape-magnit	3	0.016	0.019	0.33	**0.41**
$X(t)$, $Y(t)$ same position	Shape-magnit	4	0.034	0.037	0.34	**0.41**
$X(t)$, $Y(t)$ different position	Shape-magnit	1	0.019	0.024	**0.18**	0.002
$X(t)$, $Y(t)$ different position	Shape-magnit	2	0.021	0.026	**0.18**	0.04
$X(t)$, $Y(t)$ different position	Shape-magnit	3	0.013	0.019	0.12	**0.19**
$X(t)$, $Y(t)$ different position	Shape-magnit	4	0.119	0.127	**0.22**	0.11

Bold values indicate the greatest variation

of the association measure with respect to its value in the uncontaminated data set. We compare our results with those obtained by the dynamical correlation and canonical correlation. We can see that functional $\hat{\tau}_1$ and $\hat{\tau}_2$ coefficients are not affected by the presence of shape outliers, while the dynamical correlation and canonical correlation coefficients are sensitive to them. For magnitude outliers and shape–magnitude outliers, our coefficients present small variations unlike the other coefficients, which present variations up to 40 percent of the original value. The results are given in Tables 10, 11 and 12, where the values in bold are those that provide the largest variation in each of the cases. We can see that the functional $\hat{\tau}_1$ as well as the functional $\hat{\tau}_2$ do not present a significant variation, while $\hat{\rho}_d$ and $\hat{\rho}_c$ give the largest variations in almost all cases.

7 Conclusions

We have introduced a new numerical dependence measure between two sets of functional data. Our technique is a natural extension of the Kendall τ coefficient when the data are curves. In order to build this new coefficient, we have also introduced the concordance concept between pairs of functional data. We have presented examples of applications showing the usefulness of the new coefficients introduced for both simulated and real data.

We have compared the performance of our measure with other coefficients, such as dynamical correlations and canonical correlations. The coefficients presented here allow us to identify the global dependence between two groups of functional data regardless of the shape of their realizations. These coefficients are presented as a new alternative to find dependence between functional data. In addition, this coefficient's implementation is straightforward.

An interesting example with real data is studied. The data set corresponds to a microarray time series from a human T-cell experiment. We obtain the partial functional $\hat{\tau}_2$ for each pair of genes and construct a gene network.

We also study the sensitivity of our coefficients and conclude that these coefficients present good stability with respect to sample size and the number of points taken to discretize the functions. However the coefficient defined with the preorder of the maximun presents less stability when the data are smoothed. In terms of robustness, our coefficients can be considered quite stable in the presence of functional outliers in comparison with the measures used as a benchmark.

As one of the referees suggested, it would be very interesting to carry out a deeper inference study on these new coefficients. In this work, a first approximation with the boostrap intervals is presented, but an analysis based on hypothesis testing will be the subject of future work.

References

Borovskikh Y (1996) U-statistics in Banach space. VSP BV, Oud-Beijerland
Cardot H, Ferraty F, Sarda P (1999) Functional linear model. Stat Probab Lett 45:11–22
Cuevas A, Febrero M, Fraiman R (2004) An ANOVA test for functional data. Comput Stat Data Anal 47:111–122
Delicado P (2007) Functional k-sample problem when data are density functions. Comput Stat 22:391–410
Dubin JA, Müller HG (2005) Dynamical correlation for multivariate longitudinal data. J Am Stat Assoc 100:872–881
Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 99:96–104
Efron B (2005) Local false discovery rates. Technical report, Department of Statistics, Stanford University
Escabias M, Aguilera A, Valderrama M (2004) Principal components estimation of functional logistic regression: discussion of two different approaches. J Non Parametr Stat 16(3–4):365–384
Febrero M, Galeano P, González-Manteiga W (2008) Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels. Envirometrics 19:331–345
He G, Müller HG, Wang JL (2000) Extending correlation and regression from multivariate to functional data. In: Puri ML (ed) Asymptotics in statistics and probability. VSP, Leiden, pp 197–210
Kendall M (1938) A new measure of rank correlation. Biometrika Trust 30(1/2):81–93
Leurgans SE, Moyeed RA, Silverman BW (1993) Canonical correlation analysis when data are curves. J R Stat Soc B 55:725–740
López-Pintado S, Romo J (2007) Depth-based inference for functional data. Comput Stat Data Anal 51:4957–4968
López-Pintado S, Romo J (2009) On the concept of depth for functional data. J Am Stat Assoc 104:718–734
Opgen-Rhein R, Strimmer K (2006) Inferring gene dependency networks from genomic longitudinal data: a functional data approach. REVSTAT 4(1):53–65
Pezulli S, Silverman B (1993) Some properties of smoothed components analysis for functional data. Comput Stat 8:1–16
Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York
Rangel C, Angus J, Ghahramani Z et al (2004) Modelling T-cell activation using gene expression profiling and state-space models. Bioinformatics 20:1361–1372
Scarsini M (1984) On measure of concordance. Stochastica 8(3):201–218
Schwabik S, Guoju Y (2005) Topics in Banach space integration. World Scientific Publishing, Singapore
Taylor MD (2007) Multivariate measures of concordance. Ann Inst Stat Math 59:789–806
Taylor MD (2008) Some properties of multivariate measures of concordance. arXiv:0808.3105 [math.PR]
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, New York