UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 13-32
Statistics and Econometrics Series 28
December 2013

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

A KENDALL CORRELATION COEFFICIENT FOR FUNCTIONAL DEPENDENCE

Dalia Valencia[a], Rosa E. Lillo[b], and Juan Romo[b]

[a]Departamento de Estadística, Universidad Carlos III de Madrid, 28911, Leganés, Spain
[b] Departamento de Estadística, Universidad Carlos III de Madrid, 28903, Getafe, Spain

Abstract---
Measuring dependence is a basic question when dealing with functional observations. The usual correlation for curves is not robust. Kendall's coefficient is a natural description of dependence between finite dimensional random variables. We extend this concept to functional observations. Given a bivariate sample of functions, a robust analysis of dependence can be carried out through the functional version of a Kendall correlation coefficient introduced in this paper. We also study its statistical properties and provide several applications to both simulated and real data, including asset portfolios in finance and microarray time series in genetics.

Keywords: Dependence, functional data, concordance, Kendall's tau.

Email addresses: daliajazmin.valencia@uc3m.es (Dalia Valencia), rosaelvira.lillo@uc3m.es (Rosa E. Lillo), juan.romo@uc3m.es (Juan Romo)

A Kendall Correlation Coefficient for Functional Dependence

Dalia Valencia[*] Rosa E. Lillo[†] Juan Romo[‡]

December 12, 2013

Abstract

Measuring dependence is a basic question when dealing with functional observations. The usual correlation for curves is not robust. Kendall's coefficient is a natural description of dependence between finite dimensional random variables. We extend this concept to functional observations. Given a bivariate sample of functions, a robust analysis of dependence can be carried out through the functional version of a Kendall correlation coefficient introduced in this paper. We also study its statistical properties and provide several applications to both simulated and real data, including asset portfolios in finance and microarray time series in genetics.

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 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 functional version (Cardot et al. [2], He et al. [10]), analysis of variance (Cuevas et al. [3], Delicado [4]), principal components (Pezulli

[*] daliajazmin.valencia@uc3m.es

[†] rosaelvira.lillo@uc3m.es

[‡] juan.romo@uc3m.es

and Silverman [16]), generalized linear model (Escabias et al. [8]) and depth for functional data (López-Pintado and Romo [13], [14]) have already been extended to a functional context. Other useful methodologies can be found in Ramsay and Silverman [17]. 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. [12] 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. [10] propose an alternative way of finding the canonical correlation through the extension of multivariate analysis ideas. Opgen-Rhein and Strimmer [15] 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 Muller [5] to analyze a nonparametric method to quantify the covariation of components of multivariate longitudinal observations.

In this paper, we extend a Kendall τ correlation coefficient [11] to the functional framework. Kendall's τ allows us to measure dependence in the bivariate case through the definition of concordance, which is based on the idea of order. Since there is not total order among functions, we will 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 will show that it fulfils natural properties for a dependence measure and we will also establish the consistency of the sample version. Finally, we will illustrate with simulated and real data the performance of this new dependence measure as well as its robustness, which is a principal characteristic of the Kendall τ in its bivariate version.

We will analyze two data sets. The first one corresponds to 33 companies belonging to the IBEX35 and we calculate the functional τ for all possible pairs of the companies. This coefficient informs about companies having similar behavior over time. In finance, assets with similar dependence behavior in the same portfolio increase the portfolio's risk. Therefore, our coefficient allows us to classify the assets to build portfolios with different behavior. The second data set corresponds 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 Section 2, the functional τ is defined extending the concept of concordance for bivariate random variables. Section 3 is devoted to presenting 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 Section 4. In Section 5 we analyze with our methodology the prices of the assets in companies belonging to the IBEX35. Section 6 contains a study of dependence between genes using the genes data set. In Section 7, we present a robustness empirical study. Finally, in Section 8, we outline the main conclusions of this paper. The proofs are included in the Appendix.

2 Functional Kendall correlation coefficient

Kendall [11] introduced a correlation coefficient based on the ranks of the observations. 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

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

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

$$\widehat{\tau} = \frac{S}{\begin{pmatrix} n \\ 2 \end{pmatrix}},$$

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. For this purpose, we follow the same construction as that used for the classic Kendall coefficient. Let f and g belong to 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.*

$$f(t) \preceq_m g(t) \equiv \max_{t \in I} f(t) \leq \max_{t \in I} g(t). \tag{2}$$

$$f(t) \preceq_i g(t) \equiv \int_a^b (g(t) - f(t))dt \geq 0. \tag{3}$$

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 concordant if either $f_1 \prec f_2$ and $g_1 \prec g_2$ or $f_2 \prec f_1$ and $g_2 \prec g_1$; in the other case, they are discordant.*

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

Definition 3 *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a bivariate sample of functions in the space $C(I)$ of real continuous functions on the compact interval I. Then the functional $\widehat{\tau}$ is:*

$$\widehat{\tau} = \left(\begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum_{i<j}^{n} 2I(x_i \prec x_j \quad and \quad y_i \prec y_j) \quad + \quad 2I(x_j \prec x_i \quad and \quad y_j \prec y_i) - 1. \tag{4}$$

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

$$\tau = 2[P\{X_1 \prec X_2 \ , \ Y_1 \prec Y_2\} \ + \ P\{X_2 \prec X_1 \ , \ Y_2 \prec Y_1\}] - 1. \tag{5}$$

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

Definition 4 (UB-Statistics. Borovskikh [1], page 5.) *Let B be a real separable Banach space with a norm $\|\cdot\|$ and let B^* be the dual to space 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 $\Phi : X^m \to B$ of m variables given on X^m and taking values in B. Then, a U-statistic is*

$$U_n = \left(\begin{array}{c} n \\ m \end{array} \right)^{-1} \sum_{1 \leq i_1 < \cdots < i_m \leq n} \Phi\{(X_{i1}, \ldots X_{im})\}. \tag{6}$$

It is clear that $U_n \in B$. Hence, the U-statistic (6) with a B-values 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.

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

Theorem 1 (Borovskikh [1], page 73.) *Assume that the B-value 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ to be independent copies of the bivariate stochastic process $(X(t), Y(t))$ 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 $\widehat{\tau}$ given in (4) can be expressed as a UB-statistic,

$$U_n = \left(\begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum_{1 \leq i_1 < i_2 \leq n} \Phi\{(X_{i_1}, Y_{i_1}), (X_{i_2}, Y_{i_2})\}, \tag{7}$$

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 [19] 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 [18] 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 [[20], [21]]. The following proposition gives the properties of the functional τ. Some of them come from the axioms proposed by Scarsini [18]. Other properties of Proposition 1 are a natural extension of the well known properties of the bivariate τ itself (Kendall [11]).

Proposition 1 *Let $(X(t), Y(t))$ be a bivariate stochastic process. Then,*

1. $\tau(X(t), Y(t)) = \tau(Y(t), X(t))$. *(Symmetry).*

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 of the maximum verifies 1, 2, 4, 6 and 7, and τ with the integral preorder 1, 2, 3, 6 but 4, 5 and 7 just for affine transformations. The proof of Proposition 1 is given in the Appendix.

The consistency of functional $\widehat{\tau}$ 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 $(X(t), Y(t))$. Then,*

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

for the two preorders considered in Definition 1.

Proof
It is easy to check that the function

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

belongs to the interval $[-1, 3]$. Then, the functional $\widehat{\tau}$, given in (4) and expressed as the UB-statistic (7), 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 $\widehat{\tau}$ works in simulated functional samples with different kinds of dependence, we provide some examples. From now on, $\widehat{\tau}_1$, $\widehat{\tau}_2$ denote the functional $\widehat{\tau}$ when the maximum and integral 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) follows a bivariate standard normal distribution with correlation σ_{12} representing the random part of the processes. Each pair of curves is represented by the same color. The bivariate functional sample shown in Figure 1 was generated with a high positive value of σ_{12} close to 1. In this first case, the ordering for the maximum preorder in the first group is (red > cyan > green > blue > magenta), and for the second group 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 $\widehat{\tau}_1 = 0.6$. For the ordering to the integral preorder, in the first group are (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. Therefore $\widehat{\tau}_2 = 0.4$, whose value is smaller than for $\widehat{\tau}_1$.

On the other hand, Figure 2 shows five pairs generated from processes $X(t) = (t+Z_1)^2$ and $Y(t) = (t+Z_2)^3$ with σ_{12} close to -1. The curves are almost completely ordered in the opposite way between groups, except for the blue and black curves, which yields a strong negative dependence. In this case, our functionals $\widehat{\tau}_1$ and $\widehat{\tau}_2$ take the value of -0.8.

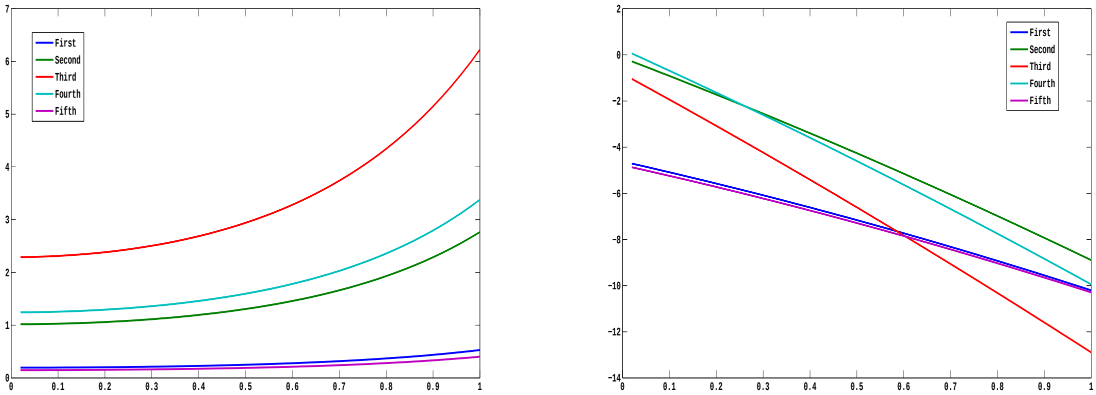

Figure 1: $\widehat{\tau}_1 = 0.6 \qquad \widehat{\tau}_2 = 0.4$

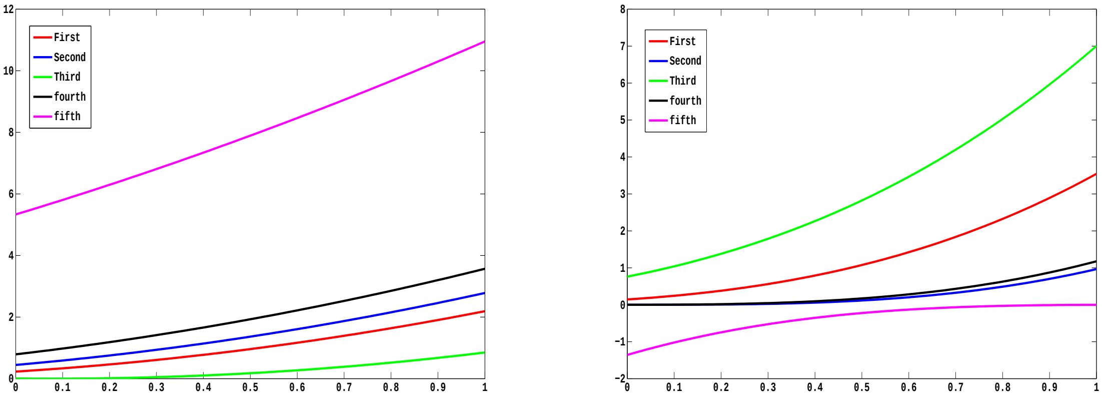

Figure 2: $\widehat{\tau}_1 = -0.8 \quad \widehat{\tau}_2 = -0.8$

4 Empirical results and comparisons

In this section, we illustrate the performance of the functional τ introduced in this work, as well as its behavior with respect to other dependency measures already introduced in the literature. We briefly describe two of them (dynamical correlation and canonical correlation). 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 Muller [5] 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 [15] proposed an estimator for the dynamical correlation considering functional data instead of longitudinal data. We will use in the paper the estimator of the dynamical correlation proposed in [15], which is a slightly revised version of the dynamical correlation introduced in Dubin and Müller [5]:

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

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.

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

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, $\widehat{\rho}_d$ is an estimator of the population dynamical correlation

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

that can been seen as an average of individual correlations.

Another well-known technique to measure functional dependency is the canonical correlation, which was defined in Leurgans et al. [12]. 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. [12] was called the canonical correlation between the two variables or groups. In a formal way, consider n observed pairs of data curves $(x_i(t), y_i(t))$ with t in a same finite interval I and all integrals are taken over I. Given canonical variates ξ and η, the canonical correlation was defined by Leurgans et al. [12] as the sample squared correlation of $\int \xi x_i$ and $\int \eta y_i$, i.e.,

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

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.

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.

As one can see, we calculate the dependence coefficient when the curves are discretized in a finite number of points. Therefore, it is necessary to define a finite dimensional version for the preorders given in Definition (1). Consider $t_1, t_2, \ldots, t_d$ to be the values of t in which the functional sample $x_1, x_2, \ldots, x_n$ is observed. Then,

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

The last expression corresponds to the composite trapezoidal rule of numerical integration, which we have used for calculating the values of the integrals.

Table 1 presents the average of the measures $\widehat{\tau}_1$ and $\widehat{\tau}_2$ as well as $\widehat{\rho}_c$ and $\widehat{\rho}_d$, 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 $\widehat{\tau}_1$ and $\widehat{\tau}_2$ in some cases take different values between them, 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 1. Indeed the value of $\widehat{\rho}_d$ 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 $\widehat{\tau}_2$ in most cases is the smallest among the other measures.

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

We can see that the canonical correlation $\widehat{\rho}_c$ 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 $\widehat{\rho}_d$ just reflects the mean of individual similarities rather than considering the set of curves as a whole. This makes the dynamical correlation to capture changes only at an individual performance level, while Kendall's coefficient detects changes at a more general level, which is one of the advantages of this coefficient.

Thus, the functional $\widehat{\tau}$ 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.

As we can see, $\widehat{\tau}$ depends on the sample size n and on the number of points to discretize the functions d. In order to assess the stability of the functional $\widehat{\tau}$, 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 is with respect to the sample size n. In this case, we move $n = 25, 50, 100, 150$ and 1000 without changing the number of points to discretize the functions, which is set as $d = 50$. This procedure is repeated 100 times and we reported their average. Table 2 shows that the changes in $\bar{\widehat{\tau}}_1$, $\bar{\widehat{\tau}}_2$ are negligible and quite stable with respect to the sample size.

Now, the same scheme is made for d, the number of points in the discretization. Fix $n = 50$, and move $d = 25, 50, 100, 150$ and 1000 points. Table 3 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 as before.

Table 2: Sensitivity to sample size

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

It is remarkable that this study of simulation was also made with smoothed data using B-spline with 13 basis functions and a smoothing parameter $\lambda = 0.01$ in the calculation of $\widehat{\tau}_{1,2}$ and the results have many similarities with those reported in this section.

Table 3: Sensitivity to the number of points in the discretization

number of points	Model 1 $\bar{\widehat{\tau}}_1$	Model 1 $\bar{\widehat{\tau}}_2$	Model 2 $\bar{\widehat{\tau}}_1$	Model 2 $\bar{\widehat{\tau}}_2$
25	0.3992	0.4168	0.2979	0.2897
50	0.4044	0.4190	0.3084	0.2774
100	0.4054	0.4135	0.2846	0.2802
150	0.4153	0.4065	0.2912	0.2801
1000	0.4089	0.4128	0.2845	0.2989

5 Ibex data

The first real data set that we use in this work corresponds to 33 companies belonging to the IBEX35. For each company we have taken a set of 108 functional observations, each one of them representing one day (108 days) in which the price of the asset has been measured every 5 minutes from 9:05 until 17:40 (104 measurements). Table 4 shows the functional $\widehat{\tau}$ coefficients, canonical correlation and dynamical correlation for some pairs of assets. Data were smoothed using cubic B-spline with 13 basis functions and a smoothing parameter $\lambda = 0.01$; recall that λ is especially used to calculate the canonical correlation. As one can see, some companies present high dependence, which can be interpreted as similar behavior of their prices in the course of time. Other companies have low dependence, whereby the prices fluctuate differently. This information given by correlation coefficients allows us to propose an alternative for organizing a portfolio of assets, which presents low risk to the investor. To carry out this methodology we will focus on the correlation coefficient $\widehat{\tau}_2$ and will use the IBEX DATA.

We construct a matrix $\mathcal{C}$ of size 33×33, whose inputs are $\widehat{\tau}_2$, in such a way that each column contains the values of the coefficient $\widehat{\tau}_2$ for a company with the other companies. In order to compare the columns of the matrix, the first component in each column will be the correlation of the company itself, i.e, the first row of the matrix will take the value 1. To classify the companies into groups depending on $\widehat{\tau}_2$, we performed a cluster analysis using the nearest neighbor technique with five groups. As results we obtain five clusters or groups where the companies are that have similar behavior in terms of the coefficient functional $\widehat{\tau}_2$. Figures 3 to 7 show the 5 groups. In each one of the groups, we plot the paths determined by the most similar columns of matrix $\mathcal{C}$.

Figure 8 shows the average correlation vectors for each group. The fact that the curves are so different could indicate that each group has a different dependence structure. The above procedure provides a good alternative for organizing a portfolio. Assets of different groups have different behavior, which can be a useful tool to avoid composing a portfolio with parallel assets, since it is well known that a portfolio with

Table 4: IBEX DATA

company 1	company 2	$\widehat{\tau}_1$	$\widehat{\tau_2}$	$\widehat{\rho}_c$	$\widehat{\rho}_d$
Antena 3 T.V.	Abertis	−0.3128	−0.3058	0.4464	−0.4338
A.C.S.	Acerinox	−0.2606	−0.2511	0.3874	−0.3664
Altadis	Acciona	0.3860	0.3918	0.4926	0.4396
B.B.V.A.	Bankinter	0.4363	0.4635	0.6759	0.6662
Cintra	Endesa	−0.1870	−0.1823	0.0808	−0.0522
Enagas	F.C.C.	−0.2464	−0.2464	0.4142	−0.39
Ferrovial	Gamesa	−0.0702	−0.0562	0.3158	−0.2056
Gas Natural	Iberdrola	0.3478	0.3511	0.4261	0.4238
Iberia	Indra A	−0.0187	0.0177	0.0668	−0.0382
Inditex	Mapfre	−0.1512	−0.1291	0.3071	−0.2927
Metrovacesa	Popular	−0.3053	−0.3406	0.4619	−0.4494
NH Hoteles	R.E.E.	−0.1193	−0.1125	0.3313	−0.3179
Repsol Y.P.F.	Sabadell	0.4846	0.4872	0.7633	0.7614
Santander.	Sogecable	0.1199	0.1131	0.1845	0.1511
Sacyr Valle	Telefónica	−0.2767	−0.2687	0.3669	−0.3553
A.G.S. Barcelona	Telecinco	−0.1431	−0.1142	0.2172	−0.2037
Unión Fenosa	Antena 3 T.V.	−0.4489	−0.4502	0.7756	−0.7697
Antena 3 T.V.	Altadis	−0.6249	−0.6690	0.7807	−0.7745
Antena 3 T.V.	F.C.C.	0.5670	0.5827	0.7718	0.7641
Antena 3 T.V.	Popular	0.6663	0.6677	0.8307	0.8354
Antena 3 T.V.	Telefónica	−0.6967	-0.7011	0.8655	−0.8628
Antena 3 T.V.	Telecinco	0.5892	0.5916	0.8032	0.7983
Abertis	Acciona	0.6296	0.6126	0.8264	0.8179
Abertis	Enagas	0.5686	0.5586	0.7699	0.7618
Abertis	Inditex	0.5953	0.5994	0.8232	0.8107
Abertis	R.E.E.	0.6147	0.6052	0.8125	0.800
Abertis	A.G.S. Barcelona	0.6969	0.7068	0.9041	0.8934
A.C.S.	Sacyr Valle	0.7132	0.7268	0.8969	0.8870
Acciona	Endesa	−0.6592	−0.6694	0.8243	−0.8130
Acciona	Iberdrola	0.7550	0.7615	0.8953	0.8908
Acciona	Santander	0.7587	0.7720	0.9273	0.9154
Acciona	Unión Fenosa	0.7587	0.7581	0.8861	0.8766
Bankinter	Sabadell	0.7941	0.8033	0.9511	0.9482
F.C.C.	Popular	0.6262	0.6310	0.8439	0.8375
Iberdrola	Unión Fenosa	0.8229	0.8195	0.9681	0.9655
Mapfre	NH Hoteles	0.6945	0.7125	0.9065	0.9008
NH Hoteles	Repsol Y.P.F.	0.7221	0.7377	0.9021	0.8982

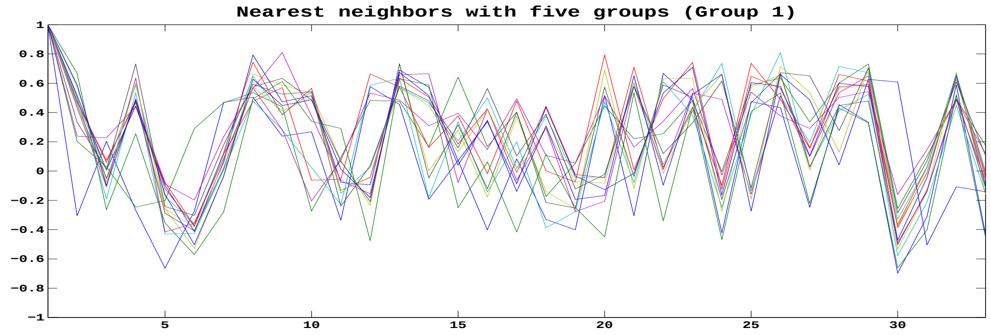

Figure 3: First group of companies

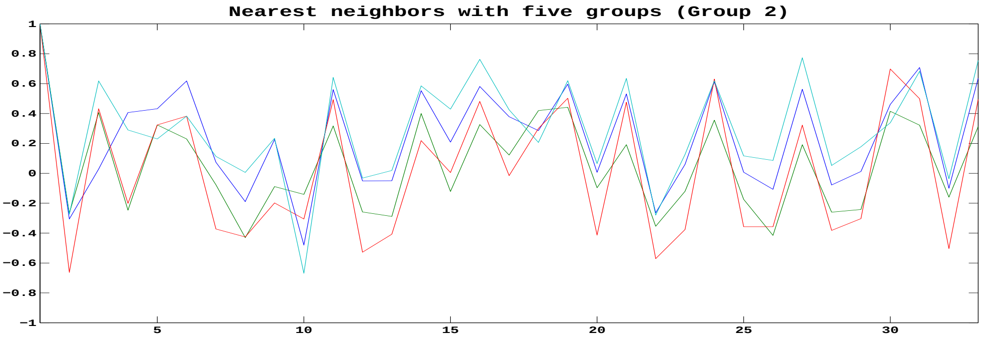

Figure 4: Second group of companies

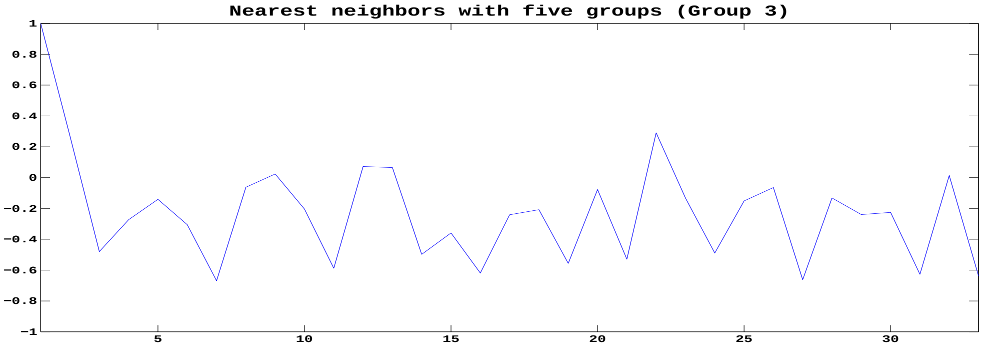

Figure 5: Third group of companies

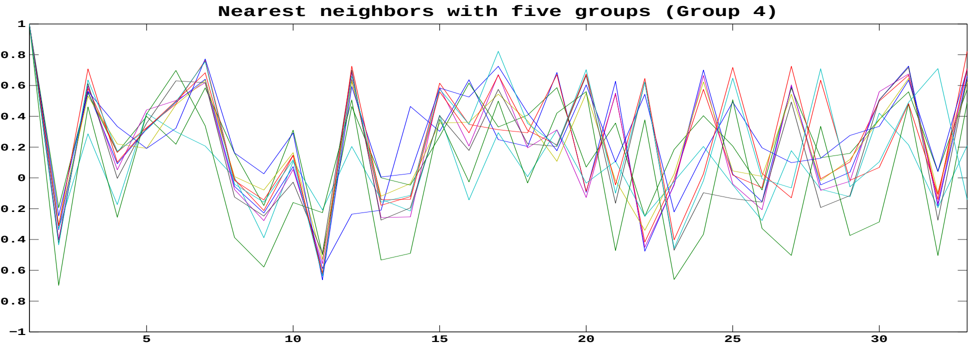

Figure 6: Fourth group of companies

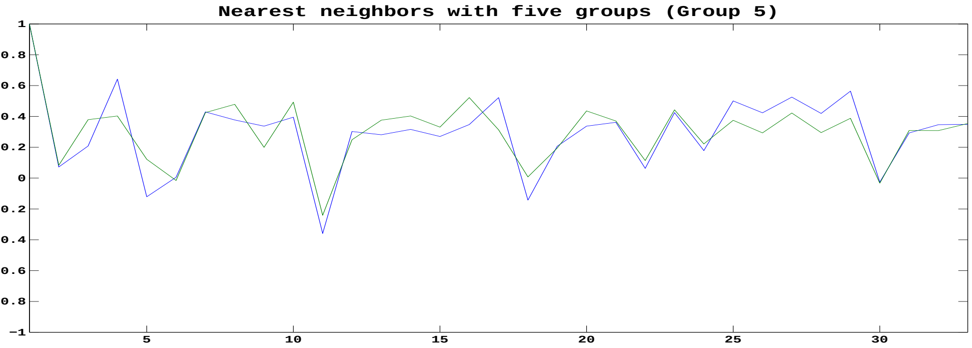

Figure 7: Fifth group of companies

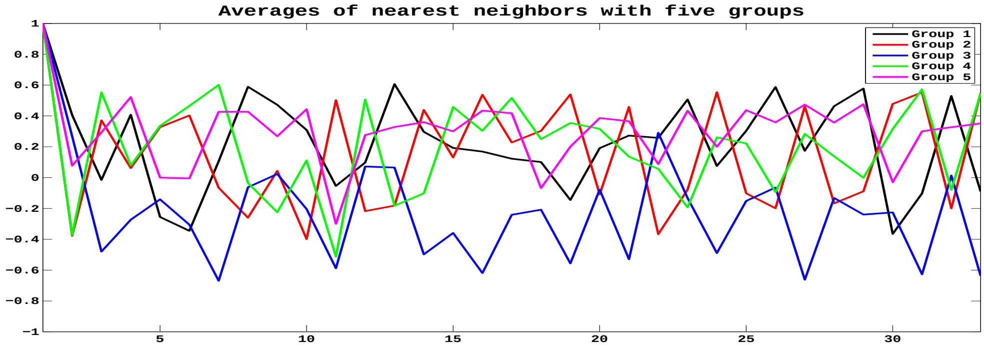

Figure 8: Average of each group

parallel assets has a very high risk.

The functional coefficient has the advantage of taking into account the temporal part of the data, i.e., the evolution of the asset over time that in this case is every five minutes. Therefore, this option works with more information for the asset. This is more meaningful and realistic than considering just the dependence between the data at the end of the day, as it is made when the dependence is measured by the usual covariance matrix.

6 Gene data

Existing relations among genes contain broad information on the structure and functioning of living beings. Therefore, 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 [22]) but their drawback is that these methods are based on the assumption of identically and independently distributed variables. Opgen-Rhein and Strimmer [15] studied the graphical Gaussian models from the perspective of functional data, where these two assumptions are not necessary.

Opgen-Rhein and Strimmer [15] 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 use the dynamical correlation introduced in Section 4. However, they do not use the dynamical correlation itself because it represents only marginal dependencies, besides including indirect interactions between two variables, since it contains information on the relations 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 $P = (\rho_{kl})$ be the correlation coefficients, and let Ω be the inverse relationships

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

then the partial correlations are given by

$$\widetilde{\rho}_{kl} = \frac{-w_{kl}}{\sqrt{w_{kk}w_{ll}}} \Rightarrow \widetilde{P} = (\widetilde{\rho}_{kl}).$$

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 [6],[7]. 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 $\widehat{\tau}_2$ and applying the “local fdr” to investigate valid relations. In order to illustrate our procedure, we use a microarray time series data set. These data were used in Opgen-Rhein and Strimmer [15]. 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, we obtain 58 genes measured across 10 time points with 44 replications. Table 5 shows the correlation coefficients including the canonical correlation $\widehat{\rho}_c$ and dynamical correlation $\widehat{\rho}_d$ for some pairs of genes. Data were smoothed with lineal B-spline, taking four basis functions and a smoothing parameter $\lambda = 0.00001$. Note how the correlations vary depending on the coefficient used, which was considered when we analyze simulated data in Section 4.

In order to compare our results with those obtained by Opgen-Rhein and Strimmer, we calculate the partial correlation matrix from the correlations matrix found with the functional $\widehat{\tau}_2$ 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 9 and 10 show the network proposed by Opgen-Rhein and Strimmer [15] 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 $\widehat{\tau}_2$ contains 22 nodes and 12 edges. In both figures, the edges in red represent negative correlation and the nodes in red represent the common nodes in both networks (CASP8, SOD1, MAPK9, CDC2, CCNA).

The advantage of using functional $\widehat{\tau}_2$ instead of the dynamical correlation studied in Opgen-Rhein and Strimmer [15] 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 6 and 7 show the results of partial correlation with dynamical correlation and partial correlation with functional $\widehat{\tau}_2$ respectively, which were found through the

Table 5: Gene data

GEN 1	GEN 2	$\widehat{\tau}_1$	$\widehat{\tau}_2$	$\widehat{\rho}_c$	$\widehat{\rho}_d$
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

GEN 1	GEN 2	$\widehat{\tau}_1$	$\widehat{\tau}_2$	$\widehat{\rho}_c$	$\widehat{\rho}_d$
JUND	RBL2	-0.5032	-0.5370	0.9556	-0.8009
LCK	CCNC	0.3499	0.6660	0.9499	0.8214
PPSGKA1	FYB	-0.0159	-0.8161	0.9582	-0.6983
CASP8	PIG3	0.6755	0.6321	0.9420	0.7787
CASP8	CSF2RA	0.50	0.6660	0.9868	0.8401
CASP8	IFNAR1	0.2886	0.3848	0.9602	0.7518
PDE4B	JUNB	0.5081	0.5370	0.8908	0.7173
IL16	EGR1	0.3319	0.0751	0.6167	0.6823
IL16	SOD1	-0.1290	-0.0106	0.7217	0.0573
APC	FYB	0.1332	0.6829	0.9736	0.2170
TCF12	CSF2RA	-0.3552	-0.6469	0.9837	-0.7988
PIG3	NFKBIA	0.5328	0.5476	0.8739	0.4362
CASP4	RBL2	-0.4440	-0.4355	0.9438	-0.7186
CSF2RA	NFKBIA	0.6047	0.6448	0.9417	0.5810

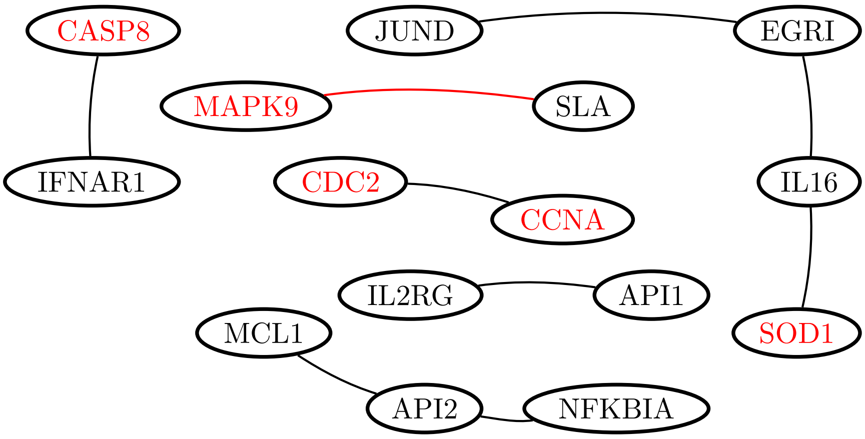

Figure 9: Gene dependence network using dynamical correlation

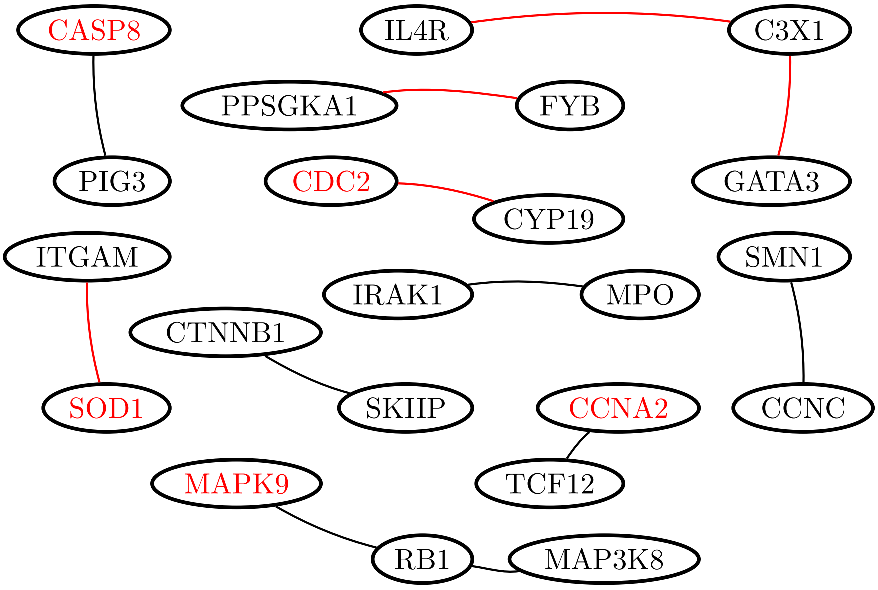

Figure 10: Gene dependence network using functional $\widehat{\tau}_2$.

"local fdr" algorithm. Also, we can see the *p-value* for each of the correlations as well as the nodes included in the networks.

Table 6: Partial correlation with dynamical correlation

Correlation	node1	node2	pval	prob
0.5196239	JUND	EGRI	$4.549748e - 09$	0.9821273
0.3971803	CDC2	CCNA2	$1.490676e - 05$	0.9821273
0.3888355	API2	$NFKBIA$	$2.325541e - 05$	0.9821273
0.3817253	CASP8	IFNAR1	$3.365286e - 05$	0.9778470
0.3749201	IL16	EGRI	$4.755512e - 05$	0.9317983
-0.3543562	MAPK9	SLA	$.291719e - 04$	0.9317983
0.3503031	IL16	SOD1	$1.560555e - 04$	0.9317983
0.3477015	IL2RG	API1	$1.759564e - 04$	0.9079010
0.3414533	MCL1	API2	$2.337537e - 04$	0.8790107

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

7 Robustness

As commented in the Introduction, we analyze the robustness of our coefficients $\widehat{\tau}_1$ and $\widehat{\tau}_2$ and compare them with the results obtained with the dynamical and canonical correlation ($\widehat{\rho}_d$ and $\widehat{\rho}_c$, respectively). We contaminate the dataset with outliers, defining a functional outlier as in Febrero et al. [9]: a "curve [that] has been gen-

Table 7: Partial correlation with functional $\widehat{\tau}_2$

Correlation	node1	node2	pval	prob
-0.3235028	PPS6KA1	FYB	$2.286947e-05$	0.9599103
0.3029697	IRAK1	MPO	$7.744064e-05$	0.9599103
0.3019622	SMN1	$CCNC$	$8.202942e-05$	0.9599103
0.2990471	RB1	MAP3K8	$9.678107e-05$	0.9400666
0.2932716	RB1	MAPK9	$1.336132e-04$	0.9287469
-0.2842216	ITGAM	SOD1	$2.184800e-04$	0.9287469
-0.2839907	CDC2	CYP19	$2.211905e-04$	0.8543381
-0.2687344	IL4R	C3X1	$4.880864e-04$	0.8543381
-0.2680201	GATA3	C3X1	$5.059491e-04$	0.8543381
0.2628164	CASP8	PIG3	$6.554510e-04$	0.8543381
0.2627168	CTNNB1	SKIIP	$6.586726e-04$	0.8543381
0.2600964	TCF12	CCNA2	$7.488866e-04$	0.8543381

erated 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.

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.)

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

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 $X(1 - t) + k$.

We use different ways to contaminate the data:

1. Contaminating a group.
2. Contaminating two groups in the same position.
3. Contaminating two groups in different positions.

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 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 $\widehat{\tau}_1$ and $\widehat{\tau}_2$ coefficients are invariant to 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 8, 9 and 10, where the values in red are those that present the largest variation in each of the cases. We can see that the functional $\widehat{\tau}_1$ as well as the functional $\widehat{\tau}_2$ do not present a significant variation, while $\widehat{\rho}_d$ and $\widehat{\rho}_c$ present the largest variations in almost all cases.

Table 8: Contamination with Shape Outliers

Contaminated Groups	Type of Outliers	№ outl	$\widehat{\tau}_1$	$\widehat{\tau}_2$	$\widehat{\rho}_d$	$\widehat{\rho}_c$
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**

Table 9: Contamination with Magnitude Outliers

Contaminated Groups	Type of Outliers	№ outl	$\widehat{\tau}_1$	$\widehat{\tau}_2$	$\widehat{\rho}_d$	$\widehat{\rho}_c$
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

Table 10: Contamination with Shape-magnitude Outliers

Contaminated Groups	Type of Outliers	№ outl	$\widehat{\tau}_1$	$\widehat{\tau}_2$	$\widehat{\rho}_d$	$\widehat{\rho}_c$
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

8 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 also have 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 dependency between two groups of functional data regardless of the shape of their realizations. Also, this coefficient's implementation is straightforward.

Two interesting examples with real data are studied. The first one corresponding to 33 companies belonging to the IBEX35 coefficient informs about companies having similar behavior over time. In finance, assets with similar dependence behavior in the same portfolio increase its risk. Therefore, our coefficient allows us to classify the assets to build portfolios with different behavior. The second data set corresponds to a microarray time series from a human T-cell experiment. We obtain the partial functional $\widehat{\tau}_2$ 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 to the number of points taken to discretize the functions. 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.

9 Appendix

Proof Proposition 1

The properties 1 and 2 are immediate from the expression (5) of functional τ.

Property 3.

Proof

Let (X_1, Y_1) (X_2, Y_2) be identically distributed copies of a bivariate stochastic process $(X(t), Y(t))$, and let $\preceq_i$ be the preorder from equation (3).
Denote $\widetilde{X}_i = \int_a^b X_i(t)dt$ and $\widetilde{Y}_i = \int_a^b Y_i(t)dt$.

$$\begin{aligned}
\tau_2(-X(t), Y(t)) &= 2[P(-X_1 \prec -X_2\ ,\ Y_1 \prec Y_2) + P(-X_2 \prec -X_1\ ,\ Y_2 \prec Y_1)] - 1. \\
&= 2[P(-\widetilde{X}_1 < -\widetilde{X}_2\ ,\ \widetilde{Y}_1 < \widetilde{Y}_2) + P(-\widetilde{X}_2 < -\widetilde{X}_1\ ,\ \widetilde{Y}_2 < \widetilde{Y}_1)] - 1 \\
&= 2[P(\widetilde{X}_2 < \widetilde{X}_1\ ,\ \widetilde{Y}_1 < \widetilde{Y}_2) + P(\widetilde{X}_1 < \widetilde{X}_2\ ,\ \widetilde{Y}_2 < \widetilde{Y}_1)] - 1 \\
&= 2[1 - \{P(\widetilde{X}_1 < \widetilde{X}_2\ ,\ \widetilde{Y}_1 < \widetilde{Y}_2) + P(\widetilde{X}_2 < \widetilde{X}_1\ ,\ \widetilde{Y}_2 < \widetilde{Y}_1)\}] - 1 \\
&= -\{2[P(\widetilde{X}_1 < \widetilde{X}_2\ ,\ \widetilde{Y}_1 < \widetilde{Y}_2) + P(\widetilde{X}_2 < \widetilde{X}_1\ ,\ \widetilde{Y}_2 < \widetilde{Y}_1)] - 1\} \\
&= -\{2[P(X_1 \prec X_2\ ,\ Y_1 \prec Y_2) + P(X_2 \prec X_1\ ,\ Y_2 \prec Y_1)] - 1\}. \\
&= -\tau_2(X(t), Y(t))
\end{aligned}$$

□

Property 4.

Proof

Let $\preceq_m$ be the preorder from equation (2) and let g be a monotone increasing function. Then,

$$\begin{aligned}
\tau_1(X(t), g(X(t))) &= 2[P\{\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)\}\ ,\ \{\max_{t\in[a,b]} g(X_1(t)) < \max_{t\in[a,b]} g(X_2(t))\}] \\
&+ 2[P\{\max_{t\in[a,b]} X_2(t) < \max_{t\in[a,b]} X_1(t)\}\ ,\ \{\max_{t\in[a,b]} g(X_2(t)) < \max_{t\in[a,b]} g(X_1(t))\}] - 1.
\end{aligned}$$

Since g is a monotone increasing function,

$$\begin{aligned}
\tau_1(X(t), g(X(t))) &= 2[P\{\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)\}\ ,\ \{\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)\}] \\
&+ 2[P\{\max_{t\in[a,b]} X_2(t) < \max_{t\in[a,b]} X_1(t)\}\ ,\ \{\max_{t\in[a,b]} X_2(t) < \max_{t\in[a,b]} X_1(t)\}] - 1 \\
&= 1
\end{aligned}$$

□

The functional preorder $\preceq_i$ from equation (3) in general, is not invariant to increasing transformations. For example: Let $f(t) = t + 1$ and $g(t) = 2t$ be continuous functions in the compact interval $[0, \frac{3}{2}]$. Then $g(t) \prec f(t)$ since

$$\int_0^{\frac{3}{2}} g(t)dt = 2.25 \quad \text{and} \quad \int_0^{\frac{3}{2}} f(t)dt = 2.625$$

Now, let $\alpha(t) = \exp(t)$ be an increasing function, then $\alpha(f(t)) = \exp(t+1)$ and $\alpha(g(t)) = \exp(2t)$

$$\int_0^{\frac{3}{2}} \exp(t+1)dt = 9.454 \quad \text{and} \quad \int_0^{\frac{3}{2}} \exp(2t)dt = 9.54$$

then,

$$g(t) \prec_i f(t) \quad \text{but} \quad \alpha(f(t)) \prec_i \alpha(g(t)).$$

Thus, the ordering is not preserved. However, for increasing affine transformations the preorder is invariant. Suppose that $\alpha(t) = ct + d$ being $c > 0$ and

$$f_i(t) \prec_i f_j(t) \Leftrightarrow \int_a^b f_i(t)dt < \int_a^b f_j(t)dt$$

$$\begin{aligned}
&\rightarrow \int_a^b cf_i(t)dt < \int_a^b cf_j(t)dt \rightarrow \int_a^b cf_i(t)dt + d(b-a) < \int_a^b cf_j(t)dt + d(b-a) \\
&\rightarrow \int_a^b (cf_i(t) + d)dt < \int_a^b (cf_j(t) + d)dt \rightarrow \int_a^b \alpha(f_i(t))dt < \int_a^b \alpha(f_j(t))dt.
\end{aligned}$$

Property 6.

Proof

Let (X_1, Y_1) and (X_2, Y_2) be identically distributed copies of a bivariate stochastic process $(X(t), Y(t))$, $X(t)$ and $Y(t)$ independent stochastic processes and

$$\tau = 2[P(X_1 \prec X_2 \ , \ Y_1 \prec Y_2) \ + \ P(X_2 \prec X_1 \ , \ Y_2 \prec Y_1)] - 1.$$

Then,

$$\begin{aligned}
\tau_1 &= 2[P(X_1 \prec X_2) \times P(Y_1 \prec Y_2)] \ + \ 2[P(X_2 \prec X_1) \times P(Y_2 \prec Y_1)] - 1 \\
&= 2[P(\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)) \times P(\max_{t\in[a,b]} Y_1(t) < \max_{t\in[a,b]} Y_2(t))] \\
&\quad + 2[P(\max_{t\in[a,b]} X_2(t) < \max_{t\in[a,b]} X_1(t)) \times P(\max_{t\in[a,b]} Y_2(t) < \max_{t\in[a,b]} Y_1(t))] - 1.
\end{aligned}$$

Also

$$\begin{aligned}
&P(\max_{t\in[a,b]} X_1(t) > \max_{t\in[a,b]} X_2(t)) = 1 - P(\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)), \\
&P(\max_{t\in[a,b]} Y_1(t) > \max_{t\in[a,b]} Y_2(t)) = 1 - P(\max_{t\in[a,b]} Y_1(t) < \max_{t\in[a,b]} Y_2(t)) \\
\text{and} \quad &P(\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)) = P(\max_{t\in[a,b]} Y_1(t) < \max_{t\in[a,b]} Y_2(t)) = \frac{1}{2} \\
&\tau_1 = 2[\frac{1}{2} \times \frac{1}{2}] + 2[(1 - \frac{1}{2}) \times (1 - \frac{1}{2})] - 1 = 0.
\end{aligned}$$

Analogously for the preorder $\preceq_i$, from equation (3).

$$\begin{aligned}\tau_2 &= 2[P(X_1 \prec X_2) \times P(Y_1 \prec Y_2)] \; + \; 2[P(X_2 \prec X_1) \times P(Y_2 \prec Y_1)] - 1\\ &= 2\left[P\left(\int_a^b X_1(t)dt < \int_a^b X_2(t)dt\right) \times P\left(\int_a^b Y_1(t)dt < \int_a^b Y_2(t)dt\right)\right]\\ &+ 2\left[P\left(\int_a^b X_2(t)dt < \int_a^b X_1(t)dt\right) \times P\left(\int_a^b Y_2(t)dt < \int_a^b Y_2(t)dt\right)\right] - 1.\end{aligned}$$

Finally,

$$\begin{aligned}&P\left(\int_a^b X_1(t)dt > \int_a^b X_2(t)dt\right) = 1 - P\left(\int_a^b X_1(t)dt < \int_a^b X_2(t)dt\right),\\ &P\left(\int_a^b Y_1(t)dt > \int_a^b Y_2(t)dt\right) = 1 - P\left(\int_a^b Y_1(t)dt < \int_a^b Y_2(t)dt\right)\\ &\text{and } \; P\left(\int_a^b X_1(t)dt < \int_a^b X_2(t)dt\right) = P\left(\int_a^b Y_1(t)dt < \int_a^b Y_2(t)dt\right) = \frac{1}{2}\\ &\tau_2 = 2[\frac{1}{2} \times \frac{1}{2}] + 2[(1 - \frac{1}{2}) \times (1 - \frac{1}{2})] - 1 = 0.\end{aligned}$$

□

Property 7

Proof

Let α and β be strictly increasing and continuous functions. For the functional preorder $\preceq_m$ from equation (2), we have:

$$\begin{aligned}\max_{t\in I} \alpha(x_i(t)) = \alpha(\max_{t\in I}(x_i(t))) \quad \text{and} \quad \max_{t\in I} \alpha(x_j(t)) = \alpha(\max_{t\in I}(x_j(t)))\\ \to \max_{t\in I} \alpha(x_i(t)) \preceq \max_{t\in I} \alpha(x_j(t)) \to \alpha(x_i(t)) \preceq \alpha(x_j(t))\end{aligned}$$

The same idea can be used for β and $Y(t)$. According to Definition 2 the number of concordant pairs is the same, therefore

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

□

References

[1] Borovskikh, Y. (1996). *U-statistics in Banach space*. VSP BV, Netherlands.

[2] Cardot, H., Ferraty, F. Sarda, P. (1999). Functional Linear Model. *Statistics and Probability Letters* 45, pp 11-22.

[3] Cuevas, A., Febrero, M. and Fraiman, R. (2004). An ANOVA Test for Functional Data. *Computational Statistics and Data Analysis* 47, pp 111-122.

[4] Delicado, P. (2007). Functional k-Sample Problem when Data are Density Functions. *Computational Statistics* 22, pp 391-410.

[5] Dubin, J. A. and Müller, H. G. (2005). Dynamical Correlation for Multivariate Longitudinal Data. *Journal of the American Statistical Association* 100, pp 872-881.

[6] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* 99, pp 96-104.

[7] Efron, B. (2005). Local false discovery rates. Technical Report, Dept. of Statistics, Stanford University.

[8] Escabias, M. Aguilera, A. and Valderrama, M. (2004). Principal Components Estimation of Functional Logistic Regression: Discussion of Two Different Approaches. *Journal of non Parametric Statistics* 16 (3-4), pp 365-384.

[9] Febrero, M. Galeano, P. and González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels. *Envirometrics* 19, pp 331-345.

[10] He, G. Müller, H. G. and Wang, J. L. (2000). Extending Correlation and Regression from Multivariate to Functional Data. *Asymptotics in Statistics and Probability.* pp 1-14.

[11] Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika Trust* 30 N$^{\underline{o}}$1/2 pp. 81-93.

[12] Leurgans, S.E., Moyeed, R.A., and Silverman, B.W. (1993). Canonical correlation analysis when data are curves. *Journal of the Royal Statistical Society B* 55, pp 725-740.

[13] López-Pintado, S. and Romo, J. (2007). Depth-based inference for Functional data. *Computational Statistics and Data Analysis* 51, pp 4957-4968.

[14] López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* 104, pp 718-734.

[15] Opgen-Rhein, R. Strimmer, K. (2006). Inferring Gene Dependency Networks from Genomic Longitudinal Data: a Functional data Approach. *REVSTAT* 4 (1), pp 53-65.

[16] Pezulli, S. and Silverman, B. (1993). Some Properties of Smoothed Components Analysis for Functional Data. *Computational Statistics* 8, pp 1-16.

[17] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis 2nd Edition*, New York, Springer Verlag.

[18] Scarsini, M. (1984). On measure of concordance. *Stochastica*8 (3), pp 201-218.

[19] Schwabik, S. Guoju, Y. (2005). *Topics in Banch Space Integration*. World Scientific Publishing, Singapore.

[20] Taylor, M. D. (2007). Multivariate measures of concordance. *Annals of the Institute of Statistical Mathematics* 59, pp 789-806.

[21] Taylor, M. D. (2008). Some properties of multivariate measures of concordance. arXiv:0808.3105 [math.PR].

[22] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York, Wiley.