UNIVERSIDAD CARLOS III DE MADRID

Ph.D. Dissertation

Dependence for Functional Data

Dalia Jazmín Valencia García

Advisors:

Rosa E. Lillo
Juan Romo

Department of Statistics

Leganés, April 2014

Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Dependence for Functional Data

Autor:

Dalia Jazmín Valencia García

Directores:

Rosa E. Lillo y Juan Romo

DEPARTAMENTO DE ESTADÍSTICA

Leganés, abril de 2014

Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

DEPENDENCE FOR FUNCTIONAL DATA

Autor: *Dalia Jazmín Valencia García*

Directores: **Rosa E. Lillo y Juan Romo**

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Secretario:

Calificación:

Leganés, de 2014

This dissertation was written in the Department of Statistics at Universidad Carlos III de Madrid under the advise of the Professors Rosa E. Lillo Rodríguez and Juan Romo. The author was supported by a scholarship for master studies (UC3M) and was subsequently hired as a teaching and research staff (PIF). Besides she had the partial support of the following research projects: UNIVERSIDAD CARLOS III DE MADRID 2010/00190/001, DIRECCIÓN GENERAL DE UNIVERSIDADES DE LA COMUNIDAD DE MADRID 2008/00059/002, DIRECCIÓN GENERAL DE UNIVERSIDADES DE LA COMUNIDAD DE MADRID 2008/00059/001 COMUNIDAD DE MADRID-UC3M 2011/00068/001, MINISTERIO DE CIENCIA E INNOVACIÓN 2012/00084/001.

Writing and Editing of Thesis

with LaTeX 2_ε, *WinEdt*

and Operating System

Windows 7

♡ ♡ ♡ ♡

A mi esposo e hijos

Mi tesoro

♡ ♡ ♡ ♡

Porque tuyo es el reino, y el poder, y la gloria, por todos los siglos.

Mateo 6:13

Nada te turbe, nada te espante todo se pasa, Dios no se muda, la paciencia todo lo alcanza, quien a Dios tiene nada le falta sólo Dios basta.

Santa Teresa de Jesús

Un poco de ciencia aleja de Dios, pero mucha ciencia devuelve a Él.

Louis Pasteur

Si tu intención es describir la verdad, hazlo con sencillez y la elegancia déjasela al sastre.

Albert Einstein

Agradecimientos

En este momento cuando termina este largo camino que emprendí hace seis años, se hace necesario mirar atrás y hacer un balance de lo bueno y no tan bueno de esta experiencia. Como en todo en la vida, debemos quedarnos solo con lo positivo, es por eso que en estas cortas líneas quiero expresar mis agradecimientos a todas aquellas personas que hicieron parte de esta "aventura" tanto en lo académico como en lo personal.

En primer lugar mis más sinceros agradecimientos a mis directores de tesis Rosa E. Lillo y Juan Romo, que a pesar de las dificultades y obstáculos que encontré en el camino, permanecieron allí, dirigiendo en todo momento mis pasos y llenos de paciencia y con su experiencia me fueron ayudando a transitar por la gran autopista del conocimiento; sé que me falta todavía mucho por aprender pero los cimientos ya están construidos y dos grandes arquitectos los han diseñado.

Quiero agradecer al Departamento de Estadística por la financiación y por proporcionarme los medios necesarios para llevar a cabo la tesis. A Gema, Susana y Paco por su amabilidad y disponibilidad cada vez que necesité de ellos. A todos y cada uno de los profesores y compañeros del Departamento, en especial a Elisa M. de quién siempre recibí un caluroso saludo y un sincero interés por mis asuntos, a Johanna, Esdras, Gabriel, Alberto, Leonardo, Sofia, Miguel Ángel, Nuria y Ángel, cada uno en su momento hizo que mi estancia aquí fuese más agradable.

A nivel personal mi gratitud absoluta a mi esposo y compañero de lucha Henry, su apoyo incondicional en todos los aspectos: personal, académico y familiar fueron fundamentales para llevar a término este proyecto; he aprendido mucho de ti, gracias por estar ahí siempre que te he necesitado. A pesar de las dificultades que hemos encontrado en este arduo navegar siempre has sabido ser el capitán de este barco y nos has dirigido a las mejores aguas llegando siempre a buen puerto. Dios te bendiga y espero pasar contigo el resto de mi vida. Al motor de mi existencia, mis hijos: Nicolás, Fabiana y Helena, gracias por existir y permitir que disfrute de sus sonrisas y compañía; nunca olviden que son mi prioridad y siempre lo serán. A los cuatro, los amo hasta el cielo.

A mi madre mil gracias, a pesar de la distancia, sus oraciones y palabras de aliento me han ayudado a salir a flote cuando creí que me hundía. A Shirley Rodas gracias por cuidar de mis hijos con tanto amor y esmero. A Diana torres (Nana), Diana Restrepo y Diana Vásquez, un Dios les pague. A Edwin y su esposa Yanitza, Marcel y su esposa Érica gracias por haber compartido conmigo y mi familia unos muy buenos momentos. **A TODOS MIL GRACIAS**.

Contents

List of Figures

Index of tables

Dependence for functional data

Ph.D. Dissertation

Abstract

Dalia Jazmín Valencia García

Department of Statistics

Universidad Carlos III de Madrid

Measuring dependence is a basic question when dealing with functional observations. It is of great interest to know the effect that one or more functional variables can have on other ones, and even predict values of one variable from another. Although, in the functional context, this theory has not been as extensively studied, some techniques to measure dependence in functional data have already been implemented, providing a single value which represents the degree of relation between the sets of curves. However, these measures are usually not robust, which makes them less stable in the presence of outliers. Therefore, it is interesting to develop robust techniques that ensure high stability of the statistics. This thesis is motivated by the above issues and aims to provide measures of dependence for sets of curves that are more robust than those used so far. Hence, we extend non-parametric bivariate coefficients, such as Kendall's τ and Spearman's coefficient, to functions, i.e. to situations where the observed data are curves generated by a stochastic process. These coefficients are based on the natural data ordering, but when we work in the context of functional data, there is no such thing as a natural order between functions, meaning that we need to provide for an ordering of curves. Thus, our first task is to consider suitable ways to sort the observations. For this, we use different functional preorders, which allow us to define the coefficients in a way similar to the bivariate case. The aforementioned coefficients provide an univariate measure of the dependence between two sets of curves, which leads us to propose in the final chapter a new functional correlation coefficient that yields a representative curve of dependence between two sets of functional data. This coefficient is based on the cross-correlation function studied in the literature of functional data, which is the classic Pearson coefficient between the values of the curves in different time instants. We adapt the concept of MAD and comedian to measure dependence between two sets of functions and, through them, introduce a robust alternative to the cross-correlation function, which we will call correlation median for functions.

The thesis is organized as follows. In Chapter 1 we start defining what is understood as complex data in this work and show several examples. These data will be treated as functional data. Then, a review of the different approaches to analyze functional data is provided. We also offer a brief review of some of the most common measures of dependence between random variables, focusing on those where we make our contribution. This chapter also analyzes some techniques that have been extended to the functional context for calculating the dependence between two sets of curves in order to compare our results. Finally, we study the principal

ordering measures for functional data which are necessary to sort the curves, and thus define the coefficients in the functional setting.

In Chapter 2 we define the Kendall τ coefficient for functional observations based on the concept of functional concordance, also new in this dissertation. We study its statistical properties and provide some applications to real data, including asset portfolios in finance and microarray time series in genetics.

In Chapter 3 we present a notion of Spearman's coefficient for functional data that extends the classic bivariate concept to situations where the observed data are points belonging to curves generated by a stochastic process. Since Spearman's coefficient for bivariate samples is based on the natural data ordering in dimension one, we need to consider a data order in the functional context. The development uses a pre-order inspired in the depth definition, but considering a down-up ordering instead of a center-outward ordering of the sample, allowing us to introduce the notion of grade for functions to properly define the Spearman coefficient. We show some of the main characteristics of Spearman's coefficient for functions and propose an independence test with a bootstrap methodology. We illustrate the performance of the new coefficient with both simulated and real data.

The results of Chapter 4 concern a new functional correlation coefficient that is more robust than the cross-correlation function. The pair $(median, MAD)$ is known to be a robust alternative to the pair (mean, standard deviation). Using the idea underlying the calculation of the MAD, Falk [19] defined a robust estimator for the covariance called comedian. In this chapter we adapt these concepts, the MAD and the comedian, to functional data. These measures allow us to define a robust alternative to the cross-correlation function studied in the literature of functional data, which we will call the correlation median for functions. Since the most natural extension of median in the functional context has been performed through depth measurements, the functional MAD and comedian will also be constructed via depth. These concepts are illustrated with simulated and real data.

Finally, in Chapter 5, we present some general conclusions and summarize the main contributions of the dissertation.

Dependencia para Datos Funcionales

Tesis Doctoral

Resumen

Dalia Jazmín Valencia García

Departamento de Estadística

Universidad Carlos III de Madrid

Medir la dependencia es un aspecto muy importante cuando tratamos con observaciones funcionales. Es de gran interés conocer el efecto que una o más variables funcionales pueden tener sobre otras, e incluso predecir valores de una por medio de los valores de otra. Aunque en el contexto funcional esta teoría no ha sido tan ampliamente estudiada, existen algunas técnicas para medir la dependencia en datos funcionales que ya han sido implementadas, proporcionando un solo valor, que representa el grado de relación entre los conjuntos de curvas. Sin embargo, estas medidas introducidas en la literatura no son generalmente robustas ante la presencia de observaciones atípicas. Por lo tanto, es de interés desarrollar técnicas robustas que nos garanticen una alta estabilidad de los estadísticos. Esta tesis está motivada por las cuestiones antes mencionadas y su principal objetivo es proporcionar medidas de dependencia para conjuntos de curvas que sean más robustas que las usadas hasta ahora. Basicamente el trabajo se enfoca en extender algunos coeficientes bivariantes no paramétricos, tales como el coeficiente τ de Kendall y el coeficiente de Spearman al campo funcional, es decir, a situaciones donde los datos observados son puntos pertenecientes a curvas generadas por algún proceso estocástico subyacente. Estos coeficientes se basan en el orden natural de los datos, pero cuando se trabaja en el contexto funcional hay una dificultad mayor y es que allí no hay un orden natural entre funciones. Esto motiva la búsqueda de metodologías para comparar funciones, algunas de ellas ya han sido estudiadas por diversos autores, pero en algunos casos concretos se propone en la tesis nuevas ordenaciones que son más adecuadas para extender los coeficientes de dependencia al escenario de funciones. Por lo tanto, el primer objetivo es investigar las formas adecuadas para ordenar las observaciones. Para ello, se utilizan diferentes preórdenes funcionales que permitirán definir los nuevos coeficientes de una forma similar al caso bivariante. Los coeficientes que se han mencionado definen una medida de respuesta escalar de dependencia entre dos conjuntos de curvas. Además, en la tesis también se propone en el último capítulo un nuevo coeficiente de correlación que proporciona una curva representativa de la dependencia entre dos conjuntos de datos funcionales. Este coeficiente está basado en la función de correlación cruzada estudiada en la literatura de datos funcionales cuya definición no es más que el clásico coeficiente de correlación de Pearson entre los valores de las curvas en diferentes instantes de tiempo. En este trabajo tambien se extienden los conceptos de desviación absoluta de la mediana MAD y la comedian, para medir dependencia entre dos conjuntos de funciones y a través de estos dos conceptos en sus versiones funcionales

se introduce una alternativa robusta de la función de correlación cruzada, que se se llamará correlación mediana para funciones.

La tesis está desarrollada con la siguiente estructura: En el Capítulo 1 se introduce lo que se ententerá, en este trabajo, como un dato complejo y se ilustran algunos ejemplos de ellos en diferentes contextos. Estos datos serán tratados como datos funcionales. Por lo tanto, en este capítulo se hace una breve revisión de algunos enfoques para analizar este tipo de datos. Se describen, además, algunas de las medidas más comunes de dependencia entre variables aleatorias, haciendo énfasis en aquellas en las que esta tesis contribuye a la literatura por su extensión a variables funcionales. En este capítulo también se hace una revisión de algunas técnicas de medición de la dependencia que ya han sido extendidas al contexto funcional, con el objetivo de comparar los resultados obtenidos. Finalmente, se analizan las principales metodologías de ordenación para datos funcionales que son necesarias para ordenar las curvas y definir los coeficientes en el ambiente funcional.

En el Capítulo 2 se introduce una versión novedosa del coeficiente τ de Kendall para observaciones funcionales. Este coeficiente se construye a través de un concepto llamado cocncordancia, cuya versión para funciones se desarrolla en el capítulo. Se estudian sus propiedades estadísticas y se proporcionan algunas aplicaciones a datos reales, incluyendo carteras de activos en finanzas y microarray de series de tiempo en genética.

En el Capítulo 3 se presenta la noción del coeficiente de Spearman para datos funcionales que extiende el concepto clásico bivariante a situaciones donde los datos observados son puntos pertenecientes a curvas generadas por un proceso estocástico. Como el coeficiente de Spearman para muestras bivariantes está basado en la ordenación natural de los datos en dimensión uno, es necesario un orden para los datos en el contexto funcional. Este desarrollo utiliza un pre-orden inspirado en la definición de profundidad, pero considerando una ordenación de abajo hacia arriba en lugar del orden del centro hacia a fuera de la muestra. El orden de funciones induce la noción de grados para curvas que permiten definir naturalmente el coeficiente de Spearman. Se presentan algunas de las principales características del coeficiente de Spearman para funciones y se propone un test de independencia con una metodología bootstrap y se ilustra su buen funcionamiento con datos simulados y reales.

Los resultados del Capítulo 4 se refieren a un nuevo coeficiente de correlación funcional más robusto que la función de correlación cruzada. La pareja $(mediana, MAD)$ es bien conocida como una alternativa robusta a la pareja (media, desviación estandar). Utilizando la idea subyacente al cálculo de la MAD, Falk [19] definió un estimador robusto para la covarianza llamado comedian. En este capítulo se adaptan estos conceptos, MAD y comedian, a datos funcionales. Estas medidas permiten definir una alternativa robusta a la función de correlación cruzada estudiada en la literatura de datos funcionales, que se llamará correlación mediana para funciones. Como la extensión más natural para la mediana en el contexto funcional se ha realizado a través de las medidas de profundidad, la MAD y la comedian funcional

se construirán también a través de la noción de profundidad. Estos conceptos también se ilustran con datos simulados y reales.

Finalmente, en el Capítulo 5, se presentan algunas conclusiones generales y se resumen las principales contribuciones de la tesis.

CHAPTER 1

Introduction and background

Nowadays, the statistical analysis of large size databases in high dimensions is experiencing a notable growth for application in different fields of science such as medicine, finance, meteorology, criminology, quality control, to name a few. Statistical surveys, over all, in high dimensionality data, lead to rethinking that the classic statistical methodologies commonly implemented until now for this purpose are increasingly limited and inefficient, and they simply cannot be used for this kind of data. For example, if each variable is observed at many different points through time, a multivariate analysis would not be valid, even if the data are observed at the same time points. In these cases, a standard multivariate analysis could not be computationally feasible due to the curse of dimensionality, since there are data where the dimension is often significantly higher than the number of variables observed, leading to possibly having ill-posed problems. Therefore, different alternatives have been introduced in recent years to analyze and study these large masses of data, such as interpolation or smoothing techniques that allow us to build functions to represent the data that facilitate the analysis and also its interpretation. However, many of the technological and industrial processes usually deliver observations that may already be considered directly as functions, avoiding the smoothing processes. The arrays, or high-dimensional vectors, are other data examples of how large information could be gathered. This kind of data, which by their nature require special statistical treatment, are those referred to in this context as "complex data".

Now we will illustrate some situations where data of a complex profile arise in terms of high dimensionality and large size:

- The analysis of the growth of a large number of children at different times, where the growth curve for each child is taken as an observation.

- The study of the evolution of the temperature of a city over a long period of time taken in different places. Each observatory has a temperature curve which is taken as a single datum.

- Functions that represent the price of different assets over time in the continuous stock market.

- A color image can be decomposed into the various matrices that form the image itself, and each matrix can be analyzed independently as a complex observation.

- In the field of genetics, micro-arrays are used to perform various analysis, these vectors are large-scale complex data.

A graphic representation of a set of complex data can be seen in Figure 1.1 and Figure 1.2. The former illustrates an example of curves that represent the daily prices of assets of two companies during 108 days; these prices are measured every 5 minutes, whereas the latter refers to the monthly temperature in two cities of Canada during 20 years.

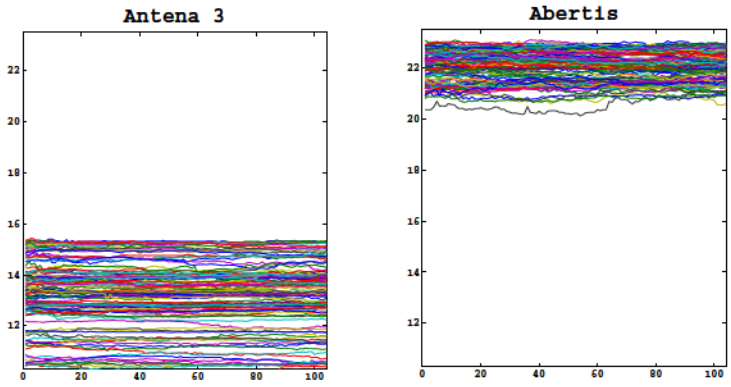

Figure 1.1: Daily prices of assets during a period of 108 days.

Observe that the data in Figure 1.1 are represented without any kind of statistical treatment; they are simply the graphs of points observed, while the data in Figure 1.2 have already had a statistical treatment because the temperatures in each year have been smoothed. Therefore, each year is represented by a function and this is where the function set can be considered as a set of functional data which are the type of data for which this dissertation offers its main contributions. Thus, we follow this introductory exposition, presenting two well-known approaches for performing statistical treatment of functional data. The first one of them is the set up of Ramsay and Silverman [45], which is based on representing the functions through a finite number of basis elements. The second one is the non-parametic vision of Ferraty and View [23] where a discretized representation of the functions is made.

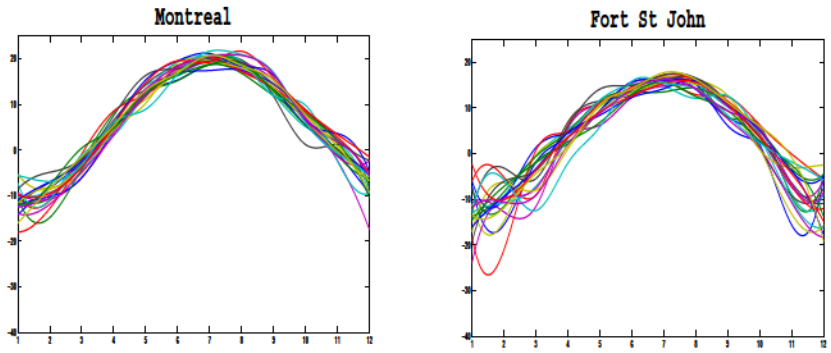

Figure 1.2: Temperatures in two cities of Canada.

We can see that by its nature the functional data have infinite dimension. This sometimes hinders its representation and especially the application of statistical methods for its analysis. In practice, so far, it has been necessary to represent the data in finite dimension, trying to lose the least amount of information possible. To carry out the transition from infinite dimension to that of a finite one, two procedures are considered:

- **The choice of basis functions**. This procedure consists of obtaining the coordinates of the projection of the function in some functional sub-space of finite dimension. Generally speaking, fixed basis functions are considered, for example, the Fourier basis, the B-splines basis, wavelets and so on. To obtain a finite number of terms it is necessary to truncate the development in a number K of basis elements. That is, $x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$, where $\{\phi_k\}_{k\in\mathbb{N}}$ are basis functions, and c_i are the coefficients in the new basis. To implement this methodology, one must be careful in the choice of the number of basis functions K, as well as the basis functions in each case, since the representation of the data in the new finite dimensional space will be influenced by selected the basis functions. In addition, the degree of smoothing of the function will also depend on the parameter K.

- **The discretization**. This procedure consists of taking a partition of the interval where the functions are defined. Let $x(t)$ be a function in $t \in I = [a, b]$. The simplest partition on time will be, $a \leq t_0 < t_1 < \cdots < t_n \leq b$, where for all i, $t_i - t_{i-1}$ have the same value. However, there are other types of partition that can also be useful, such as a random selection of the points t_i, or also considering non-regular partitions, where the length $t_i - t_{i-1}$ must be smaller in those points t_i with relevant information. The discretization of a function $x(t)$ will be the sequence given by $\{x(t_i)\}_{i=0}^{n}$.

As we have said before, these two important approaches have been analyzed by Ramsay and Silverman [45] and Ferraty and View [23], respectively. In order to be clearer in both

aspects, we will give a more detailed introduction of both methods, but omitting some formal aspects which can be found in the respective references.

The first one is the approach of Ramsay and Silverman [45], who work under a perspective in which the functional data are represented through smooth functions. Their approach takes into account different methodologies for the smoothing and interpolating of functions. We present the representations most used by them: the Fourier and the B-splines basis.

- **Fourier basis**
  This basis is periodic, therefore, it is useful for stable functions, without large changes and which show a certain periodicity. The basis expansion is provided by the Fourier series:
  $$\hat{x} = c_0\phi_0 + \Sigma_r(c_{2r-1}\phi_{2r-1}(t) + c_{2r}\phi_{2r}(t)),$$
  where
  $$\phi_0 = \frac{1}{\sqrt{T}},\ \ \phi_{2r-1}(t) = \frac{\sin(rwt)}{\sqrt{\frac{T}{2}}},\ \phi_{2r}(t) = \frac{\cos(rwt)}{\sqrt{\frac{T}{2}}},$$
  form a periodic basis with period $T = \frac{2\pi}{w}$. This basis will be orthogonal if the $\{t_j\}$ are equally spaced in $[0,T]$. An important feature of this type of basis is its easy differentiability.

- **B-splines basis**
  The B-splines basis is the most widely used approximation in the case of non periodic data. Its success lies in the fact that it combines the computational efficiency of polynomials with greater flexibility. For constructing a B-splines basis $\phi_k(t)$, it is necessary to divide the interval over which a function is going to be approximated into L subinterval separated by values ς, called breakpoints or knots. On each interval a spline is defined (polynomial of specified order m). Thus, the basis will have the following properties.

  1. Each element of the basis $\phi_k(t)$ will be a spline function, as defined by a order m and a knots sequence ς.
  2. The linear combination of these basis functions is also a spline function.
  3. Any spline functions defined by m and ς can be expressed as a linear combination of these basis functions.

  If we take the notation $B_k(t,\varsigma)$ representing the k-th basis element over the ς partition in the instant t, the spline function $S(t)$ is defined as:
  $$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t,\varsigma),$$
  where c_k are the coefficients in the basis. In Figure 1.3 we can see the representation of thirteen B-splines of order four, with eleven knots.

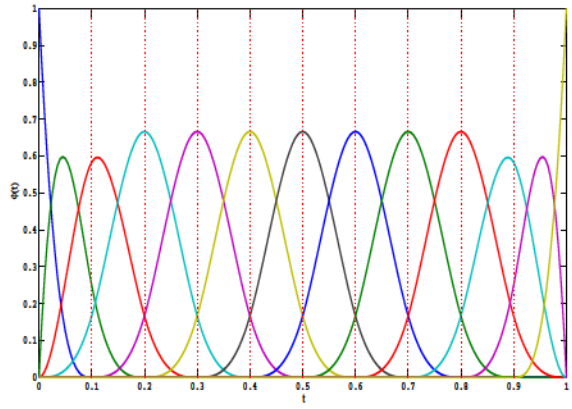

Figure 1.3: B-spline basis.

In the representation of the functional data as smoothed functions, there are a set of statistical techniques and methodologies in the literature that have been developed for its analysis. These techniques are therefore based on the basis functions chosen for the representation, so that they take into account the estimation of the parameters necessary for the smoothing.

The second approach for carrying out a representation of functions in a finite dimensional space is the point of view of Ferraty and View [23], which is based on the study of functional data in a nonparametric way, i.e., the proposed statistical theory where there is free-distribution, free-parameters, free-linearity, and free-discretization and it is focussed on nonparametric models which are very general. In this approach, the functional data are observed over a grid, which can be as fine as desired. Therefore, it is not necessary to use the basis functions for the representation and analysis of the curves. Accordingly, the approach develops a mathematical background and asymptotic properties that are independent of the number of points taken for the discretization of functions. The results of this dissertation are based on the representation of the data through the discretization approach.

The statistical analysis of functional data, in both cases of representation (basis or discretization), requires mathematical analysis tools, since it is necessary to take into account some specific features of this type of data, as for instance, the dimensionality that in theory is infinite and the space where the functions belong. Therefore, the functional data analysis is characterized by having a strong theoretical support, so that many methodologies and techniques of multivariate analysis have been extended to the functional context based on the two main approaches described previously. We present several works that have had a high impact on the development of new statistical methodologies for functional data. For example, a regression functional version can be seen in (Cardot et al. [2], He et al. [27]), analysis of variance in (Cuevas et al. [6], Delicado [10]), principal components in (Pezulli and Silverman [44]), generalized linear model in (Escabias et al. [17]) and depth for functional data in

(Fraiman and Muniz [24], López-Pintado and Romo [37], [38]). Other useful methodologies can also be found in Ramsay and Silverman [45] and Ferraty and View [23]. However, it is worth pointing out that there are still some statistical concepts that have not been fully explored for functional data, among them measures of association and dependence structures between set of curves.

Hence, this thesis will focus on dealing with the problem of measuring the dependence between sets of curves. We extend the classical bivariate concepts to situations where the observed data are curves generated by a stochastic process. The principal coefficients that we consider are Kendall's τ and Spearman's coefficient, which are based on the natural data ordering. These coefficients provide an univariate measure of the dependence between two sets of data, and our proposal here is basically the functional version of them. We also propose, in the last chapter, a new functional correlation coefficient that yields a representative curve of dependence between two sets of functional data. This coefficient is based on the cross-correlation function studied in Ramsay and Silverman [45], which is the classic Pearson coefficient between the values of the curves in different time instants.

We follow this introduction by gathering in Section 1.1 a brief historical review of the main definitions and measures of dependence that are applied to bivariate data set. We will especially focus on those that this dissertation will extend to the functional field. Later on, in Section 1.2 we analyze some techniques from the literature that have already been extended to the functional context for calculating the dependence between two sets of curves. Finally, in Section 1.3 we present some procedures for ordering curves, highlighting the kind of orders that will be used for constructing the new measures of dependence proposed in this work.

1.1 Dependence measures

In this section, we recall the concept of dependence between random variables and show the measures commonly used in order to capture this dependence. We also present a brief historical review of the measures that we aim to generalize to the functional context.

The dependence is the relationship between two or more random variables. The measures of dependence provide a value that summarizes the size of the association between two variables, and in some cases these relations or associations may be very limited or weak, while in other cases they may be strong associations. Such relations can occur in three ways: (i) when the values of one variable increase, so do the other's -positive association; (ii) when the values of one variable increase, the values of the other one decreases -negative association, and (iii) there is not consistent behavior of one variable with respect to the another -independence. Most measures of association are scaled in the same way so that they reach a maximum numerical value of 1 when the two variables have a perfect relationship with each other. They are also scaled so that they have a value of 0 when there is no relationship between two variables.

Other measures have a range from -1 to +1, which provide a means of determining whether the two variables have a positive or negative association with each other. To determine the significance of the value given by some association measure, tests of significance are provided for many of the measures of association. These tests begin by hypothesizing that there is no relationship between the two variables, and that the measure of association is equal to 0. The researcher calculates the observed value of the measure of association, and if the measure is different enough from 0, the test shows that there is a significant relationship between the two variables. Although two of the measures of association introduced in this thesis are defined for a bivariate sample of curves, they are also scaled in the interval $[-1, 1]$ and whose interpretation of the value is the same as that previously noted. A test of significance based on a bootstrap methodology for one of them is also introduced. Another contribution of the dissertation is a dependence measure between groups of functions whose response is a function instead of a single value.

There are several ways to determine the association between random variables; for this reason, it is essential to start any analysis taking into account the nature of the data, the scale of measurement of the variables and a logical reason that gives meaning to the association. The variables can be: **qualitative**, where nominal scales are used (a natural order between categories cannot be defined), and/or ordinal scales (an order or hierarchy of categories can be set); for such variables there are some measures that capture dependence such as: *Cramer's V coefficient*, λ *coefficient*, *Pearson's C coefficient*, *Kendall's* $\tau - B$ *coefficient*, *Somers's d coefficient*, and so on. The variables can be also **quantitative**, discrete or continuous, used in interval scales and ratio scales (ordinal scales can also be used); for these variables, measures such as *Pearson's coefficient*, *Spearman's coefficient*, *Kendall's* τ *coefficient*, *and Quadrant dependence*, can be used to find dependence. This section will analyze in depth measures of dependence for quantitative random variables.

When there are quantitative variables, what is usually done to interpret dependence is to determine a coefficient of correlation between variables. The decision of what coefficient to use depends on several factors, such as the type of measurement scale in which each variable is expressed, the nature of the distribution (continuous or discrete) and if the dependence sought is linear or nonlinear. The Pearson coefficient can be used whether the random variables are continuous or discrete, and whether they are measured in intervals or ratios. Although the Pearson coefficient is widely employed, it is not completely satisfactory to measure the dependence between random variables, as it provides limited information about their dependence structure overall in presence of non-linear dependence. The absence of correlation is equivalent to independence in very rare cases, such as when the random variables are Gaussian distributed. The Spearman and Kendall's τ coefficients are used when the data are sorted according to their rank. They are able to measure dependence when a nonlinear structure exists between the random variables, while the correlation coefficient only measures linear dependence between random variables. Inspired in these three coefficients, we have developed

each one of their versions as a initial alternative to explore the dependence in a bivariate sample of functional data. Accordingly, we will present a brief historical review of association measures to be discussed in this thesis: the Pearson correlation coefficient, the Kendall τ and the Spearman coefficients.

As we already pointed out, measuring the dependency between two random variables has been an important issue in statistical analysis. It is of great interest to know the effect that one or more variables can have on the other, and even predict values of one variable from another. To measure these relationships or associations among variables, various procedures have been implemented, most of them having their beginnings in the latter part of the $19th$ century. The first notions of the concept of correlation were derived from studies in biology, biometrics and eugenics. Authors such as *Adolphe Quetelet* (1796-1874) and *Augusto Bravais* (1811-1863) contributed to the development of this theory from two different fields. *Quetelet* performed some association analysis through the study of anthropometric measures, while *Bravais* also studied dependence, but through the analysis of spatial measures. Moreover, Sir *Francis Galton*, developed important statistical concepts through the study of the variability of human characteristics. He was a pioneer in explaining the meaning and usefulness of correlation and regression, not only in the context of inheritance, but also in general terms, giving rise to a wide range of applications which fall under the laws of correlation.

Many of the most brilliant ideas of Galton were collected by prestigious authors, among them, *Edgeworth*, *Pearson*, *Yule*, and *Sheppard*, who developed these ideas to construct many of the statistical concepts that are still widely applied in several disciplines. For example, *Pearson* continues the work of *Galton* and develops the well-known correlation coefficient that carries his name, which is given by

$$\rho_p = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{1.1.1}$$

where σ_x, σ_y are standard deviations of X and Y, respectively.

This definition was formally introduced in Pearson [43], where a general theory of correlation for n variables and "the best value" of the correlation coefficient was presented. This work inspired other authors, who together with *Pearson*, developed other theories also referring to correlation, such as the concepts of partial and spurious correlation and correlation ratio. Although *Pearson*, *Gosset* and *Fisher*, jointly attempted to deal with the problem of finding the distribution of the sample correlation coefficient as an estimator for the true population correlation ρ_p, the problem being solved solely by *Fisher*, with the transformation $r = tanh(Z)$, where Z is a random variable approximately normally distributed.

It is clear that from its beginning, that the correlation coefficient has been a powerful tool for multivariate statistical analysis. However, we must be cautious because there are some limitations to be applied. For instance:

- *Galton* and *Pearson* assumed that the sample comes from a bivariate normal distribution.
- The correlation coefficient measures only linear relationships.
- The correlation is not an invariant measure under strictly monotone transformations.
- Weak correlations do not necessarily imply low dependence.
- Independence always implies zero correlation, but the converse is only true in the multivariate Gaussian case.
- σ_x^2, σ_y^2 have to be finite.

Due to these drawbacks, the ordinal measures of association were introduced only a few years after *Galton* and *Pearson* had implemented the correlation coefficient as a statistical tool. Galton, for example, was the first to attempt the correlation of ranks or grades, but he discarded that approach by working in favor of what later became the standard bivariate normal correlation theory.

Most of the first papers on ordinal measures of association were developed to be applied to sample values. It is only in relatively recent years that much attention has been paid to the population meaning of these kinds of measures. The Kendall τ coefficient is one of the most well-known ordinal measures. The essential idea behind Kendall's τ coefficient was suggested by *Fechner* in 1897, although it was mainly concerned with association between two time-series. The educational psychologist, *G. Deuchler*, carried out studies on τ coefficient and considered the exact distribution of the estimator under the hypothesis of independence, obtaining virtually the same recursion formula that Kendall developed later. In 1924, *Esscher* suggested τ as measure of association and gave a clear statement of its population meaning. Finally, *Kendall* in 1938 (see Kendall [30]) began a series of papers dealing extensively with ordinal measures of correlation.

On the other hand, in France, Binet proposed measuring the association by a function of the ranks, basically the same function that was later called Spearman's foot-rule. A few years later in 1904, Spearman introduced an estimator as the sample correlation coefficient between the ranks (see Spearman [49]). The asymptotic distribution theory of the estimators for both Kendall and Spearman coefficients was studied in Hoeffding [28]. [1]

Now, we give a brief formal description of those association measures which we will try to extend to the functional field. We begin by stating that the ordinal measures association are based mainly on a fundamental concept called concordance; that is, two random variables are concordant if large values of a variable are associated with large values of the other, and the same is true for small values. They are discordant otherwise. If we consider, for example, two

[1] Brief historical overview of Kruskal [31], Estepa et al [18]

realizations of those random variables, then the concordance between them can be defined as follows:

Definition 1.1.1 *Let (x, y) and (x', y') be two observations of a continuous random vector (X, Y). Then, (x, y) and (x', y') are concordant if $(x - x')(y - y') > 0$ and discordant if $(x - x')(y - y') < 0$.*

Observe that concordance and discordance between observations can be compared with the sign of the line slope defined by the same observations. Therefore, if we consider some measure that quantifies the proportion of concordant pairs, we will have a good non-parametric indicator of the sign of dependence between the random variables where the sample comes from, even in the cases when the dependence is of a non-linear kind. The sample version of Kendall's coefficient is based mainly on that concept of concordance. It is defined as the number of concordant pairs minus those discordant pairs over the total of pairs of the sample. Although the Spearman coefficient is also an association measure, it works differently because its sample version is defined basically as the Pearson coefficient between the ranges of the observations, where each range refers to the position occupied by the observation when they are organized in an increasing way.

Both measures can be considered as the most common ordinal measures association. These coefficients are non-parametric measures of association between two random variables, being useful when the data are distribution free, so it is not necessary to assume normality (Pearson [43], Hauke and Kossowski [26]). It is well known that these coefficients present significant advantages over the Pearson coefficient: (1) These are more robust coefficients (less sensitive to outliers) and (2) Kendall and Spearman coefficients are better indicators than the Pearson correlation for determining whether a relationship exists between two variables when the relationship is nonlinear. We have previously given a small definition of the sample version of both coefficients. We now introduce the formal definition of their population version.

Definition 1.1.2 (Kendall's τ coefficient.) *Let (X, Y) be a bivariate random vector. Kendall's τ coefficient is the difference of the probabilities of concordance and discordance between two different realizations of a random vector (X, Y):*

$$\tau = [P(X_1 - X_2)(Y_1 - Y_2) > 0] - [P(X_1 - X_2)(Y_1 - Y_2) < 0], \tag{1.1.2}$$

where (X_1, Y_1) and (X_2, Y_2) are independent and identically distributed copies of (X, Y).

Definition 1.1.3 (Spearman's coefficient.) *Let $(X_1, Y_1), (X_2, Y_2)$ and (X_3, Y_3) be independent and identically distributed copies of (X, Y). Then Spearman's coefficient ρ_s associated to (X, Y) is defined by:*

$$\rho_s = 3[P\{(X_1 - X_2)(Y_1 - Y_3) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_3) < 0\}].$$

As we can see, Spearman's coefficient is proportional to the difference between the probability of concordance and the probability of discordance for two vectors (X_1, Y_1) and (X_2, Y_3).

An alternative definition of ρ_s is given by calculating the Pearson coefficient between the uniform random variables $U = F_X(X)$ and $V = F_Y(Y)$, that is,

$$\rho_s = \rho_p[U, V] = \frac{E(UV) - E(U)E(V)}{\sqrt{Var(U)}\sqrt{Var(V)}}, \tag{1.1.3}$$

where ρ_p denotes the Pearson coefficient. The random variables U and V are called the "grades" of X and Y. For this reason Spearman's coefficient is also called *the grade correlation coefficient*. Observe that the grades are values always in $[0, 1]$ and they are bounded independently of the support of the random variables from which the observations came from.

Scarsini [47] studied the measures of concordance in terms of an special function that characterizes the structure of dependence between random variables which are called copula. He 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 [[50], [51]]. These properties are gathered in Xu et al. [55] and are set out below. Let γ be any dependence measure and let (X, Y) be a bivariate random vector. Then,

- $-1 \leq \gamma \leq 1$.
- If X and Y are concordant then $\gamma = 1$.
- If X and Y are discordant then $\gamma = -1$.
- If X and Y are independent then $\gamma = 0$, but if $\gamma = 0$ the variables X and Y are not necessarily independent.
- If α and β are strictly increasing functions then $\gamma[\alpha(X), \beta(Y)] = \gamma[X, Y]$.

In this dissertation, we take into account the previous set of properties when introducing the functional versions of the Kendall and Spearman coefficients. We emphasize the population version as well as the sample version and prove some desirable properties that these two coefficient must fulfil, some of them coming from those introduced by Scarsini [47]. Other interesting measures of association and dependence that satisfy this set of properties can be seen in Kruskal [31], Fernández [22] and Lehmann [32].

From the beginning of this introduction, we have focused on the advantages that the Kendall and Spearman coefficients have over the Pearson coefficient. Now, in terms of the robustness, we can affirm that another disadvantage of the Pearson correlation coefficient is that it is very sensitive to the presence of outliers, since the definition of its sample version depends on calculating sums of transformations of the data; we know that the value of a sum is sensitive to extreme data and as a consequence the mean will be also sensitive. However, observe that if for defining the Pearson coefficient we take the median instead of the mean, we will have a robust version of this same coefficient. This idea of obtaining a robust alternative to the Pearson coefficient was developed by Falk [19], and we briefly present it here.

The median of a random variable X, from now on $med(X)$, is a location measure that have advantages over the mean since it is more robust. It is well known that $med(X)$ is the value or values on the support of the random variable that separate the higher half of the probability distribution from the lower half. Therefore, it must satisfy the inequalities

$$P\left(X \leq med(X)\right) \geq \frac{1}{2} \text{ and } P\left(X \geq med(X)\right) \geq \frac{1}{2}.$$

A widely accepted definition of $med(X)$ can be made through the generalized inverse of the distribution function $F_X(x)$,

$$med(X) \equiv \inf\left\{t \in \mathbb{R} \ : \ F_X(t) \geq \frac{1}{2}\right\}.$$

The median also allows a robust alternative to standard deviation which is called median absolute deviation from the median (MAD), this is,

$$MAD(X) \equiv med(\mid X - med(X) \mid). \tag{1.1.4}$$

Based on the concept of MAD, Falk [19] proposed a robust alternative to the covariance between random variables that he called comedian of X and Y, and it is denoted by $COM(X,Y)$. The comedian between two random variables is defined as

$$COM(X,Y) \equiv med(X - med(X))(Y - med(Y)). \tag{1.1.5}$$

Observe from (1.1.5) that a robust version for the $COV(X,Y)$ is given just by always imposing the operator $med(\cdot)$ instead of the expectation $E(\cdot)$. A very important advantage of the comedian over the covariance is that it always exists, while the covariance requires the existence of the first two moments of the random variables X and Y. Falk [19] stated that its robust version of the covariance also satisfies some desired properties such as:

- If X and Y are independent then $COM(X,Y) = 0$.
- $COM(X,Y) = aMAD(X)^2$, if $Y \overset{\text{a.s.}}{=} aX + b$, for some $a, b \in \mathbb{R}$.
- $COM(X, aY + b) = aCOM(X,Y)$.
- $COM(X,Y) = COM(Y,X)$.

The covariance is a relevant issue when we talk about correlation. Therefore, it is natural to define a correlation coefficient based on the comedian and the MAD which Falk [19] has called correlation median and is defined as

$$\delta(X,Y) = \frac{COM(X,Y)}{MAD(X)MAD(Y)}.$$

Note that as COM as MAD are more robust than COV and the standard deviation, respectively. Then clearly the correlation median will also be a more robust alternative than the correlation coefficient. $\delta(X,Y)$ fulfils two important properties:

- $\delta(X,Y) = 0$, if X and Y are independent.

- $\delta(X,Y) \in \{-1,1\}$, if $Y = aX + b$.

It is important to note that δ in general does not belong to the interval $[-1,1]$; it only falls into such an interval when the random variables follow bivariate elliptical distributions (see Falk [20]). In other cases, when $|\delta| > 1$, the interpretation of this value can be difficult as is stated in Falk [19]. In last chapter of the thesis, we will adapt this measure introduced by Falk [19] to characterize robust version of the Pearson coefficient for two groups of curves. To do that, we will use a definition of functional depth to obtain the deepest curve, which in our context of functions, will be the functional median.

Measuring dependence between two groups of functions has been not much explored in the literature. However, some statistical techniques have already been implemented to try to calculate the dependence between two sets of curves. In the next section we gather some of the most relevant ones which we will use as benchmark to compare its values with those introduced in this work.

1.2 Dependence measures for functional data

In the literature there are few references addressing the problem of dependence on this kind of data. Some authors have tried to extend it from multivariate analysis to the domain of functional data analysis; however, this is not a trivial task as it requires functional analysis tools. Leurgans et al. [33] 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. [27] proposed an alternative way of finding the canonical correlation through the extension of multivariate analysis ideas. Opgen-Rhein and Strimmer [42] proposed an estimator of the dynamical 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 [11] to analyze a nonparametric method to quantify the covariation of components of multivariate longitudinal observations. Li and Chow [35] provided a generalization of Pearson's correlation coefficient for functional data that allows a measure of agreement to be introduced. This measure is called the concordance correlation coefficient and was used to evaluate the reproducibility of repeated-paired curve data.

Next, we briefly summarize some dependence measures studied previously in the literature in order to have a benchmark for comparison later on.

- **Canonical correlation**

An extension of canonical correlation to functional data has been proposed in Leurgans et al. [33], who pointed out the need for regularization in order to provide more interpretability of the results and useful information from the data. As Ramsay and Silverman [45] argue, canonical correlation analysis seeks to investigate which modes of variability in two sets of curves are most associated with one another. As usual, assume that n observed pairs of data curves (x_i, y_i) are available for argument t in some finite interval I, and all integrals are taken over I. The problem is finding a pair of functions (ξ, η), called canonical variates weight (see Figure 1.4), which maximizes the following penalized squared sample correlation, defined as

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

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. The functions ξ and η, may be the components of variation in the two curves that most account for the interaction between the two groups of curves. Having a pair of canonical variables with fairly smooth weight functions and correlations that are not excessively low can be achieved by selecting the appropriate smoothing parameter. This parameter can be chosen either subjectively or through a cross-validation score if an automatic procedure is required. This technique is carried out using basis functions for the functions (x_i, y_i) and for the weight functions ξ, η. Figure 1.4 shows the canonical variate weight functions of two functional data sets.

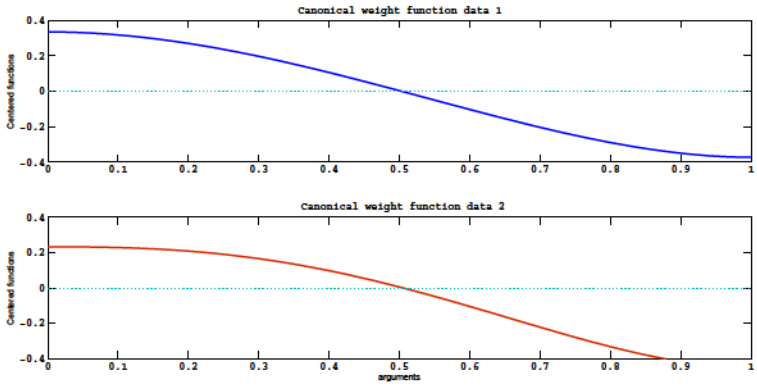

Figure 1.4: Canonical variate weight functions for two sets of curves, $\rho_c = 0.5449$.

- **Dynamical correlation**

A technique used to find the correlation between groups of functions is the dynamical correlation, which is a measure of similarity between two curves, introduced by Dubin and Muller [11] as a simple and efficient non-parametric correlation measure for multivariate longitudinal data. They interpret dynamical correlation as a measure of the average concordant or discordant behavior of pairs of random trajectories, in the sense that "*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 the dynamical correlation is negative*". Opgen-Rhein and Strimmer [42] study the dynamical correlation under a functional perspective. This approach provides a similarity score for pairs of groups of randomly sampled curves. Hence, the dynamical correlation between two exactly known curves will be $\left\langle x^S(t), y^S(t) \right\rangle$.

Thus, the dynamical correlation between two functional variables X and Y is given by

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

where $X^S(t) = \frac{X^c(t)}{\sqrt{E\langle X^c(t), X^c(t)\rangle}}$, $Y^S(t) = \frac{Y^c(t)}{\sqrt{E\langle Y^c(t), Y^c(t)\rangle}}$, $X^c(t) = X(t) - \langle EX(t), 1\rangle$, $Y^c(t) = Y(t) - \langle EY(t), 1\rangle$ and $\langle\cdot\rangle$ means the usual inner product for functions $\langle X(t), Y(t)\rangle = \int_I X(t)Y(t)dt$, which can been viewed as an average of individual correlations.

We will use in this dissertation the following estimator of the dynamical correlation proposed in Opgen-Rhein and Strimmer [42], which is a slightly revised version of the dynamical correlation introduced in Dubin and Muller[11], but for functional data,

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

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

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

As we can see, $\widehat{\rho}_d$ is a estimator of the population dynamical correlation ρ_d.

Opgen-Rhein and Strimmer [42] used this estimator to find the correlation between pairs of genes. It allows us to compute the partial dynamical correlations, which will represent the edges of a gene association network. The strength of these coefficients indicates the presence or absence of a direct association between each pair of genes.

Figure 1.5 illustrates two negatively dependent variables (genes). For each variable there are two measured curves, and there are three slightly different ways in which the

sampled curves relate to each other. The dynamical correlations for the three cases are -0.946, -0.982 and -0.947, respectively. This example is taken from Opgen-Rhein and Strimmer [42].

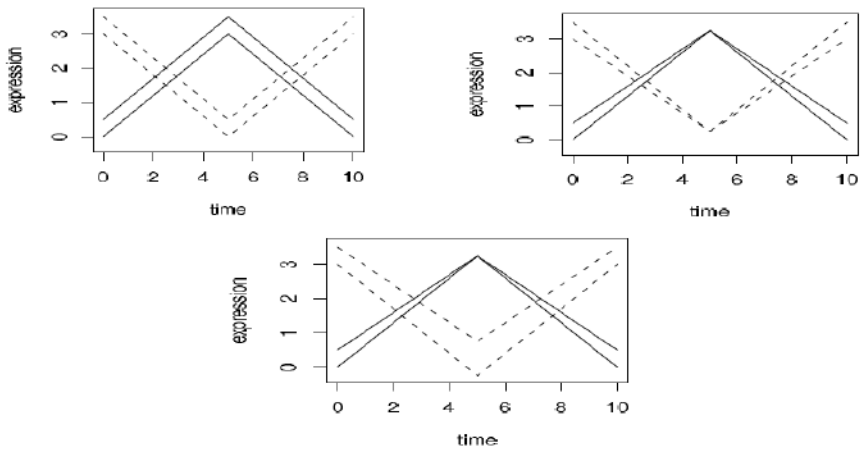

Figure 1.5: Example to illustrate the concept of dynamical correlation between two variables (genes).

- **Pearson's coefficient for functions**

Li and Chow [35] provided a generalization of Pearson's correlation coefficient for functional data that allows to be introduced a measure of agreement. This measure is called the concordance correlation coefficient and was used to evaluate the reproducibility of repeated-paired curve data.

Let $X(t)$ and $Y(t)$ be two stochastic processes. The Pearson's correlation coefficient for $X(t)$ and $Y(t)$ is

$$\rho_p(X(t), Y(t)) = \frac{\langle X(t) - E(X(t)), Y(t) - E(Y(t))\rangle}{\|X(t) - E(X(t))\|\|Y(t) - E(Y(t))\|}, \tag{1.2.1}$$

where the inner product is defined as

$$\langle X(t), Y(t)\rangle = E\int X(t)Y(t)w(t)dt,$$

and the norm is induced by the inner product. The weight function $w(t)$ allows us to assign importance to different parts of t. A subjective approach to calculate a weight function is used when some prior information on the importance of different time intervals is available. When there is no prior information, an objective approach is necessary. Therefore, t should be regarded as a random variable defined on the interval I and the

density function of t chosen based on the data $t_1, \ldots, t_N$, as the weight function. This density function can be estimated via a kernel estimator,

$$\widehat{w}(t) = \frac{1}{Nh} \sum_{i=1}^{N} K\left\{\frac{(t_j - t)}{h}\right\},$$

where $K(\cdot)$ is a kernel density function, such as the Gaussian density function, and h is a bandwidth to be chosen. A rule of thumb suggests taking $h = 1.06 s_t N^{-\frac{1}{5}}$ for the Gaussian kernel, where s_t is the sample standard deviation of $t_1, \ldots, t_N$. In our study, we set $w(t) = 1$, assigning the same weight for each t.

The estimator for calculating the Pearson correlation coefficient when we have n observed pairs of data curves (x_i, y_i) is given by

$$\widehat{\rho}_p = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{N}\{x_i(t_j) - \overline{x}(t_j)\}\{y_i(t_j) - \overline{y}(t_j)\}w(t_j)\Delta_j}{\{\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{N}(x_i(t_j) - \overline{x}(t_j))^2 w(t_j)\Delta_j\}^{\frac{1}{2}}\{\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{N}(y_i(t_j) - \overline{y}(t_j))^2 w(t_j)\Delta_j\}^{\frac{1}{2}}},$$

where $\Delta_j = t_{j+1}$-t_j, the gap size between t_{j+1} and t_j, $\overline{x}(t_j) = \frac{1}{n}\sum_{i=1}^{n} x_i(t_j)$ and $\overline{y}(t_j) = \frac{1}{n}\sum_{i=1}^{n} y_i(t_j)$ are the sample means of $x_i(t_j)$ and $y_i(t_j)$, respectively.

- **Cross-correlation function**

In functional data analysis, it is possible to measure the dependence between two sets of curves through the cross-correlation functions, discussed in Ramsay and Silverman [45] (p.24). Assume n pairs of curves (x_i, y_i), for $i = 1, \ldots, n$, from a bivariate random process $(X(t), Y(t))$ which are defined on the same interval $I = [a, b]$. Then the cross-covariance function is given by

$$\mathrm{COV}_{XY}(t_1, t_2) \equiv E[\{X(t_1) - E(X(t_1))\}\{Y(t_2) - E(Y(t_2))\}],$$

and the cross-correlation function is

$$\mathrm{CORR}_{XY}(t_1, t_2) \equiv \frac{E[\{X(t_1) - E(X(t_1))\}\{Y(t_2) - E(Y(t_2))\}]}{\sqrt{E\{X(t_1) - E(X(t_1))\}^2 E\{Y(t_2) - E(Y(t_2))\}^2}}.$$

Therefore, the samples version are given by:

$$\widehat{\mathrm{COV}}_{XY}(t_1, t_2) \equiv (n-1)^{-1} \sum_{i=1}^{n} \{x_i(t_1) - \overline{x}(t_1)\}\{y_i(t_2) - \overline{y}(t_2)\}, \tag{1.2.2}$$

where $\overline{x} = n^{-1}\sum_{i=1}^{n} x_i(t)$ and $\overline{y} = n^{-1}\sum_{i=1}^{n} y_i(t)$. And the cross-correlation function is naturally defined as:

$$\widehat{\mathrm{CORR}}_{XY}(t_1, t_2) \equiv \frac{\mathrm{COV}_{XY}(t_1, t_2)}{\sqrt{\mathrm{VAR}_X(t_1)\mathrm{VAR}_Y(t_2)}}, \tag{1.2.3}$$

where $\widehat{\mathrm{VAR}}_X(t) = (n-1)^{-1}\sum_{i=1}^{n}(x_i(t) - \overline{x}(t))^2$.

The procedure for calculating the cross-correlation consists basically of calculating Pearson's coefficient between the values of the functions for each $t \in I$, i.e., we are analyzing the way that one function depends on another in each instant of time. Note that this coefficient works with the mean of the data as a location measure, which can lead to a more sensitive procedure under the presence of outliers, as in the case of Pearson's coefficients for bivariate data.

This methodology is basically graphic. If we compute the Pearson coefficient between the values obtained from evaluating the two groups of functions in each $t_1, t_2 \in I$, then we obtain a surface. Figure 1.6 shows the surface and contour plot for the cross-correlation of two sets of curves.

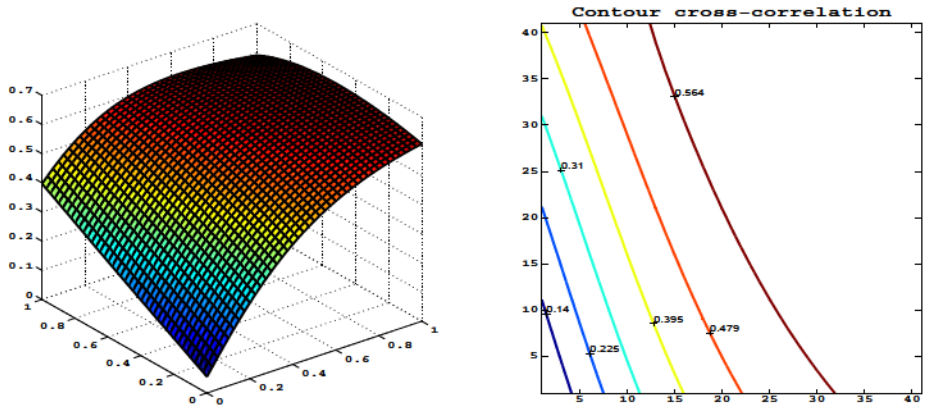

Figure 1.6: Surface and contour plot for the cross-correlation.

We only consider the case where $t_1 = t_2$; hence in the remainder of this dissertation, we will call the cross-correlation function in $t_1 = t_2$ **correlation function**, which represents a curve and can be easier to interpret than the whole surface. (see Figure 1.7).

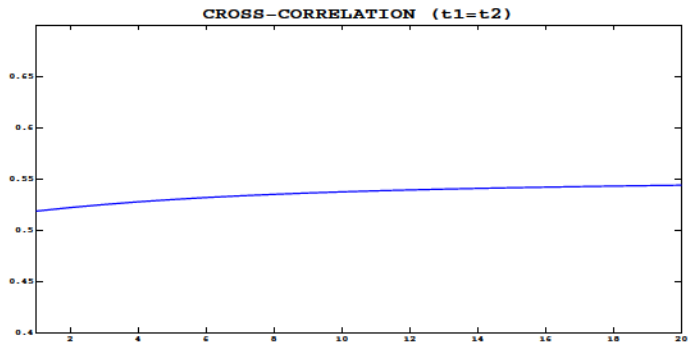

Figure 1.7: Cross-correlation for $t_1 = t_2$.

1.3 Ordering functions

The main purpose of this dissertation is to extend to functional data some of the measures of dependence introduced previously. In particular, we will focus on measures that are based on some order of the data, such as Kendall and Spearman coefficients. The major problem that we have found in giving a generalization of these dependence measures has been the difficulty of defining an order among functions that performs well. It is worth pointing out that the classic order among functions, which compares the functions pointwise, does not work here because if two curves cross then they are not comparable. Hence, we face the problem of ordering functions with different approaches that have been chosen under an exhaustive study of simulation, which allows us to identify the kind of function ordering that presents a good performance in aspects such as interpretation, practical sense and computational implementation for each one of the dependence measures introduced in this thesis. We will now give a brief description of the orders for functions that will be used to develop our contributions. One first idea of ordering functions is through data segmentation methodologies that have their origins in the multivariate analysis and that recently have also been extended to the functional data analysis. We give the basic notions to some depth functions in the multivariate setting as well as the notions of the same concept for functions. A second idea that we use to sort functions is observing the proportion of time that one curve is above another. This methodology leads to a down-up order instead of a center-outwards order induced by a depth order. The way these orders work is commented on briefly below.

1. **Center-outwards order**

   A multivariate depth notion allows us to measure the centrality or outlyingness of a point from the sample with respect to the multivariate sample or to its underlying distribution. It provides a natural center-outward order for the sample data, which allows us to extend a wide range of statistical univariate techniques to the multivariate setting such as multivariate goodness of fit, location measure, scatter estimates and risk measurement. A recent review on the depth function and its several applications can be found in Cascos [3].

   A depth function is defined by a mapping $D : \mathbb{R}^n \longmapsto [0, 1]$, which satisfies the properties of *affine invariance, vanishing at infinity, monotonicity with respect to the deepest point and maximality at center*. Here, we describe briefly two classic depth functions called *halfspace depth and simplicial depth*. The first one was proposed by Tukey [52] in a data analysis contest. Given a multivariate sample, the halfspace depth of a point $\mathbf{x} \in \mathbb{R}^d$ is the smallest fraction of data points in a closed halfspace containing $\mathbf{x}$, or also the smallest fraction of data points that should be deleted so that $\mathbf{x}$ lies outside the convex hull of the remaining data points. The sample version of the halfspace depth for a point

$x \in \mathbb{R}^d$ with respect to a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ in $\mathbb{R}^d$ is defined by

$$HD_n(\mathbf{x}) = n^{-1} \inf_{\mathbf{u}\in\mathbb{R}^n} \#\left\{i \ : \ \langle \mathbf{X}_i, \mathbf{u}\rangle \geq \langle \mathbf{x}, \mathbf{u}\rangle\right\},$$

and the population version of the halfspace depth of a point $\mathbf{x}$ in $\mathbb{R}^n$ with respect to the probability distribution F is

$$HD(\mathbf{x}, \mathbf{P}_F) = \inf\left\{\mathbf{P}_F(H) \ : \ \mathbf{x} \in H \ \text{ closed halfspace}\right\}.$$

Note that in the univariate case, the halfspace depth can be expressed as

$$HD(x, \mathbf{P}) = \min\left\{\mathbf{P}(X \leq x), 1 - \mathbf{P}(X \leq x)\right\},$$

which is maximized by the well-known univariate median.

The second one is the simplicial depth, which was introduced by Liu [36], based on random simplices. The simplicial depth of a point $\mathbf{x} \in \mathbb{R}^d$ is given by the probability that the point $\mathbf{x}$ is contained inside a random simplex whose vertices are $p+1$ independent observations. For the sample case it can be defined by

$$SD_n(\mathbf{x}) = \begin{pmatrix} n \\ p+1 \end{pmatrix}^{-1} \sum_{1\leq i_1<i_2<\cdots<i_n} \mathbb{I}(\mathbf{x} \in \mathrm{co}\{\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \ldots, \mathbf{X}_{i_{p+1}}\}),$$

where $\mathbb{I}$ and co mean the indicator function and the convex hull, respectively. The population definition of the simplicial depth is

$$SD(\mathbf{x}, \mathbf{P}_F) = \mathbf{P}_F\{\mathbf{x} \in \mathrm{co}\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{p+1}\}.$$

Observe that in the univariate case for continuous absolutely distribution, the simplicial depth can be expressed as

$$SD(x, P_F) = 2F(x)(1 - F(x)),$$

where F is the distribution function of the random variables X.

The center-outward order for the sample data induced by the empirical versions of the depth functions leads to the introduction of multivariate generalizations of the univariate sample median and also L-statistics. However, it is well known that they have the drawback of not being feasible computationally in high dimension, hence the multivariate order induced through depth functions is quite limited for dimensions greater than three. However, López-Pintado and Romo [38] introduced a depth notion for functional data and a finite-dimensional version of this concept of depth that can also be considered as a new notion of depth for multivariate data that verifies essentially all the properties established in Zuo and Serfling [56] (e.g. monotonicity with respect to the deepest observation, maximization at the center of symmetry, etc). In addition, it has

the advantage of being computationally less intensive than other multivariate depths, which makes it adequate for analyzing high-dimensional data.

The depth notion for functions has an important role in the analysis of functional data since the concept allows us to define functional versions of robust statistics such as the median curve or trimmed mean as well as provide a natural ordering within a sample of curves, thus making the definition of order statistics and the assignment of ranks to each one of the curves of the sample possible. To our knowledge, the first idea of depth concept for functional observations was introduced by Fraiman and Muniz [24] where they consider a set of n curves $\{x_1(t), x_2(t), \ldots, x_n(t)\}$ defined on an interval $[T_1, T_2]$. Then the value of the depth for any curve $x_i(t)$ is given by

$$D[x_i(t)] = \int_{T_1}^{T_2} D_1[x_i(t)]dt,$$

where $D_1[x_i(t')]$ is the univariate depth of the point defined by the curve x_i in t' respect to other $n-1$ points defined by the curves $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$, also evaluated in t'. The depth studied by Fraiman and Muniz is a useful tool to define robust estimators in the functional case and it is easily adaptable to the multivariate analysis by using a appropriate summation instead of the integral.

Another notion of functional depth was defined in Cuevas et al.[7]. Let $\{x_1(t), x_2(t), \ldots, x_n(t)\}$ be a set of n curves; according to this method, the h-modal depth of the function $x_i(t)$ is given by the expression:

$$hD_n(x_i, h) = \sum_{k=1}^{n} \frac{K(\|x_i - x_k\|)}{h}.$$

Where h should be interpreted as a bandwidth, K is a kernel function defined on the real positive numbers and $\|.\|$ is the norm L^2.

These same authors define in Cuevas et al.[8], two more measures of depth based on some ideas of Cuesta-Albertos et al.[4], [5], which combine random projections of the functions in different directions with a bivariate data depth that is used to order the corresponding results. More precisely, given $\{x_1(t), \ldots, x_n(t)\}$ and a random direction a, the sample depth of x_i is defined as the univariate depth of the corresponding one-dimensional projection. When the sample is made of functional data, the x_i belongs to the Hilbert space $L^2[0,1]$ so that the projection of a datum x is given by the standard inner product $\langle a, x\rangle = \int_0^1 a(t)x(t)dt$. It is clear that this definition leads to a random measure of depth, as it is based on the rank of the projections along a random direction; this method is called random projection (RP). The second idea is to use the method of random projections simultaneously for the functions and their derivatives, thus incorporating the information on the function smoothness provided, which is relevant in some practical applications. The sample of functions $\{x_1(t), \ldots, x_n(t)\}$ is reduced to

a sample in R^2 defined by $(\langle a, x_1\rangle, \langle a, x_1'\rangle), \ldots, (\langle a, x_n\rangle, \langle a, x_n'\rangle)$ where a is a randomly chosen direction. Now, depending on the treatment of this bi-dimensional sample, there are several alternative possibilities. The random projection method could be used again for the bi-dimensional projections $(\langle a, x_1\rangle, \langle a, x_1'\rangle), \ldots, (\langle a, x_n\rangle, \langle a, x_n'\rangle)$. This method is denoted by $RP2$.

Finally, we consider the notion of functional depth introduced in López-Pintado and Romo [38], which is based on the graphic representation of the curves and the bands that they determine in the plane. We pay special attention to this measure depth since one of the contributions of the dissertation is based on this measure. This proposal of depth follows a graph-based approach and although it is widely explained in López-Pintado and Romo [38], for the reader's convenience, we repeat some relevant definitions from there, thus making our exposition self-contained. Let $x_1(t), \ldots, x_n(t)$ be a sample of curves belonging to $C(I)$. The graph of a function x is the subset of the plane $G(x) = \{(t, x(t)), t \in I\}$. The band in $\mathbb{R}^2$ defined by the curves $x_{i1}, \ldots, x_{in}$ is

$$B(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \left\{(t, y) : t \in I, \min_{r=1,\ldots,k} x_{i_r}(t) \leq y \leq \max_{r=1,\ldots,k} x_{i_r}(t)\right\}.$$

In Figure 1.8 we show a band region for two and three curves, López-Pintado and Romo [38].

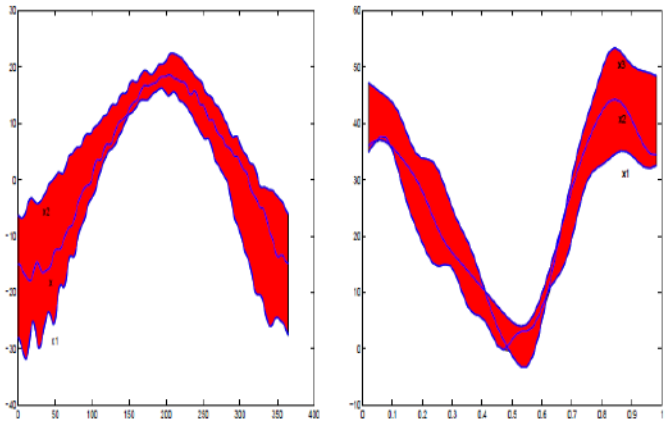

Figure 1.8: The band defined by two curves x_1, x_2 and a third curve x belonging to the band. Right: the band determined by three curves x_1, x_2 and x_3.

The proportion of bands $B(x_{i_1}, x_{i_2}, \ldots, x_{i_j})$ given by j different curves $x_{i_1}, x_{i_2}, \ldots, x_{i_j}$ containing the graph of x is

$$S_n^{(j}(x) = \begin{pmatrix} n \\ j \end{pmatrix}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \ldots i_j \leq n} I\{G(x) \subset B(x_{i_1}, x_{i_2}, \ldots, x_{i_j})\}, \quad j \geq 2,$$

Therefore, the band depth of x is given by

$$S_{n,J}(x) = \sum_{2}^{J} S_n^{(j}(x), \quad j \geq 2.$$

López-Pintado and Romo [38] also have given another more flexible definition, called generalized band depth. Band depth depends strongly on the curves's shape, whereas generalized band depth is more convenient for irregular functions. For any function x in $x_1, x_2, \ldots, x_n$ let

$$A_j(x) = A(x; x_{i_1}, x_{i_2}, \ldots, x_{i_j}) = \left\{ t \in I : \min_{r=i_1,\ldots,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,\ldots,i_j} x_r(t) \right\}, \; j \geq 2,$$

be the set of points in the interval I where the function x is inside the band given by the observations $x_{i_1}, x_{i_2}, \ldots, x_{i_j}$, then

$$GS_n^{(j}(x) = \begin{pmatrix} n \\ j \end{pmatrix}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \ldots i_j \leq n} \lambda_r(A(x; x_{i_1}, x_{i_2}, \ldots, x_{i_j})), \quad j \geq 2,$$

is a generalized version of $S_n^{(j}(x)$. If λ is Lebesgue measure in $\mathbb{R}$, then

$$\lambda_r = \lambda(A_j(x))/\lambda(I)$$

will be the proportion of time that x is inside the band. Therefore, the generalized band depth (GBD) of x is given by

$$GS_{n,J}(x) = \sum_{j=2}^{J} GS_n^{(j}(x), \quad j \geq 2. \tag{1.3.1}$$

If $X_1, X_2, \ldots, X_n$ are independent copies of the stochastic process $X(t)$, the population version of $GS_n^{(j}(x)$ and $GS_{n,J}(x)$ are given by

$$GS^{(j}(x) = E\lambda_r(A(x; X_1, X_2, \ldots, X_j)), \quad j \geq 2. \quad \text{and}$$

$$GS_J(x) = \sum_{j=2}^{J} GS^{(j}(x) = \sum_{j=2}^{J} E\lambda_r(A(x; X_1, X_2, \ldots, X_j)), \; j \geq 2, \text{ respectively.} \tag{1.3.2}$$

Note that a functional median can be seen as that curve from the sample that maximizes (1.3.1).

$$\hat{m}_{n,J} = \arg \max_{x \in \{x_1, x_2, \ldots, x_n\}} GS_{n,J}(x)$$

With this functional median definition, we therefore have a tool to develop one of the contributions of the thesis that refers to a robust alternative to the cross-correlation function. The results are detailed in Chapter 4.

2. Down-up order for funtions

Recall that the classic order between two functions, $x_1(t)$, and $x_2(t)$ defined on the same interval T is given by $x_1 \leq x_2 \equiv x_1(t) \leq x_2(t)$ for all $t \in T$. This definition induces an down-up order instead of a center-outwards order induced by a functional depth. However, this method has the disadvantage of not allowing the sample of functions to be sorted if any two of them are crossed in a finite number of points. A flexible version of the classic order (point-to-point) for functions which is also based on an down-up order was introduced in Martín-Barragan et al. [40] where the concept of epigraph and hypograph of a function is applied to characterize some indexes that are useful for sorting curves in a down-up direction, even when the curves are crossed. Basically, it states that x_1 is smaller than x_2 if, and only if, the proportion of functions under the curve of x_1 is smaller than the proportion of functions under the curve of x_2. Observe that if the curves do not cross, then this order will be the classic order between the two functions mentioned previously. However, there still are some situations where this flexible version of order can also fail, for instance, when in the sample all the curves cross each other. An alternative for dealing with these types of situations is introduced in López-Pintado and Romo [39], through two concepts called the *Inferior Length* and the *Superior Length* of a curve x, which are respectively defined as

$$IL_n(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \geq x_i(t)\},$$

$$SL_n(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \leq x_i(t)\},$$

where λ stands for the Lebesgue measure on $\mathbb{R}$. Thus, the inferior [superior] length $IL_n(x)$, $[SL_n(x)]$ can be interpreted as the "proportion of time" that the curve x is above [below] any another curve of the sample.

We want to highlight that the order for functions induced by the previous expression allows ranges to be assigned each of the curves of the sample, and its population version will be useful for assigning the grade of a stochastic process. Thus, both $IL_n(x)$ and $SL_n(x)$ will allow us to give one of main contribution of this thesis, developed in Chapter 3, which refers to the extension of the Spearman coefficient for functions.

Finally, we note that for developing the functional version of Kendall's tau studied in Chapter 2, we have considered two orders for functions that have not been used so far. These orders will sort the curves by area under the graph of the function and the maximum value of the function, both orders on the full interval where the group of the functions are defined. That is:

$$x_1(t) \prec x_2(t) \equiv \begin{cases} \int_T [x_2(t) - x_1(t)]\, dt \geq 0, & \text{order of the integral.} \\ \max_{t\in T} x_1(t) \leq \max_{t\in T} x_2(t), & \text{order of maximun.} \end{cases}$$

The details and some properties of these orders will be developed in Chapter 2.

After introducing almost all of the tools and basic concepts that will be used, for performing the main contributions of this work, in the next section we will present the structure and outline of this dissertation.

1.4 Structure of the dissertation

This thesis contains five chapters. The current Chapter 1 presents a brief historical review of the development of dependence measures through time, and some classic dependence measures for bivariate data showing their principal characteristics and properties. We also study some measures that have already been analyzed in the functional context and with which we compare our results. Finally, we present some ordering in the multivariate context as well as in functional setting to define an order among curves that performs well for our goals.

The contributions of this dissertation are developed in Chapters 2, 3 and 4. In the first part of Chapter 2, the functional τ is defined using two functional pre-orders to sort the observations and extend the concept of concordance for bivariate random variables to the functional setting. In Section 2.3, the main properties, as well as the asymptotic results, are discussed. A simulation study and sensitivity analysis are given in Section 2.4. In the second part of this chapter, we present two examples with real data. The first data set consists of the prices of the assets in companies belonging to the IBEX35. The functional τ informs about companies having similar behavior over time. 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 a gene network. Finally, we present a robust empirical study and outline the main conclusions of this chapter in Section 2.8.

In Chapter 3, firstly we recall some concepts about Spearman's coefficient for bivariate samples necessary to understand the extension to the functional context. We introduce the notion of grade for functions that it is useful to develop the theoretical background necessary to properly define the Spearman coefficient. Then, we go on to discuss the main properties, as well as the asymptotic results. A simulation study and a robustness analysis are carried out in Section 3.5, while Section 3.6 provides our independence test and a simulation study. Several examples with real data are shown in Section 3.7. Finally, the main conclusions of this chapter are listed.

In Chapter 4, we develop a more robust alternative measure of dependence than the cross-correlation function studied in Ramsay and Silverman [45]. In the first sections, we consider the principal aspects that we will take into account for the definition of our coefficient, and present the definitions of MAD and comedian for functional data. The new coefficient, called the correlation median for functions, and its properties are defined in Section 4.4. A simulation

study is carried out in Section 4.5, where we also present a sensitivity study of the coefficient. We analyze the robustness of the coefficient and offer several examples with real data showing how the correlation median for functions works. Finally, we summarize the main conclusions of this chapter.

In Chapter 5, we present some general conclusions and summarize the main contributions of the thesis.

CHAPTER 2

A Kendall correlation coefficient for functional dependence

2.1 Introduction

After introducing the preliminaries concepts, notation and references in the topic of the thesis, we develop our proposal and make our contribution to the literature, providing coefficients that capture relations between functional random variables.

In this chapter, we extend a Kendall τ correlation coefficient [30] 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 chapter is organized as follows. In Section 2.2, the functional τ is defined extending the concept of concordance for bivariate random variables. Section 2.3 is devoted to proving some properties of this correlation coefficient and to studying convergence results. A simulation study and sensitivity analysis are given in Section 2.4. In Section 2.5 we analyze with our methodology the prices of the assets in companies belonging to the IBEX35. Section 2.6 contains a study of dependence between genes using the genes data set. In Section 2.7, we present a robustness empirical study. Finally, in section 2.8, we outline the main conclusions of this chapter.

2.2 Functional Kendall correlation coefficient

Kendall [30] 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

$$\tau = 2[P\{X_1 < X_2\ ,\ Y_1 < Y_2\} \ +\ P\{X_2 < X_1\ ,\ Y_2 < Y_1\}] - 1. \tag{2.2.1}$$

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

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

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 chapter 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 2.2.1 *Let f and g be in $C(I)$. Then, we consider two alternatives.*

$$f(t) \preceq_m g(t) \equiv \max_{t\in I} f(t) \leq \max_{t\in I} g(t). \tag{2.2.2}$$

$$f(t) \preceq_i g(t) \equiv \int_a^b (g(t) - f(t))dt \geq 0. \tag{2.2.3}$$

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 2.2.2 (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 2.2.2 allows us to extend Kendall's correlation coefficient to the functional case, as described in the next Definition.

Definition 2.2.3 *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:*

$$\widehat{\tau} = \binom{n}{2}^{-1} \sum_{i<j}^{n} 2I(x_i \prec x_j \quad and \quad y_i \prec y_j) \quad + \quad 2I(x_j \prec x_i \quad and \quad y_j \prec y_i) - 1. \tag{2.2.4}$$

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

$$\tau = 2[P\{X_1 \prec X_2\ ,\ Y_1 \prec Y_2\} \ +\ P\{X_2 \prec X_1\ ,\ Y_2 \prec Y_1\}] - 1. \tag{2.2.5}$$

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 2.3.2, 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 2.2.4 (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*

$$U_n = \binom{n}{m}^{-1} \sum_{1 \le i_1 < \cdots < i_m \le n} \Phi\{(X_{i1}, \ldots X_{im})\}. \tag{2.2.6}$$

It is clear that $U_n \in B$. Hence, the U-statistic (2.2.6) 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 2.2.5 (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 Definition (2.2.3) can be expressed as a UB-statistic,

$$U_n = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} \Phi\{(X_{i_1}, Y_{i_1}), (X_{i_2}, Y_{i_2})\}, \tag{2.2.7}$$

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 [48] 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.

2.3 Properties of functional τ

We analyze in this section some desirable properties of τ as a dependence measure. Scarsini [47] proposed a set of properties that a concordance measure for ordered pairs of continuous random variables should fulfill. (See Chapter 1, Section 1.1 for more detail). The following proposition gives the properties of the functional τ. Some of them come from the axioms proposed by Scarsini [47]. Other properties of Proposition 2.3.1 are a natural extension of the well known properties of the bivariate τ itself (Kendall [30]).

Proposition 2.3.1 *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.

Proof Proposition 2.3.1

The properties 1 and 2 are immediate from the expression (2.2.5) 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 (2.2.3).
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 (2.2.2) 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 (2.2.3) 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 $\alpha(g(t)) = \exp(2t)$

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

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

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

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

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

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,

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

Also

$$\begin{aligned}&P(\max_{t\in[a,b]} X_1(t) > \max_{t\in[a,b]} X_2(t)) = 1 - P(\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)),\\ &P(\max_{t\in[a,b]} Y_1(t) > \max_{t\in[a,b]} Y_2(t)) = 1 - P(\max_{t\in[a,b]} Y_1(t) < \max_{t\in[a,b]} Y_2(t))\end{aligned}$$

$$\text{and}\quad P(\max_{t\in[a,b]} X_1(t) < \max_{t\in[a,b]} X_2(t)) = P(\max_{t\in[a,b]} Y_1(t) < \max_{t\in[a,b]} Y_2(t)) = \frac{1}{2}$$

$$\tau_1 = 2[\frac{1}{2} \times \frac{1}{2}] + 2[(1 - \frac{1}{2}) \times (1 - \frac{1}{2})] - 1 = 0.$$

Analogously for the preorder $\preceq_i$, from equation (2.2.3).

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

Finally,

$$P\left(\int_a^b X_1(t)dt > \int_a^b X_2(t)dt\right) = 1 - P\left(\int_a^b X_1(t)dt < \int_a^b X_2(t)dt\right),$$

$$P\left(\int_a^b Y_1(t)dt > \int_a^b Y_2(t)dt\right) = 1 - P\left(\int_a^b Y_1(t)dt < \int_a^b Y_2(t)dt\right)$$

$$\text{and}\quad P\left(\int_a^b X_1(t)dt < \int_a^b X_2(t)dt\right) = P\left(\int_a^b Y_1(t)dt < \int_a^b Y_2(t)dt\right) = \frac{1}{2}$$

$$\tau_2 = 2[\frac{1}{2} \times \frac{1}{2}] + 2[(1 - \frac{1}{2}) \times (1 - \frac{1}{2})] - 1 = 0.$$

□

Property 7.

Proof.
Let α and β be strictly increasing and continuous functions. For the functional preorder $\preceq_m$ from equation (2.2.2), we have:

$$\max_{t\in I}\alpha(x_i(t)) = \alpha(\max_{t\in I}(x_i(t))) \quad \text{and} \quad \max_{t\in I}\alpha(x_j(t)) = \alpha(\max_{t\in I}(x_j(t)))$$
$$\rightarrow \max_{t\in I}\alpha(x_i(t)) \preceq \max_{t\in I}\alpha(x_j(t)) \rightarrow \alpha(x_i(t)) \preceq \alpha(x_j(t)).$$

The same idea can be used for β and $Y(t)$. According to Definition 2.2.2 the number of concordant pairs is the same, therefore

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

□

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

Theorem 2.3.2 *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,*

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

for the two preorders considered in Definition 2.2.1.

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 Definition (2.2.3) and expressed as the UB-statistic (2.2.7), has associated a kernel Φ such that $E\|\Phi\|$ is finite. Therefore, from Theorem 2.2.5, 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 2.3.2 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 2.1 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 2.2 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.

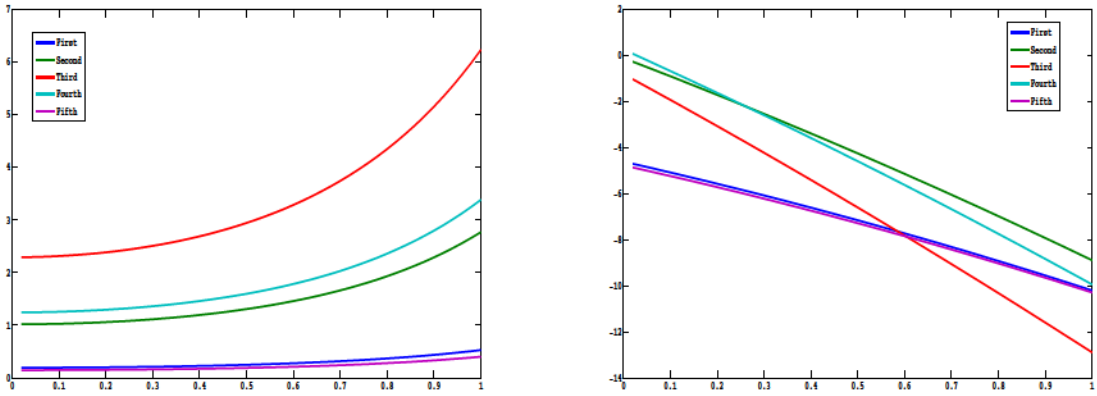

Figure 2.1: $\widehat{\tau}_1 = 0.6 \quad \widehat{\tau}_2 = 0.4$.

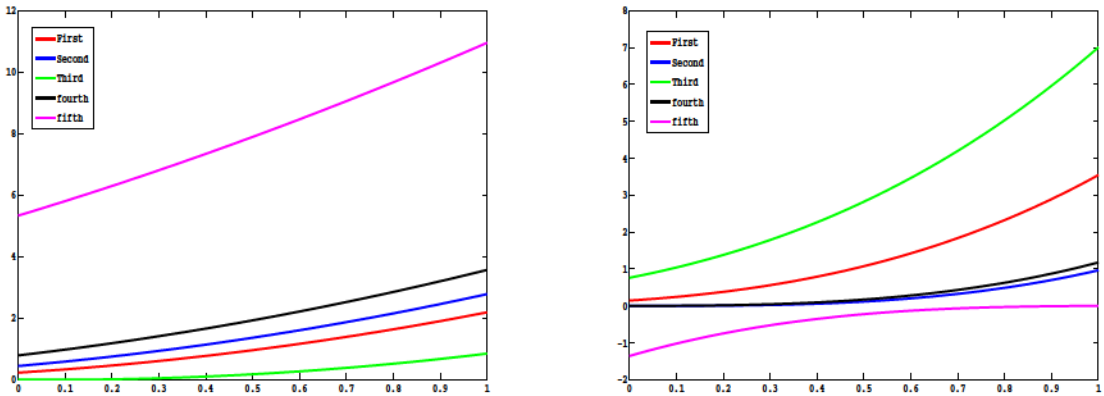

Figure 2.2: $\widehat{\tau}_1 = -0.8 \quad \widehat{\tau}_2 = -0.8$.

2.4 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 dependence measures introduced in the literature. Specifically, we are going to compare τ with dynamical correlation and canonical correlation. Recall that the dynamical correlation is a measure of similarity between two groups of curves, which is calculated through the estimator,

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

And the canonical correlation seeks to investigate which modes of variability in the two sets of curves are most associated with one another, this is, the sample squared correlation of $\int \varepsilon x_i$ and $\int \eta y_i$, i.e.,

$$ccorsq(\varepsilon,\eta) = \frac{\{cov(\int \varepsilon x_i, \int \eta y_i)\}^2}{(var\int \varepsilon x_i)(var\int \eta y_i)}.$$

These two measures were introduced in Chapter 1, Section 1.2 for more details.

Through a simulation exercise, we show the behavior of the measure introduced in this chapter 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 (2.2.1). 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,

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

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

Table 2.1 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 2.1. 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.

Tables 2.1: 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 2.2 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 2.3 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 chapter, since we obtain the same conclusions as before.

It is remarkable that this study of simulation were 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.

2.5 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 2.4 shows the functional

Tables 2.2: Sensitivity to sample size

sample size	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	0.4035 (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)

Tables 2.3: 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

$\widehat{\tau}$ 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 2.3 to 2.7 show the 5 groups. In each one of the groups, we plot the paths determined by the most similar columns of matrix $\mathcal{C}$.

Tables 2.4: 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

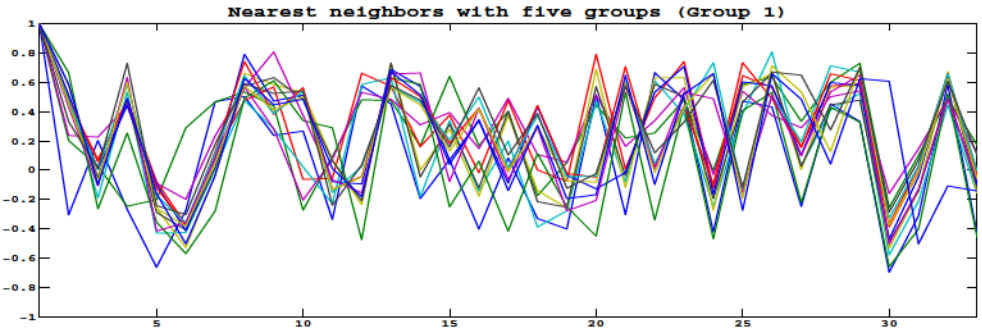

Figure 2.3: First group of companies.

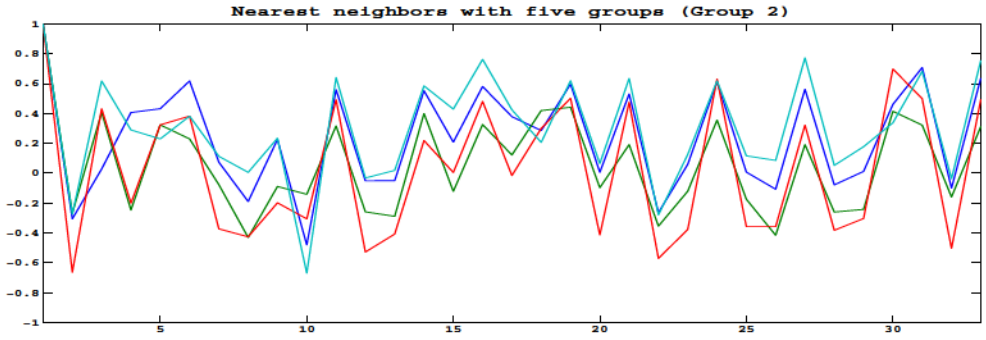

Figure 2.4: Second group of companies.

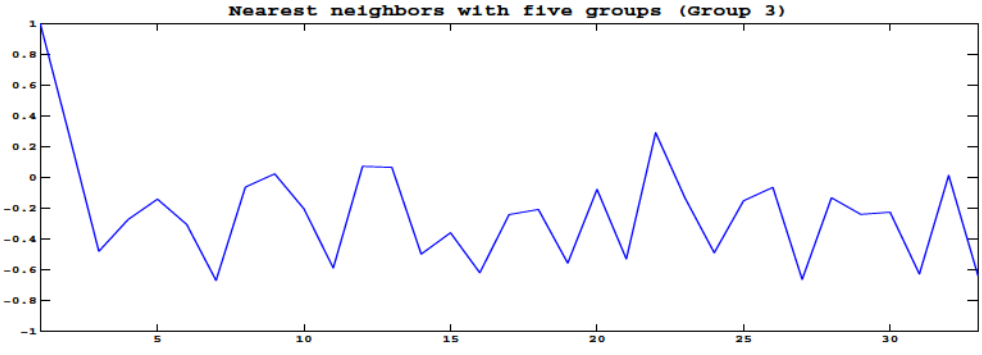

Figure 2.5: Third group of companies.

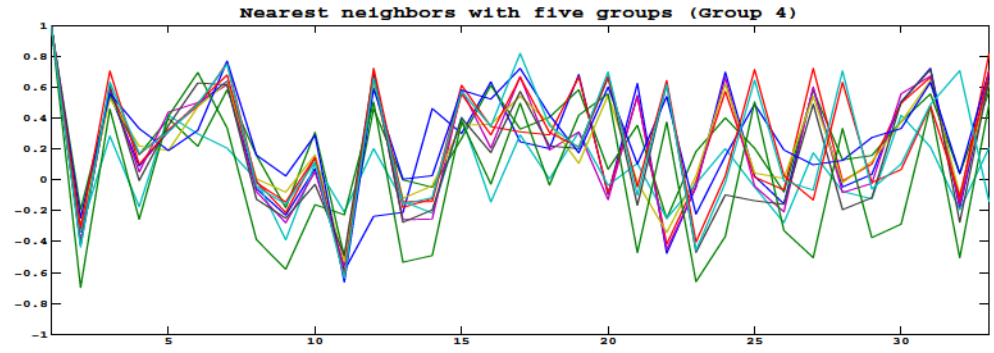

Figure 2.6: Fourth group of companies.

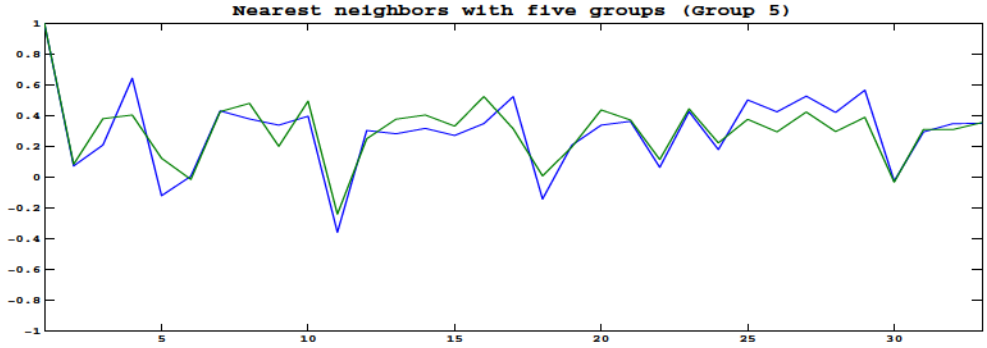

Figure 2.7: Fifth group of companies.

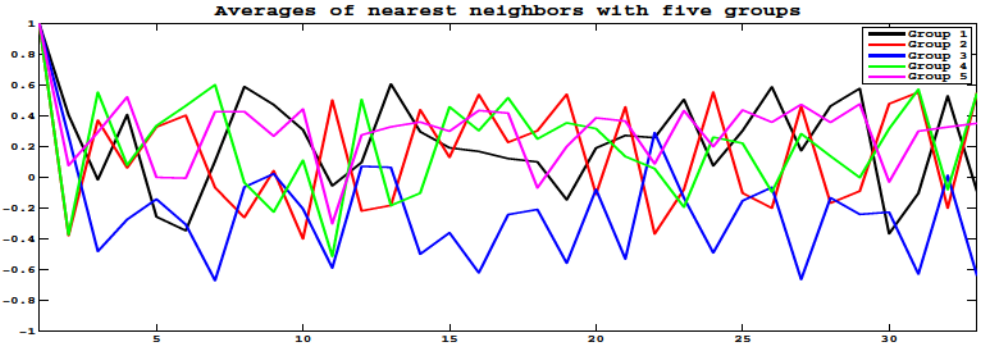

Figure 2.8: Average of each group.

Figure 2.8 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 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.

2.6 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 [53]) but their drawback is that these methods are based on the assumption of identically and independently distributed variables. Opgen-Rhein and Strimmer [42] studied the graphical Gaussian models from the perspective of functional data, where these two assumptions are not necessary.

Opgen-Rhein and Strimmer [42] 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 Chapter 1. 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 [13],[14]. 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 [42]. 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 2.5 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 2.4.

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 2.9 and 2.10 show the network proposed by Opgen-Rhein and Strimmer [42] 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).

Tables 2.5: 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

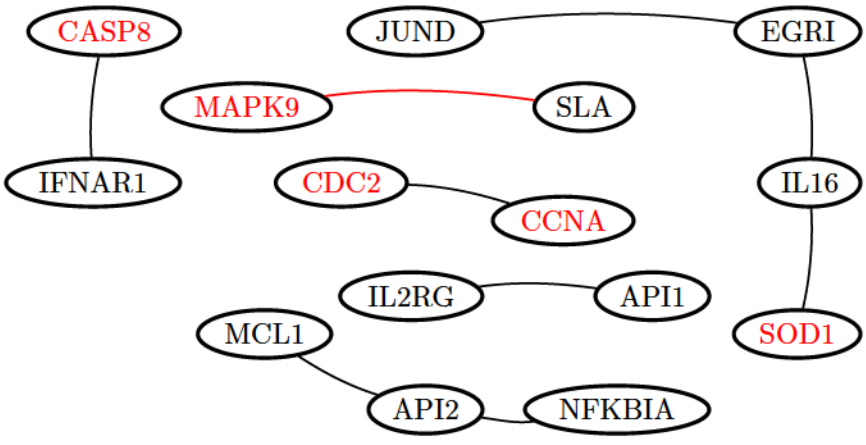

Figure 2.9: Gene dependence network using dynamical correlation.

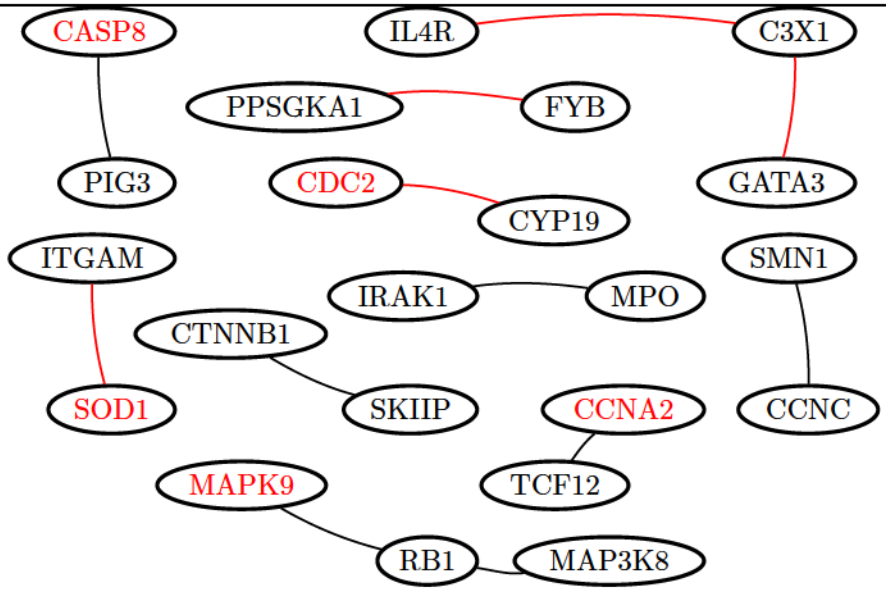

Figure 2.10: Gene dependence network using functional $\widehat{\tau}_2$.

The advantage of using functional $\widehat{\tau}_2$ instead of the dynamical correlation studied in Opgen-Rhein and Strimmer [42] 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 2.6 and 2.7 show the results of partial correlation with dynamical correlation and partial correlation with functional $\widehat{\tau}_2$ respectively, which were found through the "local fdr" algorithm. Also, we can see the *p-value* for each of the correlations as well as the nodes included in the networks.

Tables 2.6: 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.

Tables 2.7: 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

2.7 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. [21]: 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.

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 2.8, 2.9 and 2.10, 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.

Tables 2.8: Contamination with shape outliers

Contaminated Groups	Type of Outliers	Nº 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**

Tables 2.9: Contamination with magnitude outliers

Contaminated Groups	Type of Outliers	Nº 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

Tables 2.10: Contamination with shape-magnitude outliers

Contaminated Groups	Type of Outliers	Nº 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

2.8 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 the dynamical correlation and the canonical correlation. The coefficients presented here allow us to identify the global dependence 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.

CHAPTER 3

Spearman dependence coefficient for functions

3.1 Introduction

In this Chapter, we focus on another numeric dependence measure, the Spearman coefficient. The first contribution will be the definition of a Spearman coefficient for functional data that extends the classical bivariate concept, based on the ranks of the observations of the sample. Therefore, our first task is to consider a suitable way to sort the observations depending on the relative position of the curve within the sample. There are some alternatives for sorting the curves; one of them is based on the notion of depth that measures the centrality of a curve with respect to the group to which it belongs, so depth provides a way of ordering data from the center outwards. Different notions of depth have been studied for functional data (see for example, Fraiman and Muniz [24], Cuevas et al. [8], López-Pintado and Romo [38],) and each definition gives rise to different ways of ordering the curves. However, alternative definitions of ordering can also be interesting; for example, in Chapter 2 the functions are compared depending on their maximum values or on their total area below the curves. In this chapter, we have used the pre-order introduced in López-Pintado and Romo [39] and the way of sorting the functions used in Martín-Barragan et al. [40], who provided a way of sorting the data in a down-up direction based on the concepts of hypograph and epigraph of a function. This pre-order takes into consideration more the particular structure of the data. We also introduce the notion of grade for functions that it is useful to develop the theoretical background necessary to properly define the Spearman coefficient. The main properties of this coefficient as a well-defined dependence measure are also derived. To our knowledge, an independence test for functional data has not been proposed in the literature. Here, we try to fill this gap and present an independence test based on a bootstrap methodology suitable to be applied with some of the numeric dependence coefficients previously introduced in the

literature.

This Chapter is organized as follows. In Section 3.2, we recall some concepts about Spearman's coefficient for bivariate samples necessary to understand the extension to the functional context. Section 3.3 presents the main definitions that allow functions to be sorted. In Section 3.4, we introduce Spearman's coefficient for functions and study its properties. A simulation study and a robustness analysis is carried out in Section 3.5. In Section 3.6 the independence test is provided as well as a simulation study. Several examples with real data are shown in Section 3.7. Finally, Section 3.8 gathers the main conclusions.

3.2 Preliminaries

Spearman's coefficient is a non-parametric measure of association between two random variables. It is defined as the Pearson correlation coefficient between the ranks of the sample, being useful when the data are distribution free, so it is not necessary to assume the assumption of normality (Pearson [43], Hauke and Kossowski [26]). It is well known that it presents significant advantages over the Pearson coefficient: (1) It is a more robust coefficient (less sensitive to outliers) and (2) Spearman's coefficient is a better indicator than the Pearson correlation for determining whether a relationship exists between two variables when the relationship is nonlinear.

One of the definitions of the Spearman coefficient between two random variables is given by Definition 1.1.3 in the Chapter 1. Therefore, Spearman's coefficient is proportional to the difference between the probability of concordance and the probability of discordance for two vectors (X_1, Y_1) and (X_2, Y_3). The Kendall τ is also based on the concordance probability and it is well known that both coefficients measure non linear dependence from a non-parametric point of view. (For further details see Nelsen [41]).

However, we are interested in the equivalent definition of ρ_s given by calculating the Pearson coefficient between the uniform random variables $U = F_X(X)$ and $V = F_Y(Y)$; that is,

$$\rho_s = \rho_p[U, V] = \frac{E(UV) - E(U)E(V)}{\sqrt{Var(U)}\sqrt{Var(V)}}, \tag{3.2.1}$$

where ρ_p denotes the Pearson coefficient. The random variables U and V are called the "grades" of X and Y and the realizations u of U and v of V can be obtained evaluating the realizations x of X and y of Y in the distribution functions F_X and F_Y, respectively. Therefore, $u = F_X(x)$ and $v = F_Y(y)$ can also be called the grades of x and y. These grades can be seen as the population definition analogs of ranks (see Nelsen [41], page 169). If the distribution functions are unknown, then the grades of x and y can be estimated through the empirical distribution, i.e., $\hat{u} = \hat{F}_X(x)$ similar to $\hat{v}$ and hence we can calculate the sample version of this coefficient by calculating the sample version of the Pearson coefficient between the estimated

grades. For this reason, Spearman's coefficient is also called *the grade correlation coefficient.* Observe that the grades are values that are always in $[0, 1]$ and they are bounded independently of the support of the random variables. Therefore, an estimation of the Spearman coefficient is less sensitive in the presence of outliers than an estimation of the Pearson coefficient and, most importantly, ρ_s is well defined for all pairs of random variables, whereas ρ_p needs the random variables to have a finite second moment.

The definition of ρ_s based on grades inspires the development provided in this chapter: defining a Spearman coefficient for functions extending the definition of grades for functions. This is done in the following section.

Spearman's coefficient satisfies some general and intuitive properties required for any reasonable dependence measure. For example, the sign of ρ_s indicates the direction of association between X and Y, so that if Y increases when X increases, Spearman's coefficient will be positive. Now, if Y tends to decrease when X increases, Spearman's coefficient is negative. A Spearman's coefficient with value zero indicates that there is not a clear tendency for Y to either increase or decrease when X increases and its value is zero if the variables are independent. Spearman's coefficient increases in magnitude as X and Y become closer to being perfect monotone functions of each other. When X and Y are perfectly monotonically related (positive perfect dependence), Spearman's coefficient becomes 1. Therefore, Spearman's coefficient informs about the dependence, either positive or negative, between the random variables.

3.3 Grades for functional data

The possible concept of grade for functions may be linked to the relative position of a curve in the sample which implicitly implies defining an ordering among functions. There are some alternatives to sorting curves, we analyze some of them in Chapter 1. Recall that some of the most used are based on the notion of depth that measures the centrality of a curve with respect to the group to which it belongs; thus, depth provides a way of ordering data from center outwards. Different notions of depth have been studied for functional data (see for example, Fraiman and Muniz [24], Cuevas et al.[8], López-Pintado and Romo [38],) and each definition leads to different ways of ordering the curves. However, alternative definitions of ordering can also be interesting; for example, in Chapter 2 of this thesis the curves are ordered depending respectively on values of their maximum or their area below the curves in order to define a Kendall tau coefficient for functions. Martín-Barragan et al. [40] apply the concept of epigraphs and hypographs of a function to define some indexes that are useful for sorting curves in a down-up direction, even when the curves cross.

To define the grades of the curves, we will follow some concepts introduced in López-Pintado and Romo [39] and the idea of ordering implemented in Martín-Barragan et al.[40].

In López-Pintado and Romo [39], two concepts called the *Inferior Length* and the *Superior Length* of a curve, are defined as the foundation of a depth definition and these concepts are used to introduce a new boxplot for functional data in Martín-Barragan et al. [40]. In order to make the chapter self contained, we briefly define the previous concepts.

Let $C(I)$ be the space of the continuous functions defined in a compact interval I. Consider a stochastic process $X(t)$ with distribution P and whose sample paths are in $C(I)$. Let $x_1(t), \ldots, x_n(t)$ be a sample of curves from P. The graph of a function x is the subset of the plane $G(x) = \{(t, x(t)), t \in I\}$. The hypograph, written as *hyp*, and the epygraph, written as *epi*, of a function x in $C(I)$ are given respectively by

$$hyp(x) = \{(t, y) \in I \times \mathbb{R} : y \leq x(t)\},$$
$$epi(x) = \{(t, y) \in I \times \mathbb{R} : y \geq x(t)\}.$$

A natural form of ordering curves is pointwise, which means that a curve x is greater than another curve y if, and only if, $hyp(y) \subset hyp(x)$ or $epi(x) \subset epi(y)$, for all $t \in I$. However, in practical situations the curves in a sample can be crossed and hence the natural ordering in these cases does not work. An alternative for ordering curves can be developed by using two concepts, the Inferior Length and the Superior Length of a curve with respect to a stochastic process $X(t)$:

$$IL(x) = \frac{1}{\lambda(I)} E[\lambda\{t \in I : x(t) \geq X(t)\}],$$
$$SL(x) = \frac{1}{\lambda(I)} E[\lambda\{t \in I : x(t) \leq X(t)\}],$$

where λ stands for the Lebesgue measure on $\mathbb{R}$. The inferior length $IL(x)$ can be interpreted as the "proportion of time" that the stochastic process $X(t)$ is smaller than x and the superior length $SL(x)$ is the "proportion of time" that the stochastic process $X(t)$ is greater than x.

These notions are behind the definitions of the grades of a stochastic process $X(t)$ with respect to another process $\widetilde{X}(t)$, which we define as follows:

Definition 3.3.1 *Let $X(t)$ and $\widetilde{X}(t)$ be two stochastic processes. Then,*

$$IL\text{-}grade(X(t))_{\widetilde{X}(t)} = \frac{1}{\lambda(I)} E_{\widetilde{X}(t)}[\lambda\{t \in I : X(t) \geq \widetilde{X}(t)\}],$$
$$SL\text{-}grade(X(t))_{\widetilde{X}(t)} = \frac{1}{\lambda(I)} E_{\widetilde{X}(t)}[\lambda\{t \in I : X(t) \leq \widetilde{X}(t)\}].$$

Observe that *IL-grade* or *SL-grade* assigns a value between [0,1] to each process. We note that if the $X(t)$ and $\widetilde{X}(t)$ have the same distribution, we then eliminate $\widetilde{X}(t)$ from the definitions of *IL-grade* and *SL-grade* to avoid hard notation.

If we consider a sample of functional data, $x_1(t), \ldots, x_n(t)$ and fix any curve $x = x(t)$ of the data set, the sample version of both *IL-grade* and *SL-grade* can be easily obtained by substituting the expectation by the sample mean, respectively

$$IL_n\text{-}grade(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \geq x_i(t)\},$$
$$SL_n\text{-}grade(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \leq x_i(t)\}.$$

It is noteworthy that IL_n-*grade*(x) or SL_n-*grade*(x) has been viewed as the relative position of a curve with respect to the sample. Also, note that the curves can be ordered by sorting the values of IL_n-*grade* or SL_n-*grade* for each one of them. That is,

Definition 3.3.2 *Consider functional observations $x_1(t), \ldots, x_n(t)$ of a stochastic process $X(t)$. Then,*

$$x_i(t) \preceq x_j(t) \equiv IL_n\text{-}grade(x_i) \leq IL_n\text{-}grade(x_j).$$

A similar definition can be obtained by replacing the IL_n-*grade* with SL_n-*grade*.

The relation given in Definition 3.3.2 meets important properties such as reflectivity and transitivity, but, unfortunately, it does not satisfy the antisymmetry property. Therefore, the relation is a pre-order, which is less restrictive than a partial order and allows us to compare any pair of functions in the sample. Observe that if the curves do not cross each other, Definition 3.3.2 corresponds to the pointwise order.

To illustrate this pre-order, observe the example in Figure 3.1 that shows the IL_n-*grade* assigned to each function in a sample of four functions. The blue curve has the smallest IL_n-*grade* because the proportion of time that it is above any other curve is smaller than the value assigned to any curve in the same sample. The black curve has the largest IL_n-*grade* value assigned, since in this case the time proportion is greater than any other. The proportions assigned to each curve are what we call the grade of the curve regarding the sample. Note that the largest functional grade in the sample may not be one unless the curve with the highest functional grade does not cross with any other, which means that it will be largest point-to-point than them. Once the grades are introduced, we can define Spearman's coefficient for functions in a parallel way to (3.2.1).

3.4 Spearman's coefficient for functional data

In this section, we define the concept of Spearman's coefficient in the functional context in order to quantify the dependence in a bivariate data set of functions. Taking into account Definition (3.2.1), we define a Spearman coefficient for two stochastic processes as the Pearson coefficient between the random variables *IL-grade*$(X(t))$ and *IL-grade*$(Y(t))$; that is,

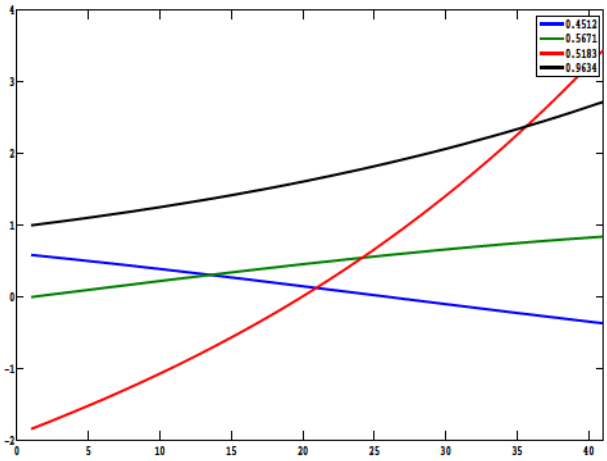

Figure 3.1: Grades for functions.

Definition 3.4.1 (Spearman coefficient for stochastic processes.) *Let $(X(t), Y(t))$ be bivariate stochastic process whose paths are continuous functions on an interval $I \subset \mathbb{R}$. Then, Spearman's coefficient of $(X(t), Y(t))$ is:*

$$\rho_s(X(t), Y(t)) \equiv \rho_p(IL\text{-}grade(X(t)), IL\text{-}grade(Y(t))), \tag{3.4.1}$$

where ρ_p denotes the Pearson correlation coefficient and IL-$grade(\cdot)$ is the grade associated to a stochastic process given in Definition 3.3.1.

In the same way, the sample version of ρ_s is the following:

Definition 3.4.2 (Spearman's coefficient for functions.) *Let*

$$(\mathbf{x}, \mathbf{y}) = \{(x_1(t), y_1(t)), \ldots, (x_n(t), y_n(t))\}$$

be a bivariate functional sample from $(X(t), Y(t))$. Then, the Spearman coefficient related to the data set and denoted by $\widehat{\rho}_s$ is defined by

$$\widehat{\rho}_s \equiv \widehat{\rho}_p(IL_n\text{-}grade(\mathbf{x}), IL_n\text{-}grade(\mathbf{y})), \tag{3.4.2}$$

where,

$$IL_n\text{-}grade(\mathbf{x}) = \{IL_n\text{-}grade(x_1), IL_n\text{-}grade(x_2), \ldots, IL_n\text{-}grade(x_n)\}$$
$$IL_n\text{-}grade(\mathbf{y}) = \{IL_n\text{-}grade(y_1), IL_n\text{-}grade(y_2), \ldots, IL_n\text{-}grade(y_n)\}.$$

Another definition of Spearman's coefficient for functions can be obtained by replacing IL_n-$grade$ by SL_n-$grade$. In order to illustrate how the Spearman coefficient works, we have taken a small bivariate set of four curves and calculated the corresponding coefficient. Figure 3.2 shows the pairs of curves, each pair represented by its own color. We can see that

the curves in a group are organized in a different way than their respective partner in the other group. Observe that the order of the curves in first group seems to have a more or less opposite direction with respect to the other group. Therefore, Spearman's coefficient for functional data is small, indicating to us that the association between the groups of curves is weak and negative.

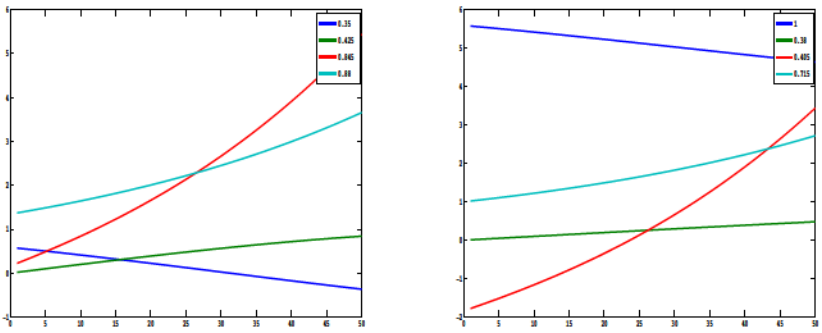

Figure 3.2: Spearman's coefficient for functional data, $\widehat{\rho}_s = -0.2994$.

3.4.1 Properties of Spearman's coefficient for functional data

As commented in Section 3.2, Spearman's coefficient for bivariate data satisfies certain desirable properties required for a dependence measure (see Xu et al. [55]). In this section, we prove that Spearman's coefficient for stochastic processes also possesses such properties. Let $(X(t), Y(t))$ be a bivariate stochastic process and ρ_s be Spearman's coefficient as in Definition 3.4.1. Then ρ_s satisfies the following properties:

1. $\rho_s(X(t), Y(t)) = \rho_s(Y(t), X(t))$. (Symmetry).
2. $-1 \leq \rho_s(X(t), Y(t)) \leq 1$.
3. $\rho_s(X(t), g(X(t))) = 1$, for any monotone increasing function g.
4. $\rho_s(X(t), g(X(t))) = -1$, for any monotone decreasing function g.
5. Spearman's coefficient for functions is invariant under strictly increasing transformations of the functional variables; that is,

   $$\rho_s(\alpha(X(t)), \beta(Y(t))) = \rho_s(X(t), Y(t)).$$

   For any α and β being strictly increasing functions.
6. If $X(t)$ and $Y(t)$ are stochastically independent then $\rho_s(X(t), Y(t)) = 0$.

7. The sample Spearman's coefficient is a consistent estimator of the population coefficient.

The proofs of properties 1 and 2 are trivial from the definition of ρ_s. The proof of properties 3, 4 and 5 are based on the following:

$$\begin{aligned}\text{IL-grade}(g(X(t)))_{\widetilde{X}(t)} &= \frac{1}{\lambda(I)} E_{\widetilde{X}(t)}[\lambda\{t \in I : g(X(t)) \geq g(\widetilde{X}(t))\}] \\ &= \frac{1}{\lambda(I)} E_{\widetilde{X}(t)}[\lambda\{t \in I : X(t) \geq \widetilde{X}(t)\}] \\ &= \text{IL-grade}(X(t))_{\widetilde{X}(t)},\end{aligned}$$

for any monotone increasing function g. The proof of property 6 is based on that, if $X(t)$ and $Y(t)$ are independent then $IL\text{-}grade(X(t))$ and $IL\text{-}grade(Y(t))$ are also independent. Therefore, $\rho_s(X(t), Y(t)) = \rho_p(\text{IL-grade}(X(t)), \text{IL-grade}(Y(t))) = 0$ by the well known property of the Pearson coefficient. The last property holds since, as n goes to infinity,

$$\frac{\sum_{i=1}^{n} IL_n\text{-}grade(x_i)}{n} \xrightarrow{\text{a.s}} E[IL\text{-}grade(X(t))],$$

where $x_1, \ldots, x_n$ is a sample from $X(t)$. Finally, since $\widehat{\rho}_p$ is a consistent estimator, also $\widehat{\rho}_s$ is.

3.5 Simulation study

In this section we show how Spearman's coefficient works in several simulated data sets and we establish comparisons with other dependence measures introduced previously in the literature. Specifically, we consider the canonical correlation, the dynamical correlation, Pearson's coefficient for functional data studied in Chapter 1, Section 1.2 and Kendall's τ for functions defined in Chapter 2 of this thesis.

To illustrate the different dependence measures, we have calculated them for the data given in Figure 3.2.

$$\widehat{\tau}_1 = 0, \quad \widehat{\tau}_2 = -0.33, \quad \widehat{\rho}_c = 0.83, \quad \widehat{\rho}_d = -0.13, \quad \widehat{\rho}_p = -0.2374$$

Note that Kendall's tau built with the pre-order of maximum and denoted as $\widehat{\tau}_1$ is zero since there are as many concordant pairs as discordant pairs. The canonical correlation $\widehat{\rho}_c$ has a very large and positive value since it is always positive and does not allow the direction of the dependency to be identified. The dynamical correlation $\widehat{\rho}_d$, Kendall's tau built with the pre-order of the integral $\widehat{\tau}_2$ and Pearson's correlation coefficient for functional data $\widehat{\rho}_p$ have negative values that reflect the direction of weak dependence shown in the data set as well as Spearman's coefficient ($\widehat{\rho}_s = -0.2994$).

We have simulated 50 realizations from different processes $X(t) = f_1(t, Z_1)$ and $Y(t) = f_2(t, Z_2)$, where (Z_1, Z_2) represents the random part of the processes, which was defined in

Chapter 2, and we have taken $d = 50$ points to discretize the functions. For each pair (f_1, f_2), we use a different correlation σ_{12}.

Table 3.1 shows the sample means of different association measures for the simulated samples with $n = d = 50$ and 100 replications. We have also included the standard deviation (between parenthesis). We can see that both coefficients, the Spearman and Kendall, properly reflect the cases where the pairs of functions present perfect co-monotonicity or counter-monotonicity, (see rows 3, 4 and 5 in Table 3.1). As we know, the canonical correlation is always positive, i.e., it does not capture the direction of the dependence. Note from the definition of the dynamical correlation that, it just reflects individual changes between the pairs of functions rather than among groups. On the other hand, Pearson's coefficient does not work well when the dependence relations are not lineal, as in cases 4 and 5.

We have also analyze the sensitivity of $\widehat{\rho}_s$ with respect to the size n. We will use the following two pairs of stochastic processes that correspond with row 1 in Table 3.1 with $\sigma_{12} = 0.8$ and $\sigma_{12} = 0.1$:

$$X(t) = (t + Z_1)^3 + (t + Z_1)^2 + 3(t + Z_1), \quad Y(t) = (t + Z_2)^2 + \frac{7}{8}(t + Z_2) - 10$$

We have considered $n = 25,\ 50,\ 100,\ 150$ and 1000 with $d = 50$. Table 3.2 shows that the changes in $\widehat{\rho}_s$ are negligible and it is stable with respect to the sample size. Table 3.3 illustrates the sensitivity with respect to d. Now, fix $n = 50$, and move $d = 25,\ 50,\ 100,\ 150$ and 1000 points. It is noteworthy that the coefficients present good stability with respect to the number of points taken to discretize the functions. We point out that we have made the sensitivity analysis with other models, but the conclusions are the same for the models reported.

3.5.1 Robustness

Spearman's coefficient is a more appropriate association measure than Pearson's correlation when the data are ordinal or non-normally distributed or a tiny fraction of outliers exists. In this section, we analyze this last point. That is, we check if Spearman's coefficient for functions fulfills the robustness property by contaminating a sample with the three different types of outliers commonly used in the functional context: shape outliers, magnitude outliers and shape-magnitude outliers. The method to contaminate data is the same implemented in Chapter 2 where the objective was to show the robustness of Kendall's τ for functions. We have simulated fifty paths of the stochastic processes,

$$X(t) = \exp(t + Z_1), \ \ Y(t) = (t + Z_2)^3 + (t + Z_2)^2 + 3(t + Z_2), \ \sigma_{12} = 0.6, \tag{3.5.1}$$

and the types of outliers to be considered are:

- Shape outliers.

Tables 3.1: Dependence measures in simulated data

	$X(t) = f_1(t, Z_1)$	$Y(t) = f_2(t, Z_2)$	σ_{12}	$\bar{\bar{\rho}}_s IL$	$\bar{\bar{\rho}}_s SL$	$\bar{\bar{\tau}}_1$	$\bar{\bar{\tau}}_2$	$\bar{\bar{\rho}}_c$	$\bar{\bar{\rho}}_d$	$\bar{\bar{\rho}}_p$
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	0.667 (0.0811)	0.6596 (0.0882)	0.4861 (0.0657)	0.4874 (0.0711)	0.7448 (0.0898)	0.7098 (0.1139)	0.6943 (0.1055)
2	$\sin(t+Z_1)$	$\cos(t+Z_2)$	-0.7	0.4354 (0.1244)	0.445 (0.1407)	0.3084 (0.0923)	0.2774 (0.0835)	0.5367 (0.1004)	0.3605 (0.11)	0.4022 (0.1189)
3	$(t+Z_1)^2$	$(t+Z_1)^4$	1	1 (0)	1 (0)	1 (0)	1 (0)	0.9566 (0.0118)	0.922 (0.0125)	0.9179 (0.0127)
4	$(t+Z_1)^2 + 7(t+Z_1) + 2$	$((t+Z_2)^2 + 7(t+Z_2) + 2)^3$	1	0.9997 (0.0029)	1 (0)	1 (0)	1 (0)	0.9989 (0)	0.7779 (0.0347)	0.7688 (0.0278)
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)	-1 (0)	-1 (0)	0.999 (0.0009)	-0.78 (0.0275)	-0.7644 (0.0285)
6	$\exp(t+Z_1)$	$(t+Z_2)^3 + (t+Z_2)^2 + 3(t+Z_2)$	0.6	0.5802 (0.0967)	0.5546 (0.1072)	0.4047 (0.0811)	0.4138 (0.0751)	0.5098 (0.1431)	0.5682 (0.1301)	0.5193 (0.1559)
7	$\exp(t+Z_1)^2$	$\cos(t+Z_2)$	-0.8	0.4417 (0.1195)	0.4430 (0.1198)	0.3097 (0.0922)	0.2982 (0.1035)	0.3101 (0.07)	0.0408 (0.1458)	0.0846 (0.1697)
8	$\sin(t+Z_1)$	$(t+Z_2)^2$	0.4	0.1706 (0.1331)	0.1458 (0.1307)	0.1080 (0.1035)	0.1059 (0.1021)	0.3382 (0.1132)	0.1647 (0.0916)	0.1173 (0.1175)
9	$(t+Z_1)^2 + 9(t+Z_1) - 5$	$\cos(3t+Z_2)$	1	-0.935 (0.0176)	-0.9327 (0.0199)	-0.7198 (0.0853)	-0.9476 (0.0358)	0.9334 (0.0458)	-0.7244 (0.0562)	-0.6976 (0.0708)
10	$\exp(t^2+Z_1)$	$(t+Z_2)^2 - 8t + Z_2$	0.9	0.7743 (0.0634)	0.7892 (0.0608)	0.3621 (0.1078)	0.5991 (0.0706)	0.8544 (0.0485)	0.4620 (0.1215)	0.8309 (0.0616)
11	$\exp(t+Z_1)$	$\sin(t+Z_2)$	0	0.05 (0.1467)	0.0051 (0.1508)	-0.0076 (0.1004)	0.0087 (0.0883)	0.1438 (0.0861)	0.0560 (0.1275)	-0.0209 (0.1221)

Tables 3.2: Sensitivity to sample size

sample size	Model 1 $\bar{\hat{\rho}}_s IL$	Model 1 $\bar{\hat{\rho}}_s SL$	Model 2 $\bar{\hat{\rho}}_s IL$	Model 2 $\bar{\hat{\rho}}_s SL$
25	0.6492 (0.1270)	0.6612 (0.1301)	0.077 (0.2030)	0.0781 (0.2137)
50	0.6697 (0.0881)	0.6748 (0.0686)	0.0732 (0.1426)	0.0993 (0.1369)
100	0.6709 (0.0559)	0.6534 (0.0617)	0.0883 (0.0945)	0.0754 (0.0998)
150	0.6598 (0.0448)	0.6668 (0.0495)	0.0626 (0.0847)	0.0685 (0.0789)
1000	0.6699 (0.0177)	0.6724 (0.0204)	0.0767 (0.0341)	0.0807 (0.0348)

Tables 3.3: Sensitivity to the number of points in the discretization

numbers of points	Model 1 $\bar{\hat{\rho}}_s IL$	Model 1 $\bar{\hat{\rho}}_s SL$	Model 2 $\bar{\hat{\rho}}_s IL$	Model 2 $\bar{\hat{\rho}}_s SL$
25	0.6542	0.6542	0.0647	0.0647
50	0.6542	0.6542	0.0648	0.0648
100	0.6546	0.6546	0.0648	0.0648
150	0.6548	0.6548	0.0646	0.0646
1000	0.6548	0.6548	0.0648	0.0648

- Magnitude outliers, with $k = 60$.
- Shape-magnitude outliers.

Figure 3.3 shows a data set generated from stochastic process $X(t) = \exp(t + Z_1)$ and the same data set but contaminated with different types of outliers, which is represented with a black curve.

Contaminated data are considered in processes (3.5.1), but introducing outliers in the following way:

1. Contaminating just the group of curves that comes from $X(t)$.
2. Contaminating both groups of curves $(X(t), Y(t))$ in the same position.
3. Contaminating both groups of curves that come from $X(t)$ and $Y(t)$ but in different positions.

Table 3.4 shows the variation of the coefficients when the outliers are introduced. Each measure is calculated before contaminating the data (row 1). Once the data are contaminated, we report the relative variation of the association measure with respect to its value in the uncontaminated data set. We can see that Kendall's τ is the most robust coefficient in most cases. However, Spearman's coefficient also exhibits a good degree of robustness, even being

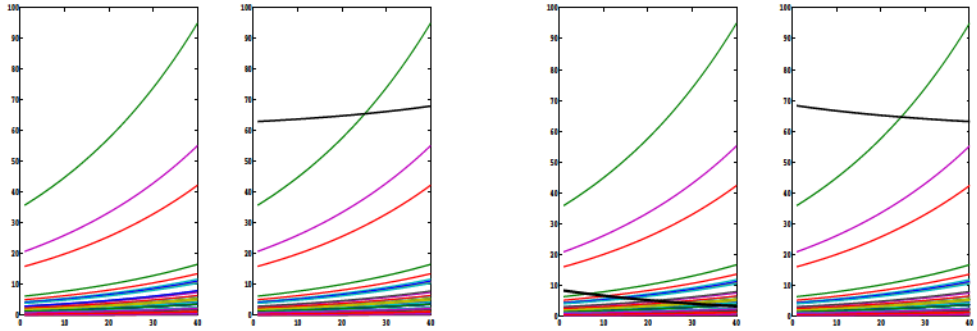

Figure 3.3: Original data, a magnitude outlier, a shape outlier, a shape-magnitude outlier.

Tables 3.4: Variation of the coefficients in presence of a different number of outliers

Contaminated Groups	Type of Outliers	Nº outlyers	$\widehat{\rho}_{IL}$	$\widehat{\rho}_{SL}$	$\widehat{\tau}_1$	$\widehat{\tau}_2$	$\widehat{\rho}_d$	$\widehat{\rho}_c$	$\widehat{\rho}_p$
none	none	0	0.6213	0.6213	0.4547	0.4547	0.5491	0.5449	0.5367
$X(t)$	Shape	1	0.0067	0.0067	0	0	0.00036	0.0027	0.0007
$X(t)$	Shape	2	0.0069	0.0069	0	0	0.042	0.1213	0.0007
$X(t)$, $Y(t)$ same position	Shape	1	0.010	0.010	0	0	0	0.006	0.0015
$X(t)$, $Y(t)$ same position	Shape	2	0.0094	0.0094	0	0	0	0.7511	0.0018
$X(t)$, $Y(t)$ different position	Shape	1	0.0086	0.0086	0	0	0.0009	0.0011	0.00037
$X(t)$, $Y(t)$ different position	Shape	2	0.0072	0.0072	0	0	0.046	0.1477	0.0005
$X(t)$	Magnitude	1	0.045	0.045	0.035	0.039	0.28	0	0.313
$X(t)$	Magnitude	2	0.039	0.039	0.025	0.028	0.066	0.035	0.5446
$X(t)$, $Y(t)$ same position	Magnitude	1	0.053	0.053	0.0646	0.075	0.227	0.6505	0.2457
$X(t)$, $Y(t)$ same position	Magnitude	2	0.055	0.055	0.078	0.086	0.47	0.7414	0.3547
$X(t)$, $Y(t)$ different position	Magnitude	1	0.074	0.074	0.072	0.082	0.418	0.008	0.436
$X(t)$, $Y(t)$ different position	Magnitude	2	0.079	0.079	0.075	0.086	0.383	0.017	0.7315
$X(t)$	Shape-magnitude	1	0.045	0.045	0.035	0.039	0.2811	0.001	0.312
$X(t)$	Shape-magnitude	2	0.039	0.039	0.025	0.028	0.092	0.043	0.5438
$X(t)$, $Y(t)$ same position	Shape-magnitude	1	0.053	0.053	0.064	0.075	0.227	0.689	0.2467
$X(t)$, $Y(t)$ same position	Shape-magnitude	2	0.055	0.055	0.086	0.086	0.4775	0.7973	0.3551
$X(t)$, $Y(t)$ different position	Shape-magnitude	1	0.074	0.074	0.072	0.082	0.419	0.0014	0.4373
$X(t)$, $Y(t)$ different position	Shape-magnitude	2	0.079	0.079	0.075	0.086	0.404	0.034	0.730

more robust in general than the canonical correlation, dynamical correlation and the Pearson correlation coefficient for functions. We highlight that the robustness analysis has been made with other models $(X(t), Y(t))$ and the same conclusions can be drawn.

3.6 Independence test for functional data

In the literature on association measures, it is usual to provide an independence test to check if the corresponding coefficient used to measure dependence can be considered zero or not (see for example Gibbons [25] and Wilcox [54] for more details). This section deals with the design of a test when data are curves and the hypotheses are:

$$H_0 : \rho_s = 0.$$
$$H_1 : \rho_s \neq 0.$$

Since the asymptotic distribution for ρ_s is not known when the data set are functions, an alternative methodology is necessary to find the critical region associated with the statistics $\widehat{\rho}_s$. We will use a bootstrap approach to estimate the statistics distribution, (see Efron [12], Efron and Tibshirani [15], Davison and Hinkley[9], for more information).

Given a sample of functions $(\mathbf{x}, \mathbf{y})$ of size n, B bootstrap samples of size n are obtained by resampling from $(\mathbf{x}, \mathbf{y})$ under the null hypothesis; that is, there is no association between the components of the stochastic process $(X(t), Y(t))$ that generated the data set $(\mathbf{x}, \mathbf{y})$. The steps necessary to obtain the *p-value* of the test are summarized in Table 3.5, where $\widehat{\rho}_s(\mathbf{x}, \mathbf{y})$ is the sampled value of ρ_s and $\widehat{\rho}_s(\mathbf{x}^*, \mathbf{y}^*)$ is its corresponding value for the bootstrap sample. The decision rule is to reject H_0 if *p-value* $\leq \alpha$, where α is the significance level. We fix $\alpha = 0.05$ in the following.

Tables 3.5: Bootstrap test

1. Input: a sample of functions $(\mathbf{x}, \mathbf{y})$ from a stochastic process (X, Y) and α-level.
2. Find $\widehat{\rho}_s(\mathbf{x}, \mathbf{y})$.
3. Obtain under H_0 a bootstrap sample $(\mathbf{x}^*, \mathbf{y}^*)$ of size n from $(\mathbf{x}, \mathbf{y})$.
4. Calculate $\widehat{\rho}_s(\mathbf{x}^*, \mathbf{y}^*)$.
5. Repeat 3 and 4 a sufficient number of times (B).
6. Find *p-value* $= \frac{\sum_{i=1}^{B} \mathbf{I}[\widehat{\rho}_s(\mathbf{x}_i^*, \mathbf{y}_i^*) \geq \widehat{\rho}_s(\mathbf{x}, \mathbf{y})]}{B}$.
7. Output: Reject H_0, if *p-value* $<$ *α-level*.

To illustrate the results of the bootstrap test, we come back with the simulated data of in Table 3.1. We fix a sample of size $n = 50$ and apply the previous test with $B = 2500$. For each case, both $\widehat{\rho}_s IL$ and the *p-value* are displayed in Table 3.6. Note how the test is consistent when the simulated models are curves generated from stochastic processes with positive or negative perfect dependence. In these cases, the test produces *p-values* equal to zero. We can also observe that when the groups of curves have a high correlation coefficient the *p-value* is smaller than 0.05 so that the null hypothesis is rejected. Likewise, when the groups of curves have a low correlation coefficient, the *p-value* is larger than 0.05 and then the null hypothesis is not rejected.

Tables 3.6: Hypothesis test

$X(t)$	$Y(t)$	σ_{12}	$\widehat{\rho}_s IL$	*p-value*	$\widehat{\tau}_1$	*p-value*	$\widehat{\tau}_2$	*p-value*
$(t+Z_1)^2$	$(t+Z_1)^4$	1	1	0	1	0	1	0
$(t+Z_1)^2+7(t+Z_1)+2$	$((t+Z_1)^2+7(t+Z_1)+2)^3$	1	0.9996	0	0.9967	0	0.9967	0
$(t+Z_1)^3+(t+Z_1)^2+3(t+Z_1)$	$(t+Z_2)^2+\frac{7}{8}(t+Z_2)-10$	0.8	0.7197	0	0.4645	0	0.4645	0
$\exp(t+Z_1)$	$(t+Z_2)^3+(t+Z_2)^2+3(t+Z_2)$	0.6	0.6213	0	0.4547	0	0.4547	0
$\sin(t+Z_1)$	$\cos(t+Z_2)$	-0.7	0.4840	0.0002	0.3763	0.0002	0.3127	0.0012
$\sin(t+Z_1)$	$(t+Z_2)^2$	0.4	0.3178	0.0241	0.2212	0.0230	0.2180	0.0244
$\cos(t+Z_1)$	$(t+Z_2)^2-9(t+Z_2)$	0.2	0.0583	0.6813	0.0351	0.7244	0.0351	0.7122
$\exp(t+Z_1)^2$	$5(t-Z_2)^3-3(t+Z_2)+9$	-0.2	0.0442	0.7587	0.0155	0.8812	0.0155	0.8826
$(t+Z_1)^3$	$(t+Z_2)^2+4(t+Z_2)-7$	-0.5	-0.6804	0	-0.4906	0	-0.4906	0
$(t+Z_1)^3+(t+Z_1)^2$	$(t+Z_2)^2-2(t+Z_2)$	-0.9	-0.8815	0	-0.5527	0	-0.7012	0
$(t+Z_1)^2+7(t+Z_1)+2$	$1-((t+Z_1)^2+7(t+Z_1)+2)^3$	1	-0.9938	0	-0.9837	0	-0.9755	0
$(5/9)(t+Z_1)^3$	$48-(5/9)(t+Z_1)^3$	0	-1	0	-1	0	-1	0

In order to make comparisons, Table 3.6 also shows the results of applying the same hypothesis test but considering the statistics $\widehat{\tau}_1$ and $\widehat{\tau}_2$, defined previously. The canonical correlation, dynamical correlation and Pearson correlation coefficient for functions are not considered because these coefficients show a very wide casuistry for which they equal zero. Hence, simulating bootstrap samples under the null hypothesis (independence) is not a good strategy for these coefficients where many anomalies are observed.

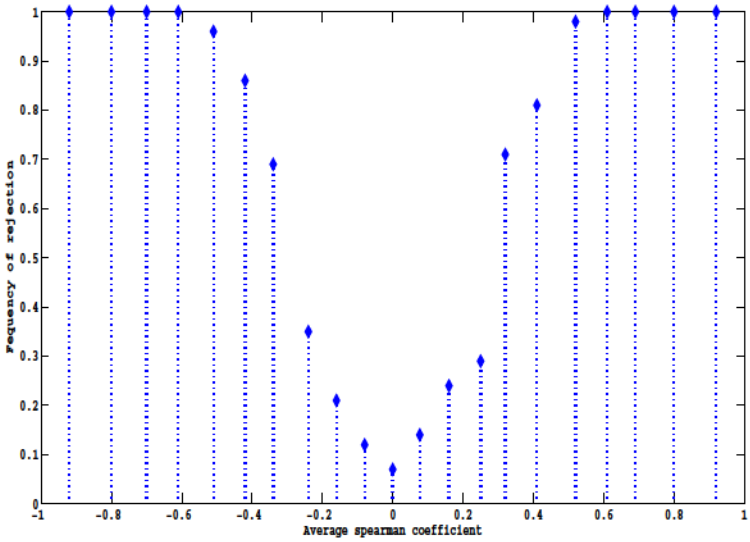

Figure 3.4: Power test.

Tables 3.7: Relationships between the coefficients, frequency of rejection (fr) and σ_{12}

σ_{12}	-1	-0,9	-0,8	-0,7	-0,6	-0,5	-0,4	-0,3	-0,2	-0,1	0
$\bar{\widehat{\rho}}_s$	-0,92	-0,8	-0,7	-0,61	-0,51	-0,42	-0,34	-0,24	-0,16	-0,08	0
fr	1	1	1	1	0,96	0,86	0,69	0,35	0,21	0,12	0,07
σ_{12}	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	
$\bar{\widehat{\rho}}_s$	0,077	0,16	0,25	0,32	0,41	0,52	0,61	0,69	0,8	0,92	
fr	0,14	0,24	0,29	0,71	0,81	0,98	1	1	1	1	

We now analyze the power of the test with a simulation study. First, we consider a bivariate sample of 50 curves generated from the process $[\exp(t + Z_1), \sin(t + Z_2)]$, being (Z_1, Z_2) a normal bivariate with zero mean and correlation σ_{12}. Given that there exists a certain relationship between σ_{12} and $\widehat{\rho}_s$, we consider different values of σ_{12}) in the interval $[-1, 1]$ in order to obtain values of $\widehat{\rho}_s$ over all the interval too. For a given σ_{12}, we generate 100 times a sample of $[\exp(t+Z_1), \sin(t+Z_2)]$, calculate $\widehat{\rho}_{si}$, $i = 1, \ldots, 100$ and its corresponding mean $\bar{\widehat{\rho}}_s$. Finally, we show in Figure 3.4 the frequency of rejection of the null hypothesis versus $\bar{\widehat{\rho}}_s$. The bootstrap part of each iteration is made with $B = 2500$. Table 3.7 shows the

relationship between the coefficient $\bar{\widehat{\rho}}_s$, σ_{12} and the frequency of rejection for the test. As we can see, as $|\bar{\widehat{\rho}}_s|$ increases, the frequency of rejection also increases which ensures the reliability of the test.

To end this section, we have carried out a sensitivity analysis of the test with respect to the bootstrap sample size B and the number of points used to discretize the functions. Table 3.8 shows the rejection frequency of the null hypothesis for different values of the $\bar{\widehat{\rho}}_s$. We can conclude that the size of bootstrap samples does not significatively affect the frequency of rejection, whereas the power test improves as d increases, which is due to more information being available about the original process.

Tables 3.8: Sensitivity analysis with respect to B and d

$\widehat{\rho}_s$	size of the bootstrap sample					number of points			
	500	1000	1500	2000	2500	25	50	100	150
0.5989	1	0.99	1	1	1	0.99	0.97	1	1
-0.2447	0.43	0.33	0.35	0.36	0.51	0.12	0.51	0.63	0.89
0.012	0.05	0.06	0.04	0.03	0.02	0	0.02	0.23	0.3
0.2516	0.47	0.5	0.38	0.42	0.36	0.13	0.36	0.69	0.73
0.6968	1	1	1	1	1	1	1	1	1
0.7969	1	1	1	1	1	1	1	1	1

Finally, In Table 3.9 we present the coefficients proposed in Chapter 2, and 3 for the simulated data of the Section 3.5. We also include the functional $\widehat{\tau}$ with the pre-order IL and SL which were used in the definition of the Spearman coefficient.

Note that the results obtained with functional $\widehat{\tau}$ for IL and SL are similar to those obtained with pre-orders of Definition 2.2.1. This indicates that the order of the curves are similar for the different pre-orders. We conclude that the choice of pre-order is an important issue to calculate the coefficients however, the results will be very close.

3.7 Application to real data sets

We consider three real data sets. The first one is composed of daily temperature and precipitation per year in 35 Canadian weather stations (see Ramsay and Silverman[45]). We also have the same data set by months. The sample size is 35. The objective in this first example is to measure the association between temperature and precipitation. The second data set corresponds to monthly temperatures in four cities of Canada from 1985 until 2004 (taken from the web page http://www.tutiempo.net/clima/Canada/CA.html). The data consist of 20 curves (one per year/city) with 12 observation points per curve where we are interested in analyzing the possible pattern of spatial correlation among cities in relation to their temperatures. Finally, the third data set is part of the original data from the web page http://www-stat.stanford.edu/ tibs/ElemStatLearn/. It consists of five groups of phonemes

Tables 3.9: Our dependence measures

	$X(t) = f_1(t, Z_1)$	$Y(t) = f_2(t, Z_2)$	σ_{12}	$\widehat{\bar{\rho}}_s IL$	$\bar{\hat{\rho}}_s SL$	$\bar{\hat{\tau}}_1$	$\bar{\hat{\tau}}_2$	$\bar{\hat{\tau}}_{IL}$	$\bar{\hat{\tau}}_{SL}$
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	0.667 (0.0811)	0.6596 (0.0882)	0.4861 (0.0657)	0.4874 (0.0711)	0.4844 (0.0699)	0.4825 (0.0714)
2	$\sin(t+Z_1)$	$\cos(t+Z_2)$	-0.7	0.4354 (0.1244)	0.445 (0.1407)	0.3084 (0.0923)	0.2774 (0.0835)	0.2906 (0.0774)	0.2941 (0.0776)
3	$(t+Z_1)^2$	$(t+Z_1)^4$	1	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
4	$(t+Z_1)^2 + 7(t+Z_1) + 2$	$((t+Z_2)^2 + 7(t+Z_2) + 2)^3$	1	0.9997 (0.0029)	1 (0)	1 (0)	1 (0)	1 (0)	0.9995 (0.0049)
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)	-1 (0)	-1 (0)	-1 (0)	-1 (0)
6	$\exp(t+Z_1)$	$(t+Z_2)^3 + (t+Z_2)^2 + 3(t+Z_2)$	0.6	0.5802 (0.0967)	0.5546 (0.1072)	0.4047 (0.0811)	0.4138 (0.0751)	0.4044 (0.0803)	0.4138 (0.0824)
7	$\exp(t+Z_1)^2$	$\cos(t+Z_2)$	-0.8	0.4417 (0.1195)	0.4430 (0.1198)	0.3097 (0.0922)	0.2982 (0.1035)	0.2924 (0.0860)	0.2963 (0.1057)
8	$\sin(t+Z_1)$	$(t+Z_2)^2$	0.4	0.1706 (0.1331)	0.1458 (0.1307)	0.1080 (0.1035)	0.1059 (0.1021)	0.0898 (0.0902)	0.0966 (0.0852)
9	$(t+Z_1)^2 + 9(t+Z_1) - 5$	$\cos(3t+Z_2)$	1	-0.935 (0.0176)	-0.9327 (0.0199)	-0.7198 (0.0853)	-0.9476 (0.0358)	-0.8097 (0.0365)	-0.8142 (0.0368)
10	$\exp(t^2+Z_1)$	$(t+Z_2)^2 - 8t + Z_2$	0.9	0.7743 (0.0634)	0.7892 (0.0608)	0.3621 (0.1078)	0.5991 (0.0706)	0.5866 (0.0581)	0.5894 (0.0627)
11	$\exp(t+Z_1)$	$\sin(t+Z_2)$	0	0.05 (0.1467)	0.0051 (0.1508)	-0.0076 (0.1004)	0.0087 (0.0883)	-0.0060 (0.0950)	0.0052 (0.1082)

SH, IY, DCL, AA, and AO; each group contains 400 log-periodograms (functions) discretized in 150 frequencies (points). Each of the log-periodograms corresponds to a different speaker. In this example, we look for possible associated phonemes. These three data sets have been extensively used in the literature in functional data analysis (Epifanio-López [16], Jacques and Preda [29], Li and Yu [34], López-Pintado and Romo [38], and in particular, Epifanio-López [16], Li and Yu [34], for other purposes such as classification.

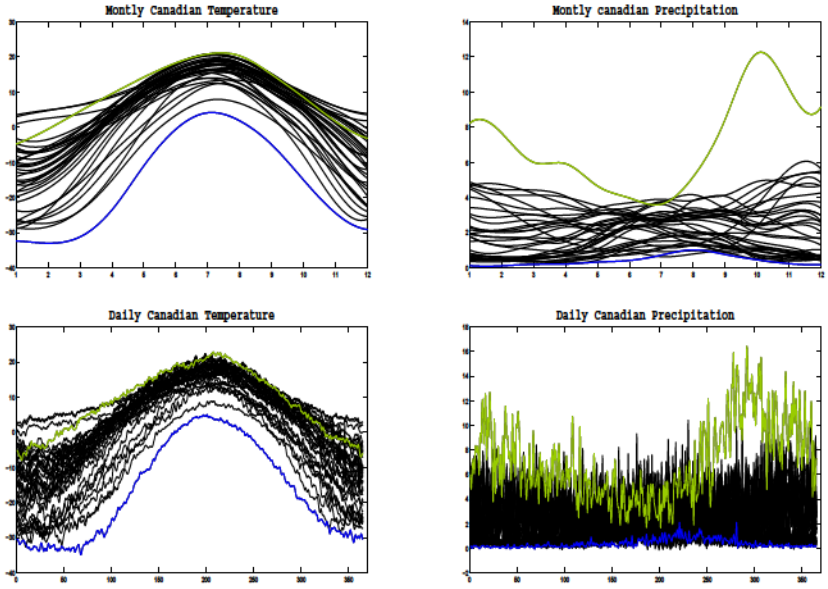

Figure 3.5: Monthly and daily temperature and precipitation of Canada.

Tables 3.10: Association test for temperature and precipitation data

Data 1	Data 2	$\widehat{\rho}_s IL$	*p-value*	Decision	$\widehat{\tau}_1$	*p-value*	$\widehat{\tau}_2$	*p-value*
Daily temperature	Daily precipitation	**0.6043**	**0.0002**	reject H_0	0.0807	0.5050	**0.4958**	**0**
Monthly temperature	Monthly precipitation	**0.5764**	**0.0004**	reject H_0	0.1378	0.2438	**0.4622**	**0.0002**
Montreal	Resolute	**0.6041**	**0.0050**	reject H_0	0.2368	0.1468	**0.3316**	**0.0394**
Montreal	Prince Rupert	−0.0612	0.7940	accept H_0	0.0632	0.6914	−0.026	0.883
Montreal	Fort San John	0.1160	0.6220	accept H_0	−0.0579	0.7398	0.0684	0.6902
Resolute	Prince Rupert	−0.1836	0.4322	accept H_0	−0.2316	0.1620	−0.1158	0.4850
Resolute	Fort San John	0.0168	0.9516	accept H_0	0.0895	0.59	0.0316	0.8668
Prince Rupert	Fort San John	0.3280	0.1474	accept H_0	0.0842	0.6092	0.1789	0.2780

Figure 3.5 shows monthly and daily data of temperature and precipitation in Canada. Green curves are the highest and the blue curves are the smallest in the sense of the IL_n-*grade* ordering. Table 3.10 shows the values of the Spearman coefficient $\widehat{\rho}_s IL$, the *p-value* related to the association test with 10000 bootstrap samples and the corresponding decision with $\alpha = 0.05$. The association test for the other coefficients is also shown in Table 3.10. As

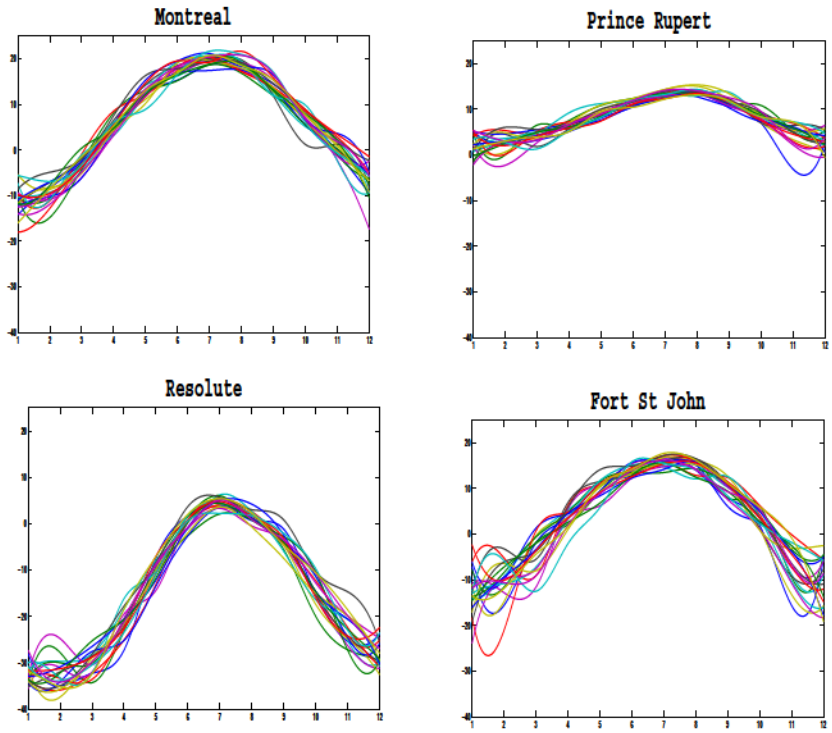

Figure 3.6: Temperatures of 4 cities in Canada.

we can see, the null hypothesis is rejected for all cases except when using $\widehat{\tau}_1$. Remember that $\widehat{\tau}_1$ is based on the pre-order induced by the maximum of the curves, which is more sensitive to outliers and reflects worse than the other pre-order a summary of the curves shapes. Therefore, we can say that the temperature and the precipitation in Canada have a significant association, which was expected because they are strongly linked to climatological phenomena.

In relation to the data sets of Canadian cities, only Montreal and Resolute present significant dependence for both Spearman's and Kendall's τ (with pre-order of the integral) coefficients. We have tried to find a physical explanation for this fact but these two cities do not share the same kind of weather, nor do they have a similar latitude or other factors that directly relate them, so the significant dependence may be due to the similarity with respect to shape and position of the curves per year (see Figure 3.6). However, the positive association between Montreal and Resolute does not hold when we pass the same test with $\widehat{\tau}_1$. Hence, spatial correlation is not observed for these four cities.

Table 3.11 shows the results of the association tests for the phoneme data. We include also the *p-value* for each test. Note that, in general, the dependence between the phonemes is very small for all measures, being only statistically significative for the phonemes AA and SH with the coefficients $\widehat{\rho}_s IL$, $\widehat{\tau}_1$, and $\widehat{\tau}_2$. This may be due to the position and shape of the

curves. We can see that the shape of the curves of the phoneme SH is in general different when compared to other phonemes. Indeed, it can be easily observed that a certain negative dependence could exist (see Figure 3.7). This fact is reflected in the sign of the coefficients since they are negative in most cases where the phoneme SH is evaluated. It can be seen that in this case, the shape of the two groups of curves exhibits opposite behavior.

Tables 3.11: Phoneme data

Phoneme 1	Phoneme 2	$\widehat{\rho}_s IL$	p-value	Decision	$\widehat{\tau}_1$	p-value	$\widehat{\tau}_2$	p-value
AA	AO	0.078	0.1144	accept H_0	0.0257	0.4536	0.0604	0.0692
AA	SH	**−0.100**	**0.0464**	reject H_0	**−0.0675**	**0.048**	**−0.0763**	**0.0192**
AA	IY	0.058	0.2664	accept H_0	0.0004	0.9624	0.0459	0.1504
AA	DCL	0.010	0.791	accept H_0	−0.0186	0.6056	0.003	0.9174
AO	SH	−0.040	0.422	accept H_0	0.0079	0.8088	−0.0245	0.4744
AO	IY	0.010	0.845	accept H_0	0.0386	0.2336	0.0086	0.7944
AO	DCL	−0.020	0.696	accept H_0	−0.0053	0.8920	0.00045	0.9840
SH	IY	−0.025	0.64	accept H_0	−0.0479	0.1592	−0.0179	0.5832
SH	DCL	0.027	0.547	accept H_0	0.0188	0.5832	0.0109	0.7616
IY	DCL	−0.019	0.691	accept H_0	0.0271	0.4256	−0.0079	0.8320

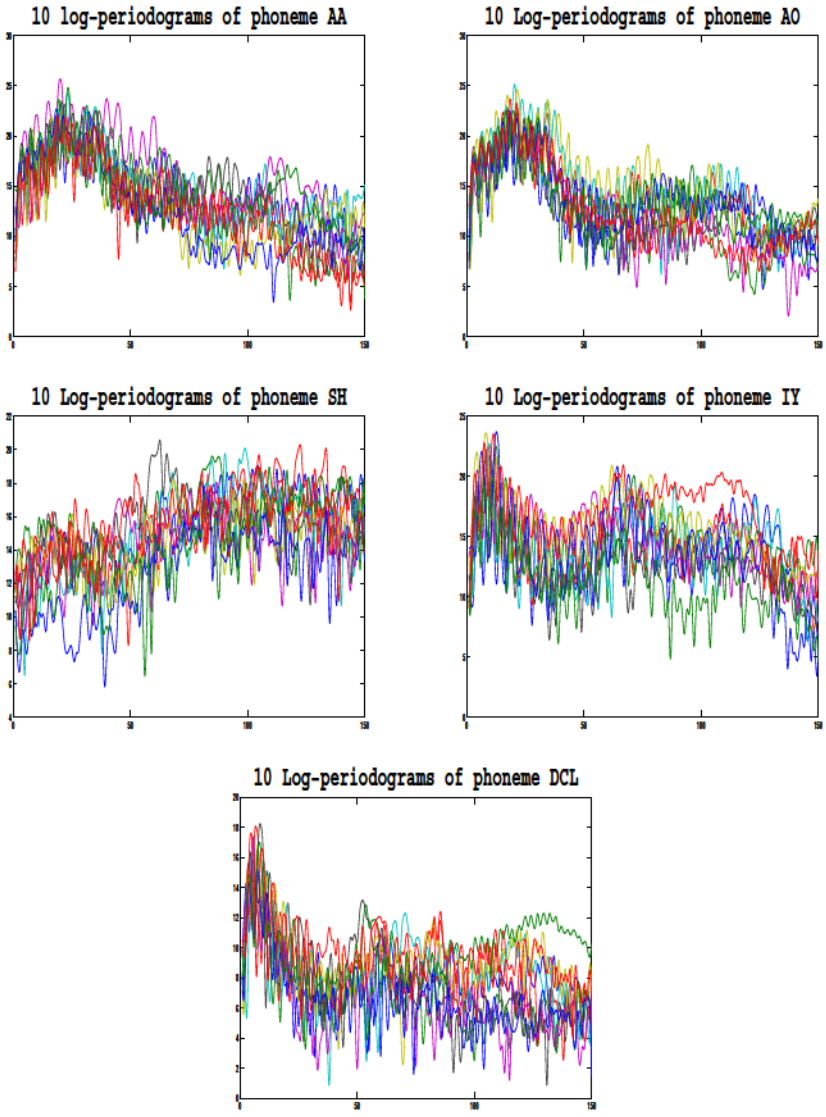

Figure 3.7: Log-periodograms of phonemes AA, AO, SH, IY and DCL.

3.8 Conclusions

We have introduced a new association coefficient to measure dependence between functions when a bivariate sample of functional data is considered. Specifically, a natural extension of the usual Spearman coefficient is provided by ranking the functions using two kinds of ordering for the curves: the Inferior Length and the Superior Length. These orderings among curves allowed us to adapt the definition of grade examined in Nelsen [41] but for the functional context and so, Spearman's coefficient can be defined as usual is, the Pearson correlation among grades. We have also proved that Spearman's coefficient has a good theoretical and practical properties. The simulation study and real examples provided in the chapter show the good performance of the Spearman coefficient as well as its robustness.

We have also introduced a bootstrap independence test to assess the significance of the association between two groups of curves. Tests of this type also allow us to quantify the statistical significance of some conjectures made on the basis of exploratory analysis. We have illustrated with simulated data the power of this test.

We focused in this chapter on an univariate dependence measure, but it could be of interest to explore other possible options such as a functional dependence measure as an alternative to the functional correlation introduced in Ramsay and Silverman [45]. In addition, note how all the univariate dependence measures are linked to a curves ordering. Thus, other pre-orders for curves can provide alternative dependence measures that can be useful for visualizing association in data sets.

CHAPTER 4

Correlation median for functions

4.1 Introduction

Coefficients already studied in this dissertation refer to a univariate measure that reflects the degree of dependence between two sets of curves in terms of a unique number. In this chapter we introduce a new functional correlation coefficient that yields a representative curve of dependence between two sets of functional data. This functional coefficient is analogous the cross-correlation studied in Ramsay and Silverman [45], which consists basically of calculating Pearson's coefficient between the values of the functions in the two groups for each $t \in I$. Accordingly, the cross-correlation does not consider the functional essence of the observations, since the method relies on the calculation of the classic Pearson coefficient for bivariate data. Furthermore, the cross-correlation is defined through the mean and variance of the data, which leads to a procedure more sensitive to the presence of outliers, as in the bivariate case. To avoid this drawback, we extend the concepts of median absolute deviation from the median (MAD) and comedian to the functional context using the idea of depth. These two alternatives, studied in Falk [19], are more robust than standard deviation and covariance. In terms of MAD and comedian, we define the correlation median for functions, which is a functional correlation coefficient that is more robust than the cross-correlation function.

The comedian and MAD are constructed using the median of the data instead of the mean of the data. In functional data we have some alternatives for calculating the median of a set of curves; most of them are based on the concept of depth (see Fraiman and Muniz [24], Cuevas et al.[8], López-Pintado and Romo [38]). Depth provides center-outward ordering of the data, where the median is considered as the deepest curve. To define the comedian and MAD for functions, we use the concept of depth studied in López-Pintado and Romo [38],

who consider that a function is deep if it is contained in many bands among all the bands that can be formed with functions of the sample; therefore, the median is the curve from the sample with highest depth value, i.e, the curve contained in the largest number of bands. Nevertheless, any other measure of depth can be used to define the functional median. The definitions provided in this chapter can be used with any functional depth measure. (Fraiman and Muniz [24], Cuevas et al.[8]).

This chapter is organized as follows. In Section 4.2, we consider the background and preliminary aspects necessary to introduce our coefficient. Section 4.3 presents the definitions of MAD and comedian for functional data. Correlation median for functions and its properties are defined in Section 4.4. A simulation study is carried out in Section 4.5 where we also carry out a sensitivity study of the coefficient. The robustness of the coefficient is analyzed in Section 4.6. In Section 4.7, real data examples are discussed, showing how the correlation median for functions performs. Finally, in Section 4.8 we summarize the main conclusions of this chapter.

4.2 Preliminaries

In functional data analysis it is possible to measure the dependence between two sets of curves through the cross-covariance and cross-correlation functions, discussed in Ramsay and Silverman [45] (p.24). This methodology has already been introduced in Chapter 1, but we will recall the principal definitions.

Assume n pairs of curves (x_i, y_i), for $i = 1, \ldots, n$ which are defined on the same interval $I = [a, b]$. Then the cross-covariance function is given by

$$\widehat{\mathrm{COV}}_{XY}(t_1, t_2) \equiv (n-1)^{-1} \sum_{i=1}^{n} \{x_i(t_1) - \overline{x}(t_1)\}\{y_i(t_2) - \overline{y}(t_2)\}, \tag{4.2.1}$$

where $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i(t)$ and $\overline{y} = n^{-1} \sum_{i=1}^{n} y_i(t)$. Therefore, the cross-correlation function is:

$$\widehat{\mathrm{CORR}}_{XY}(t_1, t_2) \equiv \frac{\widehat{\mathrm{COV}}_{XY}(t_1, t_2)}{\sqrt{\widehat{\mathrm{VAR}}_X(t_1)\widehat{\mathrm{VAR}}_Y(t_2)}}, \tag{4.2.2}$$

where $\widehat{\mathrm{VAR}}_X(t) = (n-1)^{-1} \sum_{i=1}^{n} (x_i(t) - \overline{x}(t))^2$.

The aim of this chapter is to introduce a more robust technique which is more related to functional data to measure the dependence between two sets of curves by extending the concepts of median absolute deviation (MAD) and comedian, discussed in Falk [19], to the functional setting. Recall that (MAD) is defined as

$$MAD(X) \equiv med(\mid X - med(X) \mid), \tag{4.2.3}$$

where $med(\cdot)$ is the median of the data and is a robust alternative to standard deviation. The MAD exhibits interesting advantages over other scale measures, such as, for example,

a breakdown point of 50% and an bounded influence function. Rousseeuw and Croux [46], proposed two scale estimators, which may be competitive with the MAD since they present similar properties. The first one is defined by

$$S_n = c\, med_i\{med_j \mid x_i - x_j \mid\}, \tag{4.2.4}$$

where c is a constant of consistency. The estimator S_n can be seen as an analog of Gini's average difference when replacing averages by medians. The second estimator is

$$Q_n = d\{\mid x_i - x_j \mid : i < j\}_{(k)}, \tag{4.2.5}$$

where d is a constant factor and $k = \binom{h}{2} \approx \binom{n}{2}/4$, where $h = [n/2]+1$ and n is the sample size. The estimator Q_n takes the kth order statistic of the $\binom{n}{2}$ interpoint distances.

Both estimators possess a breakdown point of 50%, but unlike MAD and S_n, the estimator Q_n possesses a smooth influence function. In addition, S_n and Q_n do not presuppose a symmetric model distribution as does MAD. The biggest difference among these concepts consists in their Gaussian efficiency, which is 37% for MAD, 58% for S_n and 82% for Q_n. In this work, we extend these concepts to the functional case. However, we focus on the MAD in order to define a correlation coefficient between two sets of curves. A similar analysis can be carried out considering the functional version of S_n and Q_n.

Based on the concept of MAD, Falk [19] proposed the comedian $COM(X,Y)$ as a robust alternative to the covariance between random variables. That is,

$$COM(X,Y) \equiv med((X - med(X))(Y - med(Y))). \tag{4.2.6}$$

Some features and properties of the comedian can be seen in Chapter 1, Section 1.1. Thus, considering these two concepts, MAD and comedian, Falk [19] introduced the correlation median as:

$$\delta(X,Y) = \frac{COM(X,Y)}{MAD(X)MAD(Y)}. \tag{4.2.7}$$

We will adapt this measure to introduce a robust version of the Pearson coefficient for two groups of curves. In order to do this, in the next section we extend the concepts of MAD and comedian to functions. Note that both $MAD(X)$ and $COM(X,Y)$ from expressions (4.2.3) and (4.2.6), need the median of the data. Hence, to introduce these concepts for a bivariate functional sample, we need a definition of median for functions. For this purpose, we use the definition of median for functional data based on the depth function studied in López-Pintado and Romo [38], (see Chapter 1, Section 1.3).

4.3 MAD and comedian for functions

In this section we define the concepts of MAD and comedian for sets of curves in order to introduce a measure of dependence between two sets of functions. In the next section, we

start by defining the functional version for the MAD and the two alternatives proposed in Rousseeuw and Croux [46] for the MAD, S_n and Q_n to compare its performance. We will also define the comedian for functional data analyzing its properties and present some comparative examples with the cross-covariance function.

4.3.1 Functional MAD

The first step to extend this concept to functional data is to calculate the median of a set of curves. There are several ways to calculate the median in a set curves in the literature. The majority of these procedures are based on depth concept. The depth notion comes from the multivariate analysis aforementioned in Chapter 1, Section 1.3. For example, Fraiman and Muniz [24] have obtained the functional median as the curve that maximizes the average of the one-dimensional depths in each point of the interval I where the curves are defined. The functional depth allows us to rank the functions from the deepest (functional median) to the farthest (outer surface). To calculate the functional median in this chapter, we use the concept of generalized band depth studied in López-Pintado and Romo [38] which we summarize briefly.

For any function x in $x_1, x_2, \ldots, x_n$ let

$$A_j(x) = A(x; x_{i_1}, \ldots, x_{i_j}) = \left\{ t \in I : \min_{r=i_1,\ldots,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,\ldots,i_j} x_r(t) \right\}, \; j \geq 2,$$

be the set of points in the interval I where the function x is inside the band given by the observations $x_{i_1}, x_{i_2}, \ldots, x_{i_j}$, then

$$GS_n^{(j}(x) = \begin{pmatrix} n \\ j \end{pmatrix}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \ldots i_j \leq n} \lambda_r(A(x; x_{i_1}, x_{i_2}, \ldots, x_{i_j})), \;\; j \geq 2,$$

where λ is the Lebesgue measure in $\mathbb{R}$, and $\lambda_r = \lambda(A_j(x))/\lambda(I)$ will be the proportion of time that x is inside the band. Therefore, the generalized band depth (GBD) of x is given by

$$GS_{n,J}(x) = \sum_{j=2}^{J} GS_n^{(j}(x), \;\; j \geq 2.$$

If $X_1, X_2, \ldots, X_n$ are independent copies of the stochastic process $X(t)$, the population version of $GS_n^{(j}(x)$ and $GS_{n,J}(x)$ are given by

$$GS^{(j}(x) = E\lambda_r(A(x; X_1, X_2, \ldots, X_j)), \;\; j \geq 2, \quad \text{and}$$

$$GS_J(x) = \sum_{j=2}^{J} GS^{(j}(x) = \sum_{j=2}^{J} E\lambda_r(A(x; X_1, X_2, \ldots, X_j)), \;\; j \geq 2, \quad \text{respectively.} \quad (4.3.1)$$

We will use the equation 4.3.1 with $J = 2$ to calculate the deepest function, which will be the functional median.

$$\widetilde{X}(t) = med(X(t)) \equiv \arg \max_{x \in C(I)} GS_J(x; X_1, X_2, \ldots, X_n). \tag{4.3.2}$$

Once the functional median is defined, we can extend the concept of MAD to functions.

Definition 4.3.1 (Functional MAD.) *Let $X_1, \ldots, X_n$ be independent copies of the stochastic process $X(t)$, then*

$$MAD(X(t)) \equiv med|X_i - \widetilde{X}(t)|.$$

Where $\widetilde{X}(t)$ is as in (4.3.2).

As in the univariate case, the functional MAD is less sensitive than the standard deviation function to extreme functional observations, since functional MAD is defined through the median of the data and does not depend on calculating sums of transformations of the data, as in the case of standard deviation function.

Now, we also extend the concepts of S_n and Q_n proposed by Rousseeuw and Croux [46] to the functional field. This is of interest since these concepts are two alternatives to MAD that possess important properties, already mentioned in the Section 4.2, especially regarding Gaussian efficiency.

Definition 4.3.2 (S_n and Q_n for functions.) *Let $X_1, \ldots, X_n$ be independent copies of the stochastic process $X(t)$. Then $S_n(X(t))$ and $Q_n(X(t))$ are:*

$$S_n \equiv med_i\{med_j \mid X_i - X_j \mid\}.$$

$$Q_n\{\mid X_i - X_j \mid : i < j\}_{(k)}.$$

Where $k = \binom{h}{2} \approx \binom{n}{2}/4$, with $h = [n/2]+1$. Observe that we do not include the constants of consistency since we do not define them as estimators of the standard deviation.

Figure 4.1 shows the curves that represent the MAD, S_n, Q_n and the standard deviation for four different groups of curves. In this thesis we will use the functional MAD as the scale measure for defining a new correlation coefficient for functions.

4.3.2 Functional comedian.

Following Falk [19], we define the comedian for functions based on the functional MAD, to measure dependence between two groups of curves.

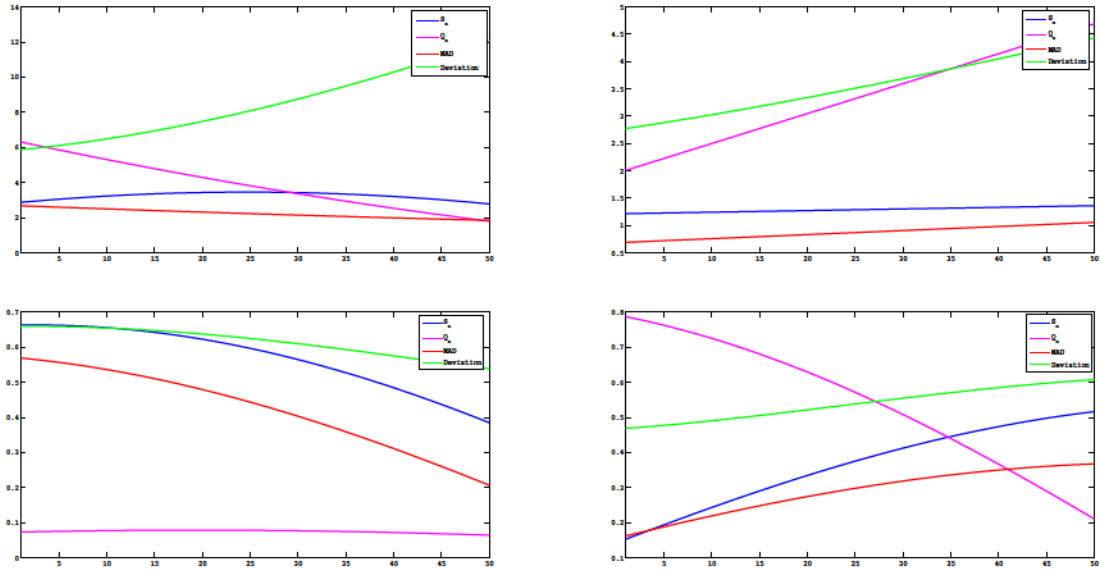

Figure 4.1: S_n, Q_n, MAD and standard deviation for functional data.

Definition 4.3.3 (Functional comedian.) *Let* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be independent copies of the bivariate stochastic process* $(X(t), Y(t))$. *The functional comedian is:*

$$COM(X(t), Y(t)) \equiv med\{(X_i - \widetilde{X}(t))(Y_i - \widetilde{Y}(t))\}.$$

where $\widetilde{X}(t)$ *and* $\widetilde{Y}(t)$ *are the functional medians of* $X(t)$ *and* $Y(t)$, *respectively.*

Recall that the functional median can be calculated in different ways depending on the definition of depth used; for our purposes, we use the generalized band depth. Figure 4.2 shows the curves of covariance and comedian for three different bivariate samples. They come from processes that are generated from $X(t) = f_1(t, Z_1)$ and $Y(t) = f_2(t, Z_2)$, where (Z_1, Z_2) represents the random part of the processes. In this chapter the data will be simulated in the same way as in Chapters 2 and 3.

- The first sample has been generated from processes $X(t) = (t+Z_1)^3 + (t+Z_1)^2 + 3(t+Z_1)$, $Y(t) = (t + Z_2)^2 + (7/8)(t + Z_2) - 10$ and $\sigma_{12} = 0.8$.
- The second from processes $X(t) = \sin(t + Z_1)$, $Y(t) = \cos(t + Z_2)$ and $\sigma_{12} = -0.7$.
- The last one from $X(t) = \exp(t^2 + Z_1)$, $Y(t) = (t + Z_2)^2 - 8t + Z_2$ and $\sigma_{12} = 0.9$.

We can see that the curves are similar. However, the scale of the covariance curve is larger than the scale of the comedian curve. This is due to the fact that the functional comedian is defined through the median, which comes from the groups of curves, while the covariance function is defined through the mean, and therefore it is an average of the curves.

The functional comedian meets some important properties:

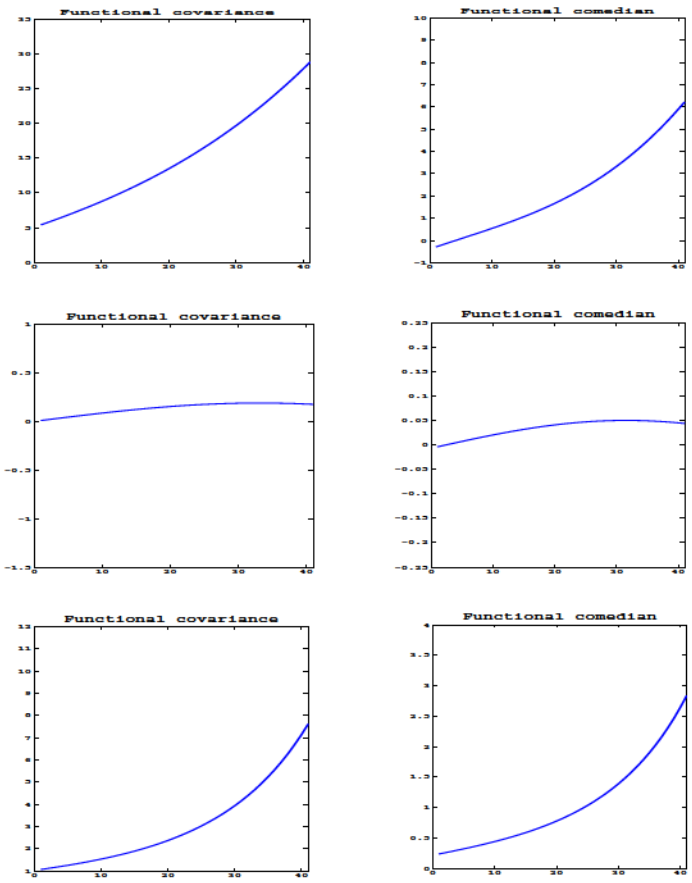

Figure 4.2: Functional covariance and functional comedian.

1. if $Y(t) \overset{\text{a.s.}}{=} aX(t) + b$ then $COM(X(t), Y(t)) = aMAD(X(t))^2$, for some $a, b \in \mathbb{R}$.
2. $COM(X(t), aY(t) + b) = aCOM(X(t), Y(t))$.
3. $COM(X(t), Y(t)) = COM(Y(t), X(t))$.

Property 1.

Proof.

$$\begin{aligned} COM(X(t), Y(t)) &= COM(X(t), aX(t) + b) \\ &= med[(X(t) - med(X(t)))(aX(t) + b - med(aX(t) + b))] \\ &= med[(X(t) - med(X(t)))(aX(t) + b - amed(X(t)) - b)] \\ &= med[(X(t) - med(X(t)))(a(X(t) - med(X(t))))] \\ &= amed|X(t) - med(X(t))|^2 \\ &= aMAD(X(t))^2. \end{aligned}$$

□

Property 2.

Proof.

$$\begin{aligned} COM(X(t), aY(t) + b) &= med[(X(t) - med(X(t)))(aY(t) + b - med(aY(t) + b))] \\ &= med[(X(t) - med(X(t)))(aY(t) + b - amed(Y(t)) - b)] \\ &= med[(X(t) - med(X(t)))(a(Y(t) - med(Y(t))))] \\ &= amed[(X(t) - med(X(t)))(Y(t) - med(Y(t)))] \\ &= aCOM(X(t), Y(t)). \end{aligned}$$

□

Property 3 it is straightforward.

We have already defined the MAD and comedian for functions as two alternatives to the standard deviation function and the cross-covariance function. Now, we will define a correlation coefficient for functions based on these new measures.

4.4 Correlation median for functions

The aim of this chapter is to provide a more robust alternative to the cross-correlation function studied in Ramsay and Silverman [45]. We focus from now on $t_1 = t_2$ and, we will call the cross-correlation function in $t_1 = t_2$ as **the correlation function**. In the previous section, we extended the concepts of comedian (COM) and MAD to the functional field. Therefore,

the correlation median defined in Falk [19] can be also extended to functions using the above concepts.

Definition 4.4.1 (Correlation median for functions.) *Let* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be independent copies of the bivariate stochastic process* $(X(t), Y(t))$. *We define the correlation median for functions as follows:*

$$\dot{\delta}(X(t), Y(t)) = \frac{COM(X(t), Y(t))}{MAD(X(t))MAD(Y(t))} = \frac{med\{(X_i - \widetilde{X}(t))(Y_i - \widetilde{Y}(t))\}}{med|X_i - \widetilde{X}(t)|med|Y_i - \widetilde{Y}(t)|},$$

where $\widetilde{X}(t)$ *and* $\widetilde{Y}(t)$ *are the median of* $X(t)$ *and* $Y(t)$.

Figure 4.3 shows the correlation function and correlation median for functions for three pairs of groups of curves generated from processes previously defined in Subsection 4.3.2.

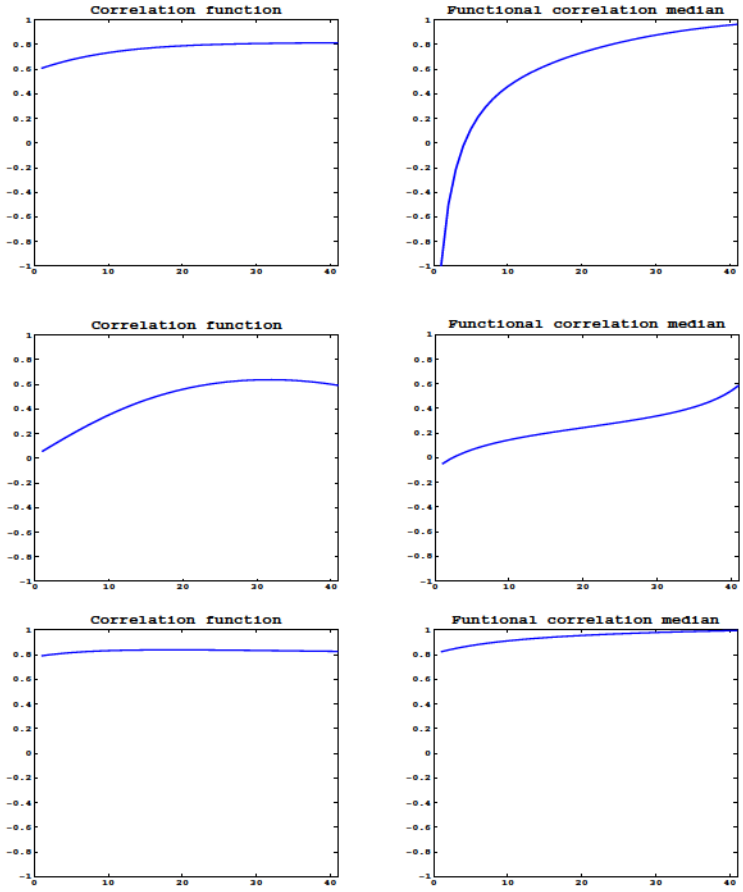

Figure 4.3: Correlation function and correlation median for functions for different pairs of groups of curves.

In some cases, the absolute value of the correlation median for functions can also be larger than 1, as in the case of bivariate data. In such cases, it is quite difficult to interpret the results; therefore we propose an alternative correlation median for functions, obtained by dividing the coefficient over the maximum between its maximum and minimum values i.e,

$$\delta(X(t),Y(t)) = \frac{\dot{\delta}(X(t),Y(t))}{\max\{|\max_{t\in T}\dot{\delta}(X(t),Y(t))|,|\min_{t\in T}\dot{\delta}(X(t),Y(t))|\}}. \tag{4.4.1}$$

The curve that represents the correlation median for functions will be in the interval $[-1,1]$, which makes it simpler to interpret. This coefficient can be useful when we need a graphic way of representing the dependence between two sets of curves.

4.4.1 Properties

As was mentioned in Chapter 1, Section 1.1, the correlation median meets two important properties. We prove these two properties for functions and other that can be inferred from being a correlation coefficient.

1. $\delta(X(t),Y(t)) = \delta(Y(t),X(t))$. (Symmetry).
2. $-1 \leq \delta(X(t),Y(t)) \leq 1$.
3. $\delta(aX(t)+b,cY(t)+d) = \begin{cases} \delta(X(t),Y(t)), & ac>0; \\ -\delta(X(t),Y(t)), & ac<0. \end{cases}$
4. $\delta(X(t),aX(t)+b) = \begin{cases} 1, & a>0; \\ -1, & a<0. \end{cases}$

To prove the above properties, we will need the following two results.

Lemma 4.4.2 *Let $X_1,X_2,\ldots,X_n$ be independent copies of the stochastic process $X(t)$ with observations $x_1,x_2,\ldots,x_n$. Then,*

$$GS_J(|x|^p,|X_1|^p,|X_2|^p,\ldots,|X_n|^p) = GS_J(|x|,|X_1|,|X_2|,\ldots,|X_n|); \quad p\geq 1,$$

where $GS_J(x;X_1,X_2,\ldots,X_n)$ be the functional depth of the curve x as in 4.3.1.

Proof.
Observe

$$\begin{aligned} A(|x|^p;|X_{i1}|^p,|X_{i2}|^p,\ldots,|X_{ij}|^p) &= \left\{t\in I : \min_{r=i_1,\ldots,i_j}|x_r(t)|^p \leq |x(t)|^p \leq \max_{r=i_1,\ldots,i_j}|x_r(t)|^p\right\} \\ &= \left\{t\in I : [\min_{r=i_1,\ldots,i_j}|x_r(t)|]^p \leq |x(t)|^p \leq [\max_{r=i_1,\ldots,i_j}|x_r(t)|]^p\right\} \\ &= \left\{t\in I : \min_{r=i_1,\ldots,i_j}|x_r(t)| \leq |x(t)| \leq \max_{r=i_1,\ldots,i_j}|x_r(t)|\right\} \\ &= A(|x|,|X_{i1}|,|X_{i2}|,\ldots,|X_{ij}|). \end{aligned}$$

Therefore, it can be easily seen that,

$$\begin{aligned} GS_J(|x|^p;|X_1|^p,|X_2|^p,\ldots,|X_n|^p) &= \sum_{j=2}^{J} \frac{E\lambda[A(|x|^p,|X_{i1}|^p,|X_{i2}|^p,\ldots,|X_{ij}|^p)]}{\lambda[I]} \\ &= \sum_{j=2}^{J} \frac{E\lambda[A(|x|,|X_{i1}|,|X_{i2}|,\ldots,|X_{ij}|)]}{\lambda[I]} \\ &= GS_J(|x|,|X_1|,|X_2|,\ldots,|X_n|). \end{aligned}$$

□

Proposition 4.4.3

$$[med|X(t)|]^p = med|X(t)|^p,; \quad p \geq 1.$$

Proof.

$$\begin{aligned} med(X(t)) &= arg_{x\in C(I)} \max GS_J(|x|;|X_1|,|X_2|,\ldots,|X_n|) \\ &= \{arg_{x^p\in C(I)} \max GS_J(|x|;|X_1|,|X_2|,\ldots,|X_n|)\}^{\frac{1}{p}} \\ &= \{arg_{x^p\in C(I)} \max GS_J(|x|^p;|X_1|^p,|X_2|^p,\ldots,|X_n|^p)\}^{\frac{1}{p}} \quad \text{(see Lemma 4.4.2).} \end{aligned}$$

which implies that

$$\begin{aligned} &\{arg_{x\in C(I)} \max GS_J(|x|;|X_1|,|X_2|,\ldots,|X_n|)\}^p \\ &= arg_{x^p\in C(I)} \max[GS_J(|x|^p;|X_1|^p,|X_2|^p,\ldots,|X_n|^p)] \\ &[med|X(t)|]^p = med|X(t)|^p. \end{aligned}$$

□

Now, we will prove the properties stated previously. The proof of property 1 is straightforward.

Property 2.

Proof.

$$\begin{aligned} &\max\{|\max \dot{\delta}(X(t),Y(t))|,|\min \dot{\delta}(X(t),Y(t))|\} \geq |\min \dot{\delta}(X(t),Y(t))| \\ &\geq -\min \dot{\delta}(X(t),Y(t)) \geq -\dot{\delta}(X(t),Y(t)) \text{ then,} \\ &-\max\{|\max \dot{\delta}(X(t),Y(t))|,|\min \dot{\delta}(X(t),Y(t))|\} \leq \dot{\delta}(X(t),Y(t)), \end{aligned}$$

and then

$$-1 \leq \frac{\dot{\delta}(X(t),Y(t))}{\max\{|\max \dot{\delta}(X(t),Y(t))|,|\min \dot{\delta}(X(t),Y(t))|\}}.$$

Observe now that

$$-\max\{|\max\dot{\delta}(X(t),Y(t))|,|\min\dot{\delta}(X(t),Y(t))|\}\leq|\max\dot{\delta}(X(t),Y(t))|$$
$$\leq\max\{|\max\dot{\delta}(X(t),Y(t))|,|\min\dot{\delta}(X(t),Y(t))|\}.$$

Hence,

$$\frac{\dot{\delta}(X(t),Y(t))}{\max\{|\max\dot{\delta}(X(t),Y(t))|,|\min\dot{\delta}(X(t),Y(t))|\}}\leq 1.$$

□

Property 3.

Proof.

$$\dot{\delta}(aX(t)+b,cY(t)+d)=\frac{COM(aX(t)+b,cY(t)+d)}{MAD(aX(t)+b)MAD(cY(t)+d)}.$$
$$=\frac{med\{[aX(t)+b-med(aX(t)+b)][cY(t)+d-med(cY(t)+d)]\}}{med|aX(t)+b-med(aX(t)+b)|med|cY(t)+d-med(cY(t)+d)|},$$
$$=\frac{med\{[aX(t)+b-amed(X(t))-b][cY(t)+d-cmed(Y(t))-d]\}}{med|aX(t)+b-amed(X(t))-b|med|cY(t)+d-cmed(Y(t))-d|},$$
$$=\frac{ac\ med\{[X(t)-med(X(t))][Y(t)-med(Y(t))]\}}{|ac|\ med|X(t)-med(X(t))|med|Y(t)-med(Y(t))|},$$
$$=\frac{ac\ COM(X(t),Y(t))}{|ac|\ MAD(X(t))MAD(Y(t))}$$
$$=\frac{ac}{|ac|}\dot{\delta}(X(t),Y(t))=\begin{cases}\dot{\delta}(X(t),Y(t)), & ac>0,\\ -\dot{\delta}(X(t),Y(t)), & ac<0.\end{cases}$$

Therefore,

$$\delta(aX(t)+b,cY(t)+d)=\frac{\dot{\delta}(aX(t)+b,cY(t)+d)}{\max\{|\max\dot{\delta}(X(t),Y(t))|,|\min\dot{\delta}(X(t),Y(t))|\}}.$$

If $ac>0$, obviously $\delta(aX(t)+b,cY(t)+d)=\delta(X(t),Y(t))$. Now, if $ac<0$,

$$\begin{aligned}\delta(aX(t)+b,cY(t)+d)&=\frac{-\dot{\delta}(X(t),Y(t))}{\max\{|\max-\dot{\delta}(X(t),Y(t))|,|\min-\dot{\delta}(X(t),Y(t))|\}}\\&=\frac{-\dot{\delta}(X(t),Y(t))}{\max\{|-\min\dot{\delta}(X(t),Y(t))|,|-\max\dot{\delta}(X(t),Y(t))|\}}\\&=\frac{-\dot{\delta}(X(t),Y(t))}{\max\{|\max\dot{\delta}(X(t),Y(t))|,|\min\dot{\delta}(X(t),Y(t))|\}}\\&=-\delta(X(t),Y(t)).\end{aligned}$$

□

Property 4.

Proof.

$$\text{We have } \dot{\delta}(X(t), aX(t)+b) = \frac{COM(X(t), aX(t)+b)}{MAD(X(t))MAD(aX(t)+b)} = \frac{aMAD(X(t))^2}{|a|MAD^2(X(t))}$$
$$= \frac{aMAD^2(X(t))}{|a|MAD^2(X(t))}, \quad \text{from Proposition 4.4.3}$$
$$= \frac{a}{|a|} = \begin{cases} 1, & \text{a>0;} \\ -1, & \text{a<0.} \end{cases}$$

Therefore,

$$\delta(X(t), aX(t)+b) = \frac{\dot{\delta}(X(t), aX(t)+b)}{\max\{|\max_{t\in T}\dot{\delta}(X(t), aX(t)+b)|, |\min_{t\in T}\dot{\delta}(X(t), aX(t)+b)|\}}.$$

If $ac > 0$, obviously $\delta(X(t), aX(t)+b) = 1$. Now, if $ac < 0$ then $\delta(X(t), aX(t)+b) = -1$. □

4.5 Simulation study

In this section we carry out a simulation study in order to show how the correlation median for functions works and compare it with the correlation function previously considered. We also simulate cases where a process is an affine transformation of another process and the case where the processes are independent. Specifically, we simulate 50 realizations from different processes $X(t) = f_1(t, Z_1)$ and $Y(t) = f_2(t, Z_2)$, where (Z_1, Z_2) represents the random part of the processes. We assume (Z_1, Z_2) to be a normal bivariate with correlation σ_{12} and for each pair $(f_1,\ f_2)$ we use a different correlation σ_{12}. Also, we have discretized each curve with 50 points, taken 100 replications of each sample, and calculated the correlation median for functions and correlation function for each of them. Functional data were generated from the following bivariate processes:

- $X(t) = (t+Z_1)^3 + (t+Z_1)^2 + 3(t+Z_1),\ \ Y(t) = (t+Z_2)^2 + (7/8)(t+Z_2) - 10,\ \ \sigma_{12} = 0.8.$
- $X(t) = \sin(t+Z_1),\ \ Y(t) = \cos(t+Z_2),\ \ \sigma_{12} = -0.7.$
- $X(t) = exp(t^2 + Z_1),\ \ Y(t) = (t+Z_2)^2 - 8t + Z_2,\ \ \sigma_{12} = 0.9.$
- $X(t) = (t+Z_1)^2 + 3,\ \ Y(t) = (t+Z_2)^2,\ \ \sigma_{12} = 0.5.$
- $X(t) = 25(t+Z_1)^2,\ \ Y(t) = 30t^{3/2}(1-t) + Z_2,\ \ \sigma_{12} = -0.4.$
- $X(t) = 4(t+Z_1)^2,\ \ Y(t) = (t+Z_2)^3,\ \ \sigma_{12} = 0.1.$
- $X(t) = 4(t+Z_1)^2,\ \ Y(t) = (t+Z_2)^3,\ \ \sigma_{12} = -0.8.$

Figure 4.4 shows the curves that represent the mean and deviation of the correlation function and the correlation median for the 100 replications of each process. We can see that

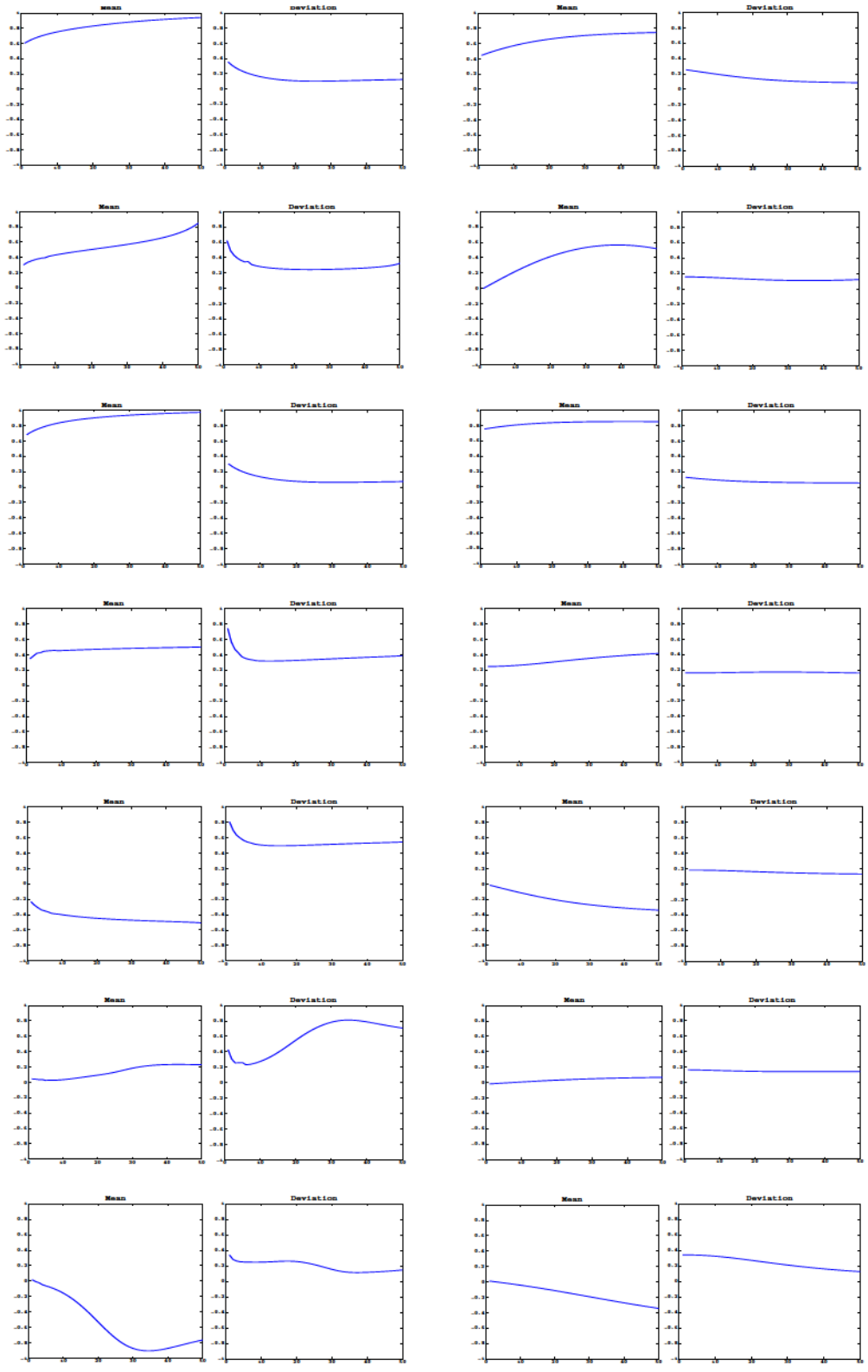

Figure 4.4: The left panel gives correlation median for functions

and the right panel contains correlation function.

the difference among the curves is obvious due to the use of the median instead of the mean, and the use of MAD instead of standard deviation. We think that the correlation median for functions makes use of more information from the data set than the correlation function, which is more focused on observing data point to point avoiding the functional structure of the data.

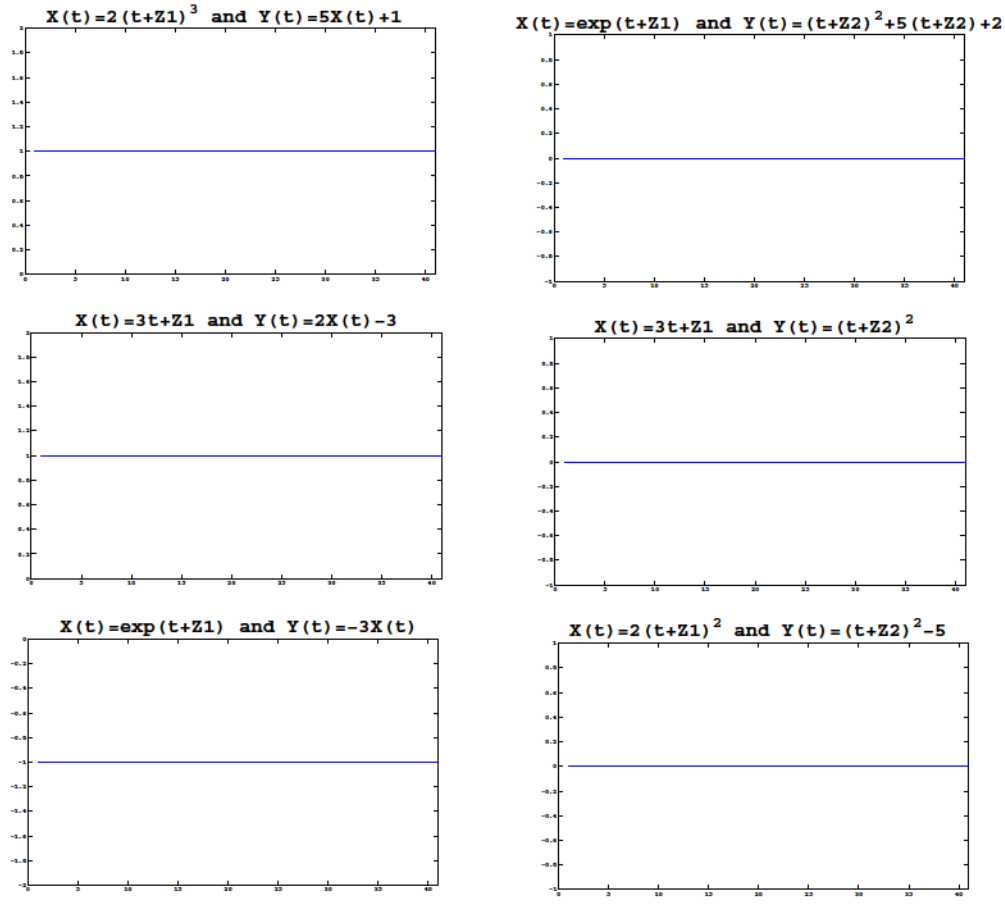

Figure 4.5: Affine transformations. Right: independent processes.

Figure 4.5 shows the correlation median for functions related to processes where one is an affine transformation of another and when the processes are independent. We can see that for affine transformations the correlation median for functions is equal to one, and for the case where the processes are independent, it is equal to zero. Figure 4.6 shows that Property 5 does not hold for other types of transformations.

Now, we analyze the sensitivity of δ with respect to the sample size n and with respect to d, the number of points taken to discretize the functions, in order to determine the steadiness of our coefficient. We will use the following pair of stochastic processes $X(t) = 4(t+Z_1)^2$, $Y(t) = (t+Z_2)^3$, $\sigma_{12} = -0.8$. We have used $n = 25,\ 50,\ 100,\ 150$ with $d = 50$. Figure 4.7 shows that the changes in the curves that represent the correlation median for functions are very small and

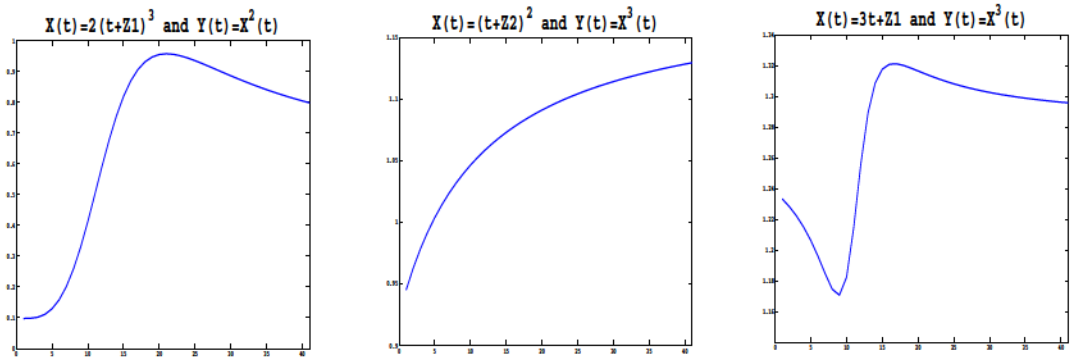

Figure 4.6: Other transformations.

$\hat{\delta}$ is stable with respect to the sample size. Now, fix $n = 50$, and consider $d = 25,\ 50,\ 100,\ 150$ points. Figure 4.8 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 point out that we have carried out the sensitivity analysis with other models, and the conclusions are the same for the model reported.

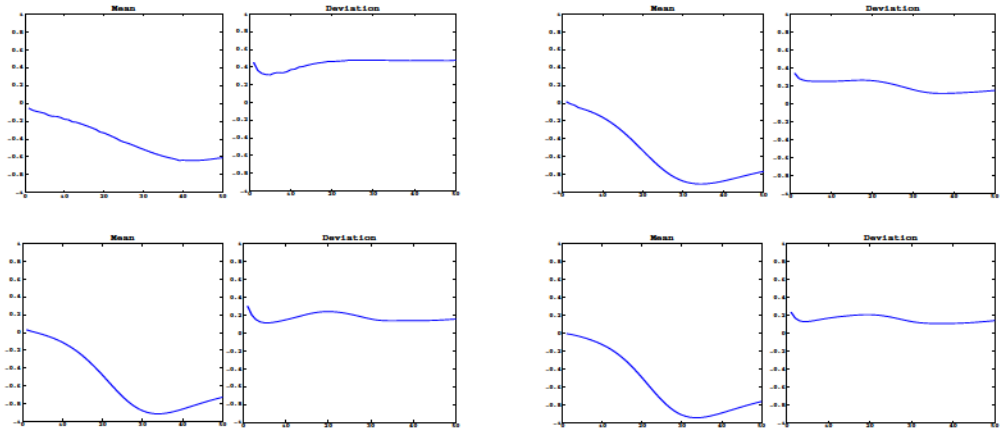

Figure 4.7: Sensitivity to sample size.

In the next section we provide a robustness study to analyze how outliers affect our coefficient.

4.6 Robustness

As stated earlier, our aim is to propose a more robust coefficient of correlation than the cross-correlation function in Ramsay and Silverman [45]. Therefore, in this section we study the robustness of our coefficient and compare it with that of the correlation function. We have

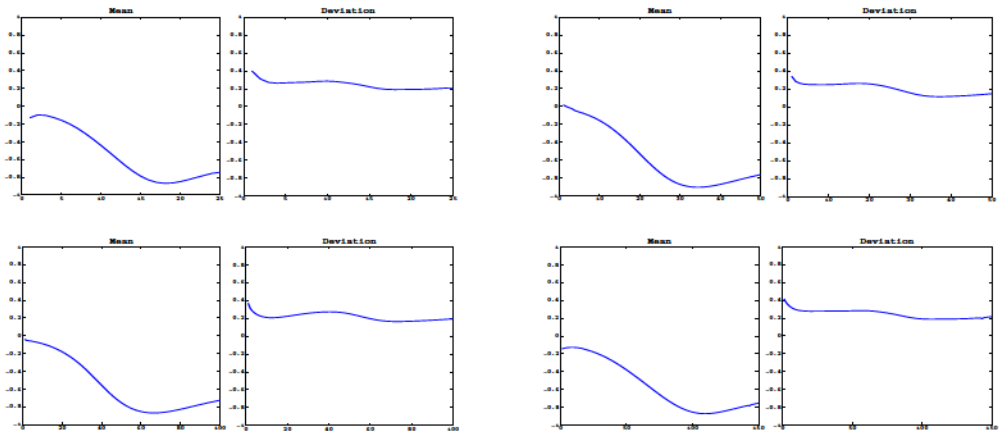

Figure 4.8: Sensitivity to the number of points in the discretization.

simulated 100 curves and taken 50 points to discretize each curve of the following processes:

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

2. $X(t) = sin(t + Z_1),\ \ Y(t) = cos(t + Z_2),\ \ \sigma_{12} = -0.7$

3. $X(t) = exp(t^2 + Z_1),\ \ Y(t) = (t + Z_2)^2 - 8t + Z_2,\ \ \sigma_{12} = 0.9$

To contaminate the samples, we have used three different types of outliers: shape outliers, magnitude outliers and shape-magnitude outliers; the structure of these outliers was summarized in Chapters 2, Section 2.4. We introduce the outliers only in the group of curves that comes from $X(t)$ in a progressive way starting with one, three and five of each.

Figure 4.9 shows the variation of the coefficients when shape outliers are introduced. Each measure is calculated before contaminating the data (row 1). The following rows contain the variation of the curves after being contaminated with one, three and five shape outliers, respectively. We can observe that neither the representative curve of correlation function nor the representative curve for correlation median for functions, present significant variation when different numbers of shape outliers are introduced. Note that in Figure 4.10, the curves show changes when we introduce magnitude outliers, and these variations are more significant in the case of the correlation function than in the case of the correlation median for functions. Finally, when we consider shape-magnitude outliers, the variation for both curves is similar to the variation for magnitude outliers, since shape-magnitude outliers are a mix between shape outliers and magnitude outliers, as can be seen in Figure 4.11.

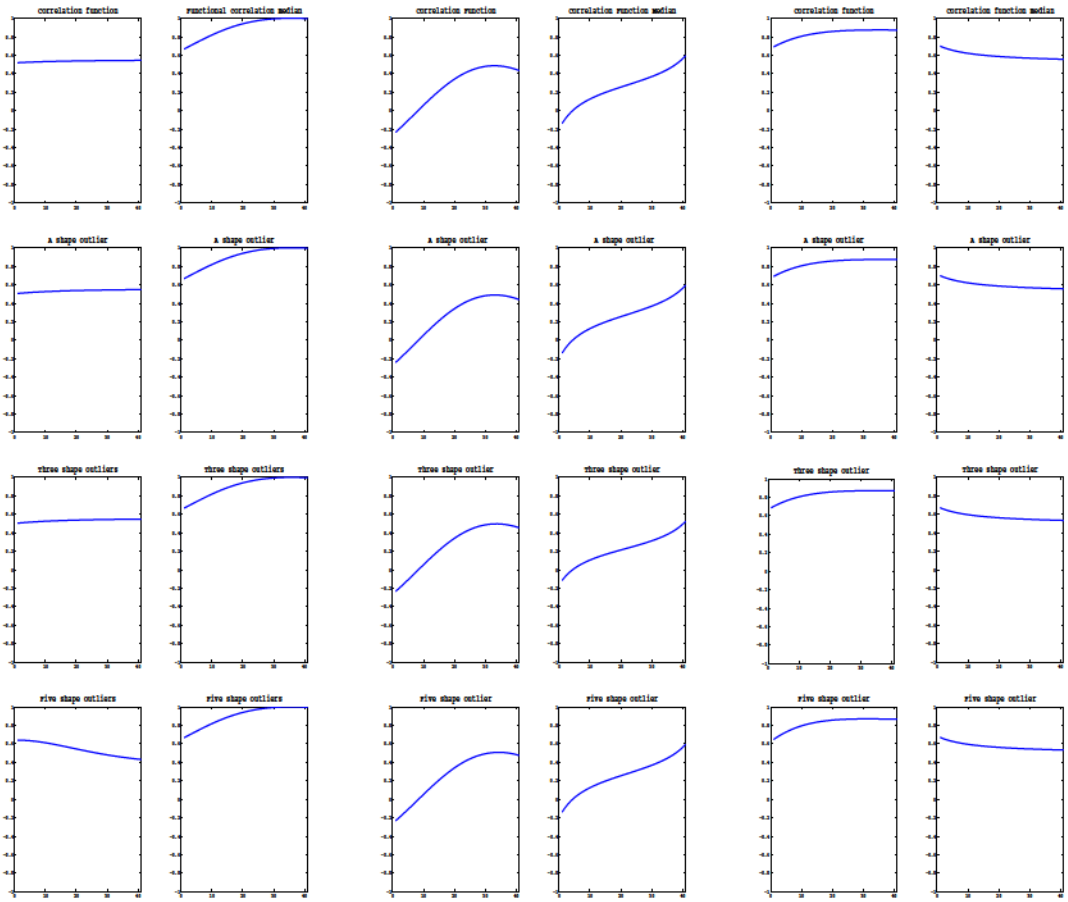

Figure 4.9: Shape outliers 1, 3, 5.

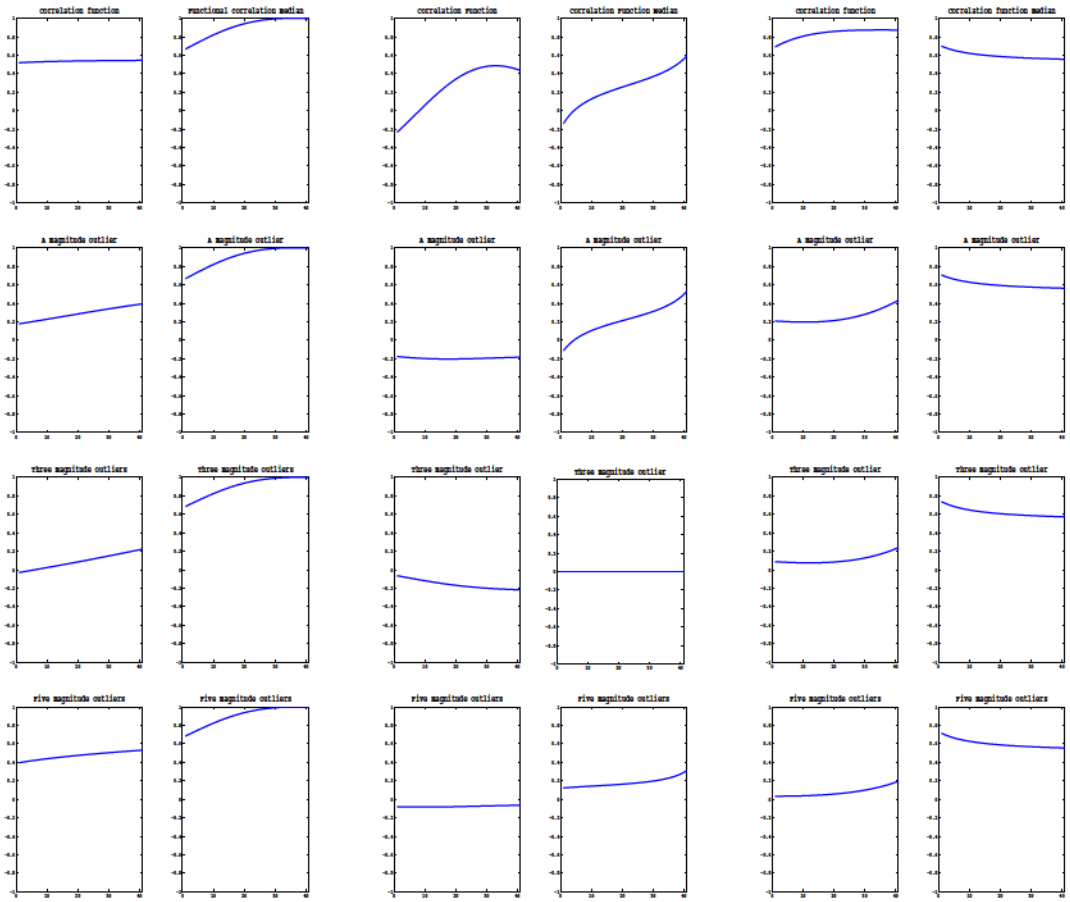

Figure 4.10: Magnitude outliers 1, 3, 5.

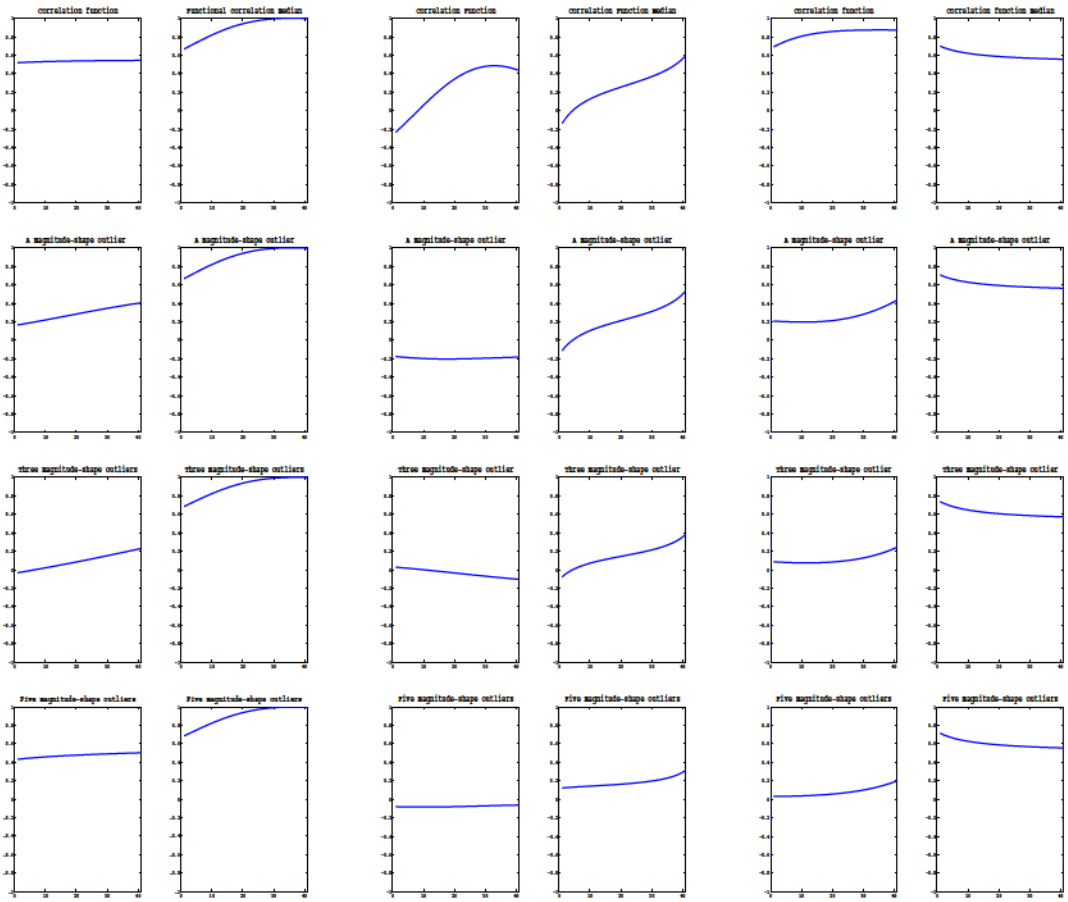

Figure 4.11: Magnitude-shape outliers 1, 3, 5.

4.7 Real Data

We consider four real data sets to assess our procedure and compare it with the correlation function. The first one is composed of monthly temperature and precipitation in 35 Canadian weather stations (see Ramsay and Silverman[45]). We have 12 points per curve i.e., the temperature and precipitation each month. In this example we calculate the curves (correlation function and correlation median for functions) that represent the dependence between temperature and precipitation during the year. The second data set, studied in Leurgans et al.[33], consists of the angular rotations in the sagittal plane of the hip and knee of 39 normal 5-year-old children. The observations are taken over a gait cycle consisting of one double step taken by each child, and time is measured in terms of the cycle. In all cases the cycle has been discretized (mathematically) to a regular grid of 20 points. In this case we can study the change of the dependence between angular rotations of the hip and knee through the gait cycle. The third data set 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 points per curve); these data were used in Chapter 2 to calculate the Kendall's τ coefficient for functions. We analyze some pairs of companies in order to study the relationship between the prices in that time period. The last data set corresponds to a micro-array time series. These data characterize the response of a human T-cell line (Jirkat) to treatment with PMA and ioconomin. The data consist of 58 genes measured across 10 time points with 44 replications. Opgen-Rhein and Strimmer [42] used this data set to calculate the dynamical correlation between functional data in order to construct genetic networks. We calculate the curve of correlation median for functions between some pairs of genes and we look for possible associated genes.

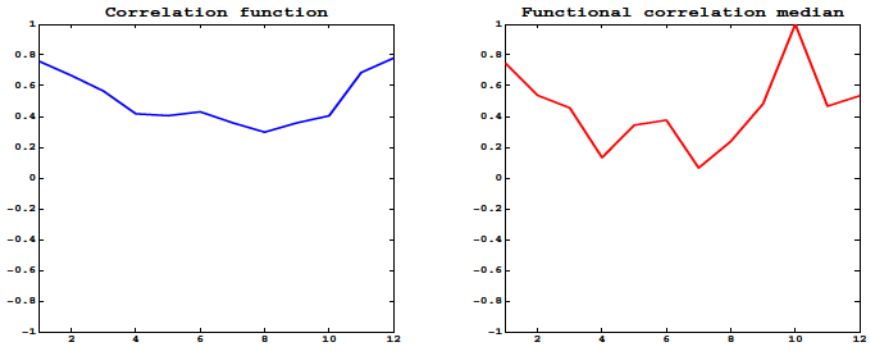

Figure 4.12: Correlation function and correlation median for functions for monthly temperature and precipitation.

Figure 4.12 presents the curves that give the correlation function and correlation median

for functions for the data of temperature and precipitation. The correlation function shows high dependence between the temperature and precipitation in the winter and a very low dependence during the rest of the year. The correlation median for functions identifies additional relationships, for example, high dependence in midwinter and early fall, and very low dependence in spring and summer. In the data gait cycle (Figure 4.13), we see that the curve of correlation median for functions presents peaks around important values of the cycle (5, 10 and 15), while the correlation function is a smoother curve and does not show abrupt changes in the dependence between the angular rotations of the hip and knee. The curve of correlation function for the data of the companies (see Figure 4.14) oscillates around the same value in all cases, while the correlation median for functions presents high variability. This fact shows us how the relationships between the asset prices vary throughout the day. Finally, in Figure 4.15 we can see the different behavior of the curves that represent the association between genes.

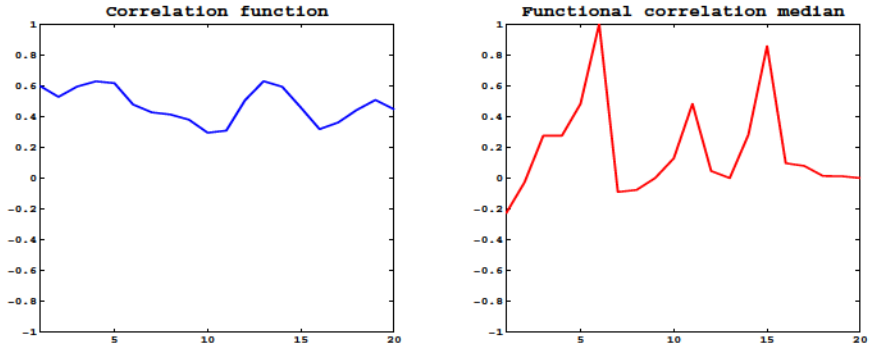

Figure 4.13: Correlation function and correlation median for hip and knee.

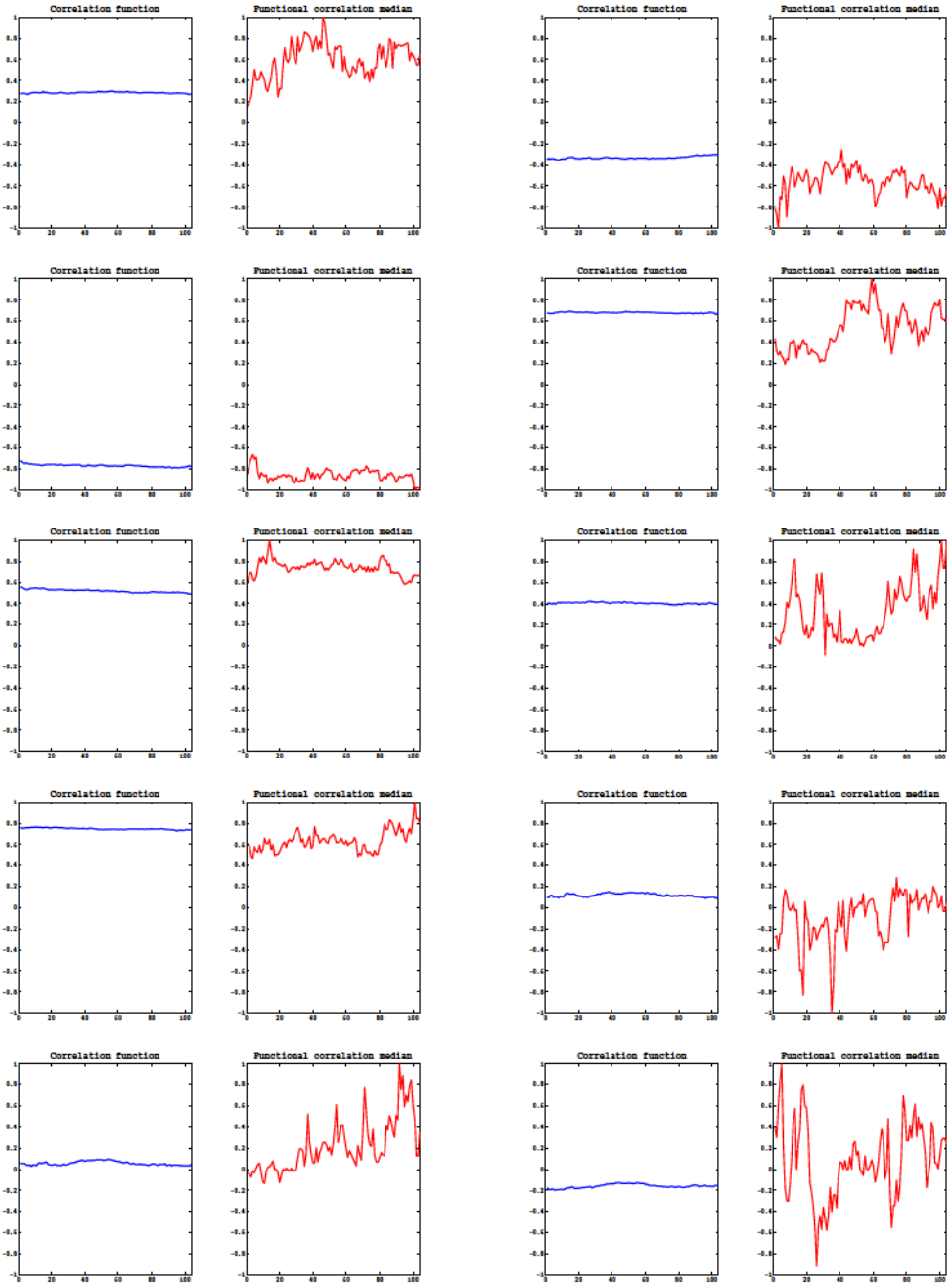

Figure 4.14: Correlation function and correlation median for functions for assets.

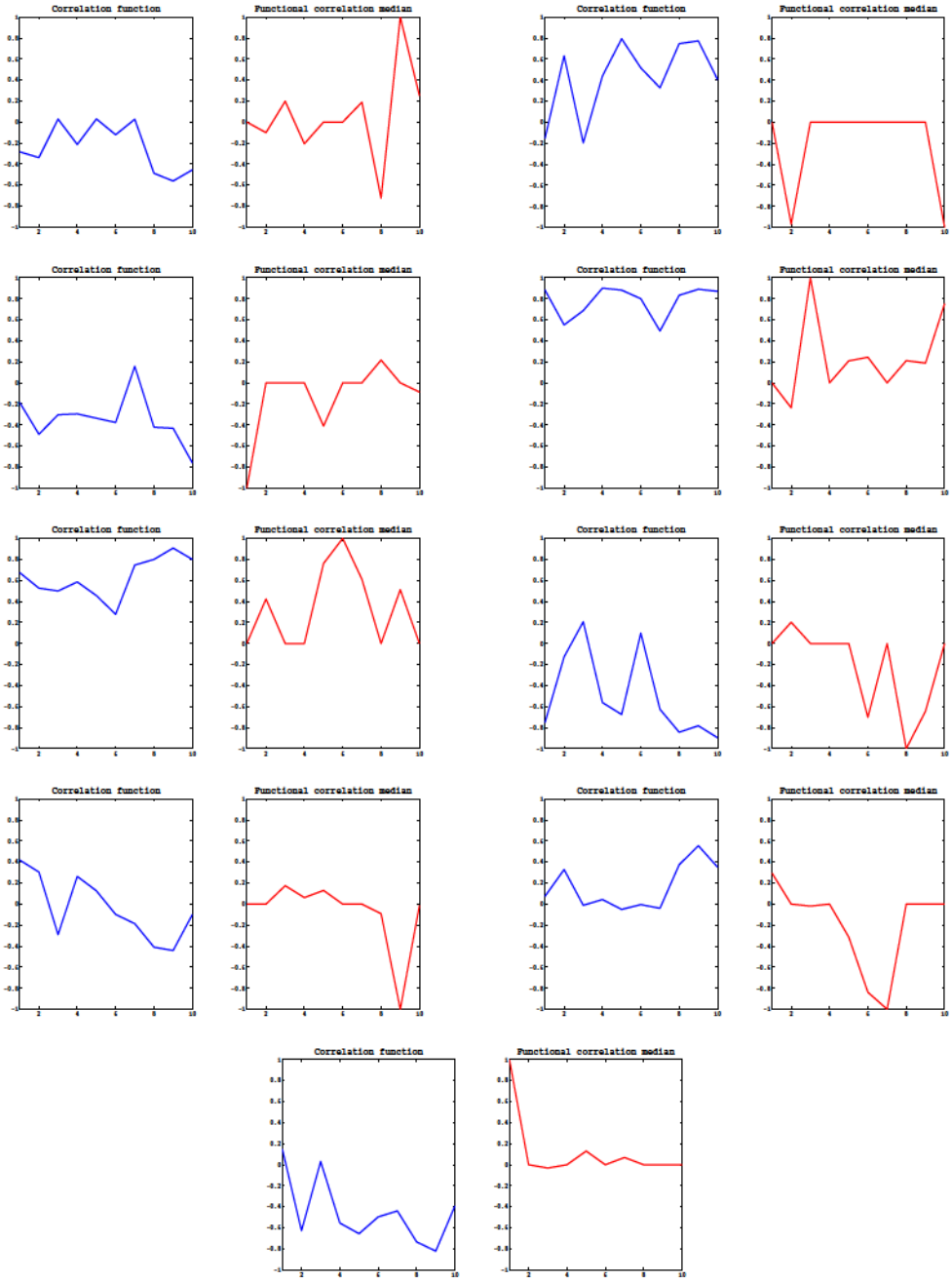

Figure 4.15: Correlation function and correlation median for functions for genes.

4.8 Conclusions

We have introduced a new correlation coefficient for functions providing a representative curve of dependence between two sets of curves. This coefficient is based on the cross-correlation function studied in Ramsay and Silverman [45], which is the classic Pearson coefficient between the values of the curves in different moments of time. To carry out this task, we extend the concepts of comedian and MAD analyzed in Falk [19] to the functional field. Since these concepts are based on the median of the data set, we considered the definition of median for functional data proposed in López-Pintado and Romo [38]. These new notions satisfy the usual properties, and appear to be more robust than other standard measures. The robustness of the new coefficient is illustrated with a simulation study. We present several simulated and real examples in order to show how the correlation median for functions works and compare the results obtained with the correlation function. We also study the sensitivity of our coefficient and conclude that it is stable with respect to sample size and to the number of points taken to discretize the functions.

A difficulty that we find is that in some cases the absolute value of the correlation median for functions can be larger than 1. Given that in such cases it is quite difficult to interpret the results, we propose an alternative by dividing the coefficient over the maximum between its maximum and minimum values. However, it would be interesting to look for alternatives that are related to the nature of the data. Another interesting question would be to study other measures of depth to calculate the functional median, and analyze which of them is more appropriate in terms of robustness to calculate correlation median for functions.

CHAPTER 5

Conclusions and main contributions

This chapter summarizes the main contributions of the thesis. We have developed new ways of measuring dependence in a bivariate sample of curves, inspired in some classic measures of dependence that are commonly used in bivariate data analysis. Basically, we have extended to the functional field the versions of three well-known coefficients for measuring dependence. They are Kendall and Spearman coefficients and a robust version of the Pearson correlation coefficient. Given a bivariate sample of functional data, the methodology implemented with the first two coefficients provides a single value that represents the degree of relationship between the sets of curves, while another provides a curve which will characterize the dependence on the whole interval where the two sets of curves are defined. Each of the contributions introduced in this dissertation are novel, and below we present the principal aspects developed in each chapter.

- Firstly, in Chapter 2, a functional version of Kendall's coefficient has been introduced which allows us to identify if there is some kind of dependence between two sets of curves. In the construction of this coefficient we have presented an alternative version of the concordance concept for two pairs of functions. Also, we define suitable orders for sorting functions. We highlight some relevant aspects of this new coefficient.

  i. The functional τ developed allows us to identify the global dependency between two groups of functional data, regardless of the shape of their paths.

  ii. We have taken into account some results of the functional analysis that have allowed us to see this new coefficient as a UB−statistics. Therefore, we have been able to provide some asymptotical results of the statistics considered for the sample version. This new way of measuring dependence between sets of functions satisfies

some classic properties that a dependence measure should fulfil; these properties have been proved.

iii. We show with simulated data that the functional τ works well in a set of curves where they are constructed with known dependence, even in the cases where the simulated data have been contaminated with different types of outliers. Also, two interesting examples with real data are studied. The first one corresponds to 33 companies belonging to the IBEX35; the functional τ allows us to obtain information about companies having similar behavior over time. The second data set corresponds to a micro-array time series from a human T-cell experiment. In this case, we obtain the partial functional τ for each pair of genes and a gene network can be constructed as an alternative to the ones existing in the literature.

- Secondly, in Chapter 3, we have developed a functional version of a rank correlation coefficient. Basically, this contribution is the extension of Spearman's coefficient in the functional context. As is well known, this measure in the bivariate context is defined by the Pearson coefficient among the rank of the data. Hence, in order to generalize it to functional data, we have introduced one way of assigning ranks to each one of the functions, which are based on some orderings for functions already studied in the literature. However, a main contribution here is the implementation of these orderings to the population case, allowing the assignation of grades to the stochastic processes where the curves come from. Therefore, a population version of the Spearman coefficient for stochastic processes is also introduced. The main results of this chapter are listed below.

  i. We provided a natural extension of the Spearman coefficient, which summarizes in one single-value the dependence between two sets of functions. It works in a way similar to the usual bivariate case, i.e., it calculates the Pearson coefficient between the ranks of the functions, which are obtained through two kinds of ordering for functions introduced in López-Pintado and Romo [39].

  ii. In Nelsen [41] the grades for random variables are defined. In this chapter we adapt the definition of those grades for stochastic processes. The new grades allowed us to introduce the population version of the Spearman coefficient for functions. We also present the asymptotical results of this new coefficient as well as its main properties.

  iii. Another contribution of this chapter is the definition of an independence test which is based on the bootstrap approach. It allows us to assess the significance of the association between two groups of curves and to quantify the statistical significance of some conjectures made on the basis of exploratory analysis. We have evaluated the power of this new test with simulated and real data and good conclusions can be drawn.

- Both functional versions of Kendall and Spearman provide just a single value for summarizing the dependence in two groups of curves. Although we have proved that they satisfy good properties, we know that in many cases one single value can be poor for measuring the dependence. Therefore, in Chapter 4, we tried to partially solve this problem by introducing a new measure of dependence between functions whose response is also a function. Our idea was inspired in a robust measure introduced in Falk [19] called comedian. We enumerate the main results of this chapter.

  i. We have introduced a correlation coefficient for functions that provides a representative curve of dependence between two sets of curves. The aim in this chapter was to introduce a robust alternative to the cross-correlation studied in Ramsay and Silverman [45]. Therefore, to do so, we generalized previously to the functional case some robust measures such as the MAD and the comedian, studied in Falk [19], which were presented as robust alternatives for the standard deviation and the covariance, respectively.

  ii. Some properties that in Falk [19] are considered relevant for his robust versions also are proved by us in the functional version introduced. An empirical study with contaminated functional data showed that the correlation median for functions has a more robust behavior than the correlation function.

  iii. An advantage of using the correlation median for functions instead of the correlation function is that, as we take the deepest curve of the product of all the curves centered around the median, the shape of the correlation median for functions curve represents the relation between the sets of curves better than the correlation function.

Future research lines

We now present some of the issues considered as future research lines and extensions of the work presented in this thesis.

- The methodologies implemented in this thesis to measure the dependence between set of curves were based mainly on some ordering definition for functions. However, those orders used to give our definitions belong to the set of the pre-orders. In fact, there are many other ways of sorting functions. Therefore, a research plan could be defining different pre-orders, for example, taking into account the arc length of each curve as well as its derivative. This would allow us to define new dependence measures where the shape of the curve is going to be considered.

- A difficulty that we found when introducing the correlation median for functions, was that in some specific cases its absolute value can be larger than 1. These cases are not easy to interpret and they are against statistical intuition of dependence. Hence, we proposed a way of improving those situations. However, it would be interesting to look for a kind of standardization appropriate for the correlation median for functions related to the nature of the data.

- Another interesting question to be investigated would be the behavior of other measures of depth for calculating the deepest curve, i.e, the functional median and analyzing which of them is more appropriate in terms of robustness.

- We also consider it of interest in the future to analyze the results obtained with the correlation median for functions when it is applied to a data set taken from different fields of science, for example: genetics, medicine, biology, and so on. The curve that represents the dependence could be a powerful tool to identify strong or weak relationships between paths of genes, continuous treatment of some diseases and their respective long-term evolution; or even the dependence through time of returns of firms belonging to different financial markets.

- Also of interest would be investigating more theoretical properties of the distributions of the coefficients introduced in this thesis; in particular, finding their asymptotical distributions for which a detailed study of U-statistics in Hilbert spaces would be necessary.

- We have defined the Spearman coefficient for functions through a functional version of grades. These grades can be seen as the population definition analogs of ranks. In this thesis, we introduced a form of assigning the grade to each of curves of the sample, and from the population point of view, we also gave an alternative for assigning grades to stochastic processes. We propose using these new versions of grades for functions in order to define a new measure that extends the Wilcoxon signed-rank test when functional data are considered, allowing us to evaluate whether their population mean ranks differ.

Bibliography

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

[2] H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics and Probability Letters*, 45:11–22, 1999.

[3] I. Cascos. Data depth: multivariate statistics and geometry. In *New Perspectives in Stochastic Geometry, Oxford University (W.S. Kendall and I. Molchanov Eds.)*, pages 398–423, Oxford University Press, 2010.

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

[5] J.A. Cuesta-Albertos, R. Fraiman, and T. Ransford. A sharp form of the cramér-wold theorem. *Journal of Theoretical Probability*, 20:201–209, 2007.

[6] A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Computational Statistics and Data Analysis*, 47:111–122, 2004.

[7] A. Cuevas, M. Febrero, and R. Fraiman. On the use of bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51:1063–1074, 2006.

[8] A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classication for functional data via projection-based depth notions. *Computational Statistics*, 22:481–496, 2007.

[9] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application.* Cambridge University Press, New York, 1997.

[10] P. Delicado. Functional k-sample problem when data are density functions. *Computational Statistics*, 22(3):391–410, 2007.

[11] J. A. Dubin and H. G. Müller. Dynamical correlation for multivariate longitudinal data. *Asymptotics in Statistics and Probability*, 100:872–881, 2005.

[12] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

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

[14] B. Efron. Local false discovery rates. Technical report, Department of Statistics, Stanford University, 2005.

[15] B. Efron and R. Tibshirani. *An introduction to the bootstrap.* Chapman and Hall, New York, 1993.

[16] I. Epifanio-López. Shape descriptors for classification of functional data. *Technometrics*, 50(3):284–294, 2008.

[17] M. Escabias, A. Aguilera, and M. Valderrama. Principal components estimation of functional logistic regression: Discussion of two different approaches. *Journal of Non-Parametric Statistics*, 16(3-4):365–384, 2004.

[18] A. Estepa, M. M. Gea, G. R. Cañadas, and J. M. Contreras. Algunas notas históricas sobre la correlación y regresión y su uso en el aula. *Números*, 81:5–14, 2012.

[19] M. Falk. On mad and comedians. *Annals of the Institute of Statistical Mathematics*, 49(4):615–644, 1997.

[20] M. Falk. A note on the comedian for elliptical distributions. *Journal of Multivariate Analysis*, 67:306–317, 1998.

[21] M. Febrero, P. Galeano, and W. González-Manteiga. Outlier detection in functional data by depth measures, with application to identify abnormal no_x levels. *Envirometrics*, 19:331–345, 2008.

[22] C. Fernández Vivas. Medidas de asociación y dependencia bivariante. *Trabajos de Estadística y de Investigación Operativa*, 34(2):25–39, 1983.

[23] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis : Theory and Practice.* Springer, New York, 2006.

[24] R. Fraiman and C. Muniz. Trimmed means for functional data. *Test*, 10(2):419–440, 2001.

[25] J.D. Gibbons. *Nonparametric Measures of Association.* Sage Publications, Newbury Park (California), 1993.

[26] J. Hauke and T. Kossowski. Comparison of values of pearson's and spearman's correlation coefficient on the same sets of data. *Quaestiones Geographicae*, 30(2):87–93, 2011.

[27] G. He, H. G. Müller, and J.L. Wang. Extending correlation and regression from multivariate to functional data. *Asymptotics in Statistics and Probability*, Edited by M. Puri:197–210, 2000.

[28] W. Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[29] J. Jacques and C. Preda. Model-based clustering of multivariate functional data. Preprint (2012).

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

[31] W.H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

[32] E. L. Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, 37(5):1137–1153, 1966.

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

[34] B. Li and Q. Yu. Classification of functional data: A segmentation approach. *Computational Statistics and Data Analysis*, 52(10):4790–4800, 2008.

[35] R. Li and M. Chow. Evaluation of reproducibility for paired functional data. *Journal of Multivariate Analysis*, 93:81–101, 2005.

[36] R.Y. Liu. On a notion of data depth based on random simplices. *Annals of Statistics*, 18:405–414, 1990.

[37] S. López-Pintado and J. Romo. Depth-based inference for functional data. *Computational Statistics and Data Analysis*, 51:4957–4968, 2007.

[38] S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

[39] S. López-Pintado and J. Romo. A half-region depth for functional data. *Computational Statistics and Data Analysis*, 55:1679–1695, 2011.

[40] B. Martin-Barragan, R. Lillo, and J. Romo. Functional boxplots based on half-regions. Preprint 2012.

[41] R. B. Nelsen. *An Introduction to Copulas 2nd Edition*. Springer Series in Statistics, New York, 2006.

[42] R. Opgen-Rhein and K. Strimmer. Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65, 2006.

[43] K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Royal Society*, 187:253–318, 1896.

[44] S. Pezulli and B. Silverman. Some properties of smoothed principal components analysis for functional data. *Computational Statistics*, 8:1–16, 1993.

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

[46] P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283, 1993.

[47] M. Scarsini. On measures of concordance. *Stochastica*, 8(3):201–218, 1984.

[48] S. Schwabik and Y. Guoju. *Topics in Banch Space Integration.* World Scientific Publishing, Singapore, 2005.

[49] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[50] M. D. Taylor. Multivariate measures of concordance. *Annals of the Institute of Statistical Mathematics*, 59:789–806, 2007.

[51] M. D. Taylor. Some properties of multivariate measures of concordance. Technical report, 2008.

[52] J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531. Canad. Math. Congress, Montreal, Que., 1975.

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

[54] R. R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing, 3rd edition.* Elsevier, San Diego, 2012.

[55] W. Xu, Y. Hou, Y. Hung, and Y. Zou. Comparison of spearman's rho and kendall's tau in normal and contaminated normal models. Technical report, 2010.

[56] Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):pp. 461–482, 2000.