Trabajo Fin de Máster

Clasificación supervisada funcional aplicada al diagnóstico de la enfermedad de Parkinson

Gisselle J. Zorrilla Green

Máster en Técnicas Estadísticas

Curso 2018-2019

Propuesta de Trabajo Fin de Máster

Título en galego: Clasificación funcional supervisada aplicada ao diagnóstico da enfermidade de Parkinson

Título en español: Clasificación supervisada funcional aplicada al diagnóstico de la enfermedad de Parkinson

English title: Functional supervised classification applied to diagnosis of Parkinson's disease

Modalidad: Modalidad A

Autor/a: Gisselle J. Zorrilla Green, Universidade da Coruña

Director/a: Graciela Estévez Pérez, Universidade da Coruña

Breve resumen del trabajo:

Los métodos de clasificación supervisada o discriminación para datos funcionales tienen como principal objetivo proporcionar reglas de clasificación de curvas en uno de los k posibles grupos o categorías definidos a priori. En las dos últimas décadas, diversos métodos de discriminación han sido propuestos en la literatura, entre los que podemos destacar el de [29] que adapta la técnica clásica de análisis discriminante lineal al contexto funcional o el de [18] que proporciona una regla de clasificación Bayes basada en el estimador tipo núcleo de la probabilidad a posteriori. Más recientemente, [41] abordan el problema de clasificación supervisada para datos funcionales considerando procedimientos basados en profundidades combinados con diversas medidas de profundidad. Los objetivos de esta propuesta de TFM son los siguientes: (1) Realizar una revisión de los métodos de clasificación supervisada para datos funcionales existentes en la literatura, prestando especial atención al caso particular de los test de diagnóstico ($k = 2$); (2) Desarrollar software con el propósito de comparar los métodos anteriores mediante un estudio de simulación; (3) Abordar el problema de discriminación entre afectados y no afectados de Parkinson partiendo de datos funcionales del ritmo motor derivados del Finger Tapping Test [2].

Doña Graciela Estévez Pérez, Profesora Tirular de Universidad de la Universidade da Coruña, informa que el Trabajo Fin de Máster titulado

Clasificación supervisada funcional aplicada al diagnóstico de la enfermedad de Parkinson

fue realizado bajo su dirección por doña Gisselle J. Zorrilla Green para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, da su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, 5 de septiembre de 2019.

La directora:

ESTEVEZ PEREZ MARIA GRACIELA - DNI 34872364V

Firmado digitalmente por ESTEVEZ PEREZ MARIA GRACIELA - DNI 34872364V
Nombre de reconocimiento (DN): c=ES, o=UNIVERSIDAD DE A CORUÑA, ou=CERTIFICADO ELECTRONICO DE EMPLEADO PUBLICO, ou=MATEMATICAS, ou=486, serialNumber=IDCES-34872364V, sn=ESTEVEZ PEREZ, givenName=MARIA GRACIELA, cn=ESTEVEZ PEREZ MARIA GRACIELA - DNI 34872364V
Fecha: 2019.09.05 22:38:18 +02'00'

Doña Graciela Estévez Pérez

La autora:

Doña Gisselle J. Zorrilla Green

Agradecimientos

A mi tutora por la paciencia, las recomendaciones, la motivación y el apoyo constante; a mis maestros por compartir el conocimiento aún cuando las condiciones en alguna ocasión no fueron las mas favorables; a Coruña por la inolvidable experiencia y, sobre todo, a mis familiares que a lo largo de este camino permanecieron a mi lado sin importar la hora o lo cansados que estuvieran.

.

Índice General

Índice de Figuras

Índice de Tablas

Resumen

Resumen en español

Los métodos de clasificación supervisada o discriminación para datos funcionales tienen como principal objetivo proporcionar reglas de clasificación de curvas en uno de los G posibles grupos o categorías definidos a priori. En las dos últimas décadas, diversos métodos de discriminación han sido propuestos en la literatura, entre los que podemos destacar el de [29] que adapta la técnica clásica de análisis discriminante lineal al contexto funcional o el de [18] que proporciona una regla de clasificación Bayes basada en el estimador tipo núcleo de la probabilidad a posteriori. Más recientemente, [41] abordan el problema de clasificación supervisada para datos funcionales considerando procedimientos basados en profundidades combinados con diversas medidas de profundidad.

Los objetivos que se persiguen en este Trabajo de Fin de Master son los siguientes: (1) Realizar una revisión de los métodos de clasificación supervisada para datos funcionales existentes en la literatura, prestando especial atención al caso particular de los test de diagnóstico ($G = 2$); (2) Desarrollar software con el propósito de comparar los métodos anteriores mediante un estudio de simulación; (3) Abordar el problema de discriminación entre afectados y no afectados de Parkinson partiendo de datos funcionales del ritmo motor derivados del Finger Tapping Test [2].

Resumo en galego

Os métodos de clasificación supervisada ou discriminación para datos funcionais teñen como principal obxectivo proporcionar regras de clasificación de curvas

nun dos G posibles grupos ou categorías definidos a priori. Nas dúas últimas décadas, diversos métodos de discriminación foron propostos na literatura, entre os que podemos destacar o de [29] que adapta a técnica clásica de análise discriminante lineal ao contexto funcional ou o de [18] que proporciona unha regra de clasificación Bayes baseada no estimador tipo núcleo da probabilidade a posteriori. Máis recentemente, [41] abordan o problema de clasificación supervisada para datos funcionais considerando procedementos baseados en profundidades combinados con diversas medidas de profundidade.

Os obxectivos que se perseguen neste Traballo de Fin de Master son os seguintes: (1) Realizar unha revisión dos métodos de clasificación supervisada para datos funcionais existentes na literatura, prestando especial atención ao caso particular dos test de diagnóstico ($G = 2$); (2) Desenvolver software co propósito de comparar os métodos anteriores mediante un estudo de simulación; (3) Abordar o problema de discriminación entre afectados e non afectados de Parkinson partindo de datos funcionais do ritmo motor derivados do Finger Tapping Test [2].

English abstract

The supervised classification or discrimination for functional data methods have as main objective to provide rules for classifying curves into one of the possible G groups or categories defined a priori. In the last two decades, various methods of discrimination have been proposed in the literature, among which we can highlight that of [29] that adapts the classical technique of linear discriminant analysis to the functional context or that of [18] that provides a Bayes classification rule based on the kernel estimator of the posterior probability. More recently, [41] addresses the problem of supervised classification for functional data considering depth-based procedures combined with various depth measurements.

The aims pursued in this Master's Thesis are as follows: (1) Perform a review of supervised classification methods for functional data existing in the literature, paying special attention to the particular case of diagnostic tests ($G = 2$); (2) Develop software with the purpose of comparing the above methods through a simulation study; (3) Address the problem of discrimination between those affected and unaf-

fected by Parkinson's starting from functional motor rhythm data derived from the Finger Tapping Test [2].

Prefacio

El Trabajo de Fin de Máster que se presenta a continuación tiene como objetivo aplicar los conocimientos adquiridos a lo largo del Máster Interuniversitario en Técnicas Estadísticas organizado por la Universidade de Santiago de Compostela, Universidade da Coruña y Universidade de Vigo, aplicándolos en el campo particular del Análisis de Datos Funcionales (FDA), concretamente en la Clasificación Funcional Supervisada.

En las últimas décadas, las técnicas de FDA han experimentado un rápido desarrollo, lo qué ha permitido que el FDA haya alcanzado una madurez metodológica notable. Muchos de los métodos estadísticos usuales tienen su versión para datos funcionales: técnicas de regresión, métodosde análisis multivariante y procedimientos de análisis par series temporales, entre otros. En paralelo, sus métodos se han ido aplicando ampliamente a múltiples campos como medicina, ciencia, negocios, ingeniería, demografía, ciencias sociales, etc. Por su parte, en el ámbito de la clasificación supervisada, también denominada discriminación, para datos funcionales también se han advertido importantes avances en las dos últimas décadas. Desde trabajos como el de [29], que adapta la técnica clásica de análisis discriminante lineal al contexto funcional o el de [18], que proporciona una regla de clasificación Bayes basada en el estimador tipo núcleo de la probabilidad a posteriori, hasta los de [41] y [8] que abordan el problema de clasificación supervisada para datos funcionales considerando procedimientos basados en el concepto de profundidad.

El objetivo del presente TFM es triple: (1) Realizar una revisión de la metodología de clasificación supervisada para datos funcionales existente en la literatura; (2) Desarrollar software con el propósito de comparar distintos métodos discriminantes mediante un estudio de simulación; (3) Aplicar esta metodología a un conjunto

de datos funcionales reales. La memoria se estructura del siguiente modo. En el Capítulo 1 se introducen los conceptos generales en el estudio de datos funcionales, resumiendo los aspectos más importantes a tomarse en consideración a lo largo de la memoria. Con el objetivo de realizar una revisión de los métodos de clasificación supervisada para datos funcionales existentes en la literatura, se plantea el Capítulo 2, en el que se seleccionan aquellos procedimientos que han demostrado un mejor desempeño desde el punto de vista teórico y/o aplicado. En el Capítulo 3, se diseña un completo estudio de simulación con el propósito de comparar el desempeño de los métodos seleccionados en el capítulo anterior. El Capítulo 4 aborda el problema de discriminación entre afectados y no afectados de Parkinson partiendo de datos funcionales del ritmo motor derivados del Finger Tapping Test (FTT). El trabajo finaliza con una pequeña sección en la que se presentan las conclusiones principales que se pudieron derivar del análisis realizado.

Capítulo 1

Análisis de Datos Funcionales

En este capítulo se realiza un resumen de los conceptos básicos en Datos Funcionales y, simultáneamente, se introduce parte de la notación matemática que se utilizará a lo largo del trabajo.

1.1. Introducción al análisis de datos funcionales

Los avances tecnológicos de las últimas décadas están permitiendo observar en distintos campos, como la Economía, la Medicina o la Ingeniería, datos con estructuras complejas, consistentes en secuencias de medidas de cierta característica de interés sobre un soporte continuo, típicamente, el tiempo o el espacio. Este tipo de datos se conocen como Datos Funcionales, es decir, realizaciones de elementos aleatorios que toman valores en un espacio de dimensión infinita. De forma más precisa, en [17] se propone la siguiente definición general:

Definición 1: Una variable aleatoria (v.a.) $\mathcal{X}$ definida sobre un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ se denomina variable funcional (o variable aleatoria funcional) si toma valores en un espacio de dimensión infinita (o espacio funcional) $\mathcal{F}$. Cada observación χ de $\mathcal{X}$ se denomina dato funcional.

Nótese que la noción de variable funcional incluye el caso de curvas aleatorias pero también situaciones más complejas como superficies aleatorias u otros objetos matemáticos de dimensión infinita. A lo largo de este trabajo las variables funcionales, y por tanto los datos funcionales manejados, serán curvas aleatorias, de modo

que $\mathcal{X}$ será interpretado como un proceso estocástico $\mathcal{X} = \{\mathcal{X}(t); t \in T\}$, siendo T un intervalo de $\mathbb{R}$ (sin pérdida de generalidad consideraremos $T = [0, 1]$).

En la práctica, un dato funcional $\chi(t)$ siempre es observado en un número finito de puntos muestrales, es decir, $\chi(t) = (\chi(t_1), ..., \chi(t_L))$, donde $t = (t_1, ..., t_L)$ es el conjunto de puntos de dominio donde se ha medido $\chi(t)$. Sin embargo, tal como se indica en [41], incluso siendo $\chi(t)$ un vector, hay al menos tres razones por las que los datos funcionales no deben ser tratados como datos multivariantes:

1. El conjunto de puntos de dominio donde se miden los datos funcionales de una muestra puede variar en tamaño y/o elementos entre los distintos datos funcionales.

2. Cada curva es una realización de un proceso estocástico con una cierta estructura de dependencia (los datos funcionales normalmente son autocorrelacionados) y los procedimientos multivariantes estándar generalmente fallan en presencia de autocorrelación.

3. Las muestras funcionales pueden contener menos curvas que los puntos de evaluación (en la jerga multivariante, menos filas que columnas), situación difícil de manejar con técnicas multivariantes.

Como respuesta a esta dificultad, surge el Análisis de Datos Funcionales, conocido comúnmente por sus siglas en inglés FDA (Functional Data Analysis), como área de la Estadística que trata con las herramientas de análisis para este tipo de datos. El gran desarrollo que ha experimentado esta disciplina en las dos últimas décadas se pone de manifiesto en la vasta literatura disponible, entre la que podemos destacar los libros [38] y [17], que tratan muchos de los problemas básicos de la estadística funcional, desde un punto de vista paramétrico y no paramétrico, respectivamente. Más recientemente, [27] se centra en la inferencia y teoría asintótica para datos funcionales, mientras que [12] y [1] presentan revisiones recientes sobre el FDA y temas relacionados.

1.2. Conceptos Básicos

Una de las cuestiones fundamentales del FDA, es la delimitación del espacio funcional en el cual se pretende representar los datos, de manera que sea posible obtener la forma funcional de las curvas a partir de sus observaciones en dominio discreto.

Definición 2: Un espacio funcional $\mathcal{F}$ se dice que es un **espacio métrico** si en $\mathcal{F}$ existe una aplicación $d : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$ (denominada distancia) tal que, para cualesquiera elementos $\chi, \chi', \chi'' \in \mathcal{F}$ se cumple:

1. $d(\chi, \chi') \geq 0$ y $d(\chi, \chi') = 0$ si y sólo si $\chi = \chi'$
2. $d(\chi, \chi') = d(\chi', \chi)$ (Simetría)
3. $d(\chi, \chi') \leq d(\chi, \chi'') + d(\chi'', \chi')$ (Desigualdad Triangular)

Nótese que si la aplicación d no verifica la primera propiedad ($d(\chi, \chi') = 0 \Rightarrow \chi = \chi'$), entonces diremos que d es una semi-métrica y, en consecuencia, $\mathcal{F}$ un espacio semi-métrico.

Definición 3: Un espacio funcional $\mathcal{F}$ se dice que es un **espacio normado** sobre un cuerpo $\mathbb{K}$ si en $\mathcal{F}$ existe una aplicación $\|\cdot\| : \mathcal{F} \longrightarrow \mathbb{R}$ (denominada norma) tal que, para cualesquiera elementos $\chi, \chi' \in \mathcal{F}$ se cumple:

1. $\|\chi\| \geq 0$ y $\|\chi\| = 0$ si y sólo si $\chi = 0$
2. $\forall \lambda \in \mathbb{K}, \|\lambda\chi\| = |\lambda|\|\chi\|$
3. $\|\chi + \chi'\| \leq \|\chi\| + \|\chi'\|$ (Desigualdad Triangular)

Nótese que si la aplicación $\|\cdot\|$ sólo verifica la segunda y tercera propiedad, entonces diremos que $\|\cdot\|$ es una semi-métrica y, en consecuencia, $\mathcal{F}$ un espacio semi-normado.

Todo espacio normado se puede convertir en un espacio métrico con la distancia inducida por la norma $d(\chi, \chi') = \|\chi - \chi'\|$. Además, si el espacio es completo se dice que es un espacio de Banach.

Definición 4: Un espacio funcional $\mathcal{F}$ se dice que es un **espacio euclídeo** sobre un cuerpo $\mathbb{K}$ si en $\mathcal{F}$ existe una aplicación $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{K}$ (denominada producto escalar) tal que, para cualesquiera elementos $\chi, \chi', \chi'' \in \mathcal{F}$ se cumple:

1. $\langle \chi, \chi \rangle \geq 0$ y $\langle \chi, \chi \rangle = 0$ si y sólo si $\chi = 0$
2. $\forall a, b \in \mathbb{K}, \langle a\chi + b\chi', \chi'' \rangle = a\langle \chi, \chi'' \rangle + b\langle \chi', \chi'' \rangle$ (Linealidad por la derecha. Análogamente linealidad por la izquierda)
3. $\langle \chi, \chi' \rangle = \overline{\langle \chi', \chi \rangle}$ (Hermiticidad)

Nótese que cualquier espacio euclídeo se puede convertir en un espacio normado sin más que considerar la norma inducida por el producto escalar. Además, si el espacio es completo se dice que es un espacio de Hilbert. Normalmente interesa trabajar con espacios de Hilbert, aunque no siempre es posible. Uno de los espacios de funciones más utilizados en el contexto de FDA son los espacios $\mathcal{L}_p(T)$.

Definición 5: Dado $0 < p < \infty$, se define el espacio de funciones $\mathcal{L}_p(T)$ como el conjunto de funciones con dominio en el intervalo real T cuya potencia p-ésima es integrable:

$$\mathcal{L}_p(T) = \{\mathcal{X} : T \longrightarrow \mathbb{R}, \text{tal que} \int_T |\mathcal{X}(t)|^p \, dt < \infty\}$$

Esta clase de espacio es frecuentemente utilizado cuando el área entre las curvas puede proporcionar información útil del comportamiento de los datos. El caso particular $\mathcal{L}_2(T)$ es un espacio de Hilbert. Además, al ser separable, dispone de bases ortonormales que ofrece muchas posibilidades a la hora de diseñar procedimientos estadísticos. En general, la representación de un dato funcional en una base ortonormal servirá de puente entre la discretización del dato funcional y su forma funcional.

Por otra parte, en el contexto de datos funcionales es crucial la elección de una norma, sin embargo la consideración de espacios métricos o normados puede ser excesivamente restrictiva. En algunos casos, los espacios semi-métricos pueden resultar más apropiados que los métricos para analizar ciertos conjuntos de datos funcionales, dependiendo de la forma de los mismos, de variables exógenas o del objetivo del análisis estadístico. Por ejemplo, si se piensa que el cambio de escala es

informativo, se puede tomar $\mathcal{L}_2(T)$ como espacio de referencia, pero si la información relevante está en la curvatura puede que una semi-norma basada en derivadas resulte más útil.

A continuación se citan algunas de las familias de semi-métricas mejor adaptadas para trabajar con datos funcionales que utilizaremos en este trabajo, cuya información puede ser ampliada en la referencia [17] de la bibliografía.

Basada en Análisis de Componentes Principales Funcionales (FPCA) El análisis de componentes principales, en datos multivariantes, permite la representación de la información en un número reducido de dimensiones, mediante la construcción de nuevas variables que contienen la mayor parte de la información de los datos en estudio. Estos métodos, al ser aplicados al contexto de datos funcionales, demostraron ser eficientes para calcular distancias o proximidades entre curvas. En efecto, dado un dato funcional χ, el FPCA nos proporciona la siguiente expansión:

$$\chi = \sum_{k=1}^{\infty} \left(\int \chi(t) \upsilon_k(t) dt \right) \upsilon_k,$$

siendo los υ_k los autovectores del operador covarianza ordenados decrecientemente por autovalor. Truncando esta expansión hasta dimensión q se puede definir la siguiente familia parametrizada de semi-métricas:

$$d_q^{PCA}(\chi_i, \chi_{i'}) = \sqrt{\sum_{k=1}^{q} \left(\int \left[\chi_i(t) - \chi_{i'}(t) \right] \upsilon_k(t) dt \right)^2},$$

Puesto que en la práctica de un dato funcional χ_i únicamente se observará una versión discretizada $\chi_i(t) = (\chi_i(t_1), ..., \chi_i(t_L))$, la distancia entre las curvas χ_i y $\chi_{i'}$, $d_q^{PCA}(\chi_i, \chi_{i'})$, será aproximada por su versión empírica:

$$d_q^{PCA}(\chi_i, \chi_{i'}) = \sqrt{\sum_{k=1}^{q} \left(\sum_{l=1}^{L} w_l (\chi_i(t_l) - \chi_{i'}(t_l)) [\upsilon_k]_l \right)^2}, \tag{1.1}$$

donde q representa el número de componentes principales retenidos, J la can-

tidad de registros por los que está conformado cada curva y los w_l los pesos que permiten aproximar la integral.

Nótese que estas semi-métricas sólo pueden utilizarse con datos medidos en los mismos puntos y tomados de una partición suficientemente fina para que los estimadores empíricos sean consistentes. En cambio, permiten trabajar con curvas rugosas".

Basada en regresión múltiple de mínimos cuadrados parciales (MPLSR)

Como se explica en [17], la regresión multivariante de mínimos cuadrados parciales fue desarrollada para predecir una respuesta multivariada a partir de variables independientes cuando se está en presencia de un elevado grado de colinealidad entre los predictores y/o cuando existe un número alto de predictores en comparación con el número de observaciones, lo que sucede en el caso que nos ocupa.

Al igual que el PCA, el enfoque PLS permite construir una clase de semimétricas y así obtener una herramienta para calcular la proximidad existente entre dos curvas discretizadas. Esta distancia puede aproximarse mediante la expresión:

$$d_q^{PLS}(\chi_i, \chi_{i'}) = \sqrt{\sum_{k=1}^{p} \left(\sum_{l=1}^{L} w_l(\chi_i(t_l) - \chi_{i'}(t_l))[\upsilon_k^q]_l \right)^2}, \tag{1.2}$$

donde q el número de factores, p la cantidad de respuestas escalares o componentes retenidos, L la cantidad de registros por los que está conformado cada curva y w_l los pesos que permiten aproximar la integral y υ_l^q son los vectores $\mathbb{R}^L$ calculados por MPLSR.

Tal como se indica en [17] esta familia de semi-métricas permite obtener muy buenos resultados en problemas de regresión y clasificación supervisada. Al igual que la familia FPCA, sólo puede utilizarse con datos medidos en los mismos puntos y tomados de una partición suficientemente fina, pudiendo también emplearse para trabajar con curvas rugosas".

Basadas en derivadas Otra manera de medir la proximidad entre curvas es con-

siderar la distancia existente entre alguna de sus derivadas. Esta familia de semimétricas se puede parametrizar de la siguiente forma:

$$d_q^{deriv}(\chi_i, \chi_{i'}) = \sqrt{\int (\chi_i^{(q)}(t) - \chi_{i'}^{(q)}(t))^2 dt}, \tag{1.3}$$

donde $\chi_i^{(q)}$ denota la derivada *q-ésima* de χ_i, con $q = 0$ se obtendría la norma usual en L^2.

Este tipo de semi-métricas son apropiadas cuando interese captar información relevante de las curvas presente en su curvatura y suelen requerir un trabajo previo para la estimación de las derivadas.

1.3. Representación de Datos Funcionales

En el FDA generalmente se contará con datos discretos medidos a lo largo del tiempo, en lugar de conocer la forma explícita de cada una de las curvas. Por este motivo, es necesario utilizar alguna técnica o procedimiento que permita obtener una buena estimación de la función de la cual provienen los datos, tomando en cuenta que puede haber contaminación o distorsión de los valores ocasionada por ruido u otras fuentes de error. Además, es importante hacer notar que los datos discretos no necesariamente tienen que estar igualmente espaciados ni tiene porque tener el mismo número de registros.

Existen diferentes formas de pre-procesar los datos para construir observaciones puramente funcionales, lo que representa una cuestión importante debido al impacto que este proceso puede tener en el análisis estadístico posterior. Si se asume que los valores discretos no contienen error, el proceso a seguir es la **intepolación**, que consiste en la unión de los puntos de observación de los datos por medio de una curva suave.

Por otra parte, si en los valores discretos se ha observado algún tipo de error que sea necesario eliminar, entonces el proceso de convertir los datos a una función requerirá de alguna clase de **suavizado** (ver [38]). Mediante este procedimiento se asume que la función $\mathcal{X}$ posee una o más derivadas, lo que permite asegurar que las curvas

sean continuas durante todo el dominio de definición. En la literatura encontramos diferentes técnicas de suavizado, algunas de ellas basadas en modelos de regresión, como en [25] y [6].

Otra forma de pre-procesar los datos consiste en representarlos en una base de funciones. En efecto, dada $\{\phi_k\}_{k\in\mathbb{N}}$ una base del espacio funcional $\mathcal{F}$, cualquier función (dato funcional) $\chi \in \mathcal{F}$ puede aproximarse por una combinación lineal finita dada por:

$$\chi(t) \approx \sum_{k=1}^{K} c_k \phi_k(t); t \in T,$$

donde, K denota el número de funciones base utilizadas y c_k los coeficientes que serán escogidos de acuerdo a algún criterio, por ejemplo, mínimos cuadrados.

La elección del valor de K y del tipo de base a utilizar, que será de gran relevancia, dependerá de la naturaleza de los datos. Algunos de los sistemas de bases más utilizadas son: *Bspline*, que utilizan las funciones splines para aproximar los datos funcionales; *Base de Fourier*, cuyos elementos son funciones trigonométricas y se recomienda su uso para datos periódicos; *Base de Wavelets*, que a diferencia de la transformada de Fourier, no se asume que los datos sean periódicos y permite alcanzar una aproximación adecuada de la forma funcional de los datos utilizando muchas menos funciones básicas.

Para mayor detalle de este y otros tipos de bases, consultar [38].

A lo largo del desarrollo de los Capítulos 3 y 4 de este trabajo, se empleará el método de regresión lineal local con un ancho de banda común para todas las curvas, según lo planteado en [21], para suavizar la trayectoria de los conjuntos de datos.

1.4. Análisis Exploratorio de Datos Funcionales

Una vez realizado el pre-procesado de los datos funcionales, el siguiente paso, independientemente del objetivo del estudio estadístico, es el análisis exploratorio de los mismos, que permitirá poner de manifiesto sus características más relevantes. Además de una descripción estadística de la muestra de datos, es fundamental obte-

ner una buena visualización de las curvas o de alguna transformación de las mismas, como por ejemplo, de sus derivadas de distinto orden. Las distintas herramientas gráficas posibilitarán observar características de posición o variabiliad, relaciones con otras variables de interés, patrones de comportamiento o presencia de atípicos, por ejemplo.

A continuación se definen algunas de las medidas características de variables aleatorias, más comúnmente utilizadas en la práctica, que han sido adaptadas al marco de los datos funcionales y que serán referidas a lo largo del trabajo. Para más detalle, ver [17].

1.4.1. Medidas de centralización y dispersión

Consideremos $\mathcal{X}$ una variable aleatoria funcional que toma valores en el espacio semi-métrico $(\mathcal{F}, d)$. Sea $\mathcal{X}_1, ..., \mathcal{X}_n$ una muestra de variables independientes e idénticamente distribuidas a $\mathcal{X}$, χ un elemento fijo de $\mathcal{F}$ y $\chi_1, ..., \chi_n$ un conjunto de datos funcionales asociados a la muestra funcional $\mathcal{X}_1, ..., \mathcal{X}_n$.

Al igual que en el caso unidimensional, una de las medidas de tendencia central más simple y más popular para datos funcionales es la **media**, que se define, desde un punto de vista formal como:

$$\mathbb{E}(\mathcal{X}) = \int_{\Omega} \mathcal{X}(\omega) dP(\omega),$$

donde $(\Omega, \mathcal{A}, P)$ es el espacio de probabilidad de la v.a. $\mathcal{X}$. Como estimador de la media podemos considerar su versión empírica, que en el caso de curvas aleatorias da lugar a la noción de curva media, que debe ser utilizada con cautela dependiendo de la forma de los datos:

$$\overline{\mathcal{X}}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_i(t); \forall t \in T \subset \mathbb{R}.$$

Como en el caso univariante, uno de los problemas fundamentales de la media es su falta que robustez por lo que se necesita extender el concepto de mediana al caso funcional. Para ello, [17] introducen la siguiente definición.

La **mediana** funcional de la v.a. $\mathcal{X}$ asociada a la semi-métrica d se obtiene como solución al problema de minimización:

$$\inf_{\chi \in \mathcal{F}} \mathbb{E}(d(\chi, \mathcal{X})).$$

Un estimador empírico de la mediana funcional se obtiene mediante:

$$\mathcal{X}_{med} = \inf_{\chi \in \mathcal{F}} \sum_{i=1}^{n} d(\chi, \mathcal{X}_i).$$

Finalizamos la sección introduciendo la versión funcional de la clásica medida de dispersión, la **varianza** de la v.a. funcional $\mathcal{X}$:

$$Var(\mathcal{X}) = \mathbb{E}[(\mathcal{X} - \mathbb{E}(\mathcal{X}))^2],$$

cuya versión muestral se obtiene mediante la expresión:

$$Var(\mathcal{X}(t)) = \frac{1}{n-1} \sum_{i=1}^{n} [\mathcal{X}_i(t) - \overline{\mathcal{X}}(t)]^2; \forall t \in T \subset \mathbb{R}.$$

1.4.2. Medidas de Profundidad

La idea de profundidad de datos surgió en el contexto multivariante en un intento de extender la noción de estadísticos de orden a espacios multidimensionales. Los estadísticos de orden univariantes permiten ordenar los datos de la observación más pequeña a la más grande y evaluar el grado de centralización de un punto en relación a la distribución de probabilidad. De acuerdo con [40], una profundidad multivariante es una función que mide cuan profundo (o central) está un punto del espacio en relación con la distribución de probabilidad de la variable en dicho espacio multivariante. Diversos conceptos de profundidad multivariada han sido propuestas en la literatura, pudiendo destacar como obras de consulta [32] y [42].

En la última década, se han propuesto diversas medidas de profundidad en el contexto funcional con el propósito de estudiar el grado de centralidad de una función respecto a un conjunto de funciones o a una distribución de probabilidad funcional, lo que permitirá obtener un criterio de orden "del centro hacia fuera". A continuación

se presentan algunas de las versiones de profundidad funcional más comunes y que serán utilizadas a lo largo del presente trabajo.

Profundidad de Fraiman y Muniz (FMD) (ver [20]). Fue una de las primeras generalizaciones de profundidad para datos funcionales. Se introduce un concepto de profundidad integrada, la cual se basa en las integrales de las profundidades univariadas a lo largo del dominio de definición.

Profundidad modal (MD) (ver [11]). También denominada profundidad basada en núcleos, establece que la profundidad modal de un punto χ viene dada por la función:

$$f_h(\chi) = E(K_h(\|\chi - \mathcal{X}\|)),$$

donde $\mathcal{X}$ es la v.a. funcional, $\|.\|$ es la semi-norma definida en $\mathcal{F}$, $K_h(t)$ es un tipo de Kernel reescalado y h es un parámetro de ajuste fijo. De lo anterior se desprende que la *profundidad modal* h de X sea el valor mas profundo de z al maximizar en z *la función anterior*. En otras palabras, podemos decir que, la profundidad modal que se define como:

$$MD(x_i) = \sum_{j=1}^{n} K\left(\frac{d(x_i, x_j)}{h}\right)$$

donde K es un núcleo asimétrico y h un parámetro ventana. Según esta medida de profundidad, la curva más profunda equivaldrá a una curva de máxima densidad, es decir, a la curva "más rodeada".

Proyecciones Aleatorias (RP) (ver [11]). Se propone una medida de profundidad basada en Proyecciones Aleatorias que consiste en, dada una muestra $\mathcal{X}_1, ..., \mathcal{X}_n$ de la v.a. $\mathcal{X}$ en un espacio funcional $\mathcal{F}$ con producto escalar $\langle ., . \rangle$, seleccionar una dirección aleatoria a y sobre esta proyectar los datos, de manera que la proyección de un punto χ viene dada por $\langle a, \chi \rangle = \int_T a(t)\chi(t)dt$. El métodos RP2 consiste en utilizar el método de Proyecciones Aleatorias tanto para las funciones como para sus derivadas, de manera que se obtienen datos

funcionales multivariantes y se utiliza esta información en la función suavizada. Por otra parte, el método RPD consiste en utilizar el procedimiento de Profundidad Modal (MD) para evaluar la profundidad de los datos de la muestra bidimensional proyectados.

Otros tipos de profundidad funcional se encuentran en: [9], [34], [41] y [13].

Capítulo 2

Clasificación supervisada

El problema de clasificación en estadística consiste en dividir una colección de objetos en grupos homogéneos considerando dos enfoques diferentes. En el primero, se parte de que las categorías o grupos son conocidos y el objetivo es determinar a cual de ellos pertenece la nueva observación. A este tipo de técnicas se le conoce con el nombre de Análisis Discriminante o Clasificación Supervisada. Por otra parte, cuando los grupos de pertenencia no son establecidos a priori y el interés es definirlos mediante algún procedimiento estadístico, se emplean otros tipos de técnicas conocidas como Análisis Clúster o Clasificación No Supervisada.

Ambas técnicas de clasificación han sido ampliamente estudiadas en el contexto de datos multivariantes (los objetos pertenecen a un espacio multidimensional) desarrollándose diversos métodos. A medida que el análisis de datos funcionales fue cobrando relevancia, muchas de estas metodologías clásicas fueron extendidas al contexto de datos funcionales (los objetos pertenecen a un espacio infinito dimensional), surgiendo además nuevos métodos específicamente desarrollados en el contexto FDA.

En este trabajo nos centraremos en los métodos de Clasificación Supervisada aplicados a datos funcionales. Una vez introducida la problemática en la Sección 2.1, se realizará un resumen de las principales características de los métodos existentes en la literatura de FDA que han exhibido un mejor desempeño tanto desde un punto de vista teórico como aplicado (Sección 2.2). El propósito es hacernos con una batería de métodos competitivos para comparar su capacidad predictiva mediante un completo

estudio de simulación en el siguiente capítulo.

2.1. El problema de Clasificación Supervisada Funcional

La clasificación supervisada o discriminación de datos funcionales se refiere a situaciones en las que se observa una variable aleatoria funcional $\mathcal{X}$ y una respuesta categórica Y, la cual nos proporciona la clase de pertenencia de cada objeto funcional, con el propósito principal de construir una regla de clasificación consistente que permita predecir el grupo de pertenencia de cada nuevo dato funcional ([17]).

Sea $(\mathcal{X}_i, Y_i)$ con $i = 1, ..., n$, una muestra de n pares iid a $(\mathcal{X}, Y)$ y evaluados en $\mathcal{F} \times \bar{G} = \{1, ..., G\}$, donde $(\mathcal{F}, d)$ es un espacio vectorial semi-métrico, es decir, $\mathcal{X}$ es una variable funcional aleatoria y d una semi-métrica. A lo largo de este trabajo asumiremos que $G{=}2$ y que $n = n_1 + n_2$, siendo n_1 y n_2 el número de observaciones en cada grupo.

Si bien en el contexto de datos multivariantes existe un número importante de técnicas de clasificación supervisada, muchas de ellas vistas en [33], desde mediados de los años 90 se han desarrollado diferentes enfoques que abordan el problema de la clasificación supervisada para datos funcionales. Los primeros pasos para abordar esta situación fueron dados por Hastie, Buja y Tibshirani que, en los años 1994 y 1995, propusieron un método no paramétrico de discriminación de curvas [24] y una versión penalizada de la técnica de análisis discriminante lineal multivariante [23], respectivamente.

Posteriormente, en el año 2001, en [29] se propone un método de análisis discriminante lineal funcional que utiliza funciones de spline cúbicas para modelar las observaciones y en [22] se sugiere realizar una reducción de dimensiones por medio del análisis de componentes principales funcional y posteriormente resolver el problema de clasificación mediante análisis discriminante cuadrático o métodos kernel. En [18] Ferraty y Vieu desarrollaron un método de discriminación no paramétrico para datos funcionales que toma en consideración la naturaleza continua de los datos. Este trabajo contiene un estudio de simulación en el que comparan su propuesta

con los métodos de clasificación desarrollados en [24], [23] y [22], entre otros, demostrando que globalmente su procedimiento (**NPCD**) proporciona el comportamiento más regular. De ahí que sea uno de los métodos seleccionados en este trabajo para el posterior estudio de simulación. Será abordado en mayor profundidad en la Subsección 2.2.1.

Por otra parte, en [4] y [5] se han estudiado propiedades de consistencia del procedimiento del k-ésimo vecino más próximo (KNN) adaptado a espacios infinito dimensionales. Baillo y Cuevas en su trabajo [3] demuestran la consistencia del clasificador *KNN* estudiado por [5] y comparan el comportamiento del mismo con otros clasificadores funcionales a través de un estudio de simulación y el análisis de varios conjuntos de datos reales funcionales. Tal como se argumenta en [3], aunque no surge ningún ganador global, el rendimiento general del método **KNN**, junto con su sólida motivación intuitiva y relativa simplicidad, sugiere que podría representar un punto de referencia razonable para el problema de clasificación con datos funcionales. Así pues, este método también ha sido seleccionado para nuestro estudio de simulación.

Otro tipo de métodos de discriminación en el contexto de datos funcionales, que ha recibido gran atención en los últimos años, es el basado en medidas de profundidad. En [34] se proponen dos de estos métodos; en su primera propuesta, asignan nuevas curvas al grupo con la media recortada más cercana, mientras que en la segunda se minimiza la distancia promedio ponderada a cada elemento del grupo. Por su parte, en [11] se asigna una nueva observación al grupo en el que presente mayor profundidad y en [41] se propone la versión local de la profundidad espacial funcional, que es utilizada para proponer un método de clasificación. En este último trabajo se incluye un estudio de simulación en el que comparan estas técnicas y concluyen que su propuesta es la que globalmente proporciona los mejores resultados, especialmente en presencia de curvas atípicas.

Recientemente, en [8] se extiende el procedimiento DD-clasificador (basado en DD-plots) propuesto en [30] al contexto funcional en varias direcciones. En efecto, el nuevo método, denotado por DD^G, se puede aplicar a diversas medidas de profundidad y al caso general de $G \geq 2$ grupos; requiere el uso de un clasificador pudiendo

trabajar con los clásicos (LDA, QDA), los basados en regresión (GLM, GAM) o los no paramétricos (KNN, NP) entre otros. Se puede aplicar a datos reales multivariantes, a datos funcionales e incluso a datos funcionales multivariantes. Los autores presentan un completo estudio de simulación en el que analizan el comportamiento de su propuesta para distintas medidas de profundidad y utilizando, no sólo las trayectorias originales si no también estas combinadas con sus derivadas de primer y segundo orden. Tomando en consideración los resultados presentados en [8], hemos seleccionado las combinaciones indicadas a continuación para nuestro estudio de simulación: $DD^G.MaxD.hM$, $DD^G.Gam.RPD$, $DD^G.Gam.hM$ y $DD^G.Gam.hMw$. Información más detallada de esta metodología será dada en la Subsección 2.2.2.

Como alternativa a la clasificación supervisada, se puede considerar ajustar un modelo de regresión que permita predecir cierta variable de interés, por ejemplo, el padecimiento o no de una determinada enfermedad. En ese contexto es posible englobar el estudio de [15], donde se propone el modelo de regresión funcional *Gkam* el cual, como los autores declaran, se aplicaría para predecir problemas de respuesta binaria. Este algoritmo, también es considerado para el estudio de simulación y es tratado con más detalle en la Subsección 2.2.3.

2.2. Métodos de Clasificación Seleccionados

De los métodos de clasificación para datos funcionales comentados en la Sección 2.1, se han seleccionado los métodos: KNN, NP, $Gkam$ DD^G y sus variantes, para ser comparados en nuestro estudio de simulación. A continuación se discuten algunas de las propiedades de los mismos y se estudian en mayor profundidad.

2.2.1. Método NPCD

En el contexto de un problema de clasificación supervisada, establecido al inicio de la Sección 2.1, el método no paramétrico de discriminación de curvas (**NPCD**), propuesto en [18], propone una extensión no paramétrica de la Regla de Clasificación General o *Regla de Bayes* consistente en asignar el dato χ a la clase cuya probabilidad de pertenencia a posteriori sea máxima. Concretamente, el método propone

un estimador kernel de las G probabilidades a posteriori para la curva χ. De este modo, las probabilidades a posteriori

$$p_i(\chi) = P(Y = i/\mathcal{X} = \chi) = \mathbb{E}(I_{Y=i}/\mathcal{X} = \chi); i = 1, ..., G,$$

serán estimadas mediante

$$\hat{p}_{i,h,q}(\chi) = \frac{\sum_{i/Y_{i=g}}^{n} K(h^{-1}d_q(\mathcal{X}_i, \chi))}{\sum_{i=1}^{n} K(h^{-1}d_q(\mathcal{X}_i, \chi))}; i = 1, ..., G,$$

donde K es una función tipo Kernel real positiva decreciente en un intervalo $(0, 1)$, q indica la familia parametrizada de semi-métricas y h es el parámetro ventana o ancho de banda.

Para la aplicación de este método se requiere la implementación de una semi-métrica. En el caso de utilizar las *Semi-métricas basadas en el Análisis de Componentes Principales*, ver (1.1), se correspondería con el método propuesto en [22]. Otras semi-métricas utilizadas a lo largo del trabajo fueron las *Semi-métricas basadas en derivadas sucesivas*, ecuación (1.3), y las *Semi-métricas basadas en Regresión Multivariante de Mínimos Cuadrados Parciales(MPLSR)*,ecuación (1.2).

En [18] realizan un completo estudio de simulación en el que el estimador propuesto fue comparado con otros 7 métodos. Se demostró su buen comportamiento, superando a todos los competidores al aplicar la semi-métrica *MPLSR*. Cuando se realizó la aplicación a datos reales sobre los datos de *phoneme* y *tecator/Spectrometric* se verificó que para el primer conjunto de datos, el método *NP* arrojaba resultados similares a los obtenidos con PDA/Ridge cuando se utilizaba la semi-métrica *MPLSR*. Por otra parte, sobre los datos de *Tecator*, las mejores tasas de error en la clasificación se obtuvieron con NP_{deriv} y NP_{MPLSR}.

2.2.2. Método DD^G

El clasificador de máxima profundidad (*MD*), propuesto por [31], fue el primer intento de sustituir los datos brutos por profundidades de datos en problemas de clasificación multivariantes. Posteriormente, [30] proponen el clasificador *DD* como una propuesta de mejora basada en sustituir en el gráfico-*DD* la diagonal por una

función que proporcione una división en este gráfico de dos zonas con la mínima tasa de mala clasificación. También proponen una solución a la limitación del número G de grupos.

El trabajo de Cuesta y colaboradores ([8]) extiende el clasificador *DD* en varios sentidos: (1) permitiéndole manejar más de dos grupos; (2) permitiendo aplicar métodos de clasificación regulares como kNN, clasificadores lineales o cuadráticos, etc.; y (3) permitiendo integrar varias fuentes de información (profundidades de datos, datos funcionales multivariados, etc.) en el procedimiento de clasificación de manera unificada. La idea sería:

Dado un proceso en un espacio producto $\mathcal{X} = \mathcal{X}_1 \times, ..., \times \mathcal{X}_p$ de datos funcionales multivariantes en el cual se distinguen G grupos o clases, se debe seleccionar una medida de profundidad apropiada D^j para cada subespacio $\mathcal{X}_j$ y una proyección

$$\begin{aligned} \mathcal{X} = \mathcal{X}_1 \times, ..., \times \mathcal{X}_p &\rightarrow \mathbb{R}^H \\ \chi = (\chi_1, ..., \chi_p) &\rightarrow d = (D^1(\chi_1), ..., D^p(\chi_p)), \end{aligned}$$

donde $D^j(\chi_j)$ es el vector G-dimensional que indica la profundidad de la curva $\chi_j \in \mathcal{X}_j$ con respecto a los grupos $1, 2, ..., G$ y $H = G \times p$.

La gran versatilidad del clasificador DD^G estriba en la posibilidad de utilizar cualquier método de clasificación disponible que funcione en un espacio G-dimensional para separar los G grupos, en variar las medidas de pronfundidad a considerar y en la posibilidad de unificar datos funcionales multivariantes, es decir, diversas fuentes de información. El trabajo [8] propone un completo estudio de simulación donde se combinan tres tipos de profundidad (profundidad FM, Modal o hM y RP) con 9 diferentes clasificadores (*DD1, DD2, DD3, LDA, QDA, GLM, GAM, NP, KNN*). A groso modo, los resultados han puesto de manifiesto que la combinación *hM.w.GAM* proporcionó las mejores tasas de clasificacion.

2.2.3. Gkam

En [15] se propone el método *Gkam* (Generalized Kernel Additive Model), un algoritmo que extiende la idea de modelos aditivos generalizados para datos multivariantes al caso en el que se dispongan de covariables funcionales. La extensión consiste en adaptar la técnica mostrada en [39], de tal manera que permita la estimación no paramétrica de las funciones parciales f_j y, de ser necesario, la estimación conjunta no paramétrica de enlance inverso $g^{-1} = \mathbf{H}$, cuando las covariables son curvas.

Proponen estimar las funciones parciales en el l-ésimo paso mediante:

$$\hat{f}_j^l(\chi^j) = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^{-j,l}) K_j(\frac{d_j(\chi^j, \chi_i^j)}{h_j})}{\sum_{i=1}^n K_j(\frac{d_j(\chi^j, \chi_i^j)}{h_j})} \tag{2.1}$$

donde $\hat{Y}_i^{-j,l} = \sum_{i=1}^{j-1} \hat{f}_j^l(\mathcal{X}^i) + \sum_{i=j+1}^{p} \hat{f}_j^{l-1}(\mathcal{X}^i)$ es la predicción sin variable j, dj es la distancia (dada por la norma) en el espacio $\mathcal{F}_j$, K_j es una función núcleo asimétrica y h_j el ancho de banda.

En el trabajo [15] se realizó un estudio de simulación en el que se comparó el rendimiento del algoritmo propuesto con los metodos *GLM* y *GSAM* en tres escenarios diferentes. Bajo el primer escenario *Gkam* proporcionó resultados ligeramente peores que el de los competidores, sin embargo se observó que las diferencias entre los modelos se reducía a medida que la proporción de señales de ruido o el tamaño de la muestra incrementaban. En sus escenarios 2 y 3, *Gkam* superó a sus competidores arrojando lo mejores resultados y poniendo en evidencia su buen rendimiento cuando el tamaño de la muestra es grande. En el mismo trabajo se compararon los resultados obtenidos al aplicar los tres métodos anteriores al conjunto de datos de *Tecator*, obteniendo el procedimiento *GKAM* los mejores resultados y la variabilidad más baja.

Capítulo 3

Estudio de Simulación

Uno de los principales objetivos de este trabajo es la comparación, mediante un estudio de simulación, de los métodos de clasificación supervisada para datos funcionales existentes en la literatura que han demostrado un mejor comportamiento por sus propiedades teóricas, en estudios de simulación o en aplicaciones a datos reales. Para ello, en el presente capítulo se ha desarrollado un estudio comparativo de los métodos de clasificación seleccionados en el Capítulo 2 con la finalidad de analizar la eficiencia de los mismos y posteriormente aplicar aquellos que exhiban un mejor desempeño a un conjunto de datos reales del ritmo motor derivados del Finger Tapping Test (FTT) ([2]).

Para la aplicación de los distintos métodos de clasificación considerados, se han utilizado las correspondientes funciones de R ([36]) disponibles en el paquete fda.usc ([16]) y en la página web NPFDA (https://www.math.univ-toulouse.fr/ ferraty/-SOFTWARES/NPFDA/).

En la Sección 3.1 se indica el procedimiento seguido para la simulación de los datos funcionales, los escenarios de simulación considerados así como los parámetros y suposiciones contemplados a lo largo del estudio. En la Sección 3.2 se muestran y discuten los resultados obtenidos.

3.1. Procedimiento

Este estudio de simulación toma como modelo el realizado en [14]. Para ello, se considerarán muestras de curvas $\chi_{ij}, j = 1, ..., n_i$ para el i-ésimo grupo, $i = 1, 2$, como realizaciones independientes del proceso estocástico $\mathcal{X}_i$ definido como:

$$\mathcal{X}_i(t) = \mu_i(t) + \sum_{p=1}^{\infty} e^{-p/2} N_{pi} \psi_p(t); t \in T$$

donde $\mu_i(.)$ son las funciones de media, N_{pi} son variables aleatorias independientes e idénticamente distribuidas (iid) a una $N(0, \sigma_1)$ y $\psi_p(t) = 2^{1/2} sin\{(p-1)\pi t\}$ $(p > 1)$ y $\psi_1 \equiv 1$ son funciones de bases ortnormales. Por razones prácticas la suma infinita en truncada en $p = 150$

Además, se ha supuesto que los datos X_{ijl} satisfacen el modelo:

$$X_{ijl} = \chi_{ij}(t_l) + \epsilon_{ijl}, i = 1, 2, 1 \leq j \leq n_i, n = n_1 + n_2, 1 \leq l \leq L,$$

donde los tiempos de observación $\{t_1, t_2, ..., t_L\}$ forman una rejilla regular de puntos de soporte en I y para cada $i = 1, 2$, las observaciones de los errores ϵ_{ijl} son independientes e idénticamente distribuidos a una $N(0, \sigma)$. Se asume que los χ_{ij} y los ϵ_{ijl} son independientes.

Al igual que en [14], las curvas simuladas provienen de cuatro familias de funciones de medias μ_i, las cuales fueron generadas a partir de los siguientes modelos:

1. **Escenario 1 [F1]**.

   $\mu_1(t) = 5 + t(1-t)$ y $\mu_2(t) = 5 + t^{3/2}(1-t)^{3/2}$, para $t \in I = [0, 1]$

2. **Escenario 2 [F2]**.

   $\mu_1(t) = 5 + t^2(1-t)^4$ y $\mu_2(t) = 5 + t^3(1-t)^3$, para $t \in I = [0, 1]$

3. **Escenario 3 [F3]**.

   $\mu_1(t) = 4t$ y $\mu_2(t) = 8t - 2$, para $t \in I = [0, 1]$

4. **Escenario 4 [F4]**.

   $\mu_1(t) = 0{,}06(\sin(t) + \cos(t))$ y $\mu_2(t) = 0{,}12(\sin(t) + \cos(t))$, para $t \in I = [0, 2\pi]$

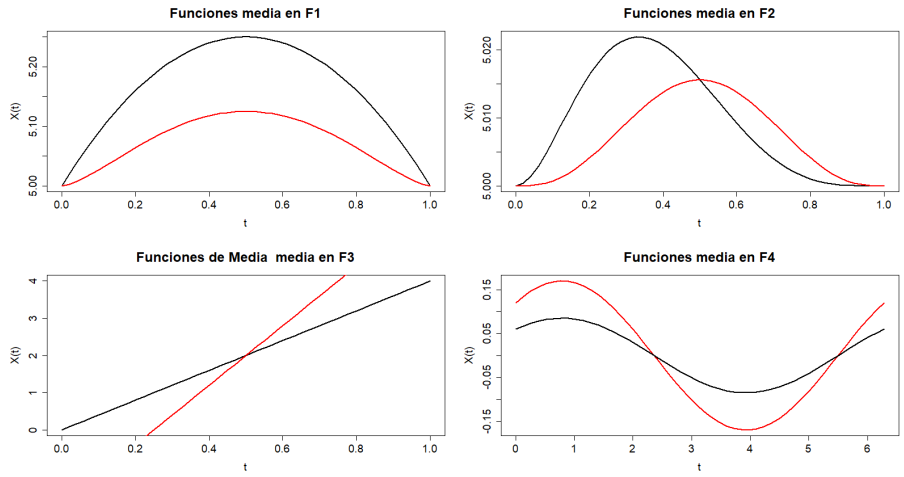

Figura 3.1: Funciones medias para los escenarios F1-F4.

En la Figura 3.1 se muestra las funciones medias para cada uno de los escenarios simulados, con μ_1 (color negro) para el grupo i=1 y μ_2 (color rojo) para el grupo i=2. En F1 se observa que las medias de los grupos están claramente separadas, por lo que se espera que los métodos puedan discriminar fácilmente entre los grupos cuando las fuentes de error consideradas no sean muy elevadas. Por otra parte, las funciones media en F2, son semejantes; en esta situación se espera mayor dificultad a la hora de clasificar a un nuevo individuo. Los escenarios F3 y F4, utilizados previamente en [41], corresponden a funciones lineales y sinusoidales, respectivamente. Es probable que, al igual que lo que se espera en F2, los métodos de clasificación en F4 encuentren mayor dificultad para discriminar entre afectado y no afectado cuando el error aumente.

Al igual que en [14], los datos brutos $(t_l, X_{ijl}); l = 1, ..., L; j = 1, ..., n_i; i = 1, 2$ fueron pre-procesados mediante suavización; concretamente, las observaciones funcionales fueron obtenidas utilizando el método de regresión local lineal con un mismo ancho de banda para todos las curvas (de acuerdo al criterio propuesto en [21]). Este parámetro equivale al promedio obtenido tras aplicar el selector de ancho de banda pluginBw del paquete `locpol` sobre cada curva simulada.

De acuerdo al modelo que hemos utilizado para simular los datos funcionales, los parámetros que controlan la variabilidad de las curvas son σ y σ_1. El primero es la desviación típica de los errores ϵ_{ijl}, cuyo efecto se traduce en el aumento o la disminución de la variabilidad dentro de las curvas. σ_1, por su parte, controla la

variabilidad que hay entre las curvas. En las Figuras 3.2-3.5 se muestra el efecto que tiene el aumento de estos parámetros sobre las curvas suavizadas en cada uno de los escenarios simulados. A la derecha, vemos que cuando los σ y σ_1 son iguales a cero, las curvas son, en efecto, la función de media. En el centro, hemos fijado en cero el parámetro σ y aumentamos σ_1, a la izquierda se ha fijado en cero σ_1 y aumentado σ. En las simulaciones se han establecido una combinación de las fuentes de variabilidad acorde a cada escenario.

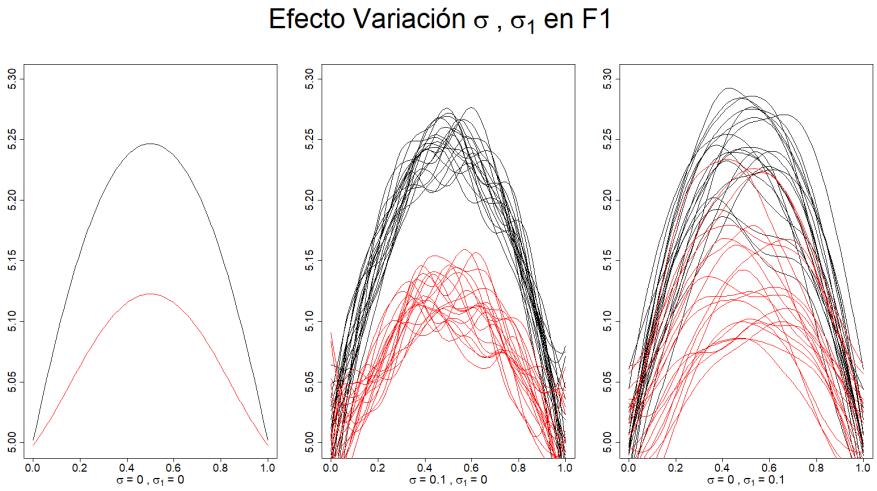

Figura 3.2: Impacto en las curvas de entrenamiento frente a cambios de σ y σ_1 en F1

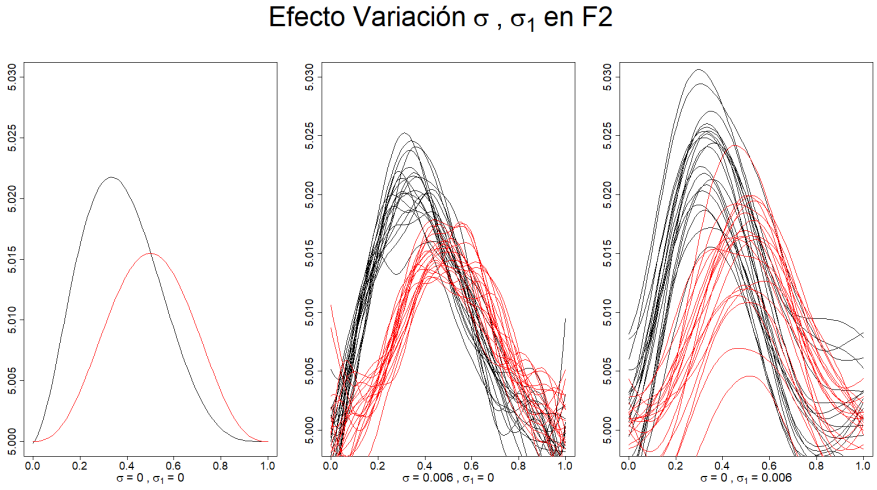

Figura 3.3: Impacto en las curvas de entrenamiento frente a cambios de σ y σ_1 en F2

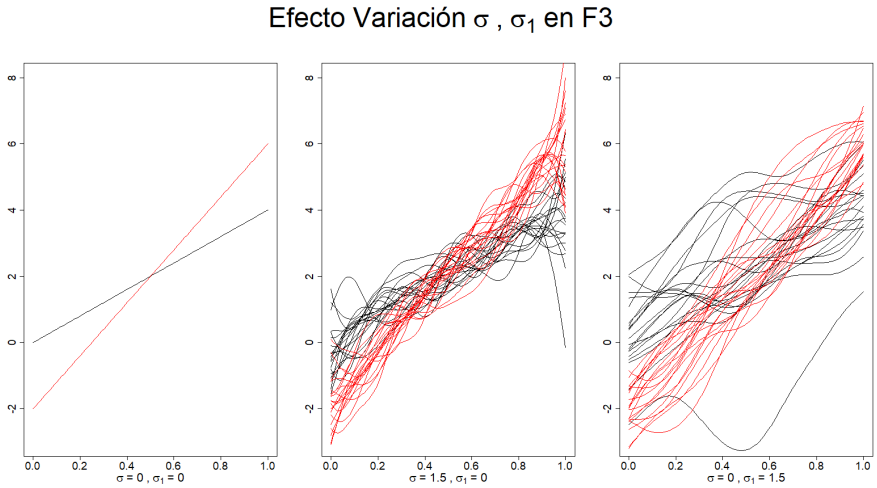

Figura 3.4: Impacto en las curvas de entrenamiento frente a cambios de σ y σ_1 en F3

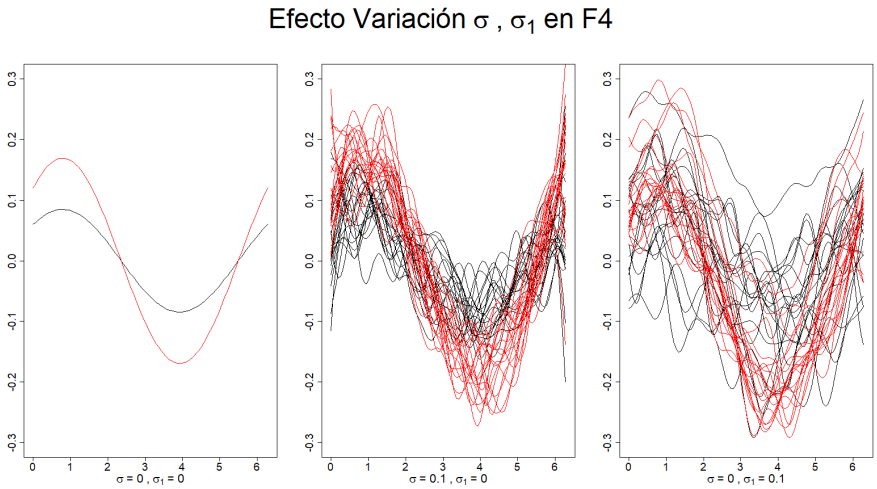

Figura 3.5: Impacto en las curvas de entrenamiento frente a cambios de σ y σ_1 en F4

Con la finalidad de evaluar el rendimiento de los métodos de clasificación seleccionados, así como el efecto de sus correspondientes parámetros bajo diferentes situaciones, se han considerado tamaños muestrales $n_1 = n_2 = 20, 30, 50$ y diversos valores para los parámetros σ y σ_1 de acuerdo a los escenarios de simulación, tal como se indica en la Tabla 3.1. En todas las situaciones el parámetro L que controla el tamaño de la discretización de las curvas se ha tomado como 50.

Escenario	**F1**	**F2**	**F3**	**F4**
σ	0.04, 0.06	0.006, 0.06	0.75, 1.5	0.075, 0.1
σ_1	0.06, 0.08	0.008, 0.06	1.0, 1.5	0.075, 0.1

Tabla 3.1: Valores para σ y σ_1 considerados en el estudio de simulación para los distintos escenarios F1-F4

Luego de agregar las fuentes de error, acorde a cada escenario, la Figura 3.6 presenta una realización de 20 curvas pertenecientes al grupo 1 y 20 al grupo 2 para los cuatro escenarios considerados.

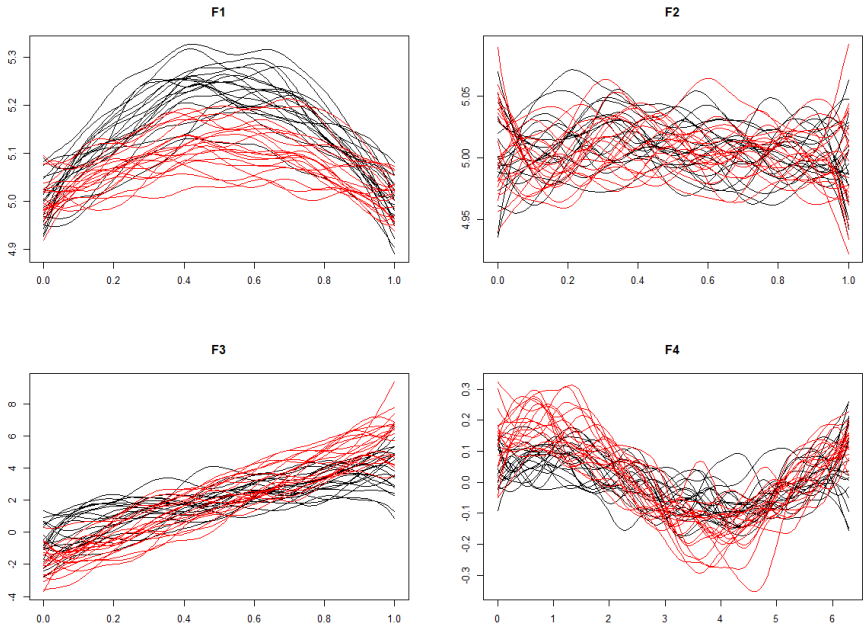

Figura 3.6: Conjunto de curvas simuladas a partir de F1, F2, F3 y F4.

Una vez conocidas las principales características de los datos funcionales simulados, se procede a aplicar los métodos de clasificación supervisada funcional seleccionados; entiéndase: El clasificador propuesto en [18] con sus distintas semimétricas (NP, NP_{PCA}, NP_{MPLSR}), las diferentes versiones del método DD^G ($DD^G.Gam.hMw$, $DD^G.MaxD$, $DD^G.Gam.RPD$) propuesto en [8], el método KNN para datos funcionales visto en [3] y el método *Gkam* estudiado en [15].

Como se comentó en la seccion anterior, todos estos métodos se encuentran programados en la libreria `fda.usc` de `R`. Para aplicar el método KNN, se ha establecido $k = 5$. Para el método NP, se han utilizado las familias paramétricas vistas en la sección 2.2, concretamente para implementar el clasificador NP con semimétrica $MPLSR$, se ha fijado un $q_{opt} = 2$, del mismo modo, para NP_{PCA}, el numero de componenres retenidos se ha fijado en $q = 2$. El ancho de banda local fue seleccionado por el procedimiento de validación cruzada y la función tipo núcleo seleccionada fue la cuadrática. Se han empleado los paramétros que vienen programados en la función *classif.gkam* para aplicar el método gkam. Por último, en las distintas versiones

del método de clasificación DD^G, se han usado la derivadas de orden 1 y 2 para $DD^G.Gam.hMw$ aunque sólo se muestran los resultados obtenidos con la combinación de los datos originales y la primera derivada, ya que se obtuvieron mejores resultados. Para aplicar el clasificador *MaxD*, se hizo combinado la profundidad hM y el método $DD^G.Gam.RPD$ se utilizó aplicando los parámetros recomendados en la ayuda que proporciona [16] para la función `classif.DD`.

3.2. Resultados

Con el propósito de valorar la capacidad de clasificación de los métodos considerados, para cada combinación de tamaños muestrales (n_1, n_2) y de pares (σ, σ_1) se han generado $R = 100$ réplicas $X_{ijl}, l = 1, ..., L; j = 1, ..., n_i; i = 1, 2$. Para cada réplica, además de las muestras de entrenamiento de tamaños $20, 30$ y 50 se han considerado muestra de prueba, cuyo grupo origen es conocido, de tamaños $20, 30$ y 50, respectivamente para cada grupo y se han calculado las media y desviaciones típicas de las tasas de mala clasificación.

Todos los resultados para el escenario F1 están resumidos en la Tabla 3.2. En la misma, podemos apreciar que los mejores resultados se obtienen con el método NP_{MPLSR} para cualquier combinación de tamaños muestrales y valores de σ y σ_1, seguido de los métodos $DD^G.Gam.hM.w$ (tomando como información adicional la primera derivada de los datos), $DD^G.Gam.RPD$ y NP_{PCA}. A medida que vamos aumentando la variabilidad dentro de cada curva (σ) se observa una reducción de la potencia discriminante de los métodos, así se aumente el número de curvas simuladas. Por otro lado, vemos que tiene mayor efecto en la capacidad de predecir correctamente a qué población pertenece un nuevo individuo un aumento en la variabilidad entre curvas (σ_1), que un aumento en la variabilidad dentro de cada curva (σ). Bajo este escenario se observa también como $DD^G.MaxD$ es el método con el desempeño más bajo en relación a los otros métodos vistos, mejorando muy poco a medida que aumentamos el tamaño de n_1 y n_2.

La Tabla 3.3 resume los resultados del escenario F2. Bajo este escenario, tal como se comentaba en la sección anterior, las medias funcionales son bastante similares lo

que provoca que los métodos tengan mayor dificultad a la hora de discriminar entre grupos cuando alguno de los parámetros que controla la variabilidad es incrementado. En este caso, un ligero incremento en σ o σ_1 provocará grandes aumentos en las tasas de error de clasificación. También se aprecia que, al igual que en F1, el método NP_{MPLSR} supera a los demás métodos en 3 de las 4 situaciones consideradas, siendo únicamente mejorado por el método $Gkam$ cuando los σ y σ_1 son 0,06 y 0,008, respectivamente. Otro método que mostró buen comportamiento en este escenario fue el $DD^G.Gam.hMw$. Los resultados que se muestran de este método han sido obtenidos combinando la información de los datos originales y la primera derivada, ya que, aunque se probaron otras combinaciones, fue con esta que se obtuvieron mejores resultados.

Escenario F1									
Métodos	$n_1 = n_2$	σ=0.04, σ_1=0.06		σ=0.04, σ_1=0.08		σ=0.06, σ_1=0.06		σ=0.08, σ_1=0.08	
KNN	20	0.0865	(0.0508)	0.4728	(0.0806)	0.4285	(0.0956)	0.4922	(0.0822)
	30	0.0773	(0.0347)	0.4800	(0.0618)	0.4298	(0.0753)	0.4847	(0.0679)
	50	0.0691	(0.0279)	0.4780	(0.0534)	0.4267	(0.0508)	0.4760	(0.0542)
NP	20	0.0918	(0.0451)	0.4775	(0.0800)	0.4358	(0.0897)	0.4900	(0.0862)
	30	0.0813	(0.0348)	0.4733	(0.0630)	0.4172	(0.0734)	0.4792	(0.0648)
	50	0.0782	(0.0285)	0.4665	(0.0537)	0.4219	(0.0540)	0.4770	(0.0542)
NP_{PCA}	20	0.0842	(0.0466)	0.4865	(0.0811)	0.4612	(0.0912)	0.4830	(0.0880)
	30	0.0745	(0.0340)	0.4815	(0.0559)	0.4465	(0.0777)	0.4893	(0.0634)
	50	0.0736	(0.0277)	0.4699	(0.0590)	0.4416	(0.0555)	0.4731	(0.0480)
NP_{MPLSR}	20	0.0568	(0.0367)	0.4432	(0.0925)	0.4260	(0.0956)	0.4585	(0.0888)
	30	0.0565	(0.0300)	0.4427	(0.0732)	0.4137	(0.0651)	0.4667	(0.0666)
	50	0.0517	(0.0222)	0.4315	(0.0519)	0.3954	(0.0539)	0.4539	(0.0550)
$Gkam$	20	0.0788	(0.0462)	0.4730	(0.0777)	0.4018	(0.0898)	0.4818	(0.0787)
	30	0.0725	(0.0352)	0.4660	(0.0627)	0.3817	(0.0599)	0.4662	(0.0694)
	50	0.0648	(0.0271)	0.4652	(0.0566)	0.3726	(0.0534)	0.4674	(0.0508)
$DD^G.MaxD.hM$	20	0.0950	(0.0541)	0.4845	(0.0778)	0.4102	(0.0910)	0.4855	(0.0770)
	30	0.0855	(0.0450)	0.4725	(0.0579)	0.3943	(0.0681)	0.4732	(0.0655)
	50	0.0863	(0.0392)	0.4666	(0.0518)	0.3835	(0.0523)	0.4714	(0.0518)
$DD^G.Gam.RPD$	20	0.1222	(0.0679)	0.4555	(0.0830)	0.4548	(0.0801)	0.4810	(0.0832)
	30	0.0862	(0.0405)	0.4528	(0.0706)	0.4280	(0.0656)	0.4700	(0.0638)
	50	0.0784	(0.0331)	0.4348	(0.0509)	0.4059	(0.0622)	0.4559	(0.0573)
$DD^G.Gam.hMw$	20	0.0838	(0.0542)	0.4435	(0.0923)	0.4218	(0.0788)	0.4838	(0.0871)
	30	0.0722	(0.0399)	0.4465	(0.0670)	0.4013	(0.0613)	0.4592	(0.0663)
	50	0.0667	(0.0300)	0.4294	(0.0531)	0.3918	(0.0587)	0.4620	(0.0548)

Tabla 3.2: Medias y desviaciones típicas (entre paréntesis) de las tasas de error de clasificación en el escenario F1

ESCENARIO F2

Métodos	$n_1 = n_2$	σ= 0.006, σ_1=0.008		σ= 0.006, σ_1=0.06		σ= 0.06, σ_1=0.008		σ= 0.06, σ_1=0.06	
KNN	20	0.0865	(0.0508)	0.4728	(0.0806)	0.4285	(0.0956)	0.4922	(0.0822)
	30	0.0773	(0.0347)	0.4800	(0.0618)	0.4298	(0.0753)	0.4847	(0.0679)
	50	0.0691	(0.0279)	0.4780	(0.0534)	0.4267	(0.0508)	0.4760	(0.0542)
NP	20	0.0918	(0.0451)	0.4775	(0.0800)	0.4358	(0.0897)	0.4900	(0.0862)
	30	0.0813	(0.0348)	0.4733	(0.0630)	0.4172	(0.0734)	0.4792	(0.0648)
	50	0.0782	(0.0285)	0.4665	(0.0537)	0.4219	(0.0540)	0.4770	(0.0542)
NP_{PCA}	20	0.0842	(0.0466)	0.4865	(0.0811)	0.4612	(0.0912)	0.4830	(0.0880)
	30	0.0745	(0.0340)	0.4815	(0.0559)	0.4465	(0.0777)	0.4893	(0.0634)
	50	0.0736	(0.0277)	0.4699	(0.0590)	0.4416	(0.0555)	0.4731	(0.0480)
NP_{MPLSR}	20	0.0568	(0.0367)	0.4432	(0.0925)	0.4260	(0.0956)	0.4585	(0.0888)
	30	0.0565	(0.0300)	0.4427	(0.0732)	0.4137	(0.0651)	0.4667	(0.0666)
	50	0.0517	(0.0222)	0.4315	(0.0519)	0.3954	(0.0539)	0.4539	(0.0550)
$Gkam$	20	0.0788	(0.0462)	0.4730	(0.0777)	0.4018	(0.0898)	0.4818	(0.0787)
	30	0.0725	(0.0352)	0.4660	(0.0627)	0.3817	(0.0599)	0.4662	(0.0694)
	50	0.0648	(0.0271)	0.4652	(0.0566)	0.3726	(0.0534)	0.4674	(0.0508)
$DD^G.MaxD.hM$	20	0.0950	(0.0541)	0.4845	(0.0778)	0.4102	(0.0910)	0.4855	(0.0770)
	30	0.0855	(0.0450)	0.4725	(0.0579)	0.3943	(0.0681)	0.4732	(0.0655)
	50	0.0863	(0.0392)	0.4666	(0.0518)	0.3835	(0.0523)	0.4714	(0.0518)
$DD^G.Gam.RPD$	20	0.1222	(0.0679)	0.4555	(0.0830)	0.4548	(0.0801)	0.4810	(0.0832)
	30	0.0862	(0.0405)	0.4528	(0.0706)	0.4280	(0.0656)	0.4700	(0.0638)
	50	0.0784	(0.0331)	0.4348	(0.0509)	0.4059	(0.0622)	0.4559	(0.0573)
$DD^G.Gam.hMw$	20	0.0838	(0.0542)	0.4435	(0.0923)	0.4218	(0.0788)	0.4838	(0.0871)
	30	0.0722	(0.0399)	0.4465	(0.0670)	0.4013	(0.0613)	0.4592	(0.0663)
	50	0.0667	(0.0300)	0.4294	(0.0531)	0.3918	(0.0587)	0.4620	(0.0548)

Tabla 3.3: Medias y desviaciones típicas (entre paréntesis) de las tasas error de clasificación para el escenario F2.

En las Figuras 3.7 y 3.8, se muestran los diagramas de caja de las tasas de

clasificación para los escenarios F3 y F4, respectivamente. Al observar la Figura 3.7, se aprecia que con niveles de ruido bajos la mediana de la tasa de error de todos los métodos se encuentra en cero. Exceptuando los métodos $Gkam$ y $DD^G.Gam.RPD$, puede verse que, como máximo, las tasas de error llegan al 5 % en alguna realización. Sin embargo, aún en esta situación, es posible visualizar como descienden estas tasas al aumentar el número de curvas en este escenario. El método NP_{MPLSR} continúa siendo el que proporciona las tasas de error más bajas, en tanto que el estimador $DD^G.Gam.RPD$ es el que proporciona en general peores resultados.

Por último, podemos apreciar en la Figura 3.8 que en el escenario F4 los métodos $DD^G.Gam.hMW$, $DD^G.Gam.RPD$ y NP, son los que proporcionan las tasas de error más elevadas. Nuevamente el estimador NP_{MPLSR} genera los mejores resultados.

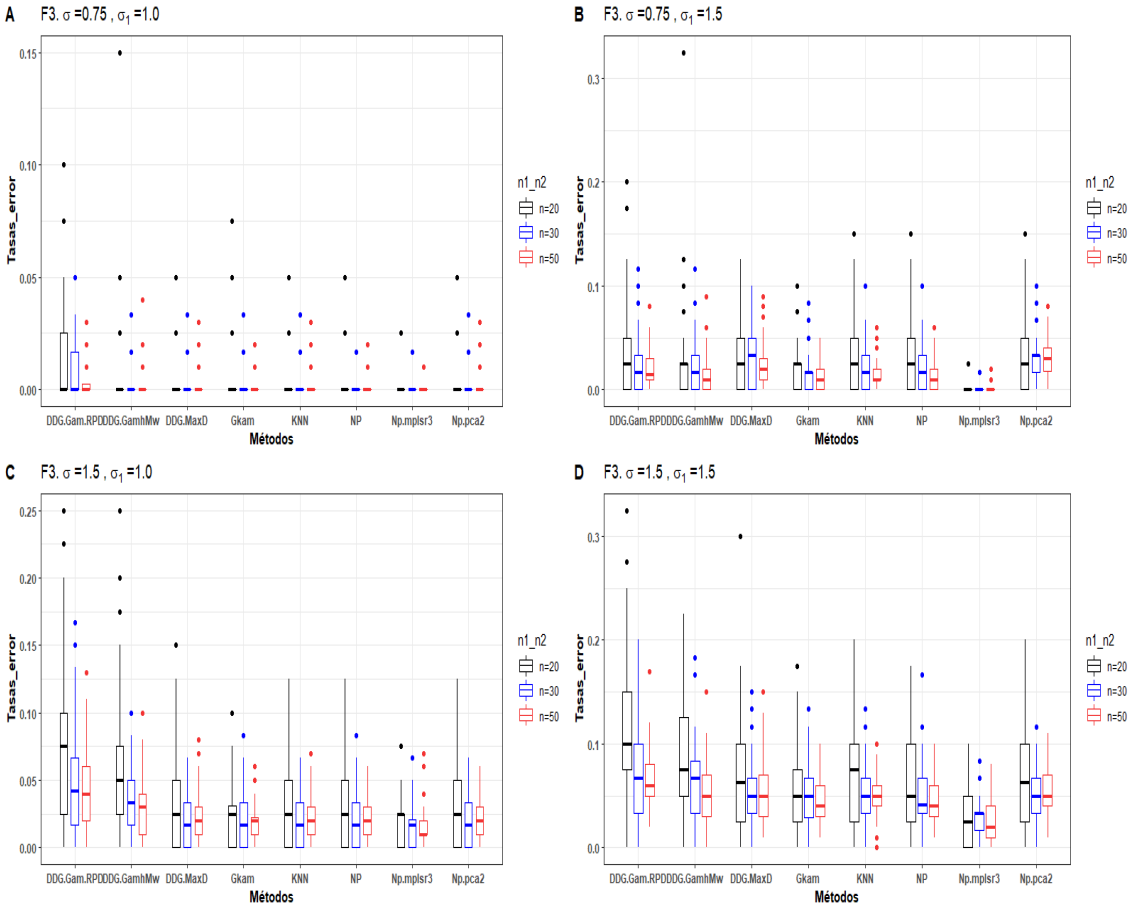

Figura 3.7: Boxplots Tasas de Mala clasificación para los distintos tamaños muestrales y niveles de variación σ y σ_1 en el Escenario F3.

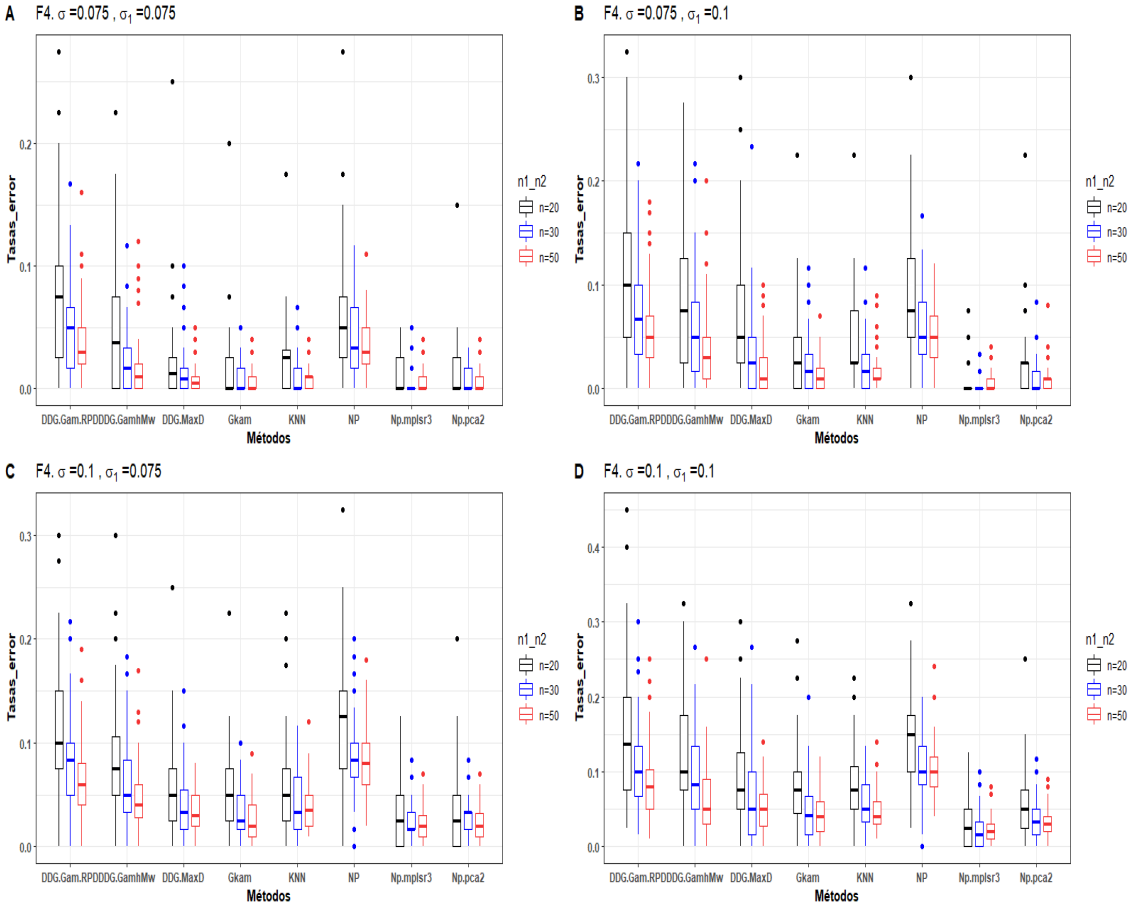

Figura 3.8: Boxplots Tasas de Mala clasificación para los distintos tamaños muestrales y niveles de variación σ y σ_1 en el Escenario F4.

Luego de observar los resultados obtenidos tras la aplicación de los distintos métodos de clasificación en los cuatro escenarios propuestos, se puede concluir que en todos ellos, a medida que se aumenta el número de curvas simuladas, salvo escasas excepciones, se ve una mejora progresiva en su desempeño. Asimismo, se concluye que el método NP_{MPLSR} proporciona mejores resultados que sus rivales, independientemente de que se estén observando tamaños muestrales pequeños ($n_1 = n_2 = 20$) o curvas con variabilidad alta. Otros métodos parecen funcionar mejor o peor, de acuerdo al escenario en donde se apliquen.

A pesar de que el método $DD^G.Gam.hMw$ permiten introducir en el proceso de clasificación más fuentes de información de los datos funcionales (como son las derivadas de las curvas), en nuestro caso concreto, al hacer distintas combinaciones de los datos simulados y sus derivadas, no se consiguió mejorar las tasas de error que se obtuvieron en las tablas presentadas.

También fueron considerados los métodos NP_{MPLSR} con $q_{opt} = 3$, NP_{PCA} con 3 y 4 componentes principales, sin embargo estos resultados no mejoraron las tasas de error al ser comparados con sus análogos.

Capítulo 4

Aplicación a Datos Reales

En este capítulo se presenta un conjunto de datos funcionales procedentes de la aplicación de una prueba neurofisiológica que evalúa una medida simple de la velocidad y del control motor, que se utiliza en distintas áreas de la Medicina. Una vez presentada la forma de obtención de los datos, se realizará un análisis exploratorio de los mismos y se aplicarán aquellos métodos de clasificación supervisada, que presentaron mejores resultados en el Capítulo 3, con el propósito de discriminar sujetos afectados o no de la Enfermedad de Parkinson.

4.1. Prueba FTT

La prueba de golpeteo con los dedos FTT (Finger Taping Test), es una herramienta que se utiliza con la finalidad de evaluar patrones de movimiento rítmico, con uso práctico en evaluaciones clínicas y como parte protocolos de investigación. Esta prueba permite el estudio de varios elementos que posteriormente se utilizan como datos complementarios para caracterizar perfiles de enfermedades ([2]). Existe un número importante de artículos que hacen referencia a los diferentes campos de aplicación de esta prueba. En [26] se hace un resumen del estado del arte del FTT y sus aplicaciones en estudios clínicos.

Arias y colaboradores han planteado en [2] un estudio cuyo objetivo fue examinar la validez y fiabilidad del FTT para estudiar las alteraciones en el ritmo motor en tres grupos diferentes: pacientes diagnosticados con la enfermedad de Parkinson,

con edad promedio de 69 años; sujetos de control sanos en edad avanzada, con edad promedio de 70 años y sujetos de control sanos jóvenes cuya media de edad era de 24 años.

Para la realización de la prueba FTT se dispone de un sistema electrónico conectado a un ordenador portátil, que contiene una placa de metal y un anillo del mismo material, que se adapta a la falange distal del dedo índice de cada sujeto, para registrar el tiempo (en mili-segundos) que tardan las piezas metálicas en ponerse en contacto a lo largo de movimientos sucesivos durante un intervalo de tiempo concreto. Para el estudio llevado a cabo en [2], la prueba se realizó en dos modos de golpeteo:*Modo cómodo*, donde los sujetos se encontraban en una posición de confort óptima para realizar la prueba y *Modo rápido*, aquí los sujetos debían golpear a su mayor velocidad posible. Los ejercicios, en cada uno de los modos se repitieron 3 veces, con 3 minutos de descanso entre las sucesivas repeticiones, en dos días diferentes.

Una vez analizados los resultados obtenidos, concluyen que la prueba aporta información relevante para distinguir, en ambos modos, los tres grupos de sujetos, sin embargo, concluye que el FTT en modo rápido no resulta una prueba válida para detectar diferencias en la formación del ritmo en los grupos estudiados, ya que la variable fatiga, no controlada en el experimento, resultó una variable de confusión. Como conclusión recomiendan el *modo cómodo* en protocolos que incluyen el FTT para evaluar la formación del ritmo.

Utilizando los datos obtenidos tras la aplicación de la prueba FTT en el modo cómodo sobre 3 poblaciones de pacientes similares a las utilizadas en [2], en este capítulo se pretende estudiar si, además de utilizar este procedimiento para evaluar la formación del ritmo motor, es posible identificar el grupo de procedencia de un sujeto, en concreto si el paciente padece o no la enfermedad de Parkinson, aplicando para ello técnicas de clasificación supervisada para datos funcionales.

4.1.1. Datos

Para aplicar las técnicas de clasificación funcional supervisada, se cuenta con un conjunto de datos funcionales producto de la realización de la prueba de FTT sobre 53 pacientes, 15 de ellos diagnosticados con la enfermedad de Parkinson sin medicar en las 12 horas previas a la prueba, 19 sujetos sanos en edad avanzada y 19 sujetos sanos de edad joven.

Puesto que el FTT se ha aplicado a todos los pacientes en dos días diferentes, para cada sujeto se dispone de dos curvas, una por día. Cada una de estas curvas es obtenida luego del pre-procesado de las 37 observaciones obtenidas, mediante el método de regresión local lineal con un ancho de banda común (de tipo Plug-in) dentro de cada grupo de sujetos. Dichas observaciones miden el tiempo que cada sujeto tarda en repetir un ciclo desde que toca con el anillo colocado en el dedo la superficie metálica de la placa.

Con el propósito de valorar la capacidad de clasificación de los distintos métodos, se precisa disponer de dos muestras de curvas, ***la muestra de entrenamiento*** **y** ***la muestra de prueba***. Dado que el tamaño muestral de los grupos de sujetos es relativamente bajo y no parece oportuno dividirlos en dos subgrupos, se ha optado por tomar como muestra de entrenamiento las 53 curvas del día 1 y como muestra de prueba, las curvas generadas a partir de las mediciones tomadas el día 2.

4.2. Análisis de los datos Reales

4.2.1. Análisis Exploratorio

En primer lugar, se realiza un análisis exploratorio de los datos para conocer las principales características de las curvas que conforman la muestra de entrenamiento.

En la Figura 4.1, a la izquierda, se muestran las curvas sin suavizar de la muestra de entrenamiento según el grupo de procedencia y a la derecha, los datos suavizados mediante el método de regresión no paramétrica. Por otra parte, en la Figura 4.2 se aprecia, a la izquierda, las curvas sin suavizar del conjunto de entrenamiento, esta vez distinguiendo entre afectados y no afectados por la enfermedad de Parkinson y,

a la derecha, las curvas suavizadas. En ambas imágenes se aprecia que las curvas se encuentran bastante superpuestas por lo que, en estas condiciones, se espera que los métodos de clasificación encuentren dificultad a la hora de discriminar entre afectado y no afectado por la enfermedad de Parkinson.

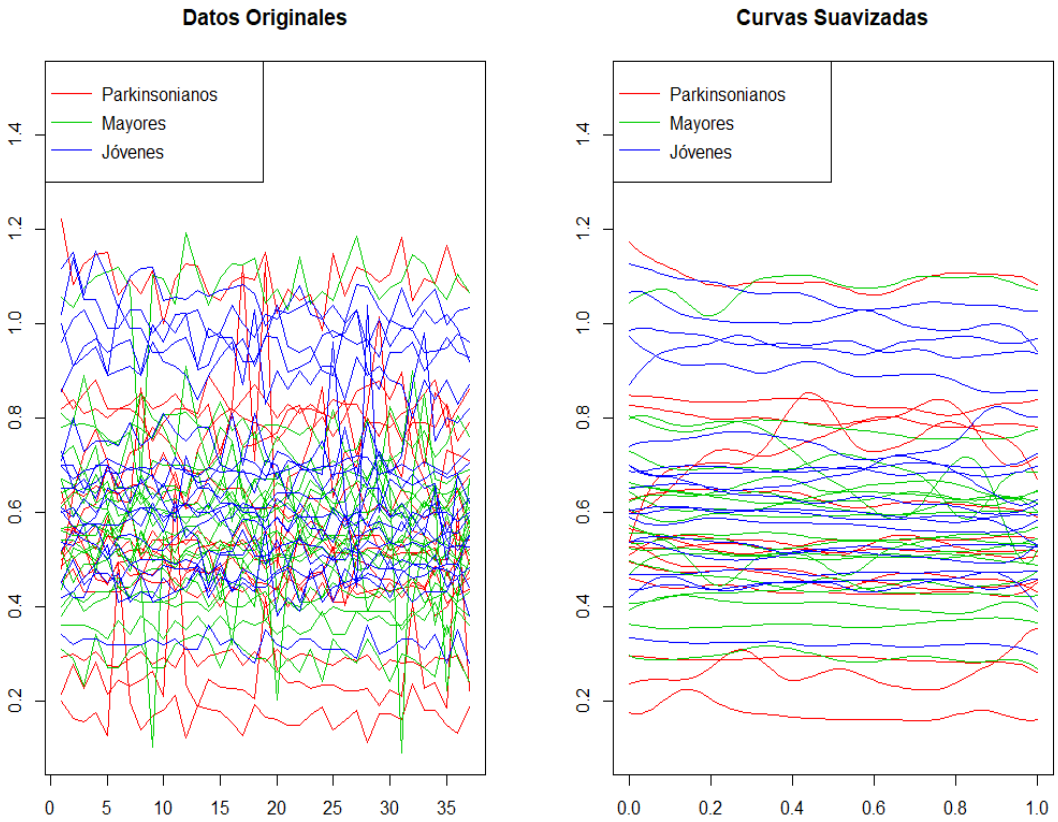

Figura 4.1: Datos FTT Originales y Suavizados según grupo de pertenencia

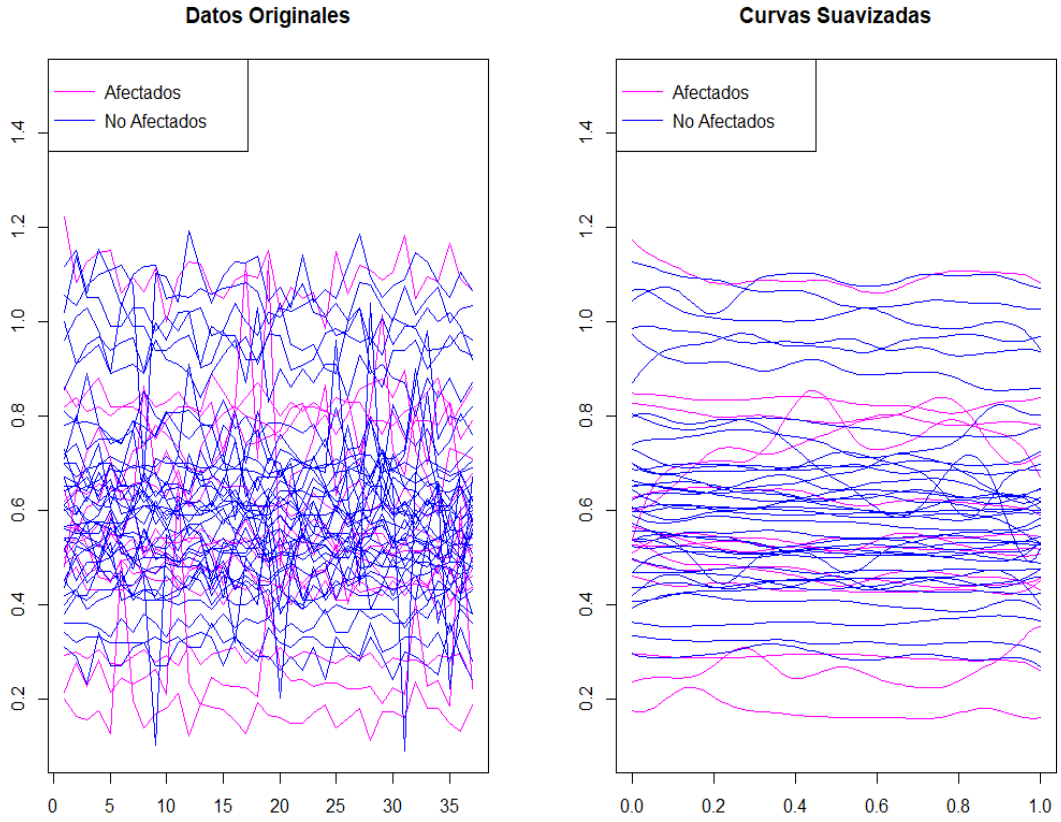

Figura 4.2: Datos FTT Originales y Suavizados según condición

Con el objetivo de explorar el comportamiento de los grupos que conforman nuestro conjunto de curvas, se aportan las Figuras 4.3 y 4.4, las cuales contienen algunas medidas de tendencia central y la varianza, atendiendo a la condición de los sujetos y al grupo en el cual se ubican.

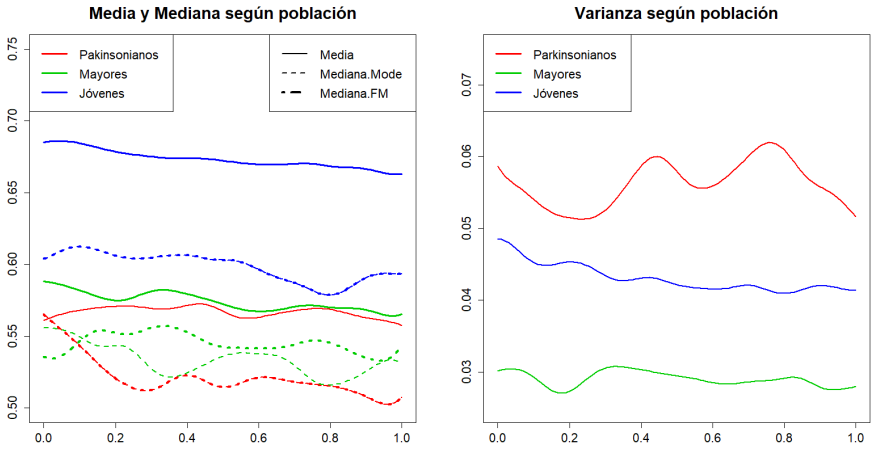

Figura 4.3: Media, Mediana y Varianza por grupo de pertenencia

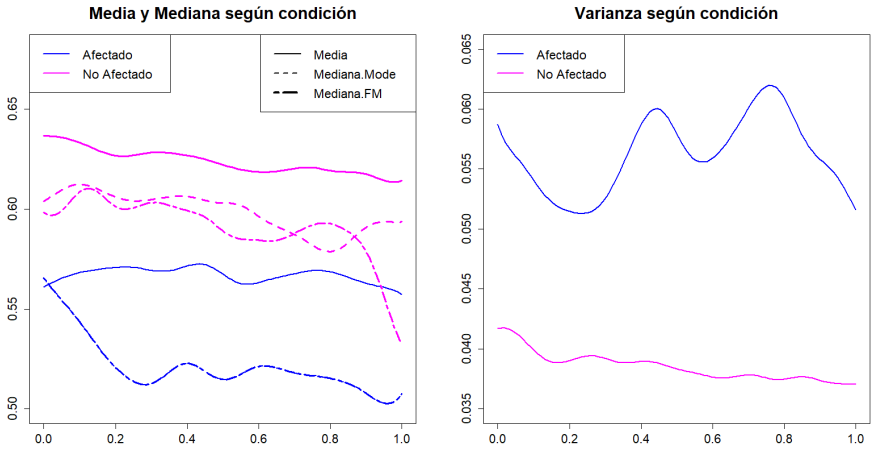

Figura 4.4: Media, Mediana y Varianza por condición

Tanto en la Figura 4.3 como en la Figura 4.4, se observa que en el grupo de sujetos diagnosticados con la enfermedad de Parkinson los tiempos de duración de los ciclos son ligeramente más bajos que en los grupos de sujetos mayores y jóvenes, pero considerablemente más variables, lo que genera la superposición de curvas que se aprecia en la Figura 4.1. Por el contrario, para los jóvenes se aprecian medidas de localización superiores, al ser comparados con los otros grupos, y menos variabilidad. Al parecer el ritmo en el cual los sujetos jóvenes parecen sentirse cómodos es más lento al mostrado por los sujetos mayores y los parkinsonianos, sin embargo, a diferencia de los afectados por la enfermedad de Parkinson, son capaces de mantenerlo a lo largo de toda la prueba, reflejándose esto en la curva de variabilidad que tiene un comportamiento relativamente constante. El ritmo de los enfermos de

Parkinson, como era de esperar, es, en comparación con los demás, más rápido y con una mayor variabilidad. Es importante destacar que, tanto las medias como las medianas, de los tiempos en individuos afectados de Parkinson y sujetos mayores son muy similares, por lo que cabe pensar que, ante la presencia de ruido, distinguir entre ellos resultará dificultoso para cualquier método de discriminación.

Otra herramienta gráfica utilizada en la exploración datos funcionales es el Boxplot adaptado a este contexto que ofrece la librería *fda*, la cual aporta una nueva perspectiva de la distribución de las curvas.

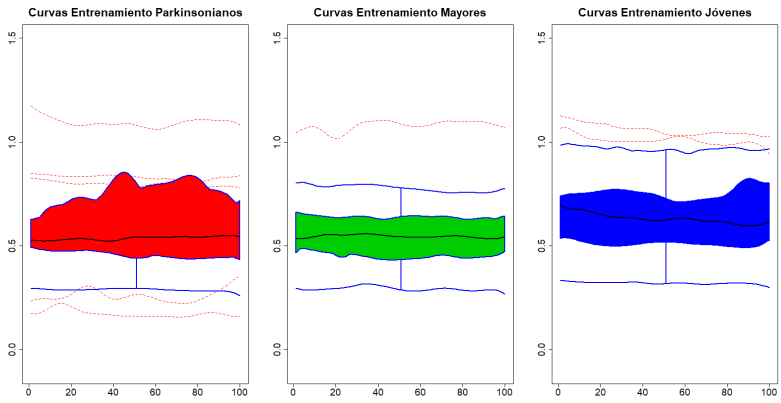

Figura 4.5: Boxplot funcional para cada uno de los grupos. A la izquierda sujetos diagnosticados con la enfermedad de Parkinson, en el centro sujetos de control de edad avazada, a la derecha sujetos de control jóvenes.

En la Figura 4.5 se puede apreciar que, producto de la variabilidad que hay dentro de los grupos, los boxplot consideran en cada uno de los grupos algunos sujetos como posibles curvas atípicas. Sin embargo, al verlas en conjunto (ver Figura 4.6), solo las curvas no. 9 y 48 (Parkinsoniano y Joven, respectivamente) fueron consideradas anómalas. Por otro lado, el analisis de deteccion de outliers no detectó curvas atípicas.

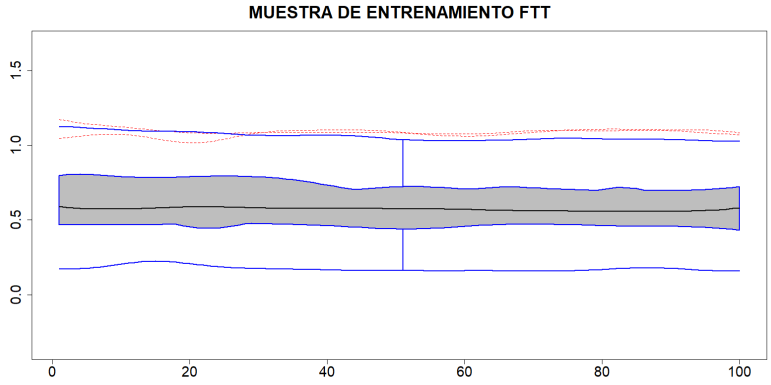

Figura 4.6: Boxplot Funcional para los datos del FTT

Se ha planteado la posibilidad de realizar alguna transformación a los datos de manera que podamos extraer nueva inofrmación de los mismos que contribuya a la posterior clasificación. En la Figuras 4.7 y 4.8, se presentan las derivadas de orden 1 y 2, respectivamente de las muestras de entrenamiento.

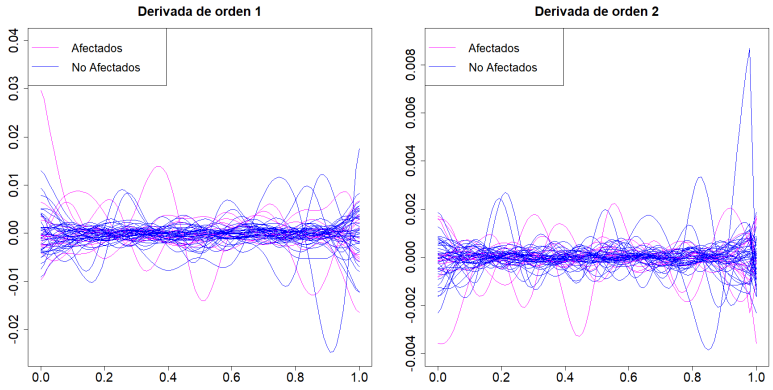

Figura 4.7: Derivadas de orden 1 y 2 de las curvas FTT según condición.

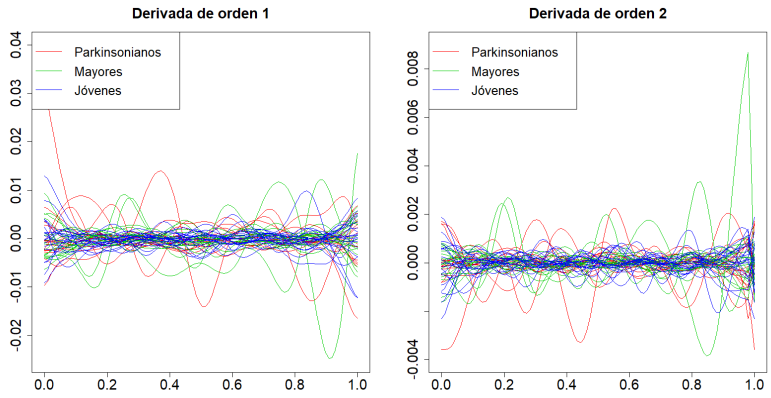

Figura 4.8: Derivadas de orden 1 y 2 de las curvas FTT según grupo de pertenencia

Al igual que sucedía con las curvas originales, vemos que los grupos continúan estando bastante entrecruzados, lo que sugiere que aún con aquellos métodos que utilizan la información procedente de las derivadas de los datos, se tengan dificultades a la hora de clasificar.

Una vez vistas las principales características de los datos, el siguiente paso consiste en determinar si existen diferencias significativas entre los sujetos afectados y no afectados con la enfermedad de Parkinson. Para ello, realizaremos el Test de Anova para datos funcionales (FANOVA), propuesto en [10].

Como se indica en [10], el estadístico en el cual está basado el FANOVA es:

$$V_n = \sum_{g<h} \sum n_g \|\bar{X}_{g\cdot} - \bar{X}_{h\cdot}\|^2,$$

donde g y h son subindices variando de 1 a G y G es el número de grupos, n_g el número de curvas en el grupo g, y $\bar{X}_{g\cdot}$ la media muestral funcional de las curvas en el grupo $g; g = 1, ..., G$. El test es unilateral a la derecha por lo que, para valores de para V_n lo suficiente grandes o equivalentemente $p.valores$ pequeños, se concluye que el test es significativo detectando evidencias estadísticas de diferencias entre alguna de las media funcionales de los G grupos.

En el caso que nos ocupa, $G = 2$, $n_1 = 15$ y $n_2 = 38$. Los resultados obtenidos tras aplicar el FANOVA (ver Figura 4.9), revelan que, con un $p - valor = 0{,}38$, no hay evidencia estadística significativa que haga pensar que la media entre los grupos de pacientes afectados y los pacientes no afectados con la enfermedad de Parkinson sean distintas.

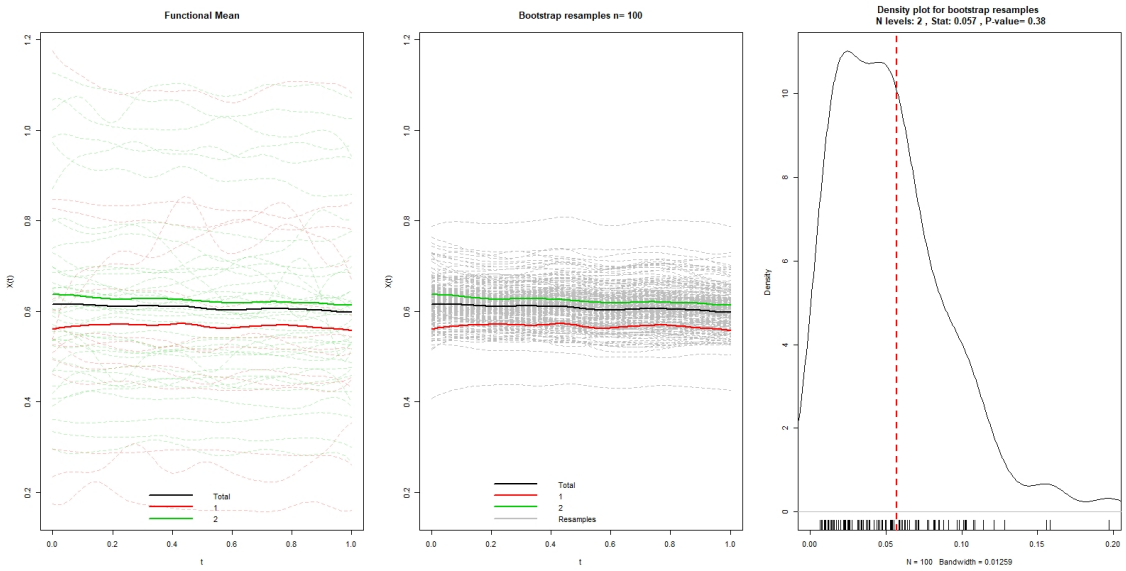

Figura 4.9: Resultado Anova Funcional para comparar tiempos entre afectados de Parkinson y no afectados

En un intento por encontrar diferencias entre las medias de los sujetos sometidos al FTT, se aplicó el FANOVA sobre los grupos de pacientes diagnosticados con Parkinson(n_1) y los individuos sanos de edad avanzada (n_2) y los diagnosticados con Parkinson(n_1) y los individuos sanos de edad joven(n_2). Aún haciendo esta nueva división, con *p.valores* de 0,963 y 0,175, respectivamente, no se hallaron evidencias en contra de la igualdad de medias tanto para Diagnosticados con Parkinson versus Mayores Sanos, como para Afectados versus Jóvenes Sanos.

4.2.2. Clasificación supervisada sobre los datos de Finger Tapping Test

Tal como se indicó en la sección previa, se dispone de una base de datos que contiene los ciclos de duración (en mili-segundos) obtenidos tras la aplicación del FTT sobre 53 sujetos (15 parkinsonianos, 19 sanos mayores y 19 sanos jóvenes) en dos días diferentes, siendo consideradas las observaciones del día 1 como las muestras de entrenamiento y las tomadas el día 2 como muestra de prueba. Para chequear la capacidad de clasificación de los métodos, se generaron aleatoriamente 50 submuestras del conjunto total de las muestras test. Cada una de estas submuestras, obtenidas independientemente unas de las otras, de tamaño 16 (30 % del tamaño

total). Aplicados los métodos de clasificación a cada submuestra se obtuvieron las correspondientes tasas de mala clasificación.

En un principio, se consideró utilizar únicamente los métodos que proporcionaron mejores resultados en las simulaciones, pero en vista de la poca información que gráficamente proporcionan las curvas, finalmente se optó por aplicar todas las técnicas implementadas en el estudio de simulación.

En primer lugar se proporcionan gráficos con los diagramas de caja para las tasas de mala clasificación, obtenidas para las 50 muestras de prueba, considerando el problema de discriminación entre enfermos parkinsonianos y sujetos sanos (independientemente de la edad). Mientras la Figura 4.10 muestra las tasas tras aplicar los métodos de clasificación a los datos originales, las Figuras 4.11 y 4.12 proporcionan los correspondientes valores al aplicar los métodos directamente sobre las derivadas de orden 1 y 2, respectivamente.

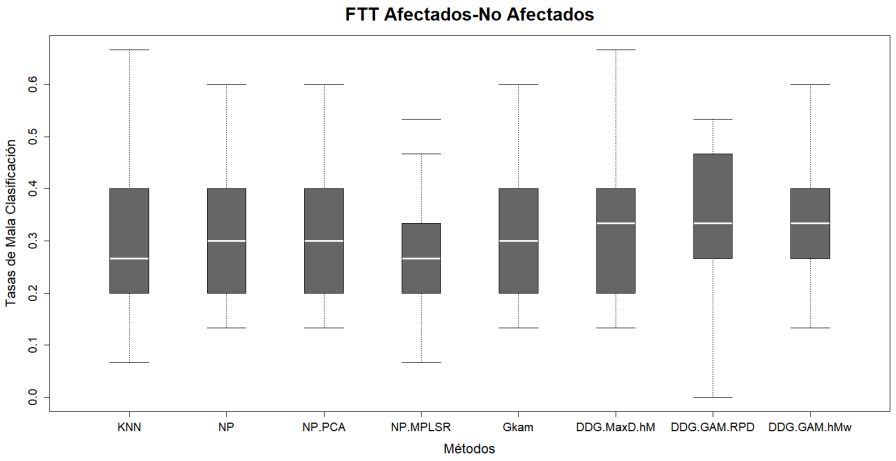

Figura 4.10: Tasas de mala clasificación para las curvas originales del FTT

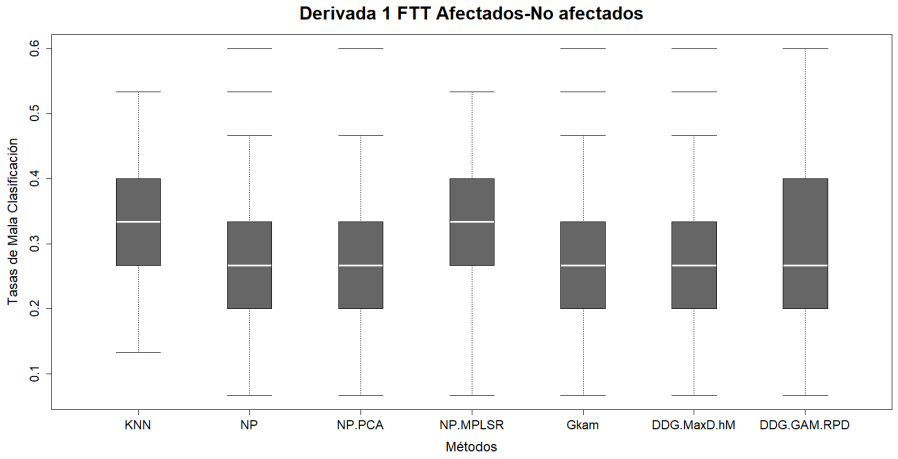

Figura 4.11: Tasas de mala clasificación para las primeras derivadas de los datos del FTT

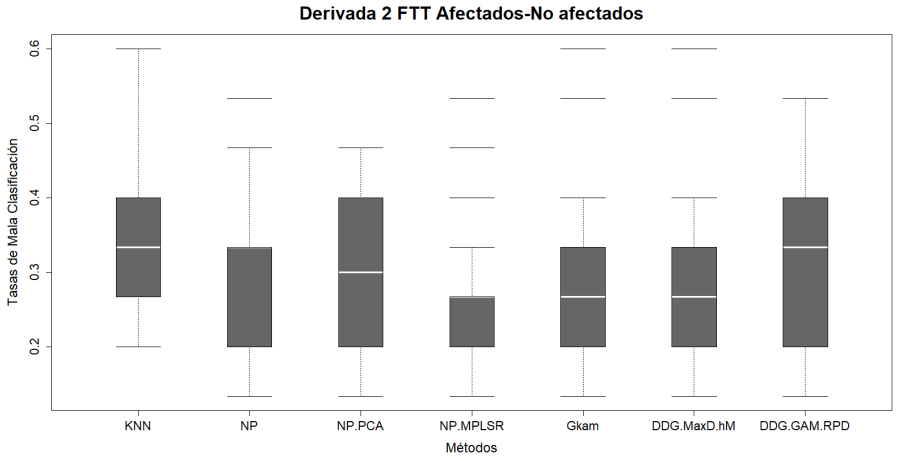

Figura 4.12: Tasas de mala clasificación para la segunda derivada de los datos del FTT

Al aplicar los clasificadores en el conjunto de datos funcionales y tratar de predecir la condición de un nuevo sujeto, se observa que la media de las tasas de mala clasificacion rondan el 30 %, consiguiéndose los mejores resultados (tasa media de 0,285 y desviación tipica de 0,11) con el método NP_{PLSR} y $q_{opt} = 3$.

Para analizar la información discriminante que puede ser extraída de las derivadas de las curvas, se aplican nuevamente los métodos de clasificación, excepto el clasificador $DD^G.Gam.hMw$. Algunos de los métodos parecen mejorar, tal es el caso de NP_{deriv}, cuya tasa de error pasa de un 30,4 % con $deriv = 0$, a un 27,7 % con $deriv = 1$; NP_{PCA} con $q = 4$ pasó de presentar una tasa de error media del 30,4 % a 26,8 %. Sin embargo, la tasa mas baja de mala clasificación fue obtenida al aplicar

el método NP_{MPLSR} con q_{opt}=2 sobre la derivada de orden dos de las curvas, con media de 0,257 y una desviación típica de 0,082.

De modo análogo, se aplicaron los métodos de discriminación de curvas tratando de separar, esta vez, las poblaciones Parkinsonianos y Mayores Sanos. Las Figuras 4.13, 4.14 y 4.15 presentan los boxplots de las tasas de mala clasificación correspondientes al aplicar los métodos a los datos originales, a las derivadas de orden 1 y orden 2, respectivamente.

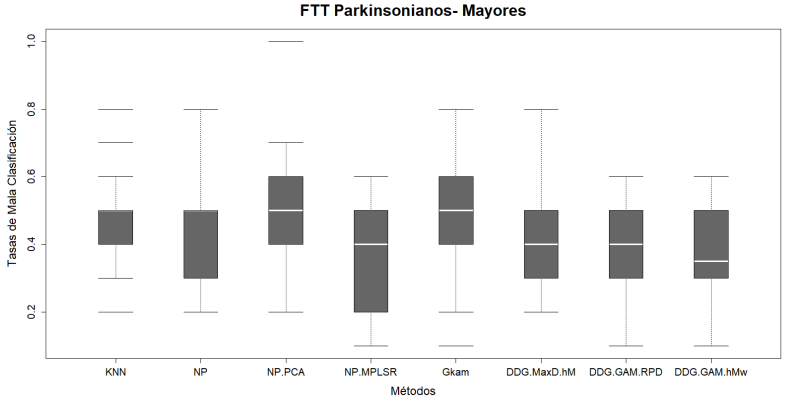

Figura 4.13: Tasas de mala clasificación de los datos del FTT en afectados de Parkinson y Mayores Sanos

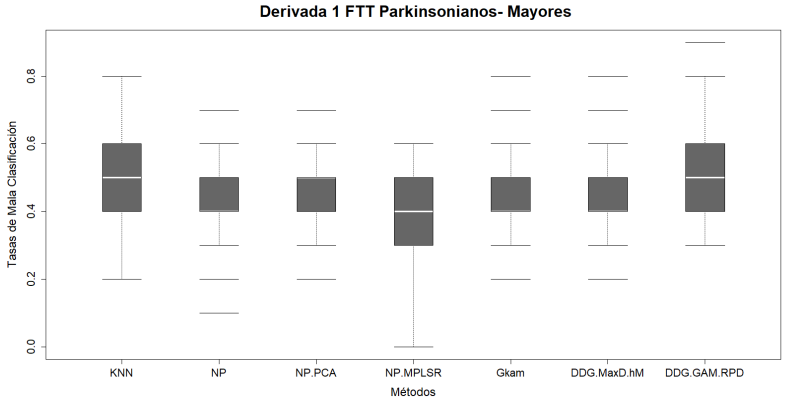

Figura 4.14: Tasas de mala clasificación para la primera derivada de los datos del FTT en afectados de Parkinson y Mayores Sanos

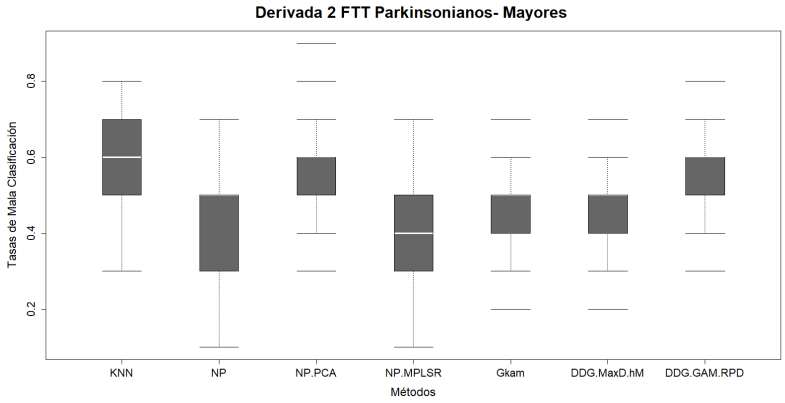

Figura 4.15: Tasas de mala clasificación para la segunda derivada de los datos del FTT en afectados de Parkinson y Mayores Sanos

Por lo que se puede apreciar en las Figuras anteriores, todos los métodos presentan dificultad a la hora de discriminar entre Afectados vs No Afectados y Afectados vs Mayores Sanos. Los niveles de variación oscilan entre el 12 % y el 15 % . En sentido general, el método NP con semimétrica MPLSR proporciona las tasas de mala clasificación más bajas, tal como ocurría en el estudio de simulación. Al utilizar las derivadas de las curvas, no se consigue una mejora considerable, lo que corrobora lo que se observó en el análisis exploratorio.

Para finalizar el análisis de datos, se muestran en la Tabla 4.1 las medias y desviaciones típicas obtenidas sobre las 50 muestras de prueba, al tratar de discriminar entre afectados de Parkinson y sujetos sanos de edad joven. Aunque entre estos dos grupos hay más diferencia, las mismas no son suficientes para permitir que nuestros métodos proporcionen tasas medias por debajo del 30 %. Nuevamente el método NP_{MPLSR} con $q_{opt} = 3$ es el que proporciona la tasa de mala clasificación mas baja, seguido del metodo KNN aplicado directamente sobre la segunda derivada de las curvas. El método $DD^G.Gam.RPD$ al ser aplicado sobre la primera derivada de las curvas, proporcionó las peores tasas de clasificación con una media del 50,6 % de error.

—Derivada	KNN	NP	NP_{PCA}	NP_{MPLSR}	$Gkam$	$DD^G.MaxD.hM$	$DD^G.Gam.RPD$	$DD^G.Gam.hMw$
0	0.358	0.366	0.334	0.292	0.360	0.336	0.412	0.362
	(0.1356)	(0.1318)	(0.1205)	(0.1337)	(0.1195)	(0.1045)	(0.1303)	(0.1259)
1	0.494	0.414	0.490	0.354	0.442	0.430	0.506	N/A
	0.1252	(0.1355)	(0.1216)	(0.1110)	(0.1213)	(0.1165)	(0.1376)	N/A
2	0.314	0.384	0.392	0.400	0.438	0.438	0.440	N/A
	(0.1125)	(0.1375)	(0.1468)	(0.1261)	(0.1227)	(0.1227)	(0.1498)	N/A

Tabla 4.1: Media y Desviación típica (en paréntesis) de las tasas de mala clasificación para los métodos aplicados sobre los datos de FTT con el propósito de discriminar entre afectados de Parkinson y sujetos sanos de edad joven

Conclusiones

Tal como se indica en la propuesta inicial de este TFM, el objetivo era triple: (1) Realizar una revisión de los métodos de clasificación supervisada para datos funcionales existentes en la literatura, prestando especial atención al caso particular de los test de diagnóstico ($k = 2$); (2) Desarrollar software con el propósito de comparar los métodos anteriores mediante un estudio de simulación; (3) Abordar el problema de discriminación entre afectados y no afectados de Parkinson partiendo de datos funcionales del ritmo motor derivados del Finger Tapping Test. Si bien los dos primeros objetivos fueron cumplidos, logrando una selección de los "mejores"métodos de discrimación, la aplicación de estos al conjunto de datos reales no rindió los resultados deseados. La falta de capacidad predictiva de los diversos procedimientos para detectar afectados de Parkinson en base a los tiempos de FTT puede ser debida a:

1. Pocos puntos de discretización para los sujetos de estudio;

2. Tamaños muestrales bajos y desequilibrados, especialmente en las muestras de aprendizaje, de los diferentes grupos (Parkinsonianos, Adultos Sanos y Jóvenes Sanos);

3. La información funcional proporcionada por estos tiempos no sea la más apropiada para diferenciar los grupos antes mencionados.

En todo cao, bajo estas condiciones, tanto en el estudio de simulación como en la aplicación a datos reales el método NP_{MPLSR} destacó por encima de los demás competidores, ofreciendo mejores resultados en la mayor parte de los escenarios vistos. No sólo se puede considerar como el método con un mejor comportamiento global, si no que ha conseguido tasas de error muy bajas en todas aquellas situaciones en las que los grupos podían ser diferenciados.

En la aplicación a datos reales, La principal conclusión que se pudo obtener tras la realización de este trabajo, ha sido que debido a la naturaleza de los datos, los métodos encontraron mucha dificultad para discriminar entre los distintos grupos considerados. El diagnóstico de la enfermedad de Parkinson no es es tarea sencilla, en Escocia, por ejemplo, según una encuesta publicado en [35], se demostró que el 56 % de los sujetos que recibían tratamiento contra la enfermedad de Parkinson se encontraban erróneamente diagnosticados. Por otra parte, [28] y [37] presentan estudios donde se han reportado tasas de error diagnóstico de hasta el 24 %.

Bibliografía

[1] Aneiros, G., Cao R., Fraiman R., Genest C., and Vieu P. (2019). Selected review on functional data analysis and related topics. *J. Multivariate Anal.* 146, 1-6.

[2] Arias, P., Robles-García, V., Espinosa, N., Corral Y. and Cudeiro, J. (2012). Validity of the finger tapping test in Parkinson's disease, elderly and young healthy subjects: Is there a role for central fatigue?. *Clinical Neurophysiology* 123, 2034 - 2041.

[3] Baíllo, A., and Cuevas, A. (2008). Supervised Classification for Functional Data: A Theoretical Remark and Some Numerical Comparisons. *In: Functional and Operatorial Statistics. Contributions to Statistics*. Physica-Verlag HD.

[4] Biau, G., Bunea, F. y Wegkamp, M.H.(2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*. DOI:10.1109/tit.2005.847705

[5] Cérou, F. y Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM Probab. Stat.* 10, 340-355.

[6] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local Regression Models. In: S, J. M. Chambers and T. Hastie, (eds) *Statistical Models in S* . Chapman and Hall, New York.

[7] Cleveland, W.S., and Loader, C. (1996). Smoothing by Local Regression: Principles and Methods. In: Härdle W., Schimek M.G. (eds) *Statistical Theory and Computational Aspects of Smoothing*. Contributions to Statistics. Physica-Verlag HD.

[8] Cuesta-Albertos, J.A., Febrero-Bande, M., and Oviedo de la Fuente, M. (2017) The DDG-classifier in the functional setting. *TEST*, 26. 119-142.

[9] Cuesta-Albertos, J.A., and Nieto- Reyes, A. (2008). The Random Tukey Depth.*The Annals of Statistics.* Volume 28, Number 2, 461-482.

[10] Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational Statistics and Data Analysis*, 47, 111-122.

[11] Cuevas, A., Febrero-Bande, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* 22, 3, 481-496.

[12] Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *J. Statist. Plann. Inference,* 147, 1-23.

[13] Cuevas, A. and Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data.*Journal of Multivariate Analysis* 100, 753-766.

[14] Estévez-Pérez, G., and Vieu, P. (2019). A new way for ranking functional data with applications in diagnostic. *TEST*. Under revision.

[15] Febrero, M., Galeano, P., and González-Manteiga, W. (2011). Generalized Additive Models for Functional Data. *Test*. 22, 91-96.

[16] Febrero-Bande, M., and Oviedo de la Fuente, M. (2009). Statistical computing in functional data analysis: The R package fda. usc. *Journal of Statistical Software* Volume 51, Issue 4.

[17] Ferraty F, and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Series in Statistics New York: Springer.

[18] Ferraty F., and Vieu, P. (2003). Curves discrimination: A Nonparametric Functional Approach. *Computational Statistics and Data Analysis* 44, 161-173.

[19] Fisher, R. A.(1936).The Use of Multiple Measurments in Taxonomic Problems. *Annals of Eugenics* 7, 179-188.

[20] Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data *Test* 10: 419. https://doi.org/10.1007/BF02595706.

[21] Hall, P. and Van Keilegomm, I. (2007). Two-Sample Test in Functional Data Analysis Starting from Discrete Data.*Statistica Sinica* 17, 1511-1531.

[22] Hall, P., Poskitt, D. and Presnell, B. (2001). A Functional Data-Analytic Approach to Signal Discrimination, *Technometrics*, 43:1, 1-9, DOI: 10.1198/00401700152404273.

[23] Hastie, T., Buja, A., and Tibshirani, R.(1995). Penalized Discriminant Analysis. *The Annals of Statistics*, 23, 73-102.

[24] Hastie T., Tibshirani R. and Buja A. (1994). Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association*, 89, 1255-1270.

[25] Härdle, W. (1990). *Applied Nonparametic Regression.* Cambridge University Press, Cambridge.

[26] Hernández, A., Morales, V., and García, V. (2011). Finger Taping Test. Precisión del diseño de medidas entre muestras de deportistas de élliite y no deportistas. *Cuadernos de Psicología del Deporte*, 11 (1), 29-43.

[27] Horváth, L., and Kokoszka, P. (2012). *Inference for functional data with applications.* Springer.

[28] Hughes, A.J., Daniel, S..E, Blankson, S., and Lees, A.J. (1993). A clinicopathologic study of 100 cases of Parkinson disease. *Arch Neurol.* 50, 140-8.

[29] James, G., and Hastie, T. (2001). Functional Linear Discriminant Analysis for Irregularly Sampled Curves. *Journal of the Royal Statistical Society.* Series B (Statistical Methodology), 63 (3), 533-550.

[30] Li, J., Cuesta-Albertos, J., and Liu R. (2012). Functional data classification using covariate-adjusted subspace projection. *Journal of the American Statistical Association*, 17, 737-753.

[31] Liu R.Y. (1990) On a notion of data depth based on random simplices. *Ann Stat*, 18, 405-414.

[32] Liu R. M., Parelius, J. and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics.* 27. DOI: 10.1214/aos/1018031260.

[33] López-Álvarez, X. (2015). *Clasificación supervisada en datos multivariantes, de alta dimensión e funcionais*. TFM, Universidade de Santiago.

[34] López-Pintado, S., and Romo, J. (2005). Depth-based classification for functional data. DIMACS Series in Discrete Mathematics, *American Mathematical Society*, 72, 103-119.

[35] Newman, E.J., Breen, K., Patterson, J., Hadley, D.M., Grosset, K.A., and Grosset, D.G. (2009). Accuracy of Parkinson's disease diagnosis in 610 general practice patients in the West of Scotland *Mov Disord*, 24, 2379-85.

[36] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[37] Rajput, A. H., Rozdilsky, B., and Rajput, A. (1991). Accuracy of clinical diagnosis in parkinsonism: a prospective study. *Can J Neurol Sci.* 18, 275-8.

[38] Ramsay, J.O., and Silverman, B.W. (2005) *Functional Data Analysis*. Second. New York: Springer.

[39] Roca-Pardiñas J., GOnzález-Manteiga W., Febrero-Bande M., Prada-Sánchez J.M. y Cadarso-Suárez C. (2004) Prediction binary time series of SO2 using Generalized Additive Models with unknown link function. *Environmetrics* 15, 1-14.

[40] Serfling, R. (2006). Depth functions in nonparametric multivariate inference. In: Liu R., Serfling R., Souvaine D.L. (eds) Data depth: robust multivariate analysis, computational geometry and applications. DIMACS Series. *American Mathematical Society*, Providence, 1-16.

[41] Sguera, C., Galeano, and P., Lillo, R. (2014). Spatial depth-based classification for functional data. Test 23, 725-750.

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