**UNIVERSIDAD DE BUENOS AIRES**
**Facultad de Ciencias Exactas y Naturales**
**Departamento de Matemática**

## Tesis de Licenciatura

# Clasificación de datos funcionales

**Ulises Bercovich Szulmajster**

**Directora: Dra. Graciela Boente Boente**

Julio de 2022

# Clasificación de datos funcionales

El problema de clasificación de datos funcionales presenta características que lo distinguen del caso multivariado. Algunas de estas son la importancia de la regularidad de un elemento aleatorio o la ausencia de la función de densidad. Como consecuencia de estas diferencias, muchos de los clasificadores utilizados en el caso multivariado dejan de ser factibles para el caso funcional. Esta tesis parte de la revisión de una selección de métodos de clasificación para ambos escenarios, así como la exposición de distintas nociones de profundidad y atipicidad. Finalmente, se presenta un nuevo método no paramétrico basado en dos herramientas, a saber, el DD-plot combinado con una medida de profundidad. El DD-plot es una herramienta de reducción de la dimensión y visualización de datos que se utilizará para pasar de la dimensión infinita a un espacio de baja dimensión. Por otra parte, la medida de profundidad considerada fue creada específicamente para el caso funcional y tiene en cuenta dos magnitudes de importancia, la distancia entre curvas y la similaridad de sus formas.

*Palabras Clave:* Clasificación; Datos Funcionales; DD-plot; Profundidad; Atipicidad direccional.

# Agradecimientos

Esta tesis no podría haber sido llevada a cabo sin tanta gente (y animales) a mi alrededor:

A mis viejxs, por tanto amor y apoyo, por grandísima juventú. En particular, para esta tesis, por el amor a las estructuras, al aprendizaje y la lectura.

A mi hermana, por haberme hecho el camino infinitamente más fácil. Por los colores y el brillo.

A mi abuelo, por mostrarme que ir al Colón o ir a la cancha a ver al Bicho al final es lo mismo.

A Luna, Nina y Oli, por tan bestial compañía. Por ser un poco como ustedes.

A mis amistades de la facu, Bruno, Checha, Chehebar, Darío, Ger, Gonza, Ivo, Jaz, Jesi, Marian, Martín, Mati, Nets, Pablito, Santi, Solcito, Tano y Vane, por poder contar siempre con ustedes. Imposible concebir mi carrera sin tenerlxs.

A mis amistades de la vida por el ocio. En particular a Berdi, Fede y Guido, por una relación indispensable que sigue (y siga) trascendiendo.

A Tina, por hacer que todo se parezca más a un bosque.

A Graciela, por presentarme un tema de tesis tan interesante. Por la ayuda con todo lo logrado.

A las juradas Marina y Dani, por la lectura y la devolución, entre muchísimas otras cosas. En especial al área de estadística del IC por hacerme sentir su amor al campo.

A toda la gente del DM y del IC, por haberme hecho disfrutar tanto la carrera con gente tan humana. A mi facultad, por ser un lugar donde siempre me voy a sentir comodx. A la UBA, por mantener su enorme calidad a pesar de todo.

# Índice

# Capítulo 1

# Introducción

## 1.1 Introducción

A lo largo de las últimas décadas, con el aumento de los datos recolectados y almacenados, se empezaron a recopilar datos que ya no provienen sólo de variables o vectores aleatorios, si no que ahora las observaciones podrían representar instancias de un objeto de dimensión infinita, por ejemplo los píxeles de una imagen o la medición de una función continua, como puede ser un proceso estocástico a lo largo del tiempo. Ante esta situación, los métodos utilizados tradicionalmente en la estadística multivariada se enfrentan a diversos problemas, tales como la alta dimensionalidad de los datos, ya que los puntos en los que se registra un proceso estocástico pueden corresponder a una grilla que puede ser arbitrariamente fina. Además, tendremos cualidades propias de los datos funcionales, como la noción de regularidad. Dicha regularidad no existe cuando se consideran datos multivariados en $\mathbb{R}^p$ y puede ser de suma importancia en el contexto funcional. Cabe también mencionar que en el caso de tratar con este tipo de datos funcionales no es razonable realizar cambios de coordenadas, práctica usual en el análisis multivariado, pero que en el caso funcional puede provocar una pérdida de información al existir un orden natural en el que los datos se registran. Por estas razones surge la rama del análisis de datos funcionales que se separa del análisis univariado y multivariado por necesitar distintas herramientas y tener tantas aplicaciones en diferentes áreas, por ejemplo dentro del reconocimiento del habla o dentro del análisis de imágenes como puede ser el reconocimiento de escritura.

Independientemente del análisis de datos funcionales, dentro de la estadística una de las áreas más importantes es la del aprendizaje supervisado. En particular, nos interesaremos en el problema de clasificación. Este problema parte de la existencia de varias poblaciones o grupos dentro de un mismo espacio. Luego, a partir de información que tenemos, como un conjunto de datos para los cuales sabemos a qué población pertenece cada uno, se pretende dar una regla de clasificación a modo de poder asignar una nueva observación a alguna de las poblaciones consideradas. Partiendo de los métodos de clasificación del extensamente desarrollado caso multivariado, se buscará dar el salto al caso funcional, con los cuidados

que este nuevo espacio de dimensión infinita necesita.

## 1.2 Estructura de la tesis

En el **Capítulo 2** se presenta la definición de un elemento aleatorio en un espacio de Hilbert. Además, se definen los conceptos de esperanza y covarianza, observando qué propiedades necesita el elemento aleatorio para que estas estén bien definidas. Se describen estas nociones con particular énfasis en el espacio $L^2(\mathcal{I})$, el espacio de funciones de cuadrado integrable sobre un intervalo $\mathcal{I} \subset \mathbb{R}$ compacto. Presentamos además el Teorema de Karhunen-Loève, que permite descomponer un elemento aleatorio a partir de las autofunciones del operador de covarianza. Describimos las hipótesis necesarias junto con la motivación de usar esta descomposición.

En el **Capítulo 3** se describe el problema de clasificación, comenzando con la presentación de la regla de clasificación Bayes. A partir de esta, se definen otros métodos clásicos tanto paramétricos como no paramétricos para el caso multivariado, que además se adaptarán para ser utilizados en el método propuesto en la última sección. Luego, se presentan diversos métodos usados actualmente para el caso funcional.

En el **Capítulo 4** se define la noción de profundidad y su relación con la atipicidad. Se presentan las características generales de dichas funciones de profundidad, pudiendo separarlas en cuatro categorías distintas. Se introducen luego nociones de profundidad para el caso univariado, de modo a extenderlas al caso multivariado donde se dan más ejemplos, viendo además si cumplen las propiedades consideradas en el inicio del capítulo. Finalmente, se analiza el caso de las profundidades para datos funcionales, donde también se describen nuevos ejemplos.

Por último, en el **Capítulo 5**, se expone el método propuesto en esta tesis para la clasificación de datos funcionales. Para esto se utilizarán herramientas trabajadas en los capítulos anteriores y se presentarán resultados de simulación que permiten comparar la propuesta dada con otras existentes.

# Capítulo 2

# Nociones previas

## 2.1 Introducción

En este capítulo daremos algunas nociones necesarias para el problema que abordaremos en los siguientes capítulos. Buscaremos, a partir de los definiciones y resultados conocidos para el caso de variables o vectores aleatorios, resumir parte de la teoría existente sobre elementos aleatorios en espacios de Hilbert reales, es decir, espacios con producto interno que además son completos con la norma asociada a dicho producto interno. Supondremos además que el espacio de Hilbert a considerar es separable, lo que nos permitirá centrarnos en el caso $\mathbb{H} = L^2(\mathcal{I})$, siendo $\mathcal{I}$ un intervalo compacto. Indicaremos respectivamente por $\langle \cdot, \cdot \rangle$ y por $\| \cdot \|$ el producto interno y la norma en dicho espacio, es decir, dadas funciones $u, v \in L^2(\mathcal{I})$, $\langle u, v \rangle = \int_{\mathcal{I}} u(t)v(t)\,dt$ mientras que $\|u\|^2 = \int_{\mathcal{I}} u^2(t)\,dt$. Finalmente, de ahora en más indicaremos con mayúsculas $X$, $Y$ a elementos aleatorios en el espacio $\mathbb{H}$, mientras que vectores aleatorios se indicarán en negrita y minúscula, es decir, como $\mathbf{x}$. Por otra parte, las matrices fijas en $\mathbb{R}^{p\times q}$ se denotarán con mayúscula y negrita, o sea, como $\mathbf{A}$, $\mathbf{B}$ por ejemplo.

En la Sección 2.2 presentaremos la definición de elementos aleatorios en espacios de Hilbert, mientras que en las Secciones 2.3 y 2.4 presentaremos la extensión al caso infinito-dimensional de la esperanza y matriz de covarianza definidas para vectores aleatorios. Finalmente, en la Sección 2.5 probaremos el Teorema de Karhunen-Loève que permite justificar el amplio uso de las componentes principales y la representación de datos funcionales en forma concisa, reduciendo su dimensión mediante la proyección en los primeros elementos de la base de autofunciones del operador de covarianza. Esta aproximación es de suma importancia en problemas de regresión, pues permite circunvalar los problemas producidos por la dimensión infinita. Para una descripción exhaustiva de estas nociones, referimos a los libros de Bosq (2000) y Horvath y Kokoska (2012).

## 2.2 Elementos Aleatorios

La Figura 2.1 muestra un ejemplo del tipo de datos con el que trabajaremos. El panel superior de este gráfico presenta observaciones correspondientes a realizaciones de un elemento aleatorio $X \in L^2([1, 18])$. Dicho elemento aleatorio consiste en la altura de una niña elegida al azar a lo largo del tiempo, más precisamente, entre su primer año de vida y los 18 años. Los datos graficados en el panel (a) corresponden a 10 observaciones del conjunto total de 54 niñas cuyas alturas se midieron sobre un conjunto de 31 edades (de 1 a 18 años) en el estudio Berkeley Growth Study. El panel (b) presenta las altura del conjunto completo de 54 niñas y los valores intermedios se interpolaron graficando las curvas. Los datos pueden encontrarse en la librería `fda` de `R`.

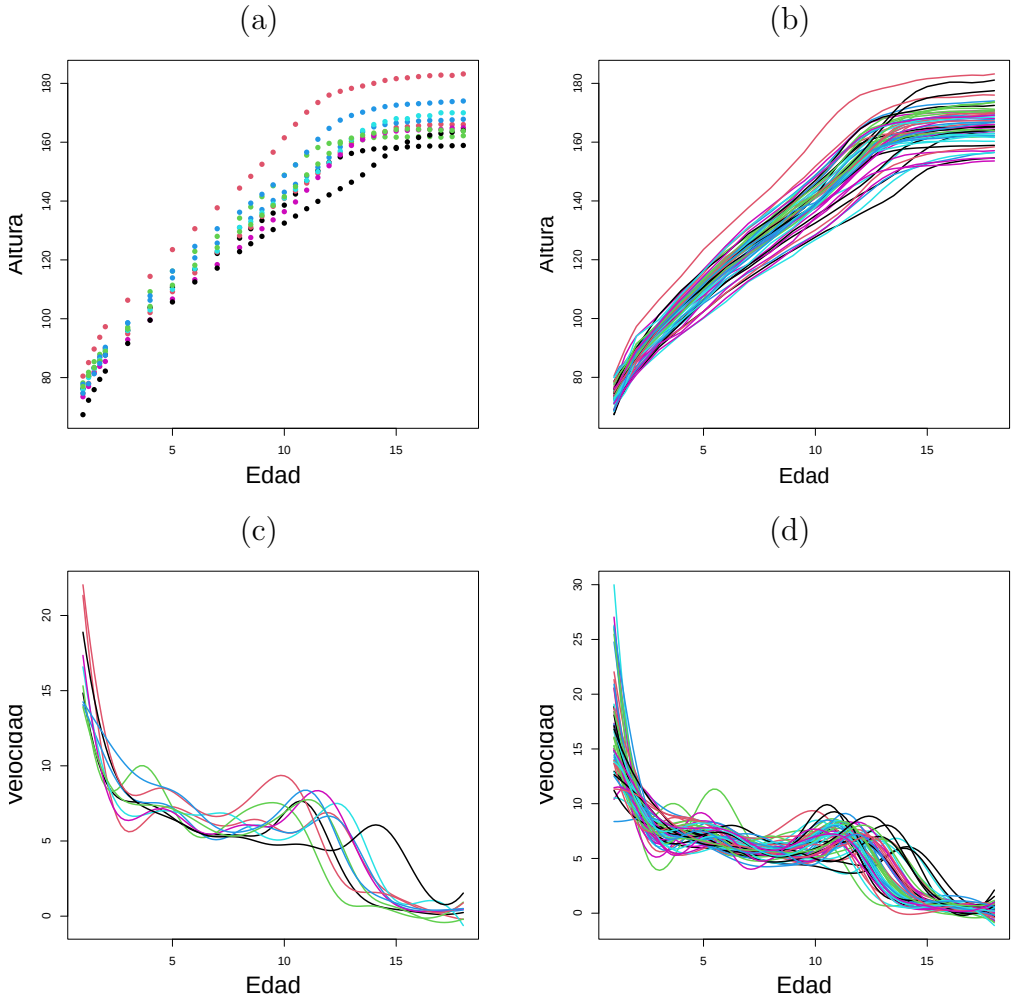


Figura 2.1: Datos de altura de niñas medidas en 31 momentos distintos desde el primer año de vida hasta los 18 (Berkeley Growth Study, disponibles en la librería fda de `R`): (a) Altura de 10 de las niñas, (b) Altura de las 54 niñas, (c) Velocidad de 10 de las niñas, (d) Velocidad de las 54 niñas.

Aunque estas mediciones son tomadas en una grilla de tiempos, no podemos dejar de

pensarlas como realizaciones de un dato funcional. Más precisamente, $X_i(t)$ representa la altura de la $i-$ésima niña, $1 \leq i \leq n$, cuando el tiempo $t$ varía en el intervalo $[1, 18]$. Nuestro dato corresponde a una observación discretizada del proceso, donde todos los datos se registran en la misma grilla de tiempos $\{t_j\}_{1 \leq j \leq p}$, con $t_1 \leq \cdots \leq t_p$. Más precisamente, observamos $X_{ij} = X_i(t_j)$, $1 \leq i \leq n$, $1 \leq j \leq p$, donde en nuestro caso $n = 54$ y $p = 31$. Si bien en muchas situaciones, por ejemplo cuando los registros corresponden a un instrumento, como un electroencefalograma o a una imagen por resonancia magnética, la grilla de puntos es usualmente equiespaciada $t_j - t_{j-1} = t_{j+1} - t_j$ para todo $j$, las técnicas usuales para datos multivariados que consisten en considerar el vector $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ no son las más adecuadas, ya que no tienen en cuenta la estructura subyacente de las curvas como su continuidad o diferenciabilidad ni el orden natural en que se realizan las mediciones. Por esta razón, se han desarrollado diversas técnicas de análisis para datos funcionales que sacan provecho del hecho de que $X$ o la cantidad a estimar presentan alguna información de regularidad que puede ser explotada de forma funcional, como la mencionada continuidad o diferenciabilidad. A partir de esta perspectiva, podremos usar herramientas que no tenemos en el caso univariado o multivariado, como el análisis de la forma de cada curva mediante las derivadas.

Para el conjunto de datos relacionado con las alturas de las niñas del estudio Berkeley Growth Study, los paneles (c) y (d) de la Figura 2.1 presentan las derivadas de la altura para las 10 niñas elegidas y para el total de las niñas medidas. Podemos notar cómo la derivada de la función de altura no es constante y se va acercando a cero a medida que termina la pubertad, por lo que a partir de la pubertad la altura en el tiempo $s$ es más fácil de predecir a partir de las alturas previas. Esta observación sugeriría que si consideramos dos tiempos $t$ y $s$ posteriores a la pubertad, la correlación entre $X(t)$ y $X(s)$ debería tener valores cercanos a 1. Este es el tipo de herramientas que usaremos para el análisis de datos funcionales y que describiremos a lo largo de este capítulo.

La definición formal de un elemento aleatorio en un espacio de Hilbert separable se puede dar de manera análoga al caso univariado de variables aleatorias.

**Definición 2.2.1.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y sea $\mathbb{H}$ un espacio de Hilbert con su $\sigma$-álgebra de Borel $\mathcal{B}(\mathbb{H})$. Dado $X : \omega \to \mathbb{H}$, diremos que es un elemento aleatorio de $\mathbb{H}$ si $X^{-1}(A) \in \mathcal{F}$ para todo $A \in \mathcal{B}(\mathbb{H})$.

## 2.3 Esperanza

La esperanza es una herramienta fundamental de la estadística para representar la centralidad de una distribución. Por lo tanto, extender esta noción al caso funcional es necesario ya que diversos métodos estadísticos usan estimadores de la esperanza en su desarrollo. En el caso de variables o vectores aleatorios, la esperanza se calcula a partir de la función de distribución o de la función de densidad cuando esta última existe. Un primer problema que surge al trabajar con elementos aleatorios en espacios de Hilbert de dimensión infinita es

la ausencia tanto de la función de distribución como de una función de densidad, como se discute en Delaigle y Hall (2010).

En esta sección, extenderemos la noción de esperanza al caso funcional. Como se mencionó en la Sección 2.1, seguiremos la descripción dada en Horvath y Kokoska (2012). Para esto presentaremos una definición más general de la esperanza, que podremos calcular usando herramientas propias de los espacios de Hilbert como el producto interno y la completitud del espacio.

Comencemos recordando el Teorema de representación de Riesz, que será de utilidad en lo que sigue.

**Teorema 2.3.1.** (Teorema de representación de Riesz) Sean $\mathbb{H}$ un espacio de Hilbert y $\mathbb{H}^*$ su espacio dual. Entonces, para cada funcional lineal continuo $L \in \mathbb{H}^*$, existe un único elemento $u \in \mathbb{H}$ tal que $L(v) = \langle u, v\rangle$.

*Demostración.* Sea $L \in \mathbb{H}^*$ un funcional lineal continuo. Si $L(x) = 0$ para todo $x$, entonces $L(x) = \langle x, 0\rangle$ para todo $x$. Supongamos entonces que $L \not\equiv 0$. Sea $\mathcal{N} = \{x : L(x) = 0\}$ el núcleo de $L$. Como $L$ es continuo, $\mathcal{N}$ es un subespacio cerrado. Por lo tanto, podemos descomponer a $\mathbb{H}$ como suma directa de $\mathcal{N}$ y $\mathcal{N}^\perp$, con $\mathcal{N}^\perp \neq \{0\}$.

Sea $z \in \mathcal{N}^\perp$ con $\|z\| = 1$. Dado $v \in \mathbb{H}$ arbitrario, definamos $\lambda = L(v)/L(z)$ y $w = v - \lambda z$, entonces se cumple que $v = w + \lambda z$ donde $w \in \mathcal{N}$. Resumiendo, hemos probado que todo elemento $v \in \mathbb{H}$ se descompone como un múltiplo de $z$ y un elemento $w \in \mathcal{N}$, lo que implica que

$$0 = \langle z, w\rangle = \langle z, v - \lambda z\rangle = \langle z, v\rangle - \lambda\|z\|^2.$$

Como $\|z\| = 1$, tomando $u = zL(z)$ se concluye que $L(v) = \langle u, v\rangle$ para todo $v \in \mathbb{H}$.

Para probar la unicidad supongamos que existen $u_1$ y $u_2$ tales que $L(v) = \langle u_1, v\rangle = \langle u_2, v\rangle$ para todo $v \in \mathbb{H}$. Luego $\langle u_1, v\rangle - \langle u_2, v\rangle = 0$, de donde tomando $v = u_1 - u_2$ se obtiene que $u_1 = u_2$. □

Este teorema será de suma importancia para demostrar la existencia y unicidad de la esperanza dada en el Teorema 2.3.2. Veamos ahora una condición necesaria para que un elemento aleatorio tenga esperanza.

**Definición 2.3.1.** Un elemento aleatorio $X$ se dice integrable si $\mathbb{E}\|X\| < \infty$.

Ahora ya podremos enunciar el teorema central de la sección sobre la existencia y unicidad de la esperanza en el caso funcional, que notaremos con la letra $\mu$.

**Teorema 2.3.2.** Si $X$ es integrable, entonces para todo $v \in \mathbb{H}$, $\mathbb{E}\langle X, v\rangle < \infty$. Además, existe un único $\mu \in \mathbb{H}$ tal que $\mathbb{E}\langle X, v\rangle = \langle \mu, v\rangle$.

*Demostración.* Supongamos que $X$ es integrable y sea $M = \mathbb{E}\|X\|$. La desigualdad de Cauchy-Schwartz implica que

$$\mathbb{E}|\langle X, v\rangle| < \mathbb{E}[\|X\|\|v\|] = M\|v\|\,. \tag{2.1}$$

Por lo tanto, el operador $L : \mathbb{H} \to \mathbb{R}$ dado $L(v) = \mathbb{E}\langle X, v\rangle$ está bien definido. La bilinealidad del producto interno y la linealidad de la esperanza implican que $L$ es lineal. Por otra parte, de (2.1) deducimos que el operador lineal $L$ es acotado. Por lo tanto, por el Teorema 2.3.1, existe un único $\mu$ tal que $\mathbb{E}\langle X, v\rangle = \langle \mu, v\rangle$, lo que concluye la demostración. □

**Definición 2.3.2.** Sea $X$ un elemento aleatorio integrable. Definimos la esperanza de $X$ como el elemento $\mu \in \mathbb{H}$ que satisface $\mathbb{E}\langle X, v\rangle = \langle \mu, v\rangle$ para todo $v \in \mathbb{H}$.

Un teorema indispensable a la hora de trabajar con integrales es el Teorema de Fubini, para ello será necesario tener algún resultado que permita acotar la esperanza $X$. Presentaremos además una de las principales propiedades de las esperanza: la conmutatividad de esta con los operadores lineales, que resulta fundamental al trabajar con integrales o productos internos.

**Proposición 2.3.1.** Si $X$ es integrable, entonces:

a) $\|\mathbb{E}\, X\| \leq \mathbb{E}\|X\|$,

b) para todo operador lineal acotado $L$, $\mathbb{E}[L(X)] = L(\mathbb{E}\, X)$.

*Demostración.* a) Sea $\mu = \mathbb{E}\, X$. Usando que si $Y$ es una variable aleatoria $|\mathbb{E}\, Y| \leq \mathbb{E}|Y|$ y que la esperanza es única, obtenemos que $|\langle \mu, v\rangle| = |\mathbb{E}\langle X, v\rangle| \leq \mathbb{E}|\langle X, v\rangle| \leq \|v\|\mathbb{E}\|X\|$. Luego, tomando $v = \mu$, deducimos que $\|\mu\|^2 \leq \|\mu\|\mathbb{E}\|X\|$, lo que concluye la demostración de a).

b) Primero veamos que $\mathbb{E}[L(X)]$ está bien definida. Efectivamente, como $L$ es acotado, tenemos que $\mathbb{E}|L(X)| \leq \mathbb{E}\big(\|L\|\,\|X\|\big) = \|L\|\mathbb{E}\|X\| < \infty$. Luego, si denotamos por $\alpha$ a $L(\mu)$, bastará ver que $\mathbb{E}[L(X)] = \alpha$. Sea $L^*$ el adjunto de $L$. Entonces, utlizando el Teorema 2.3.2 deducimos que

$$\langle \alpha, v\rangle = \langle L(\mu), v\rangle = \langle \mu, L^*(v)\rangle = \mathbb{E}\langle X, L^*(v)\rangle = \mathbb{E}\langle L(X), v\rangle\,.$$

lo que concluye la demostración. □

Consideremos el caso particular en que $\mathbb{H} = L^2(\mathcal{I})$. Por el Teorema 2.3.2, un elemento aleatorio $X \in L^2(\mathcal{I})$ será integrable y tendrá esperanza si

$$\mathbb{E}\|X\| = \mathbb{E}\left[\left(\int_{\mathcal{I}} X^2(t)dt\right)^{\frac{1}{2}}\right] < \infty\,.$$

Con el siguiente resultado buscaremos caracterizar explícitamente la esperanza de $X$.

**Proposición 2.3.2.** Sea $X \in L^2(\mathcal{I})$ integrable y sea $\mu$ su esperanza. Entonces, vale que $\mu(t) = \mathbb{E}[X(t)]$ para casi todo $t \in \mathcal{I}$.

*Demostración.* Por el Teorema 2.3.2, sabemos que para todo $v \in L^2(\mathcal{I})$

$$\langle \mu, v \rangle = \mathbb{E}\langle X, v \rangle = \mathbb{E}\left( \int_{\mathcal{I}} X(t)v(t)dt \right) .$$

Sea $\nu(t) = \mathbb{E}(X(t))$. Para obtener el resultado, bastará probar que para todo $v \in L^2(\mathcal{I})$

$$\mathbb{E}\left( \int_{\mathcal{I}} X(t)\, v(t)\, dt \right) = \int_{\mathcal{I}} \nu(t)\, v(t)\, dt = \langle \nu, v \rangle ,$$

pues en dicho caso tendríamos que $\langle \mu, v \rangle = \langle \nu, v \rangle$ para todo $v \in L^2(\mathcal{I})$ de donde $\mu(t)$ es igual a $\nu(t)$ para casi todo $t \in \mathcal{I}$.

Como $\mathbb{E}\left( \int_{\mathcal{I}} |X(t)v(t)|dt \right) \leq \mathbb{E}\left[ \left( \int_{\mathcal{I}} X^2(t)dt \right)^{\frac{1}{2}} \left( \int_{\mathcal{I}} v^2(t)dt \right)^{\frac{1}{2}} \right] = \|v\| \mathbb{E}\|X\| < \infty$, por el Teorema de Fubini podemos intercambiar la esperanza con la integral, de donde obtenemos que $\mathbb{E}\left( \int_{\mathcal{I}} X(t)v(t)dt \right) = \int_{\mathcal{I}} \mathbb{E}[X(t)\, v(t)]\, dt = \int_{\mathcal{I}} \mathbb{E}[X(t)]\, v(t)\, dt$, lo que concluye la demostración. □

Sean $X_1, \ldots, X_n$ una muestra aleatoria, con $X_i \sim X$, la medida empírica se define como

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} ,$$

donde $\delta_{x_i}$ es la medida puntual en $x_i$. Utilizando la función de probabilidad empírica y los conceptos anteriormente descriptos, podemos definir la media muestral como $\widehat{\mu} = \sum_{i=1}^{n} X_i/n$. En particular, si $X_i \in L^2(\mathcal{I})$ tenemos que $\widehat{\mu}(t) = \sum_{i=1}^{n} X_i(t)/n$. Muchas veces, las observaciones se observan sólo en una grilla de puntos $\{t_j\}_{j=1}^{p}$ por lo que tendremos solamente definidos los valores $\widehat{\mu}(t_j) = \sum_{i=1}^{n} X_i(t_j)/n$. Para obtener un estimador de la función $\mu(t)$ se utilizan técnicas de suavizado basadas en núcleos, ver por ejemplo Wang et al. (2016).

La Figura 2.2 muestra el gráfico de las alturas de las 54 niñas antes descrito en líneas punteadas al que se sobreimpuso en línea continua roja la media muestral.

## 2.4 Covarianza

La covarianza entre dos variables aleatorias es una medida de asociación entre ambas. Para el caso multivariado, se define la matriz de covarianza de un vector aleatorio $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ como la matriz simétrica semidefinida positiva $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ cuya componente $(i, j)$ es igual a $\mathrm{Cov}(x_i, x_j)$, es decir,

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Cov}(x_1, x_1) & \ldots & \mathrm{Cov}(x_1, x_p) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_p, x_1) & \ldots & \mathrm{Cov}(x_p, x_p) \end{pmatrix} .$$

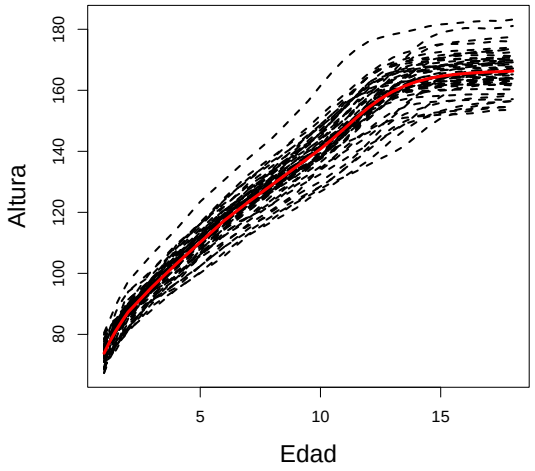


Figura 2.2: Datos de altura de niñas (Berkeley Growth Study, disponibles en la librería `fda` de `R`) en líneas punteadas negras y su media muestral graficada con una línea roja.

Para el caso de datos funcionales, esta noción se extiende al caso de dimensión infinita mediante un operador que, dado un elemento aleatorio $X$ y dos direcciones $u, v \in \mathbb{H}$, mida la covarianza entre las proyecciones de $X$ en $u$ y de $X$ en $v$. Además, buscaremos mantener las principales propiedades de la matriz de covarianza, como la simetría, la bilinealidad y el hecho de que es un operador semidefinido positivo. En Bosq (2000) y Horvath y Kokoska (2012) pueden encontrarse más detalles sobre las propiedades de este operador.

Así como necesitamos una condición para la existencia de la media, tendremos otra para que el operador de covarianza exista, observando la analogía con la condición de segundo momento finito en el caso univariado.

**Definición 2.4.1.** Un elemento aleatorio $X$ se dice de cuadrado integrable si $\mathbb{E}(\|X\|^2) < \infty$.

Definamos ahora al operador bilineal de covarianza tomando la idea comentada previamente de que buscaremos medir la covarianza entre el elemento aleatorio proyectado a dos direcciones del espacio.

**Definición 2.4.2.** Sea $X$ de cuadrado integrable, definimos el operador bilineal de covarianza $c : \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ como $c(u, v) = \mathrm{Cov}(\langle X, u\rangle, \langle X, v\rangle)$.

La siguiente proposición muestra que dicho operador es bilineal y continuo.

**Proposición 2.4.1.** Sea $X$ de cuadrado integrable, entonces $c(u, v)$ es un operador bilineal y acotado.

*Demostración.* La bilinealidad se deduce del hecho de que la covarianza entre variables aleatorias y el producto interno son bilineales. Efectivamente,

$$\begin{aligned}|c(\lambda_1 u_1 + \lambda_2 u_2, v)| &= \text{Cov}(\langle X, \lambda_1 u_1 + \lambda_2 u_2\rangle, \langle X, v\rangle) \\ &= \text{Cov}(\lambda_1\langle X, u_1\rangle + \lambda_2\langle X, u_2\rangle, \langle X, v\rangle) \\ &= \lambda_1 \text{Cov}(\langle X, u_1\rangle, \langle X, v\rangle) + \lambda_2 \text{Cov}(\langle X, u_2\rangle, \langle X, v\rangle)\,.\end{aligned}$$

La linealidad del segundo término del operador se demuestra de la misma forma.

Mostraremos ahora que $c$ es un operador acotado. La desigualdad de Cauchy-Schwartz y el hecho que $X$ tiene segundo momento finito implican que

$$\begin{aligned}|c(u, v)| = |\text{Cov}(\langle X, u\rangle, \langle X, v\rangle)| &\leq \left\{\text{Var}(\langle X, u\rangle)\right\}^{\frac{1}{2}} \left\{\text{Var}(\langle X, v\rangle)\right\}^{\frac{1}{2}} \\ &\leq \left\{\mathbb{E}(\langle X, u\rangle^2)\right\}^{\frac{1}{2}} \left\{\mathbb{E}(\langle X, v\rangle^2)\right\}^{\frac{1}{2}} \leq \|u\|\,\|v\|\,\mathbb{E}\|X\|^2 < \infty\,,\end{aligned}$$

lo que concluye la demostración. □

Al ser $c(\cdot, \cdot)$ un operador bilineal y acotado, por el Teorema de representación de Riesz, existen únicos operadores acotados $A : \mathbb{H} \to \mathbb{H}$ y $B : \mathbb{H} \to \mathbb{H}$ tales que $c(u, v) = \langle Au, v\rangle = \langle u, Bv\rangle$ para todo $u, v \in \mathbb{H}$. A partir de esto tendremos otra manera de representar al operador.

**Definición 2.4.3.** Sea $c$ el operador bilineal de covarianza de $X$ y sea $A$ tal que $c(u, v) = \langle Au, v\rangle$ para todo $u, v \in \mathbb{H}$. El operador $A$ se llama operador de covarianza y lo notaremos como $\Gamma_X$.

A lo largo de este trabajo, utilizaremos el operador de covarianza $\Gamma_X : \mathbb{H} \to \mathbb{H}$ en lugar del operador bilineal de covarianza. Veremos algunas de sus propiedades que serán de utilidad en la siguiente sección, donde describiremos el problema de reducción de la dimensión mediante componentes principales.

**Proposición 2.4.2.** El operador $\Gamma_X$ es autoadjunto (simétrico) y semidefinido positivo.

*Demostración.* La simetría de $\Gamma_X$ se deduce de la simetría del operador bilineal de covarianza $c(\cdot, \cdot)$ junto con la simetría del producto interno de $\mathbb{H}$. Por otra parte, como

$$\langle \Gamma_X(u), u\rangle = c(u, u) = \text{Cov}(\langle X, u\rangle, \langle X, u\rangle) = \text{Var}(\langle X, u\rangle) \geq 0\,,$$

obtenemos que $\Gamma_X$ es un semidefinido positivo. □

A continuación definiremos tres familias de operadores que serán de utilidad para describir al operador de covarianza.

**Definición 2.4.4.** Sea $L : \mathbb{H} \to \mathbb{H}$ un operador lineal y acotado. Decimos que $L$ es un operador compacto si existen dos bases ortonormales $\{v_j\}_{j\geq 1}$ y $\{u_j\}_{j\geq 1}$ y una secuencia de reales $\{\lambda_j\}_{j\geq 1}$ convergente a cero tales que

$$L(x) = \sum_{j\geq 1} \lambda_j \langle x, v_j\rangle u_j \tag{2.2}$$

para todo $x \in \mathbb{H}$.

**Definición 2.4.5.** Sea $L : \mathbb{H} \to \mathbb{H}$ un operador compacto. Decimos que $L$ es un operador Hilbert-Schmidt si la representación dada en (2.2) cumple que $\sum_{j\geq 1} \lambda_j^2 < \infty$.

**Definición 2.4.6.** Sea $L : \mathbb{H} \to \mathbb{H}$ un operador lineal y acotado. Definimos la traza de $L$ como $\text{TR}(L) = \sum_{j\geq 1}\langle L(v_j), v_j\rangle$, donde $\{v_j\}_{j\geq 1}$ es una base ortonormal de $\mathbb{H}$. Además, el operador $L$ es un operador de traza si $\text{TR}(|L|) < \infty$, donde $|L| := \sqrt{L^*L}$ es el valor absoluto de $L$.

**Proposición 2.4.3.** El operador $\Gamma_X$ es un operador de traza.

*Demostración.* Sea $\{v_j\}_{j\geq 1}$ una base ortonormal de $\mathbb{H}$. Como $\Gamma_X$ es semidefinido positivo y autoadjunto, basta ver que $\text{TR}(\Gamma_X) < \infty$. Usando que $\langle \Gamma_X(u), u\rangle = \text{VAR}(\langle X, u\rangle)$, obtenemos que

$$\text{TR}(\Gamma_X) = \sum_{j\geq 1}\langle \Gamma_X(v_j), v_j\rangle = \sum_{j\geq 1} \text{VAR}(\langle X, v_j\rangle) \leq \sum_{j\geq 1} \mathbb{E}\left(\langle X, v_j\rangle^2\right) = \mathbb{E}\left(\|X\|^2\right) < \infty\,,$$

donde la última igualdad es consecuencia de la igualdad de Parseval, lo que concluye la demostración. □

Teniendo en cuenta que los operadores de traza son un subconjunto de los operadores Hilbert-Schmidt y por lo tanto son compactos, se obtiene el siguiente resultado.

**Corolario 2.4.1.** El operador $\Gamma_X$ es de clase Hilbert-Schmidt y además es compacto.

Volviendo al caso de nuestro interés donde $\mathbb{H} = L^2(\mathcal{I})$, veremos en la Proposición 2.4.4 que si $X$ es cuadrado integrable, entonces el operador de covarianza se puede evaluar mediante la función $\gamma(s,t) = \text{COV}(X(s), X(t))$, que se llama núcleo de covarianza. Esto nos dará una forma explícita del operador de covarianza, que para el caso empírico va a ser de suma importancia, ya que de esta manera lo podremos aproximar al discretizar el intervalo $\mathcal{I}$.

**Proposición 2.4.4.** Sea $X \in L^2(\mathcal{I})$ de cuadrado integrable y sea $\Gamma_X$ su operador de covarianza, entonces $\Gamma_X(u)(s) = \int_{\mathcal{I}} \gamma(s,t)\, u(t)dt$, donde $\gamma(s,t) = \text{COV}(X(s), X(t))$.

*Demostración.* Para esto, por la unicidad del Teorema de representación de Riesz, será suficiente mostrar que para todo $v \in L^2(\mathcal{I})$, $c(u,v) = \langle \widetilde{\Gamma}(u), v\rangle$, donde $\widetilde{\Gamma}(u)(s) = \int_{\mathcal{I}} \gamma(s,t)\, u(t)dt$. Empezando por $\langle \widetilde{\Gamma}(u), v\rangle$,

$$\begin{aligned}
\langle \widetilde{\Gamma}(u), v\rangle &= \int_{\mathcal{I}} \left( \int_{\mathcal{I}} \gamma(s,t)\, u(t)\, dt \right) v(s)\, ds = \int_{\mathcal{I}} \int_{\mathcal{I}} u(t)\, v(s) \mathrm{Cov}(X(s), X(t))\, dt\, ds \\
&= \int_{\mathcal{I}} \int_{\mathcal{I}} u(t)v(s) \left( \int_{\Omega} (X(s) - \mu(s))(X(t) - \mu(t)) dP \right) dt\, ds \\
&= \int_{\Omega} \left( \int_{\mathcal{I}} (X(t) - \mu(t))\, u(t) dt \right) \left( \int_{\mathcal{I}} (X(s) - \mu(s)\,) v(s)\, ds \right) dP \\
&= \mathbb{E}\left( \langle X - \mu, u\rangle \langle X - \mu, v\rangle \right) = \mathbb{E}\left\{ \left( \langle X, u\rangle - \mathbb{E}[\langle X, u\rangle] \right) \left( \langle X, v\rangle - \mathbb{E}[\langle X, v\rangle] \right) \right\} \\
&= \mathrm{Cov}(\langle X, u\rangle, \langle X, v\rangle) = c(u,v) = \langle \Gamma_X(u), v\rangle\,,
\end{aligned}$$

donde en la cuarta igualdad hemos podido usar el Teorema de Fubini ya que

$$\begin{aligned}
&\int_{\Omega} \int_{\mathcal{I}\times\mathcal{I}} |((X(t) - \mu(t))u(t))\ ((X(s) - \mu(s))v(s))|\ dt\, ds\, dP \\
&\qquad = \mathbb{E}\left[ \left( \int_{\mathcal{I}} |X(t) - \mu(t)|\, |u(t)|\, dt \right) \left( \int_{\mathcal{I}} |X(s) - \mu(s)|\, |v(s)|\, ds \right) \right] \\
&\qquad \leq \mathbb{E}(\|u\| \|X - \mu\|\, \|v\|\, \|X - \mu\|) \leq \|u\|\, \|v\|\, \mathbb{E}(\|X - \mu\|^2) < \infty\,.
\end{aligned}$$

□

**Definición 2.4.7.** Sea $X \in L^2(\mathcal{I})$ de cuadrado integrable y sea $\Gamma_X$ su operador de covarianza donde $\Gamma_X(u)(s) = \int_{\mathcal{I}} \mathrm{Cov}(X(s), X(t))u(t)dt$. Definimos la función o núcleo de covarianza de $X$ como $\gamma(s,t) = \mathrm{Cov}(X(s), X(t))$.

Como en el caso de la esperanza, si $X_1, \ldots, X_n$ es una muestra aleatoria, con $X_i \sim X$, utilizando la función de probabilidad empírica podemos definir un estimador $\widehat{\gamma}$ del núcleo de covarianza como

$$\widehat{\gamma}(t,s) = \frac{1}{n} \sum_{i=1}^{n} \left[ X_i(t) - \widehat{\mu}(t) \right]\, \left[ X_i(s) - \widehat{\mu}(s) \right]\,.$$

Si las observaciones se registran en una grilla común de puntos $\{t_j\}_{j=1}^{p}$, sólo podemos definir $\widehat{\gamma}$ en los pares $(t_j, t_\ell)$. es decir,

$$\widehat{\gamma}(t_j, t_\ell) = \frac{1}{n} \sum_{i=1}^{n} \left[ X_i(t_j) - \widehat{\mu}(t_j) \right]\, \left[ X_i(t_\ell) - \widehat{\mu}(t_\ell) \right]\,,$$

por lo que nuevamente será necesario utilizar procedimientos de suavizado para obtener toda la superficie. Una descripción de procedimientos para estimar $\widehat{\gamma}$ cuando las observaciones no se registran en la misma grilla de puntos o cuando se registran con error, puede verse en Wang et al. (2016).

Para ejemplificar, la Figura 2.3 muestra en el panel (a) la función de covarianza estimada $\widehat{\gamma}$ para el conjunto relacionado con las alturas de las niñas del estudio Berkeley Growth Study. Hemos representado en el panel (b) la función $\widehat{\rho}(s,t) = \widehat{\gamma}(s,t)/\left(\widehat{\gamma}(s,s)\,\widehat{\gamma}(t,t)\right)^{1/2}$ que estima la correlación entre las alturas en dos edades $s$ y $t$. Como mencionamos en la Sección 2.2, se observa la alta correlación existente al tomar dos tiempos $t$ y $s$ posteriores a la pubertad, donde el crecimiento se empieza a detener.

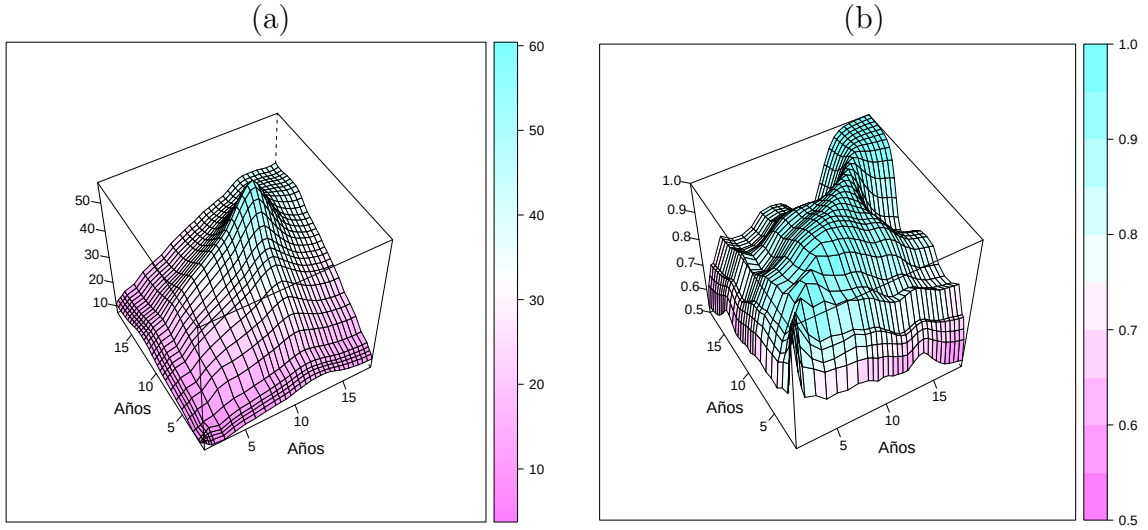


Figura 2.3: Datos de las altura de niñas del Berkeley Growth Study (disponibles en la librería `fda` de `R`): (a) Función de covarianza estimada $\widehat{\gamma}$. (b) Función $\widehat{\rho}(s,t)$.

## 2.5 Teorema de Karhunen-Loève

En esta sección veremos un teorema de descomposición para elementos aleatorios que nos servirá como herramienta para la reducción de la dimensión. Nuestro objetivo será extender el análisis de componentes principales al caso funcional. Esto lo haremos gracias al Teorema de Karhunen-Loève, que da una representación de un elemento aleatorio a través de las autofunciones de su operador de covarianza. Para esto, de ahora en más supondremos que $\mathbb{H} = L^2(\mathcal{I})$, donde $\mathcal{I} = [a,b]$ es un intervalo compacto de $\mathbb{R}$, y consideraremos el subconjunto de elementos aleatorios que definimos a continuación.

**Definición 2.5.1.** Diremos que $X \in L^2(\mathcal{I})$ es de media cuadrática continua si para toda sucesión $\{t_n\}_{n\geq 1}$ tal que $t_n \to t$ se cumple que $\lim_{n\to\infty} \mathbb{E}\left[(X(t_n) - X(t))^2\right] = 0$.

El Teorema 2.5.1 permite caracterizar los elementos aleatorios de media cuadrática continua de una forma más tangible.

**Teorema 2.5.1.** Sea $X$ de cuadrado integrable, entonces $X$ es de media cuadrática continua si y sólo si su media y su función de covarianza son ambas continuas.

*Demostración.* Comencemos viendo que si la media $\mu(t) = \mathbb{E}X(t)$ y la función de covarianza $\gamma(s,t) = \textsc{Cov}(X(s), X(t))$ son continuas, entonces $X$ es de media cuadrática continua. Como

$$\begin{aligned}\mathbb{E}\big[(X(s) - X(t))^2\big] &= \textsc{Var}\big(X(s) - X(t)\big) + \big\{\mathbb{E}\big[X(s) - X(t)\big]\big\}^2 \\ &= \gamma(s,s) + \gamma(t,t) - 2\gamma(s,t) + (\mu(s) - \mu(t))^2,\end{aligned}$$

la continuidad de la media y la función de covarianza implican que $X$ es de media cuadrática continua.

Veamos la otra implicación. Para ello supongamos que $X$ es de media cuadrática continua. Probaremos en primer lugar que $\mu(t)$ es continua. Usando la desigualdad de Jensen obtenemos que

$$(\mu(s) - \mu(t))^2 = \{\mathbb{E}\left[X(s) - X(t)\right]\}^2 \leq \mathbb{E}(X(s) - X(t))^2,$$

de donde, como $X$ es de media cuadrática continua, se deduce que $\mu$ es continua.

Por simplicidad y sin pérdida de generalidad, supongamos que $\mu \equiv 0$. Probaremos primero la continuidad de la función de covarianza en la primera variable. La desigualdad de Cauchy-Schwartz implica que

$$|\gamma(s,t) - \gamma(s',t)| = |\textsc{Cov}\left(X(s) - X(s'), X(t)\right)| \leq \{\textsc{Var}(X(t))\}^{\frac{1}{2}} \left\{\mathbb{E}\left[X(s) - X(s')\right]^2\right\}^{\frac{1}{2}}, \tag{2.3}$$

de donde se deduce que $\gamma(s,t)$ es continua en $s$ cuando la segunda variable está fija.

Ahora veamos que $\sigma^2(t) = \textsc{Var}(X(t))$ es continua en $t$. Sea $\{t_n\}_{n\geq 1}$ una sucesión que converge a $t$. Observemos que

$$\begin{aligned}|\textsc{Var}(X(t_n)) - \textsc{Var}(X(t))| &= \left|\mathbb{E}\left(X(t_n) - X(t) + X(t)\right)^2 - \textsc{Var}(X(t))\right| \\ &= \Big|\mathbb{E}\left(X(t_n) - X(t)\right)^2 + \mathbb{E}X^2(t) \\ &\quad + 2\mathbb{E}\left[X(t)(X(t_n) - X(t))\right] - \textsc{Var}(X(t))\Big| \\ &= \left|\mathbb{E}\left(X(t_n) - X(t)\right)^2 + 2\,\mathbb{E}\left[X(t)(X(t_n) - X(t))\right]\right| \\ &\leq \mathbb{E}\left(X(t_n) - X(t)\right)^2 + 2\,\left|\gamma(t_n,t) - \gamma(t,t)\right|.\end{aligned}$$

Como $X$ es de media cuadrática continua, el primer término converge a 0. Por otra parte, como $\gamma(s,t)$ es continua en $s$ para cada $t$ fijo, el segundo sumando también converge a 0, de donde deducimos que $\sigma(t)$ es continua.

Falta probar que $\gamma(s_n,t_n) \to \gamma(s,t)$ para toda sucesión $\{(s_n,t_n)\}_{n\geq 1}$ tal que $(s_n,t_n) \to (s,t)$. Una acotación directa mediante la desigualdad triangular permite obtener que $|\gamma(s_n,t_n) - \gamma(s,t)| \leq a_1 + a_2$, donde $a_1 = |\gamma(s_n,t_n) - \gamma(s_n,t)|$ y $a_2 = |\gamma(s_n,t) - \gamma(s,t)|$. Como $\gamma$ es continua en el primer argumento cuando el segundo está fijo, obtenemos que $a_2 \to 0$.

La desigualdad dada en (2.3) permite obtener que $a_1^2 \leq \text{VAR}\left(X(s_n)\right) \mathbb{E}\left[\left(X(t_n) - X(t)\right)^2\right]$. Por lo tanto, usando que $\sigma^2(s)$ es continua y que $X$ es de media cuadrática, se deduce que $a_1 \to 0$, lo que junto con el hecho que $a_2 \to 0$ implica la continuidad de $\gamma$. □

Un teorema que va a ser fundamental para esta sección es el Teorema de Hilbert-Schmidt. A partir de este teorema podremos probar la existencia de una base ortonormal de $L^2(\mathcal{I})$ formada por las autofunciones del operador de covarianza, que luego usaremos en el Teorema de Karhunen-Loève.

**Teorema 2.5.2.** (Teorema de Hilbert-Schmidt) Sea $\Gamma$ un operador lineal simétrico y compacto, entonces existe un sistema ortonormal de autofunciones $\{\varphi_j\}_{j\in J}$, $J \subset \mathbb{N}$, con autovalores asociados $\{\lambda_j\}_{j\in J}$ no nulos tales que, para cada $v \in \mathbb{H}$ se puede representar en forma única

$$v = \sum_{j\in J} \alpha_j \varphi_j + u\,, \tag{2.4}$$

donde $\alpha_j \in \mathbb{R}$ y $u \in \ker(\Gamma)$, siendo $\ker(\Gamma)$ el núcleo de $\Gamma$.

La demostración de este resultado se puede encontrar en Debnath y Mikusinski (2005).

Una consecuencia del Teorema de Hilbert-Schmidt es el Teorema de descomposición espectral que se presenta en el siguiente corolario.

**Corolario 2.5.1.** (Teorema de descomposición espectral) Sea $\Gamma$ un operador lineal simétrico y compacto. Entonces existe una base ortonormal completa de autofunciones $\{\varphi_j\}_{j\geq 1}$ con autovalores asociados $\{\lambda_j\}_{j\geq 1}$, $\lambda_1 \geq \lambda_2 \geq \ldots$ tales que, para cada $v \in \mathbb{H}$,

$$\Gamma(v) = \sum_{j=1}^{\infty} \lambda_j \langle v, \varphi_j\rangle \varphi_j\,. \tag{2.5}$$

*Demostración.* Sea $\{\varphi_j\}_{j\in J}$, $J \subset \mathbb{N}$, el sistema dado por el Teorema de Hilbert-Schmidt. Observemos que los coeficientes $\alpha_j$ en (2.4) están dados por $\alpha_j = \langle v, \varphi_j\rangle$, por lo tanto $\Gamma(v) = \Gamma\left(\sum_{j\in J} \alpha_j\varphi_j + u\right)$, de donde usando la linealidad de $\Gamma$ y el hecho que $u \in \ker(\Gamma)$, obtenemos que

$$\Gamma(v) = \sum_{j\in J} \alpha_j \Gamma\left(\varphi_j\right) = \sum_{j\in J} \lambda_j \alpha_j \varphi_j\,.$$

Si $\ker(\Gamma) = \{0\}$, $\{\varphi_j\}_{j\in J}$ es una base ortonormal completa. Si $\ker(\Gamma) \neq \{0\}$, para obtener una base completa de autofunciones, debemos extender el sistema ortonormal de autofunciones $\{\varphi_j\}_{j\in J}$ con un sistema ortonormal arbitrario del núcleo de $\Gamma$. Esta base del núcleo de $\Gamma$ tendrá autovalores asociados nulos. Por la separabilidad de $\mathbb{H}$ la base es finita o numerable. De esta manera, por la continuidad de $\Gamma$, vale la igualdad (2.5). □

Además de la base ortonormal de $L^2(\mathcal{I})$ formada por las autofunciones del operador $\Gamma$, necesitamos un teorema para caracterizar la convergencia de la serie dada por la descomposición de la función de covarianza $\gamma(\cdot,\cdot)$. Dicho resultado, conocido como el Teorema de Mercer, se enuncia a continuación.

**Teorema 2.5.3.** (Teorema de Mercer) Sea $\Gamma$ un operador lineal simétrico y compacto de $L^2(\mathcal{I})$ con $k(\cdot,\cdot)$ su núcleo, es decir, supongamos que $\Gamma(v)(t) = \int_{\mathcal{I}} k(t,s)v(s)ds$ para todo $v \in \mathbb{H}$, $t \in \mathcal{I}$, con $k$ continuo. Entonces,

$$k(s,t) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(t)\varphi_j(s)\,, \tag{2.6}$$

donde $\{\varphi_j\}_{j\geq 1}$ es la base ortonormal completa dada por descomposición espectral de $\Gamma$ en (2.5), con autovalores asociados $\{\lambda_j\}_{j\geq 1}$, tales que $\lambda_1 \geq \lambda_2 \geq \dots$. Además, la serie en (2.6) converge uniforme y absolutamente.

*Demostración.* Como $\Gamma(v)(t) = \int_{\mathcal{I}} k(t,s)v(s)ds$, tenemos que $\Gamma(\varphi_i)(t) = \int_{\mathcal{I}} k(t,s)\varphi_i(s)ds$. Por otra parte, como $\varphi_i$ es una autofunción de $\Gamma$ con autovalor asociado $\lambda_i$, tenemos que $\Gamma(\varphi_i)(t) = \lambda_i\varphi_i(t)$. Por lo tanto, para todo $i \in \mathbb{N}$, se cumple $\langle k(t,\cdot), \varphi_i\rangle = \lambda_i\varphi_i(t)$, lo que implica que podemos representar al núcleo como

$$k(t,s) = \sum_{j=1}^{\infty} \langle k(t,\cdot), \varphi_j\rangle \varphi_j(s) = \sum_{j=1}^{\infty} \lambda_j\varphi_j(t)\varphi_j(s)\,,$$

como deseábamos probar.

Mostraremos ahora que la convergencia es uniforme y absoluta. Observemos que, como $\lambda_j \geq 0$, por la desigualdad de Cauchy-Schwartz,

$$\left(\sum_{j=1}^{\infty} |\lambda_j\varphi_j(t)\varphi_j(s)|\right)^2 \leq \left(\sum_{j=1}^{\infty} \lambda_j\varphi_j^2(t)\right)\left(\sum_{j=1}^{\infty} \lambda_j\varphi_j^2(s)\right) = k(t,t)k(s,s) \leq \sup_{t\in\mathcal{I}} k^2(t,t)\,.$$

Por lo tanto, como $k$ es continuo en un compacto y por ende acotado, por el criterio de Weierstrass la convergencia de la serie es uniforme y absoluta. □

Sin pérdida de generalidad, a partir de ahora, supondremos que $\mu \equiv 0$.

Para demostrar el Teorema de Karhunen-Loève, necesitamos primero fijar notación y definir unas funciones que se usarán en lo que sigue. Para esto nos basaremos en el trabajo de Giambartolomei (2015).

Sea $\Delta$ una partición de $\mathcal{I} = [a,b]$ dada por $a = t_0 < t_1 < \cdots < t_m = b$, siendo $\mathcal{I}_i = [t_{i-1}, t_i]$ y $|\Delta| = \max_{1\leq i\leq m} |\mathcal{I}_i|$, definamos

$$I(u;\Delta) = \sum_{i=1}^{m} X(t_i) \int_{\mathcal{I}_i} u(t)\,dt\,.$$

Observemos que, de esta manera, estamos trabajando con variables aleatorias bien definidas, ya que $I(u;\Delta)$ es una suma finita de variables aleatorias multiplicadas por escalares.

En primer lugar, veamos que $I(u;\Delta)$ converge cuando $|\Delta|$ tiende a 0. Efectivamente, sean $\Delta$ y $\Delta'$ dos particiones de $\mathcal{I} = [a,b]$ dadas por $a = t_0 < t_1 < \cdots < t_m = b$ y $a = t'_0 < t'_1 < \cdots < t'_{m'} = b$, entonces

$$\begin{aligned}\mathbb{E}\left[(I(u;\Delta) - I(u;\Delta'))^2\right] &= \sum_{i=1}^{m}\sum_{j=1}^{m} \gamma(t_i,t_j) \int_{\mathcal{I}_i} u(t)\,dt \int_{\mathcal{I}_j} u(t)\,dt \\ &+ \sum_{i=1}^{m'}\sum_{j=1}^{m'} \gamma(t'_i,t'_j) \int_{\mathcal{I}'_i} u(t)\,dt \int_{\mathcal{I}'_j} u(t)\,dt \\ &- 2\sum_{i=1}^{m}\sum_{j=1}^{m'} \gamma(t_i,t'_j) \int_{\mathcal{I}_i} u(t)dt \int_{\mathcal{I}'_j} u(t)dt\,.\end{aligned}$$

Como el intervalo $\mathcal{I}$ es compacto, la función de covarianza además de continua es uniformemente continua. Luego, a medida que $|\Delta|$ y $|\Delta'|$ tienden a 0, cada uno de los tres términos converge a $\int_{\mathcal{I}\times\mathcal{I}} \gamma(s,t)\,u(s)u(t)\,ds\,dt$. Por lo tanto, como el espacio $L^2(\Omega,\mathcal{F},P)$ es completo, existe una variable aleatoria $I(u)$ para cada $u \in L^2(\mathcal{I})$ tal que $I(u;\Delta)$ converge a $I(u)$ en media cuadrática y por lo tanto en probabilidad, cuando $|\Delta|$ tiende a 0. De ahora en más indicaremos como $\int_{\mathcal{I}} X(t)u(t)dt$ a $I(u)$.

En el siguiente lema enunciaremos alguna propiedades de $I(u)$ que usaremos en la demostración del Teorema de Karhunen-Loève.

**Lema 2.5.1.** Dado $u \in L^2(\mathcal{I})$, sea $I(u)$ la variable aleatoria definida anteriormente, entonces de tiene que

(a) $\mathbb{E}[I(u)] = 0$ ,

(b) $\mathbb{E}[I(u)X(t)] = \int_{\mathcal{I}} \gamma(s,t)u(s)\,ds$ ,

(c) $\mathbb{E}[I(u)I(v)] = \int_{\mathcal{I}^2} \gamma(s,t)u(s)v(t)\,ds\,dt$ .

*Demostración.* (a) Como supusimos que $\mathbb{E}(X) = 0$, deducimos que $\mathbb{E}\left[I(u;\Delta)\right] = 0$, de donde obtenemos

$$|\mathbb{E}[I(u)]|^2 = |\mathbb{E}[I(u) - I(u;\Delta)]|^2 \leq \mathbb{E}[(I(u) - I(u;\Delta))^2]\,.$$

Como el límite del miembro derecho de la desigualdad es 0 cuando $|\Delta|$ tiende a 0, obtenemos que $\mathbb{E}[I(u)] = 0$, lo que concluye la demostración de (a).

(b) Para probar (b), observemos que

$$|\mathbb{E}[I(u)X(t) - I(u;\Delta)X(t)]| = |\mathbb{E}[(I(u) - I(u;\Delta))X(t)]| \leq (\mathbb{E}[X(t)^2])^{\frac{1}{2}} (\mathbb{E}[(I(u) - I(u;\Delta))^2])^{\frac{1}{2}}\,,$$

donde el último término del miembro derecho de la desigualdad converge a 0 cuando $|\Delta|$ tiende a 0. Por lo tanto,

$$\lim_{|\Delta|\to 0} \mathbb{E}[I(u;\Delta)X(t)] = \mathbb{E}[I(u)X(t)]\,.$$

Por otra parte, observemos que

$$\begin{aligned}\mathbb{E}[I(u;\Delta)X(t)] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} X(t_i)\int_{\mathcal{I}_i} u(s)ds\right)X(t)\right] \\ &= \sum_{i=1}^{m}\int_{\mathcal{I}_i}\mathbb{E}[X(t_i)X(t)]u(s)\,ds = \sum_{i=1}^{m}\int_{\mathcal{I}_i}\gamma(t_i,t)u(s)\,ds\,.\end{aligned}$$

Luego, $\mathbb{E}[I(u;\Delta)X(t)]$ converge a $\int_{\mathcal{I}}\gamma(s,t)u(s)ds$ cuando $|\Delta|$ tiende a 0, lo que concluye la demostración de (b).

De la misma manera se puede demostrar (c). □

Ahora podemos demostrar el teorema central de esta sección, donde presentamos una descomposición de un elemento aleatorio de media cuadrática continua utlizando las autofunciones de su operador de covarianza.

**Teorema 2.5.4.** (Teorema de Karhunen-Loève) Sea $X$ de media cuadrática continua y sean $\{\varphi_i\}_{i\geq 1}$ las autofunciones del operador de covarianza de $X$ ordenadas según sus autovalores de manera decreciente, es decir, $\lambda_1 \geq \lambda_2 \geq \ldots$ donde $\lambda_i$ es el autovalor asociado a $\varphi_i$. Definamos $S_N(t) = \sum_{i=1}^{N} I(\varphi_i)\varphi_i(t)$, entonces se tiene que

$$\lim_{N\to\infty}\sup_{t\in\mathcal{I}}\mathbb{E}\left[(X(t)-S_N(t))^2\right] = 0\,.$$

*Demostración.* El Lema 2.5.1(c) junto con el hecho que $\varphi_i$ es una autofunción del operador de covarianza de $X$ implican que $\mathbb{E}[I(\varphi_i)I(\varphi_j)] = \delta_{ij}\lambda_i$, donde $\delta_{ij} = 1$ si $i = j$ y 0 en caso contrario.

Usando el Lema 2.5.1, obtenemos

$$\begin{aligned}\mathbb{E}[(X(t)-S_N(t))^2] &= \mathbb{E}[X(t)^2] + \mathbb{E}[S_N(t)^2] - 2\mathbb{E}[S_N(t)X(t)] \\ &= \gamma(t,t) + \sum_{i=1}^{N}\lambda_i\varphi_i^2(t) - 2\sum_{i=1}^{N}\lambda_i\varphi_i^2(t) = \gamma(t,t) - \sum_{i=1}^{N}\lambda_i\varphi_i^2(t)\,.\end{aligned}$$

Por el Teorema 2.5.1, la función de covarianza es continua, por lo tanto, el Teorema de Mercer enunciado en el Teorema 2.5.3 permite concluir que la convergencia de las suma parcial $\sum_{i=1}^{N}\lambda_i\varphi_i^2(t)$ a $\gamma(t,t)$ es absoluta y uniforme, lo que concluye la demostración. □

Observemos que si $\mu = \mathbb{E}(X) \neq 0$, definiendo $Z = X - \mu$ y aplicando el Teorema 2.5.4 al elemento aleatorio $Z$, obtenemos que

$$X = \mu + \sum_{i \geq 1} I(\varphi_i)\varphi_i \,,$$

donde $I(\varphi_i) = \int_{\mathcal{I}} (X - \mu)(t)\varphi_i(t)dt$ se llama score de $X - \mu$ respecto de la autofunción $\varphi_i$.

La principal cualidad de la descomposición dada por el Teorema de Karhunen-Loève es que, si truncamos la serie y consideramos solamente los primeros $N$ términos, estos maximizan la variabilidad explicada por sobre cualquier otra elección de $N$ direcciones.

**Corolario 2.5.2.** Sea $X$ un elemento aleatorio centrado, es decir, $\mathbb{E}X = 0$, de media cuadrática continua y sea $X = \sum_{i \geq 1} I(\varphi_i)\varphi_i$ la descomposición dada por el Teorema de Karhunen-Loève. Entonces,

$$\mathbb{E}\left( \left\| \sum_{i=1}^{N} I(\varphi_i)\varphi_i \right\|^2 \right) = \max_{\mathcal{L}} \mathbb{E}(\|\pi_{\mathcal{L}}(X)\|^2) \,,$$

donde $\mathcal{L}$ es un subespacio (cerrado) de dimensión $N$ de $L^2(\mathcal{I})$ y $\pi_{\mathcal{L}}(X)$ es la proyección de $X$ en dicho subespacio.

*Demostración.* Primero observemos que

$$\mathbb{E}\left( \left\| \sum_{i=1}^{N} I(\varphi_i)\varphi_i \right\|^2 \right) = \mathbb{E}\left( \sum_{i=1}^{N} \sum_{j=1}^{N} I(\varphi_i)I(\varphi_j) \right) = \sum_{i=1}^{N} \lambda_i \,,$$

donde usamos la ortonormalidad de las autofunciones $\varphi_i$ junto con el Lema 2.5.1(c).

Sea ahora $\mathcal{L}$ un subespacio de dimensión $N$ generado por las direcciones ortonormales $\{v_i\}_{i=1}^{N}$. Entonces,

$$\mathbb{E}(\|\pi_{\mathcal{L}}(X)\|^2) = \mathbb{E}\left( \sum_{i=1}^{N} \langle X, v_i \rangle^2 \right) = \sum_{i=1}^{N} \mathbb{E}(\langle X, v_i \rangle^2) = \sum_{i=1}^{N} \textsc{Var}(\langle X, v_i \rangle) = \sum_{i=1}^{N} \langle \Gamma_X(v_i), v_i \rangle \,.$$

El miembro derecho de la ecuación anterior se maximiza al tomar $v_i = \varphi_i$ las autofunciones de $\Gamma_X$ asociadas a sus $N$ mayores autovalores, lo que concluye la demostración. □

De esta manera conseguimos un análogo para el caso funcional del análisis de componentes principales. Además podremos calcular la proporción $\sum_{i=1}^{p} \lambda_i / \sum_{i \geq 1} \lambda_i$ de la variabilidad total explicada al considerar $p$ componentes principales.

En la Figura 2.4 (a), representamos los estimadores de las primeras tres direcciones principales del conjuntos de datos de altura de las 54 niñas entre 1 año y los 18 años, obtenidos a partir del estimador $\widehat{\gamma}$ definido en la Sección 2.4.

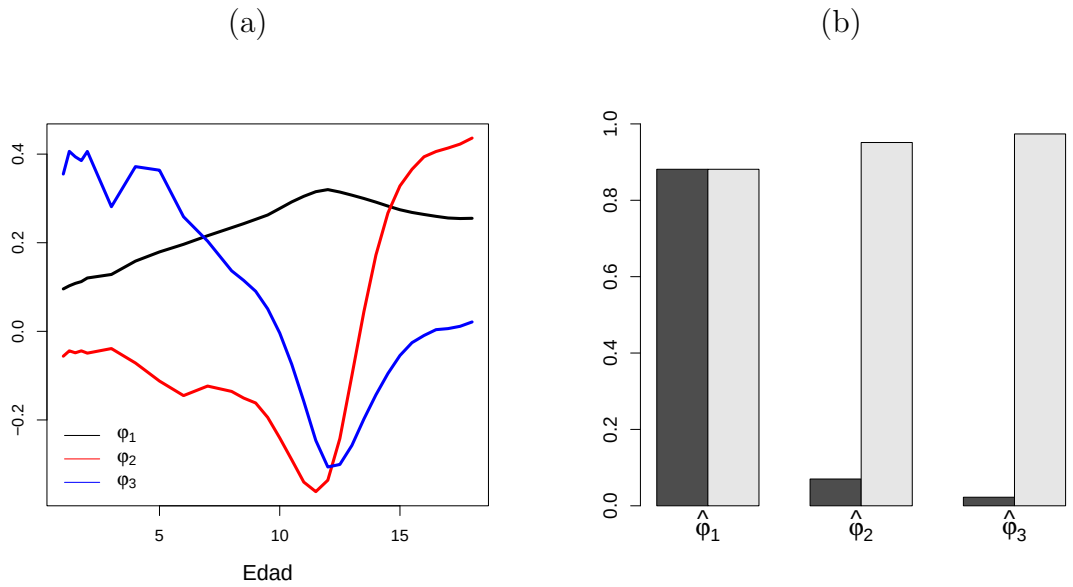


Figura 2.4: Datos de altura de niñas (Berkeley Growth Study, disponibles en la librería fda de R): (a) Gráfico de los estimadores de las tres primeras direcciones principales, en negro, rojo y azul se representan a $\widehat{\varphi}_1$, $\widehat{\varphi}_2$ y $\widehat{\varphi}_3$, respectivamente. (b) Porcentaje de varianza explicada por cada componentes principal (gris oscuro) y propoción acumulada de varianza explicada (gris claro).

A partir de los estimadores de las primeras tres direcciones principales y de sus autovalores asociados, podemos calcular la variabilidad explicada por cada autofunción y la variabilidad acumulada para las primeras $p$ componentes principales, con $p = 1, 2, 3$ que se representan en la Figura 2.4 (b).

# Capítulo 3

# Clasificación

## 3.1 Introducción

El problema de clasificación es un área de la estadística con un amplio desarrollo que considera la siguiente situación. Supongamos que tenemos una observación que proviene de uno de $M$ grupos distintos $g_i$, $1 \leq i \leq M$ dentro de una población. Queremos obtener una regla de clasificación que permita asignar dicha observación a uno de los $M$ grupos. En este capítulo se describen distintos métodos de clasificación tanto para observaciones multivariadas o datos funcionales. Para ello, en la Sección 3.2 definiremos qué es una regla de clasificación junto con un criterio clásico de comparación entre distintas reglas, a partir del cual definiremos la regla de clasificación Bayes. En las Secciones 3.3 y 3.4 describiremos diversas reglas de clasificación utilizadas en el caso multivariado y funcional, respectivamente. Para este capítulo nos basaremos en los libros Hastie et al. (2001), Efron y Hastie (2016), Scott (1992) y Seber (1984).

## 3.2 Regla de clasificación

De ahora en más indicaremos por $\mathbf{x} \in \mathbb{R}^p$ la observación que pertenece a cualquiera de los $M$ grupos $g_i$, con $1 \leq i \leq M$. Además, si $G$ es la variable aleatoria que indica la pertenencia del grupo, notaremos con $\pi_j$ a la probabilidad a priori $\mathbb{P}(G = j)$ de pertenecer al grupo $g_j$.

Una regla de clasificación determina una partición del espacio $\mathbb{R}^p$ en regiones $\mathcal{G}_j$ disjuntas y asigna al punto $\mathbf{x}$ al grupo $g_j$ si $\mathbf{x} \in \mathcal{G}_j$. Por lo tanto, la podemos identificar con una variable $C(\mathbf{x})$ tal que $C(\mathbf{x}) = j$ si $\mathbf{x} \in \mathcal{G}_j$.

Para poder comparar distintas reglas de clasificación debemos contar con algún criterio de comparación. De esta manera, se define la función de pérdida $L(G, C(\mathbf{x}))$ que penaliza cuando se clasifica una observación del grupo $g_j$ en el grupo $g_i$ con un costo $c_{i|j} > 0$, es decir,

la función de pérdida es

$$L(j,i) = \begin{cases} c_{i|j} & \text{si } i \neq j \\ 0 & \text{si } i = j\,. \end{cases}$$

Para este trabajo, tomaremos el caso particular donde $c_{i|j} = 1$ para todo $i \neq j$. Luego, el error de mala clasificación de la población $j$ se define como

$$L(j, C(\mathbf{x})) = \sum_{i \neq j} \mathbb{1}(\mathbf{x} \in \mathcal{G}_i | \mathbf{x} \in g_j)\,.$$

De esta manera, cuando $c_{i|j} = 1$, se define el riesgo de la regla de clasificación $C(\mathbf{x})$ como

$$R(G, C) = \mathbb{E}_G \mathbb{E}[L(G, C(\mathbf{x}))|G] = \sum_{j=1}^{M} \pi_j \sum_{i \neq j} \mathbb{P}(\mathbf{x} \in \mathcal{G}_i | \mathbf{x} \in g_j)\,,$$

que equivale a la probabilidad total de mala clasificación.

Observemos que $\sum_{i \neq j} \mathbb{P}(\mathbf{x} \in \mathcal{G}_i | \mathbf{x} \in g_j) = 1 - \mathbb{P}(\mathbf{x} \in \mathcal{G}_j | \mathbf{x} \in g_j)$, con lo que

$$R(G, C) = 1 - \sum_{j=1}^{M} \pi_j \mathbb{P}(\mathbf{x} \in \mathcal{G}_j | \mathbf{x} \in g_j)\,.$$

### 3.2.1 Regla de clasificación Bayes

A partir de la definición de riesgo presentada en la sección anterior, buscaremos la regla de clasificación que minimiza dicho riesgo. Esta regla se denomina regla de clasificación Bayes. Para ello indicaremos por $f_j$ a la densidad de $\mathbf{x}$ cuando $\mathbf{x}$ pertenece al grupo $g_j$, es decir, que $\mathbf{x}|_{G=j} \sim f_j$. De esta manera, la probabilidad a posteriori será

$$\mathbb{P}(G = j | \mathbf{x} = \mathbf{x}_0) = \frac{\pi_j f_j(\mathbf{x}_0)}{\sum_{\ell=1}^{M} \pi_\ell f_\ell(\mathbf{x}_0)}\,.$$

**Teorema 3.2.1.** El riesgo de clasificación $R(G, C)$ cuando $c_{i|j} = 1$ se minimiza al tomar la regla de clasificación $C(\mathbf{x})$ que asigna a $\mathbf{x}$ en el grupo $g_j$ si $\mathbf{x} \in \mathcal{G}_j$, donde

$$\mathcal{G}_j = \{\mathbf{x} \in \mathbb{R}^p : \pi_j f_j(\mathbf{x}) > \pi_i f_i(\mathbf{x}) \quad \forall i \neq j\}\,.$$

*Demostración.* Sea $\{\tilde{\mathcal{G}}_1, \ldots, \tilde{\mathcal{G}}_M\}$ una partición del espacio que define una regla de clasifi-

cación $\tilde{C}(\mathbf{x})$ que asigna a $\mathbf{x}$ en el grupo $g_j$ si $\mathbf{x} \in \tilde{\mathcal{G}}_j$.

$$\begin{aligned} R(G,\tilde{C}) &= \sum_{j=1}^{M}\sum_{i\neq j}\mathbb{P}(\mathbf{x}\in\tilde{\mathcal{G}}_i|\mathbf{x}\in g_j) = \sum_{j=1}^{M}\sum_{i\neq j}\pi_j\int_{\tilde{\mathcal{G}}_i} f_j(\mathbf{x}_0)d\mathbf{x}_0 \\ &= \sum_{j=1}^{M}\pi_j\left(1-\int_{\tilde{\mathcal{G}}_j} f_j(\mathbf{x}_0)d\mathbf{x}_0\right) = 1-\sum_{j=1}^{M}\int_{\tilde{\mathcal{G}}_j}\pi_j f_j(\mathbf{x}_0)d\mathbf{x}_0 \\ &= 1-\sum_{j=1}^{M}\int_{\tilde{\mathcal{G}}_j}\frac{\pi_j f_j(\mathbf{x}_0)}{\sum_{\ell=1}^{M}\pi_\ell f_\ell(\mathbf{x}_0)}\left(\sum_{\ell=1}^{M}\pi_\ell f_\ell(\mathbf{x}_0)\right)d\mathbf{x}_0 \\ &= 1-\int\sum_{j=1}^{M}\mathbb{1}_{\tilde{\mathcal{G}}_j}(\mathbf{x}_0)q_j(\mathbf{x}_0)f(\mathbf{x}_0)d\mathbf{x}_0\,, \end{aligned}$$

donde $q_j(\mathbf{x}_0)$ es la probabilidad a posteriori $\mathbb{P}(G=j|\mathbf{x}=\mathbf{x}_0)$, es decir, $q_j(\mathbf{x}_0)=\pi_j f_j(\mathbf{x}_0)/f(\mathbf{x}_0)$, siendo $f(\mathbf{x}_0)=\sum_{\ell=1}^{M}\pi_\ell f_\ell(\mathbf{x}_0)$. Luego, las regiones de clasificación de la regla de clasificación $C$ son

$$\begin{aligned} \mathcal{G}_j &= \{\mathbf{x}_0\in\mathbb{R}^p : \pi_j f_j(\mathbf{x}_0) > \pi_i f_i(\mathbf{x}_0) \quad \forall i\neq j\} \\ &= \{\mathbf{x}_0\in\mathbb{R}^p : q_j(\mathbf{x}_0) > q_i(\mathbf{x}_0) \quad \forall i\neq j\}\,. \end{aligned}$$

Sea

$$\tilde{h}(\mathbf{x}_0) = \sum_{j=1}^{M}\mathbb{1}_{\tilde{\mathcal{G}}_j}(\mathbf{x}_0)q_j(\mathbf{x}_0)f(\mathbf{x}_0)\,.$$

Con esta notación,

$$R(G,\tilde{C}) = 1-\int\tilde{h}(\mathbf{x}_0)d\mathbf{x}_0\,.$$

Luego, para ver que $R(G,\tilde{C})\geq R(G,C)$, basta ver que $\int\tilde{h}(\mathbf{x}_0)d\mathbf{x}_0\leq\int h(\mathbf{x}_0)d\mathbf{x}_0$, donde

$$h(\mathbf{x}_0) = \sum_{\ell=1}^{M}\mathbb{1}_{\mathcal{G}_\ell}(\mathbf{x}_0)q_\ell(\mathbf{x}_0)f(\mathbf{x}_0)\,.$$

Como $\{\mathcal{G}_\ell\}_{\ell=1}^{M}$ es una partición de $\mathbb{R}^p$, tenemos que

$$\int\tilde{h}(\mathbf{x}_0)d\mathbf{x}_0 = \sum_{\ell=1}^{M}\int_{\mathcal{G}_\ell}\tilde{h}(\mathbf{x}_0)d\mathbf{x}_0\,.$$

Entonces, bastará mostrar que $\tilde{h}(\mathbf{x}_0)\leq h(\mathbf{x}_0)$ para todo $\mathbf{x}_0\in\mathcal{G}_\ell$. Sea $\mathbf{x}_0\in\mathcal{G}_\ell$, entonces, $h(\mathbf{x}_0)=q_\ell(\mathbf{x}_0)f(\mathbf{x}_0)$. Como $\{\tilde{\mathcal{G}}_j\}_{j=1}^{M}$ también es una partición, $\sum_{j=1}^{M}\mathbb{1}_{\tilde{\mathcal{G}}_j}(\mathbf{x}_0)=1$. Por lo tanto,

$$h(\mathbf{x}_0) = q_\ell(\mathbf{x}_0)f(\mathbf{x}_0) = \sum_{j=1}^{M}\mathbb{1}_{\tilde{\mathcal{G}}_j}(\mathbf{x}_0)q_\ell(\mathbf{x}_0)f(\mathbf{x}_0)\,.$$

Como $\mathbf{x}_0 \in \mathcal{G}_\ell$, $q_\ell(\mathbf{x}_0) \geq q_j(\mathbf{x}_0)$ para todo $1 \leq j \leq M$. Luego,

$$h(\mathbf{x}_0) \geq \sum_{j=1}^{M} \mathbb{1}_{\tilde{\mathcal{G}}_j}(\mathbf{x}_0) q_j(\mathbf{x}_0) f(\mathbf{x}_0) = \tilde{h}(\mathbf{x}_0)\,,$$

lo que concluye la demostración. □

A partir del Teorema 3.2.1, obtenemos que la regla de clasificación Bayes clasifica a $\mathbf{x}$ en el grupo $g_j$ si $\mathbf{x} \in \mathcal{G}_j$, donde

$$\mathcal{G}_j = \{\mathbf{x} \in \mathbb{R}^p : q_j(\mathbf{x}) > q_i(\mathbf{x}) \quad \forall i \neq j\}\,, \tag{3.1}$$

y, en caso de haber empates entre grupos que maximizan la probabilidad a posteriori, clasificamos arbitrariamente entre dichos grupos. De esta manera, obtenemos que la regla de clasificación Bayes clasifica al punto $\mathbf{x}_0$ en el grupo para el cual se maximiza la probabilidad a posteriori.

## 3.3 Clasificación en $\mathbb{R}^p$

### 3.3.1 Caso de distribución normal

Los métodos de análisis discriminante lineal y análisis discriminante cuadrático, LDA y QDA respectivamente por sus siglas en inglés, parten del supuesto de que las densidades de $f_j(\mathbf{x})$, $1 \leq j \leq M$ son normales, que pueden ser tanto univariadas como multivariadas. De esta manera, podremos calcular las probabilidades $\mathbb{P}(G = j|\mathbf{x} = \mathbf{x}_0)$ necesarias en el clasificador de Bayes.

#### Caso en el que las matrices de covarianza son iguales

Como comentamos previamente, el método del análisis discriminante lineal surge de suponer que la densidad de cada grupo sigue una distribución normal, es decir, que

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\right],$$

donde $f_j$ es la función de densidad de la clase $j$ con media $\boldsymbol{\mu}_j$ y matriz de covarianza $\boldsymbol{\Sigma}_j$.

Supongamos además que todos los grupos tienen la misma matriz de covarianza $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_j$, con $j = 1, \ldots, M$. De esta manera, a partir de la regla de clasificación Bayes dada en el Teorema 3.2.1, buscamos la clase que maximice la probabilidad a posteriori $\mathbb{P}(G = j|\mathbf{x} = \mathbf{x}_0)$. Partiendo de (3.1), bastará calcular $q_i(\mathbf{x}_0)/q_j(\mathbf{x}_0)$. Como $q_i(\mathbf{x}_0) = \pi_i f_i(\mathbf{x}_0)/f(\mathbf{x}_0)$ y el logaritmo es una función monótona, obtenemos

$$\begin{aligned}\log\frac{\mathbb{P}(G=i|\mathbf{x}=\mathbf{x}_0)}{\mathbb{P}(G=j|\mathbf{x}=\mathbf{x}_0)} &= \log\frac{f_i(\mathbf{x}_0)}{f_j(\mathbf{x}_0)}+\log\frac{\pi_i}{\pi_j}\\ &= -\frac{1}{2}(\mathbf{x}_0-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0-\boldsymbol{\mu}_i)+\frac{1}{2}(\mathbf{x}_0-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0-\boldsymbol{\mu}_j)+\log\frac{\pi_i}{\pi_j}\\ &= \mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)-\frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)+\log\frac{\pi_i}{\pi_j}\,.\end{aligned}$$

Partiendo de la idea del cálculo anterior, sea

$$L_i(\mathbf{x})=\log\pi_i+\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_i)\,.$$

La regla de clasificación asigna $\mathbf{x}$ al grupo $g_i$ si $L_i(\mathbf{x})>L_j(\mathbf{x})$ para todo $j\neq i$, o sea

$$\mathcal{G}_i=\{\mathbf{x}\in\mathbb{R}^p : L_i(\mathbf{x})\geq L_j(\mathbf{x})\quad\forall j\neq i\}\,.$$

Sea

$$\begin{aligned}d_{ij}(\mathbf{x})=L_i(\mathbf{x})-L_j(\mathbf{x}) &= (\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2}\right)+\log\pi_i-\log\pi_j\\ &= \boldsymbol{\alpha}_{ij}^{\mathrm{T}}\left(\mathbf{x}-\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2}\right)+\log\pi_i-\log\pi_j\,,\end{aligned}$$

donde $\boldsymbol{\alpha}_{ij}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$.

Las funciones $d_{ij}$ se llaman funciones discriminantes lineales. Observemos además que $d_{ij}=-d_{ji}$ y que dichas funciones son lineales respecto de $\mathbf{x}$. Luego, podemos escribir

$$\mathcal{G}_i=\{\mathbf{x}\in\mathbb{R}^p : d_{ij}(\mathbf{x})\geq 0\quad\forall i\neq j\}\,. \tag{3.2}$$

Donde los hiperplanos $d_{ij}=0$ delimitan las regiones de clasificación. Si $M=2$ y si $\boldsymbol{\alpha}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$, entonces

$$d_{12}(\mathbf{x})=\boldsymbol{\alpha}^{\mathrm{T}}\left(\mathbf{x}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)+\log\pi_1-\log\pi_2\,.$$

$d_{12}(\mathbf{x})$ es la función discriminante lineal de Fisher que clasifica $\mathbf{x}$ en el grupo 1 si $d_{12}>0$.

En el caso donde $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}$ son desconocidos, deben estimarse a partir de las observaciones de cada población. Más precisamente, sean $\mathbf{x}_{i,j}\sim N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, $1\leq j\leq n_i, 1\leq i\leq M$, con $\mathbf{x}_{i,j}\in\mathbb{R}^p$ las observaciones correspondientes a la población $i$-ésima.

Estimadores de $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}$ pueden obtenerse como

$$\widehat{\boldsymbol{\mu}}_i=\overline{\mathbf{x}}_i\quad\text{y}\quad\mathbf{S}=\frac{1}{n-M}\sum_{i=1}^{M}\mathbf{Q}_i\,,$$

donde $n = \sum_{i=1}^{M} n_i$ y $\mathbf{Q}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)$.

Luego, una manera de obtener una regla de clasificación es reemplazar los parámetros desconocidos en (3.2) por sus estimadores. De esta forma, clasificamos a una nueva observación $\mathbf{x}$ en el grupo $g_i$ si $\mathbf{x} \in \widehat{\mathcal{G}}_i$, con

$$\begin{aligned}\widehat{\mathcal{G}}_i &= \{\mathbf{x} \in \mathbb{R}^p : \widehat{L}_i(\mathbf{x}) \geq \widehat{L}_j(\mathbf{x}) \quad \forall j \neq i\} = \{\mathbf{x} \in \mathbb{R}^p : \widehat{L}_i(\mathbf{x}) = \max_{1 \leq j \leq M} \widehat{L}_j(\mathbf{x})\} \\ &= \{\mathbf{x} \in \mathbb{R}^p : \hat{d}_{ij}(\mathbf{x}) \geq 0 \quad \forall i \neq j\}\,,\end{aligned}$$

y donde, para cada $1 \leq i \leq M$,

$$\begin{aligned}\widehat{L}_i(\mathbf{x}) &= \log \hat{\pi}_i + \widehat{\boldsymbol{\mu}}_i^{\mathrm{T}} \mathbf{S}^{-1} \left( \mathbf{x} - \frac{1}{2} \widehat{\boldsymbol{\mu}}_i \right) \\ \hat{d}_{ij}(\mathbf{x}) &= (\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j)^{\mathrm{T}} \mathbf{S}^{-1} \left( \mathbf{x} - \frac{\widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{\mu}}_j}{2} \right) + \log \hat{\pi}_i - \log \hat{\pi}_j\,,\end{aligned}$$

donde si $\pi_i$ son desconocidos se toma $\hat{\pi}_i = n_i/n$.

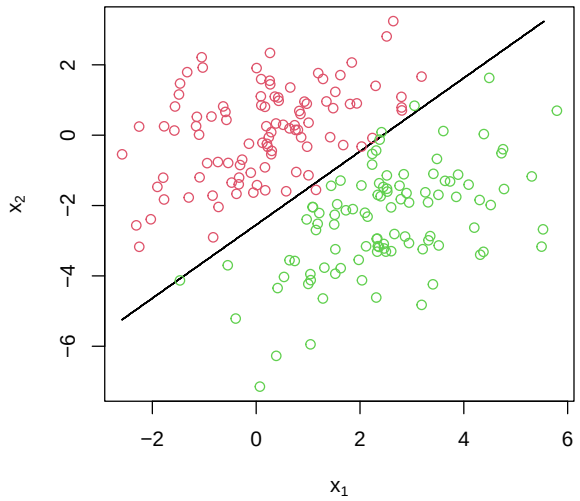


Figura 3.1: Frontera entre las regiones de clasificación (en negro). Dos muestras de normales bivariadas con igual matriz de covarianza según la regla del discriminante lineal.

Por lo tanto, vemos que la regla de clasificación es lineal respecto de $\mathbf{x}$. Geométricamente, esto nos dice que el borde de la regla de clasificación entre dos clases va a ser un hiperplano, como podemos ver en el ejemplo de la Figura 3.1, donde tenemos dos muestras provenientes de normales bivariadas con igual matriz de covarianza pero distinta media.

**Caso en el que las matrices de covarianza son distintas**

En este caso, llamado análisis discriminante cuadrático, también tendremos que las densidades son normales para cada grupo, pero con la diferencia de que ahora no se supone que tengan la misma matriz de covarianza. De la misma manera que en el caso anterior, vamos a usar que la regla de clasificación Bayes asigna un punto a la clase que maximiza la probabilidad a posteriori. Supongamos que sólo tenemos dos grupos $g_1$ y $g_2$ con distribuciones normales de parámetros $\mu_i$ y $\mathbf{\Sigma}_i$, $1 \leq i \leq 2$. De manera análoga al cálculo del discriminante lineal, definimos

$$\begin{aligned} Q(\mathbf{x}) &= \log \frac{\mathbb{P}(G=1|\mathbf{x}=\mathbf{x}_0)}{\mathbb{P}(G=2|\mathbf{x}=\mathbf{x}_0)} = \log \frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} + \log \frac{\pi_1}{\pi_2} \\ &= \log \frac{|\mathbf{\Sigma_2}|}{|\mathbf{\Sigma_1}|} - \frac{1}{2}(\mathbf{x}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}} \mathbf{\Sigma}_1^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}_0 - \boldsymbol{\mu}_2)^{\mathrm{T}} \mathbf{\Sigma}_2^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_2) + \log \frac{\pi_1}{\pi_2}\,. \end{aligned}$$

De esta forma, usando (3.1), clasificamos a una nueva observación $\mathbf{x}$ en el grupo $g_1$ si $\mathbf{x} \in \mathcal{G}_1$, donde

$$\mathcal{G}_1 = \{\mathbf{x} \in \mathbb{R}^p : Q(\mathbf{x}) > 0\}\,,$$

observando que, al ser $Q(\mathbf{x})$ cuadrática respecto de $\mathbf{x}$, la frontera de esta región ya no será un hiperplano sino que será la curva de nivel cero de una función cuadrática.

Para el caso en el que no conocemos los parámetros de cada grupo, estos deben estimarse a partir de las observaciones de cada población como en el caso del análisis discriminante lineal. Más precisamente, sean $\mathbf{x}_{i,j} \sim N(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$, $1 \leq j \leq n_i, 1 \leq i \leq 2$, con $\mathbf{x}_{i,j} \in \mathbb{R}^p$ observaciones correspondientes a cada uno de los grupos.

Estimadores de $\boldsymbol{\mu}_i$ y $\mathbf{\Sigma}_i$ pueden obtenerse como

$$\widehat{\boldsymbol{\mu}}_i = \overline{\mathbf{x}}_i \quad \text{y} \quad \mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)\,.$$

Luego, obtendremos

$$\widehat{Q}(\mathbf{x}) = \log \frac{|\mathbf{S_2}|}{|\mathbf{S_1}|} - \frac{1}{2}(\mathbf{x}_0 - \widehat{\boldsymbol{\mu}}_1)^{\mathrm{T}} \mathbf{S}_1^{-1} (\mathbf{x}_0 - \widehat{\boldsymbol{\mu}}_1) + \frac{1}{2}(\mathbf{x}_0 - \widehat{\boldsymbol{\mu}}_2)^{\mathrm{T}} \mathbf{S}_2^{-1} (\mathbf{x}_0 - \widehat{\boldsymbol{\mu}}_2) + \log \frac{\hat{\pi}_1}{\hat{\pi}_2}\,,$$

donde si $\pi_i$ son desconocidos, se toma $\hat{\pi}_i = n_i/n$.

Como ejemplo de la forma que puede tener la frontera, generamos dos muestras $\mathbf{x}_{i,j} \in \mathbb{R}^2$, $1 \leq i \leq 2$, $1 \leq j \leq n_i$ tales que $\mathbf{x}_{i,j} \sim N(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$, donde $n_i = 100$, $\boldsymbol{\mu}_1 = (0,0)^{\mathrm{T}}$, $\boldsymbol{\mu}_2 = (1.5, -1.5)^{\mathrm{T}}$, $\mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $\mathbf{\Sigma}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$. La Figura 3.2 presenta las observaciones obtenidas y la región donde $\widehat{Q}(\mathbf{x}) = 0$. Es importante mencionar que la forma de esta frontera es uno de los casos posibles, ya que una curva de nivel de una función cuadrática puede tener diversas formas.

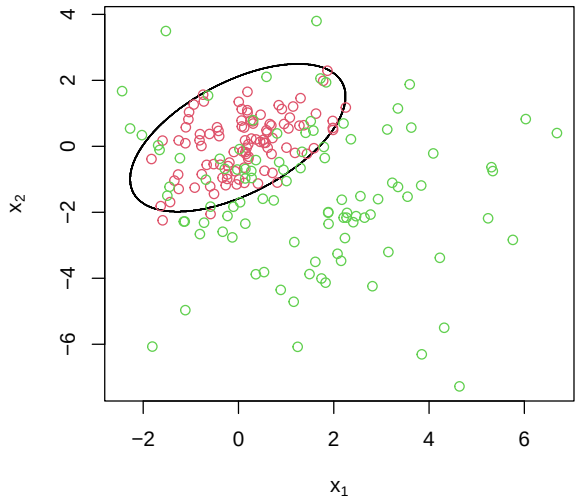


Figura 3.2: Frontera entre las regiones de clasificación cuadrática (en negro). Muestras de normales bivariadas con distinta matriz de covarianza.

### 3.3.2 Regresión logística

Este modelo surge también para intentar estimar las probabilidades a posteriori de cada clase. Comencemos viendo el método en el caso donde $M = 2$, y después generalizaremos al caso en que $M$ es arbitrario. Si $M = 2$, este modelo parte del supuesto de que el log-ratio es lineal respecto de $\mathbf{x}_0$, es decir

$$\log \frac{\mathbb{P}(G = 1|\mathbf{x} = \mathbf{x}_0)}{\mathbb{P}(G = 2|\mathbf{x} = \mathbf{x}_0)} = \log \frac{p}{1-p} = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}} \mathbf{x}_0 \,.$$

Equivalentemente, podemos escribir dicha ecuación como

$$p = \frac{\exp(\beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}} \mathbf{x}_0)}{1 + \exp(\beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}} \mathbf{x}_0)} \,,$$

donde $p = \mathbb{P}(G = 1|\mathbf{x} = \mathbf{x}_0)$ es la probabilidad a posteriori de ser clasificado en la población 1 y $\boldsymbol{\beta} = \{\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}}\}^{\mathrm{T}}$ son los parámetros del modelo.

De esta manera, podemos escribir nuestra probabilidad a posteriori con la forma

$$\mathbb{P}(G = 1|\mathbf{x} = \mathbf{x}_0) = H(\eta(\mathbf{x}_0)) \,,$$

donde $H$ es la función de enlace, que en nuestro caso es la función

$$H(t) = \frac{e^t}{1 + e^t} \,,$$

siendo $\eta(\mathbf{x}_0)$ el predictor lineal dado por $\eta(\mathbf{x}_0) = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_0$. A partir de esto, obtendremos la región de clasificación del grupo $g_1$ como

$$\mathcal{G}_1 = \{\mathbf{x} \in \mathbb{R}^p : H(\eta(\mathbf{x})) > \frac{1}{2}\}\,.$$

En el caso de no contar con $\boldsymbol{\beta}$, podremos estimarlo utilizando una muestra $\mathbf{x}_{i,j} \in \mathbb{R}^p$ con $1 \leq j \leq n_i, 1 \leq i \leq 2$ de observaciones correspondientes a cada uno de los grupos. Para reforzar la dependencia en $\boldsymbol{\beta}$, indicaremos por $\mathbb{P}(G = i|\mathbf{x} = \mathbf{x}_0; \boldsymbol{\beta})$ a $\mathbb{P}(G = i|\mathbf{x} = \mathbf{x}_0)$. Luego, elegiremos los parámetros de manera tal que maximicen la función de verosimilitud $L(\boldsymbol{\beta}) = \prod_{i=1}^{2}\prod_{j=1}^{n_i}\mathbb{P}(G = i|\mathbf{x} = \mathbf{x}_{i,j}; \boldsymbol{\beta})$ o, lo que es igual, que maximicen la función de log-verosimilitud $l(\boldsymbol{\beta}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\log\mathbb{P}(G = i|\mathbf{x} = \mathbf{x}_{i,j}; \boldsymbol{\beta})$.

En la Figura 3.3 podemos ver la forma de la función $H(\eta(x))$ para distintos $\boldsymbol{\beta}$ en el caso donde $x \in \mathbb{R}$, notando como los cambios en $\beta_0$ provocan traslaciones sobre el eje $x$ y los cambios en $\beta_1$ provocan que la curva sea más o menos empinada o si es creciente o decreciente.

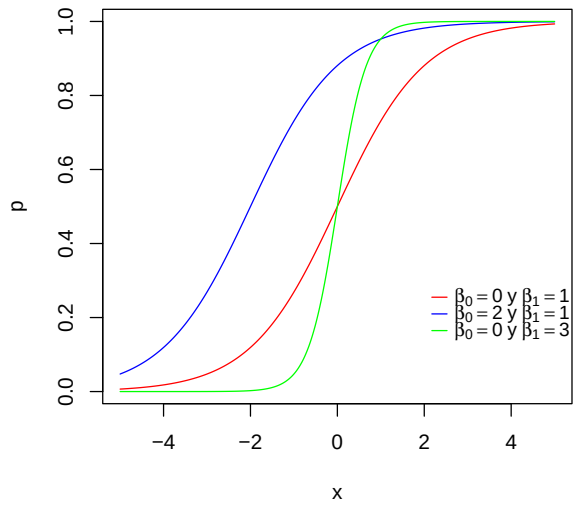


Figura 3.3: 3 casos se la función $H(\eta(x))$ con distintos parámetros: $\beta_0 = 0$ y $\beta_1 = 1$ (línea roja), $\beta_0 = 2$ y $\beta_1 = 1$ (línea azul) y $\beta_0 = 0$ y $\beta_1 = 3$ (línea verde).

Para el caso general donde $M$ es arbitrario, se tiene en cuenta además que $\sum_{i=1}^{M}\mathbb{P}(G = i|\mathbf{x} = \mathbf{x}_0) = 1$. Se parte como en el caso anterior con que el log-ratio de cualquier clase

distinta a $M$ y la clase $M$ son lineales respecto de $x$, es decir, que

$$\log \frac{\mathbb{P}(G=1|\mathbf{x}=\mathbf{x}_0)}{\mathbb{P}(G=M|\mathbf{x}=\mathbf{x}_0)} = \beta_{10} + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_0$$
$$\vdots$$
$$\log \frac{\mathbb{P}(G=M-1|\mathbf{x}=\mathbf{x}_0)}{\mathbb{P}(G=M|\mathbf{x}=\mathbf{x}_0)} = \beta_{(M-1)0} + \boldsymbol{\beta}_{M-1}^{\mathrm{T}}\mathbf{x}_0,$$

donde $\boldsymbol{\beta} = \{\beta_{10}, \boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \beta_{(M-1)0}, \boldsymbol{\beta}_{M-1}^{\mathrm{T}}\}^{\mathrm{T}}$ es un parámetro a elegir. Por lo tanto, si

$$\mathbb{P}(G=M|\mathbf{x}=\mathbf{x}_0) = \frac{1}{1+\sum_{j=1}^{M-1}\exp(\beta_{j,0}+\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_0)},$$

para $i \neq M$,

$$\begin{aligned}
&\log \frac{\mathbb{P}(G=i|\mathbf{x}=\mathbf{x}_0)}{\mathbb{P}(G=M|\mathbf{x}=\mathbf{x}_0)} = \beta_{i,0} + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}_0 \\
\iff &\frac{\mathbb{P}(G=i|\mathbf{x}=\mathbf{x}_0)}{\mathbb{P}(G=M|\mathbf{x}=\mathbf{x}_0)} = \exp(\beta_{i,0}+\boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}_0) \\
\iff &\mathbb{P}(G=i|\mathbf{x}=\mathbf{x}_0) = \exp(\beta_{i,0}+\boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}_0)\mathbb{P}(G=M|\mathbf{x}=\mathbf{x}_0) \\
\iff &\mathbb{P}(G=i|\mathbf{x}=\mathbf{x}_0) = \frac{\exp(\beta_{i,0}+\boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}_0)}{1+\sum_{j=1}^{M-1}\exp(\beta_{j,0}+\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_0)}.
\end{aligned}$$

Luego, se obtiene que las probabilidades a posteriori vienen dadas por

$$\begin{aligned}
\mathbb{P}(G=i|\mathbf{x}=\mathbf{x}_0) &= \frac{\exp(\beta_{i,0}+\boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}_0)}{1+\sum_{j=1}^{M-1}\exp(\beta_{j,0}+\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_0)}, \quad 1 \leq i \leq M-1 \\
\mathbb{P}(G=M|\mathbf{x}=\mathbf{x}_0) &= \frac{1}{1+\sum_{j=1}^{M-1}\exp(\beta_{j,0}+\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_0)}.
\end{aligned}$$

De esta manera, obtendremos la región de clasificación que clasifica a $\mathbf{x}$ en el grupo $g_i$ si $\mathbf{x} \in \mathcal{G}_i$, donde

$$\mathcal{G}_i = \{\mathbf{x} \in \mathbb{R}^p : \log \mathbb{P}(G=i|\mathbf{x}=\mathbf{x}) \geq \log \mathbb{P}(G=j|\mathbf{x}=\mathbf{x}) \quad \forall j \neq i\},$$

o, equivalentemente,

$$\mathcal{G}_i = \{\mathbf{x} \in \mathbb{R}^p : (\beta_{i,0}-\beta_{j,0}) + (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)^{\mathrm{T}}\mathbf{x}_0 \geq 0 \quad \forall j \neq i\}.$$

En el caso de no contar con $\boldsymbol{\beta}$, podremos estimarlo como en el caso de la regresión logística para dos grupos, es decir, de manera tal que maximice la función de verosimilitud $L(\boldsymbol{\beta}) = \prod_{i=1}^{M}\prod_{j=1}^{n_i}\mathbb{P}(G=i|\mathbf{x}=\mathbf{x}_{i,j};\boldsymbol{\beta})$.

### 3.3.3 Clasificación no paramétrica

La regla Bayes clasifica $\mathbf{x}$ en el grupo $g_i$ cuando $\mathbf{x} \in \mathcal{G}_i$, donde

$$\begin{aligned}\mathcal{G}_i &= \{\mathbf{x} \in \mathbb{R}^p : \pi_i f_i(\mathbf{x}) \geq \pi_j f_j(\mathbf{x}) \quad \forall j \neq i\} \\ &= \{\mathbf{x} \in \mathbb{R}^p : \frac{f_i(\mathbf{x})}{f_j(\mathbf{x})} \geq \frac{\pi_j}{\pi_i} \quad \forall j \neq i\}\,.\end{aligned}$$

En la mayoría de los casos $f_i$ es desconocida, y por lo tanto deberíamos estimarla en el punto $\mathbf{x}_0$ a clasificar a partir de una muestra aleatoria. Sean $\mathbf{x}_{i,1} \ldots \mathbf{x}_{i,n_i}$, $1 \leq i \leq M$, independientes tales que $\mathbf{x}_{i,j} \sim f_i$ para todo $1 \leq j \leq n_i$.

El estimador de núcleos de la densidad, introducido por Rosenblatt (1956) y Parzen (1962) se define como

$$\hat{f}_{i,n_i}(\mathbf{x}) = \frac{1}{n_i h^p} \sum_{j=1}^{n_i} \mathcal{K}\left(\frac{\mathbf{x}_{i,j} - \mathbf{x}}{h}\right)\,,$$

donde $h$ es la ventana y $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ es un núcleo, o sea, una función tal que $\mathcal{K}(\mathbf{u}) \geq 0$ para todo $\mathbf{u} \in \mathbb{R}^p$ y $\int_{\mathbb{R}^p} \mathcal{K}(\mathbf{u}) d\mathbf{u} = 1$. Una posible elección de $\mathcal{K}(\mathbf{u})$ es tomar

$$\mathcal{K}(\mathbf{u}) = \frac{K(\|\mathbf{u}\|)}{\int_{\mathbb{R}^p} K(\|\mathbf{z}\|) d\mathbf{z}}\,,$$

donde $K : \mathbb{R} \to \mathbb{R}$ y $K \geq 0$.

La ventana $h$ regula el compromiso entre sesgo y varianza del estimador y podría elegirse de forma diferente para cada grupo. Una descripción más detallada de estimadores de la densidad puede verse en Scott (1992).

De esta forma, el método no paramétrico de clasificación basado en núcleos, clasifica a $\mathbf{x}$ en el grupo $g_i$ si

$$\hat{f}_{i,n_i}(\mathbf{x}) = \max_{1 \leq j \leq M} \hat{f}_{j,n_j}(\mathbf{x})\,.$$

Otras elecciones de la función $\hat{f}_{i,n_i}$ son posibles. Por ejemplo, podemos definir

$$\hat{f}_{i,n_i}(\mathbf{x}) = \frac{1}{n_i h^p} \sum_{j=1}^{n_i} K\left(\frac{d(\mathbf{x}_{i,j}, \mathbf{x})}{h}\right)\,,$$

donde $d(\mathbf{u}, \mathbf{x})$ es una distancia adecuada entre $\mathbf{u}$ y $\mathbf{x}$. El ejemplo anterior correspondería a tomar $d(\mathbf{u}, \mathbf{x}) = \|\mathbf{u} - \mathbf{x}\|$. Otra posible elección es tomar la distancia de Mahalanobis, es decir, tomar una distancia que varía con cada grupo y definir

$$\hat{f}_{i,n_i}(\mathbf{x}) = \frac{1}{c_i n_i h^p} \sum_{j=1}^{n_i} K\left(\frac{(\mathbf{x} - \mathbf{x}_{i,j})^{\mathrm{T}} \mathbf{S}_i^{-1} (\mathbf{x} - \mathbf{x}_{i,j})}{h}\right)\,,$$

con $c_i$ elegido para que $\int_{\mathbb{R}^p} \hat{f}_{i,n_i}(\mathbf{u}) d\mathbf{u} = 1$.

Para ejemplificar el uso del ancho de banda, tomemos el núcleo $K(s) = \mathbb{1}_{[0,1]}(s)(1-s)$. El valor de $h$ va a restringir las observaciones que miramos de la siguiente manera: si $h = 1$, todas las observaciones de la muestra que estén a distancia menor a 1 de $\mathbf{x}$ van a afectar la clasificación con su correspondiente peso. Si en cambio $h = 10$, las observaciones que estén a distancia menor a 10 de $\mathbf{x}$ afectarán al clasificador, dando origen a un estimador de la densidad con menor varianza.

Otro método para estimar la densidad es el método de vecinos más cercanos. Fijado un entero positivo $k$, tendremos que, como el método anterior, elegir una distancia entre puntos $d(\mathbf{x}_{i,j}, \mathbf{x})$, como puede ser la distancia de Mahalanobis o la euclídea. Luego, dado un punto a clasificar $\mathbf{x}_0$, se calcula la distancia de dicho punto a todos los elementos $\mathbf{x}_{i,j}$ de nuestra muestra. A partir de todas las distancias $d_{i,j} = d(\mathbf{x}_{i,j}, \mathbf{x}_0)$, se efectúa un orden de estas de menor a mayor, quedándonos con las $k$ menores, notadas $d^{(1)}, \ldots, d^{(k)}$. Sea $D := d^{(k)}$ y sea $M_i := \#\{\mathbf{x}_{i,j} : d_{i,j} \leq D\}$, o sea, la cantidad de elementos del grupo $i$ que están entre los $k$ más próximos.

A partir de este conjunto $M_i$, podemos tener una aproximación de la densidad dada por

$$\hat{f}_{i,n_i}(\mathbf{x}) = \frac{M_i}{n_i D^p |B_1(\mathbf{0})|},$$

siendo $|B_1(\mathbf{0})|$ la medida de la bola unitaria.

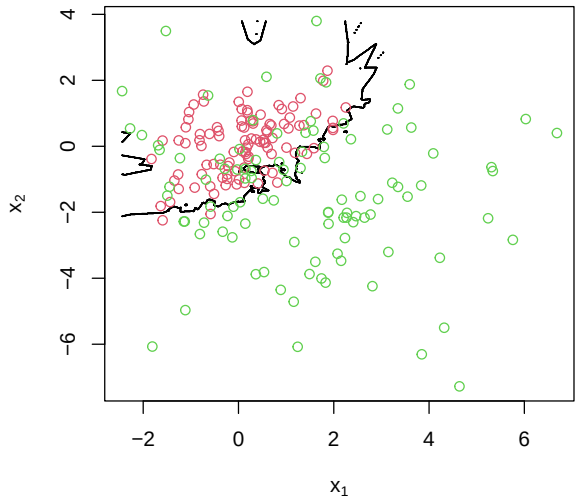


Figura 3.4: Frontera de las regiones de clasificación (en negro) dadas dos muestras pertenecientes a distintos grupos según el método de $k$ vecinos más cercanos con $k = 11$.

Luego, clasificamos a $\mathbf{x}_0$ en el grupo $g_i$ si

$$\pi_i \hat{f}_{i,n_i}(\mathbf{x}_0) = \max_{1\leq j\leq M} \pi_j \hat{f}_{j,n_j}(\mathbf{x}_0)$$
$$\iff \pi_i \frac{M_i}{n_i} = \max_{1\leq j\leq M} \pi_j \frac{M_j}{n_j}\,.$$

En el caso donde no conocemos $\pi_i$, tomamos $\hat{\pi}_i = n_i/n$. Este método se puede pensar como la regla de clasificación que asigna $\mathbf{x}_0$ al grupo que tiene mayor frecuencia, respecto de su grupo, entre los $k$ elementos de la muestra más cercanos. En la Figura 3.4 vemos un ejemplo del borde de decisión de este método.

### 3.3.4 Clasificación mediante máquinas del vector soporte

Para este modelo vamos a suponer que estamos trabajando con el problema de clasificación binaria, es decir, que sólo tenemos dos clases posibles $g_1$ y $g_2$, además trabajaremos con una notación distinta para esta sección, pero con el objetivo de facilitar la comprensión. Supongamos que tenemos una muestra aleatoria $\mathbf{x}_{i,1}\dots\mathbf{x}_{i,n_i}$, con $\mathbf{x}_{i,j} \sim f_i$ para todo $1 \leq j \leq n_i$, $1 \leq i \leq 2$. Sea $\{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{((\mathbf{x}_{1,1}, 1), \dots, (\mathbf{x}_{1,n_1}, 1), (\mathbf{x}_{2,1}, -1), \dots, (\mathbf{x}_{2,n_2}, -1))\}$. que $y_i = 1$ si $\mathbf{x}_i$ pertenece al grupo $g_1$ e $y_i = -1$ si $\mathbf{x}_i$ pertenece al grupo $g_2$.

Supongamos además que las observaciones de los dos grupos pueden ser separadas por un hiperplano. Buscaremos entonces el hiperplano que separe los grupos de forma tal que la distancia mínima entre dicho hiperplano y el punto más cercano de cada clase sea máxima. Si parametrizamos al hiperplano como $\{\mathbf{x} : \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \beta_0 = 0\}$, podemos escribir nuestro problema como un problema de maximización de la siguiente manera:

$$\begin{aligned} \max_{\boldsymbol{\beta},\beta_0} T \quad &\text{sujeto a } \|\boldsymbol{\beta}\| = 1 \\ &y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0) \geq T, \quad i = 1, \dots, n. \end{aligned} \tag{3.3}$$

De esta manera, todos los puntos están a una distancia mayor o igual a $T$ del hiperplano. Además, podemos redefinir (3.3) sacando la restricción de $\|\boldsymbol{\beta}\| = 1$ pidiendo $y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)/\|\boldsymbol{\beta}\| \geq T$, lo que redefine a $\beta_0$. Esta desigualdad es equivalente a $y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0) \geq T\|\boldsymbol{\beta}\|$. De esta manera, tenemos infinitos argumentos $\boldsymbol{\beta}$ y $\beta_0$ que maximizan la ecuación, ya que cualquier reescalamiento de estos parámetros por un valor estrictamente positivo va a hacer que se sigan cumpliendo las desigualdades sin afectar el valor de $T$. En particular, podemos asignar la norma de $\boldsymbol{\beta}$ como $1/T$. Luego, el sistema (3.3) es equivalente a

$$\min_{\boldsymbol{\beta},\beta_0} \|\boldsymbol{\beta}\| \quad \text{sujeto a } y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0) \geq 1, \quad i = 1, \dots, n. \tag{3.4}$$

En este nuevo sistema estamos trabajando con un problema de optimización convexo y podremos calcular la solución utilizando las condiciones de Karush-Kuhn-Tucker.

En la Figura 3.5 vemos la recta que separa los dos grupos dada por la clasificación mediante máquinas del vector soporte. Esta maximiza el margen dado por las líneas punteadas. Es importante destacar que, en este ejemplo, existe un hiperplano que separa las observaciones de manera tal que ninguna cae en la región delimitada por las líneas punteadas o del otro lado de la frontera marcada con la línea negra.

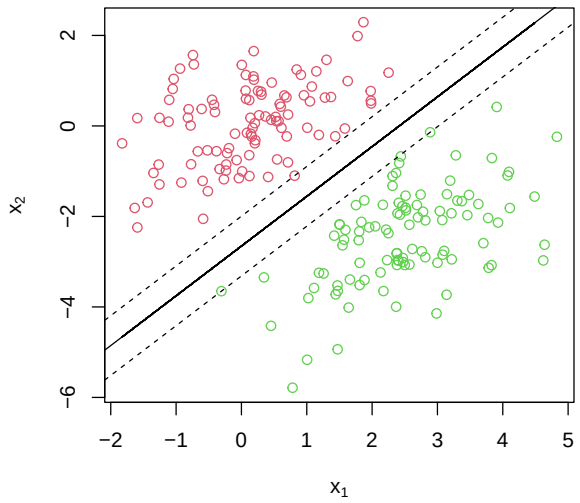


Figura 3.5: Borde de decisión de dos muestras de normales multivariadas según el método de clasificación mediante máquinas del vector soporte. Las líneas punteadas representan el margen.

Consideremos ahora el caso en el que las clases no son separables por un hiperplano, o sea que no existe un hiperplano que separe las observaciones de nuestros dos grupos. En dicha situación, la solución será alterar el sistema (3.4) agregando un margen de error de clasificación de la siguiente manera,

$$\begin{aligned} &\min_{\boldsymbol{\beta},\beta_0} \|\boldsymbol{\beta}\| \quad \text{sujeto a } y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0) \geq 1-\xi_i, \quad i=1,\ldots,n \\ &\sum_{i=1}^{n} \xi_i \leq \text{constante} \\ &\xi_i \geq 0, \quad i=1,\ldots,n\,. \end{aligned} \tag{3.5}$$

Luego, el valor $\xi_i$ en la condición $y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0) \geq 1-\xi_i$ representa cuánto permitimos que el punto $\mathbf{x}_i$ se encuentre del otro lado de la frontera. Con la condición de que $\sum_{i=1}^{n}\xi_i$ es menor a alguna constante, controlamos cuántos errores de clasificación permitimos. Al sistema (3.5)

lo podemos reescribir equivalentemente como

$$\min_{\boldsymbol{\beta},\beta_0} \|\boldsymbol{\beta}\| + C\sum_{i=1}^{n}\xi_i \quad \text{sujeto a } y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0) \geq 1-\xi_i, \quad i=1,\ldots,n$$
$$\xi_i \geq 0, \quad i=1,\ldots,n\,,$$

donde $C$ es un parámetro de costo para penalizar los errores de clasificación y la cercanía al hiperplano, reemplazando la "constante" de (3.5). En el libro de Hastie et al. (2001) se puede encontrar la solución de este problema de optimización utilizando las condiciones de Karush–Kuhn–Tucker.

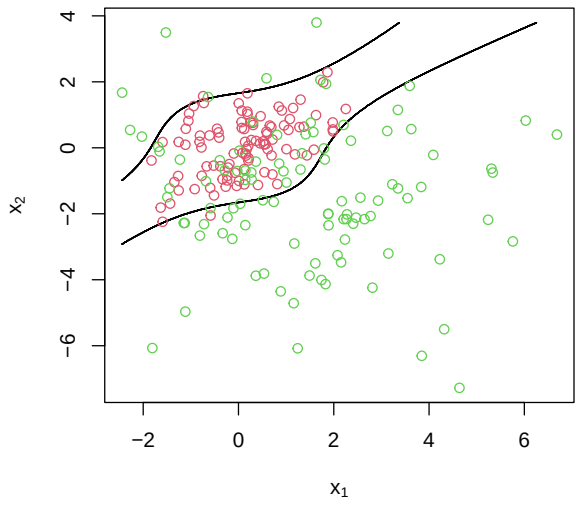


Figura 3.6: Borde de decisión de dos muestras de normales multivariadas según el método de clasificación mediante máquinas del vector soporte con un núcleo polinomial de grado 4.

Ahora buscaremos extender este método cuando se trabaja con conjuntos de datos donde la regla de clasificación dada por un hiperplano no sea posible, es decir, cuando los grupos no son linealmente separables, como en la Figura 3.6. Para resolver este problema, se puede cambiar el espacio donde se encuentran los datos para tener mayor dimensión y por ende mayor flexibilidad. Para mantener la solución encontrada al problema de optimización, tendremos que trasladarnos a espacios con producto interno y con dimensión suficientemente alta, donde sí podamos conseguir un hiperplano en este nuevo espacio que separe nuestros grupos. La transformación de un espacio a otro es a través de un núcleo $K$ que tendrá que ser simétrico y definido positivo. Por ejemplo, una familia de núcleos está dada por $K(\mathbf{x},\mathbf{y}) = (1+\mathbf{x}^{\mathrm{T}}\mathbf{y})^d$, llamados los núcleos de polinomios de grado $d$. En el caso donde el grado del polinomio es 2 y estamos trabajando originalmente en $\mathbb{R}^2$, se puede ver que

$$K(\mathbf{x},\mathbf{y}) = 1 + 2x_1y_1 + 2x_2y_2 + (x_1y_1)^2 + (x_2y_2)^2 + 2x_1y_1x_2y_2.$$

De esta manera, la observación $\mathbf{x} = (x_1, x_2)$ pasa a ser $(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2) \in \mathbb{R}^4$.

Como ejemplo, presentamos la Figura 3.6 donde tomamos el clasificador el método de clasificación mediante máquinas del vector soporte con el núcleo polinomial de grado 4. De esta manera, vemos la proyección del hiperplano separador sobre el espacio en dimensión 2, donde originalmente están nuestros datos.

## 3.4 Clasificación en $\mathbb{H}$

En esta sección, veremos distintos métodos para afrontar el problema de clasificación de datos funcionales. Para esto, en lo posible tomaremos ideas del caso multivariado y las adaptaremos al caso funcional. Debemos tener en cuenta que esto puede ocasionar distintos problemas que describiremos.

En primera instancia, los clasificadores multivariados no son sensibles al siguiente cambio de coordenadas: si estamos trabajando en $\mathbb{R}^p$, el intercambio entre dos coordenadas dado por $x = (x_1, \ldots, x_i, \ldots, x_j, \ldots, x_p) \to (x_1, \ldots, x_j, \ldots, x_i, \ldots, x_p)$ no afecta a ningún clasificador multivariado. En cambio, en el caso funcional, el cambio de coordenadas dado por $x(t) \to \tilde{x}(t)$, donde

$$\tilde{x}(t) = \begin{cases} x(t) & si\, t \notin \{t_i, t_j\} \\ x(t_i) & si\, t = t_j \\ x(t_j) & si\, t = t_i \end{cases}$$

puede provocar la pérdida de regularidad particular propia del problema a trabajar. Por ejemplo, podemos estar trabajando con un método que necesite la propiedad de tener la media cuadrática continua, como cualquier método que use la descomposición dada por el Teorema de Karhunen-Loève, o métodos que utilicen la derivada de cada observación. Por la misma razón, los métodos basados en núcleos o vecinos más cercanos que necesitan preprocesamiento de los datos a través de la estandarización van a correr el riesgo de no poderse llevar a cabo. La estandarización también va a afectar considerablemente la derivada de las observaciones, ya que esta aplanaría los datos distorsionando la derivada.

Otro problema es la ausencia de la densidad en el caso funcional, que va a provocar que métodos como la regla de clasificación lineal o cuadrática descripta en la Sección 3.3.1 no se puedan llevar a cabo. Estos métodos paramétricos obligatoriamente parten de un modelo donde la densidad del elemento aleatorio del grupo $i$-ésimo pertenece a una familia de densidades conocidas, como puede ser la densidad de una normal. A partir de esto, dichos métodos buscan estimar los parámetros de la densidad para luego proveer una aproximación al clasificador. De manera que, sin la noción de densidad para el caso funcional, estos métodos no son posibles de adaptar.

Finalmente, un problema mayor con el que no trabajaremos en esta tesis se puede dar de la siguiente manera: supongamos que tenemos dos observaciones de un elemento aleatorio $X \in L^2(\mathcal{I})$ notadas $X_1$ y $X_2$. Puede ocurrir que las mediciones de estas observaciones

a lo largo del tiempo sean llevadas a cabo en distintos momentos. Es decir que, dadas $t_1, \ldots, t_p$ y $s_1, \ldots, s_{p'}$ dos particiones del intervalo $\mathcal{I}$, conocemos solamente $X_1(t_i)$ y $X_2(s_j)$, con $1 \leq i \leq p$ y $1 \leq j \leq p'$. En esta situación, cualquier método del análisis multivariado se enfrentará al problema de que el espacio está mal definido y ni siquiera tendremos una dimensión en común. En el caso funcional, sin embargo, podremos resolver esta situación a través de la regularización de los datos con herramientas como las dadas por expansiones en bases de splines, polinomios o utilizando suavizadores basados en núcleos.

### 3.4.1 Métodos de clasificación basados en núcleos o vecinos más cercanos

Teniendo en cuenta las ideas desarrolladas en la Sección 3.3.3, para las reglas de clasificación basadas en núcleos y vecinos más cercanos cuando los datos pertenecen a $\mathbb{R}^p$, definiremos procedimientos para datos funcionales tomando una métrica o semi-métrica adecuada en la definición del procedimiento. Para esto comenzaremos describiendo distintas métricas o semi-métricas presentadas en Ferraty y Vieu (2003) y Chang et al. (2014), quienes analizan estas propuestas. La particularidad de poder trabajar con semi-métricas se debe a que la hipótesis de identificación no es necesaria, es decir, que no hace falta la condición de $d(x, y) = 0 \implies x = y$. Esto sucede ya que, en el caso en el que la semi-métrica asigne una distancia nula a dos observaciones distintas, podemos pensar que estamos en una situación donde dichas observaciones pertenecen a una misma clase de equivalencia, en la que van a compartir suficientes particularidades como para que no sea necesario distinguirlas.

*1. Métrica basada en la norma $L^2(\mathcal{I})$:*

Esta métrica es la más característica del espacio $L^2(\mathcal{I})$ y se define como

$$d(x, y) = \left( \int_{\mathcal{I}} (x(t) - y(t))^2 dt \right)^{\frac{1}{2}} .$$

De igual manera se puede usar cualquier métrica del espacio $L^p(\mathcal{I})$ definida por

$$d(x, y) = \left( \int_{\mathcal{I}} (x(t) - y(t))^p dt \right)^{\frac{1}{p}} ,$$

si $1 \leq p < \infty$ y $d(x, y) = \sup_{t \in \mathcal{I}} |x(t) - y(t)|$ si $p = \infty$. Es importante mencionar que esta versatilidad en el uso de otras métricas, es decir, de métricas de $L^p(\mathcal{I})$, $1 \leq p \leq \infty$ con $p \neq 2$, se debe a que en la práctica siempre estaremos trabajando con observaciones acotadas.

*2. Semi-métrica basada en derivadas:*

En algunos casos las observaciones corresponden a funciones diferenciables. Por esta razón, es natural tomar $\mathbb{H} = \mathbb{W}^{q,2}$ con $\mathbb{W}^{q,2} = \{x \in L^2(\mathcal{I}) : x^{(k)} \in L^2(\mathcal{I}) \, \forall k \leq q\}$ el espacio de Sobolev de funciones pertenecientes a $L^2(\mathcal{I})$ cuyas derivadas de orden menor o igual a $q$

también pertenecen a dicho espacio. De esta manera, podremos usar la métrica propia de este espacio dada por

$$d_q(x,y) = \left( \sum_{k=0}^{q} \int_{\mathcal{I}} (x^{(k)}(t) - y^{(k)}(t))^2 dt \right)^{\frac{1}{2}}.$$

A partir de esta métrica, Ferraty y Vieu (2003) proponen una semi-métrica donde solo se considera la derivada de orden $q$, es decir,

$$d_q(x,y) = \left( \int_{\mathcal{I}} (x^{(q)}(t) - y^{(q)}(t))^2 dt \right)^{\frac{1}{2}}.$$

Esta variante puede ser de utilidad en el caso donde la información necesaria para la clasificación está contenida en la derivada de orden $q$, ya que la métrica del espacio de Sobolev puede perturbar la clasificación al utilizar todas las derivadas con la misma jerarquía.

*3. Semi-métrica basada en componentes principales:*

Supongamos que $X \in L^2(\mathcal{I})$ es un elemento de media cuadrática continua, para de esta manera poder usar el Teorema de Karhunen y Loève dado en el Teorema 2.5.4 para escribir a $X$ como

$$X = \mu + \sum_{k=1}^{\infty} \langle X - \mu, \varphi_k \rangle \varphi_k,$$

donde $\mu = \mathbb{E}[X]$ y $\{\varphi_k\}_{k \geq 1}$ son las autofunciones ortonormales entre sí de la función de covarianza $\Gamma_X$ ordenadas de forma tal que, si $\lambda_k$ es el autovalor asociado a $\varphi_k$, con $\lambda_1 \geq \lambda_2 \geq \dots$. Sea ahora $\tilde{X}^{(q)} = \mu + \sum_{k=1}^{q} \langle X - \mu, \varphi_k \rangle \varphi_k$, recordando que esta aproximación, por el Corolario 2.5.2, minimiza $\mathbb{E}\{\int_{\mathcal{I}} [(X-\mu)(t) - \pi_{\mathcal{L}}(X-\mu)(t)]^2 dt\}$ entre todas las proyecciones $\pi_{\mathcal{L}}(X)$ a subespacios $\mathcal{L}$ de dimensión $q$.

Basada en las autofunciones $\{\varphi_k\}_{k \geq 1}$, Ferraty y Vieu (2003) definen la semi-norma de $u \in L^2(\mathcal{I})$ como

$$\left( \sum_{k=1}^{q} \left( \int_{\mathcal{I}} u(t) \varphi_k(t) dt \right)^2 \right)^{\frac{1}{2}} = \left( \sum_{k=1}^{q} \langle u, \varphi_k \rangle^2 \right)^{\frac{1}{2}},$$

donde $u \in L^2(\mathcal{I})$ y $q$ es un parámetro que marca la resolución al fijar con cuántas componentes principales trabajamos. De esta manera, podemos definir la semi-métrica como

$$d_q(u,v) = \left( \sum_{k=1}^{q} \left( \int_{\mathcal{I}} (u(t) - v(t)) \varphi_k(t) dt \right)^2 \right)^{\frac{1}{2}}.$$

Para definir una regla de clasificación a partir de estas nociones de distancia, utilizaremos una estimación de la probabilidad a posteriori $q_i(x) = \mathbb{P}(G = i | X = x)$ utilizando el estimador de regresión y observando que $q_i(x) = \mathbb{E}(\mathbb{1}_{(G=i)} | X = x)$. Si $X_{i,j}$, $1 \leq j \leq n_i$,

$1 \leq i \leq M$ son observaciones independientes, podemos escribir al estimador de la probabilidad a posteriori como

$$\hat{q}_i(x) = \frac{\sum_{j=1}^{n_i} K(\frac{d(x,X_{i,j})}{h})}{\sum_{k=1}^{M}\sum_{j=1}^{n_k} K(\frac{d(x,X_{k,j})}{h})}\,. \tag{3.6}$$

De esta manera, al igual que en el caso multivariado, clasificamos a $x$ en el grupo $g_i$ si

$$\hat{q}_i(x) = \max_{1\leq k\leq M} \hat{q}_k(x)$$

Cabe mencionar que si tomamos la semi-métrica basada en componentes principales, es necesario estimar las direcciones principales $\varphi_k$ y para ello, o bien suponemos que los operadores de covarianza de los distintos grupos son iguales, es decir, $\Gamma_i = \Gamma$ para todo $1 \leq i \leq M$, donde $\Gamma_i$ es el operador de covarianza de $X|_{G=i}$, o bien no hacemos ese supuesto.

En el primer caso, las direcciones $\varphi_k$ se pueden estimar considerando las autofunciones del operador

$$\widehat{\Gamma} = \sum_{i=1}^{M} \frac{n_i}{n}\widehat{\Gamma}_i\,,$$

donde

$$\widehat{\Gamma}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)\otimes(X_{i,j} - \bar{X}_i)\,.$$

En el segundo caso, la expresión de (3.6) debe adaptarse tomando

$$\hat{q}_i(x) = \frac{\sum_{j=1}^{n_i} K(\frac{d_i(x,X_{i,j})}{h})}{\sum_{k=1}^{M}\sum_{j=1}^{n_k} K(\frac{d_k(x,X_{k,j})}{h})}\,,$$

siendo

$$d_i(u,v) = \left(\sum_{k=1}^{q}\left(\int_{\mathcal{I}}(u(t)-v(t))\hat{\varphi}_{i,k}(t)dt\right)^2\right)^{\frac{1}{2}}\,,$$

con $\hat{\varphi}_{i,k}$ la $k$-ésima autofunción del operador $\widehat{\Gamma}_i$.

El método de vecinos más cercanos puede adaptarse el caso funcional mediante los mismos pasos que en el caso multivariado. Más precisamente, fijado un entero positivo $k$, una métrica o semi-métrica $d(u,v)$ y una nueva observación a clasificar $x$, se calcula su distancia a todos los elementos $X_{i,j}$ de nuestra muestra. A partir de todas las distancias $d_{i,j} = d(X_{i,j}, x)$, se efectúa un orden de estas de menor a mayor, quedándonos con las $k$ menores, notadas $d^{(1)}, \ldots, d^{(k)}$, definiendo $D := d^{(k)}$. Sea $M_i := \#\{X_{i,j} : d_{i,j} \leq D\}$, o sea, la cantidad de elementos del grupo $i$ que están entre los $k$ más próximos. En este caso no podremos tener una aproximación de la densidad como en el caso multivariado, pero igualmente podremos clasificar a nuestra observación $x$ en el grupo $g_i$ si

$$\pi_i\frac{M_i}{n_i} = \max_{1\leq k\leq M}\pi_k\frac{M_k}{n_k}\,.$$

En el caso donde no conocemos $\pi_i$, tomamos $\hat{\pi}_i = n_i/n$.

### 3.4.2 Regresión logística funcional

Este método parte del modelo lineal generalizado para clasificación binaria, donde partimos de la suposición de que

$$\mathbb{E}(G|X) = H(\eta),$$

donde $\eta = \alpha + \int_{\mathcal{I}} \beta(t)X(t)dt$ es el predictor lineal y $H(t) = e^t/(1+e^t)$ es la función de enlace.

Como en el trabajo de Leng y Müller (2005), para encontrar una expresión para el parámetro $\beta$, supondremos que se cumplen las condiciones para utilizar el Teorema de Karhunen-Loève dado en el Teorema 2.5.4, suponiendo además que $\mathbb{E}X = 0$. De esta manera podemos utilizar las descomposiciones $X = \sum_{k=1}^{\infty}\langle X, \varphi_k\rangle\varphi_k$ y $\beta = \sum_{k=1}^{\infty}\langle \beta, \varphi_k\rangle\varphi_k$. Por lo tanto,

$$\eta = \alpha + \langle \beta, X\rangle = \alpha + \sum_{k=1}^{\infty} \lambda_k\beta_k\,,$$

donde $\beta_k = \langle \beta, \varphi_k\rangle$. Esto sugiere aproximar el desarrollo utilizando solamente $q$ direcciones principales de modo de obtener estimadores de $\beta$, pasando al problema de hallar estimadores de $\{\beta_k\}_{k=1}^{q}$.

Para $q$ fijo, el parámetro $\boldsymbol{\beta} := (\alpha, \beta_1, \ldots, \beta_q)$ es estimado a partir de los estimadores de cuasi-verosimilitud o máxima verosimilitud. Supongamos que tenemos una muestra aleatoria $X_{i,1} \ldots X_{i,n_i}$ con $X_{i,j} \in L^2(\mathcal{I})$ proveniente de la clase $g_i$ para todo $1 \leq j \leq n_i$, $1 \leq i \leq 2$. Para esta sección utilizaremos la siguiente notación para facilitar la comprensión. Sea $\{(X_i, y_i)\}_{i=1}^{n} = \{((X_{1,1}, 1), \ldots, (X_{1,n_1}, 1), (X_{2,1}, 0), \ldots, (X_{2,n_2}, 0))\}$. Es decir que $y_i = 1$ si $X_i$ pertenece al grupo $g_1$ e $y_i = 0$ si $X_i$ pertenece al grupo $g_2$. Luego, utilizando que $H'(t) = H(t)(1 - H(t))$, para hallar la estimación $\widehat{\boldsymbol{\beta}}$ resolveremos la ecuación $U(\widehat{\boldsymbol{\beta}}) = 0$, donde para hallar la estimación $\widehat{\boldsymbol{\beta}}$ por máxima verosimilitud resolveremos

$$U(\widehat{\boldsymbol{\beta}}) = \sum_{j=1}^{n}(y_j - \hat{q}_j)\widehat{\boldsymbol{\xi}}_j\,,$$

donde $\hat{q}_j = H(\alpha + \sum_{k=1}^{q}\hat{\xi}_{j,k}\hat{\beta}_k)$ y $\widehat{\boldsymbol{\xi}}_j = (\hat{\xi}_{j,1} \ldots, \hat{\xi}_{j,q})^{\mathrm{T}}$, siendo $\hat{\xi}_{j,k} = \langle X_j, \hat{\varphi}_k\rangle$ con $\hat{\varphi}_k$ obtenidos de igual manera que en la Sección 3.4.1.

Si además conocemos la probabilidad a priori $\pi_1$, obtenida la estimación $\widehat{\boldsymbol{\beta}}$, podremos estimar para la observación $x$ la probabilidad a posteriori de pertenecer al grupo $g_1$ con

$$\hat{q}_1(x) = H\left(\alpha + \sum_{k=1}^{q}\langle x, \hat{\varphi}_k\rangle\hat{\beta}_k\right)\,.$$

Luego, al estar trabajando solamente con dos clases, clasificaremos la observación $x$ en el grupo $g_1$ si $\hat{q}_1(x) \geq 1/2$, y la clasificaremos en $g_2$ en el caso contrario.

### 3.4.3 Clasificación basada en profundidades y atipicidades

El método de clasificación de esta sección parte de la noción de profundidad que definiremos en el Capítulo 4. De esta manera surge una familia de clasificadores que dependerá de la profundidad elegida. Para esto supondremos que estamos trabajando con una noción de profundidad $D(x;\mathbb{P}_X)$ fija. De esta manera tendremos las funciones de profundidad $(D(x;\mathbb{P}_{X_i}))_{i=1}^M$, donde $\mathbb{P}_{X_i}$ es la medida de probabilidad del elemento aleatorio $X_i$ correspondiente al grupo $g_i$.

Como primera propuesta de clasificación tomaremos la presentada en Cuevas et al. (2007). Dada una observación $x$, buscaremos la clase donde $x$ sea más profunda, es decir, donde esta observación sea más central. Análogamente, Hubert et al. (2016) proponen calcular la atipicidad de $x$ respecto de cada clase y asignarle la clase donde la observación sea menos atípica. Luego, volviendo al caso donde usamos la profundidad, para este método calcularemos $D_i(x) := D(x;\mathbb{P}_{X_i})$ para $1 \leq i \leq M$, y le asignaremos a $x$ la clase que maximice dicha profundidad. Es importante mencionar que este método necesita una regla de clasificación en el caso de que sucedan empates. Por ejemplo, supongamos que trabajamos con profundidades como la del semi-espacio o la simplicial, donde la profundidad se vuelve nula al salirse de la región central, evento que puede suceder si estamos trabajando con pocas observaciones. Si tenemos una nueva observación suficientemente atípica para todas las clases, la profundidad será nula para todas estas y no podremos clasificar. Por esto, Hubert et al. (2016) recomiendan usar la atipicidad en lugar de la profundidad, ya que de esta manera se evita el problema de empates en observaciones atípicas para toda clase. Para ejemplificar esto, en el caso de utilizar una noción de atipicidad como puede ser la distancia de Mahalanobis respecto de cada grupo, si tenemos una nueva observación atípica para todo grupo, la probabilidad de que la distancia de Mahalanobis sea igual para todo grupo es despreciable, evitando así el problema del empate.

Por otro lado, es importante notar que la diferencia de dispersión de los datos según cada clase puede afectar negativamente al clasificador. Para ejemplificar, vemos la Figura 3.7 (a) donde tenemos dos clases de curvas de semejante forma, pero con un grupo con mayor dispersión. Estas curvas están dadas por

$$X_{1,j}(t) \equiv U_{1,j} \quad \text{y} \quad X_{2,j}(t) \equiv U_{2,j}\,, \tag{3.7}$$

donde $U_{1,j}$ son variables aleatorias independientes con distribución $U(-1/2, 1/2)$ y $U_{2,j} \sim U(-2, 2)$ independientes entre sí.

Lo que sucede en este ejemplo es que todos los elementos de la clase roja tienen una profundidad alta respecto de la clase verde, de manera que las observaciones rojas no centrales para su misma clase serán clasificadas como verdes. Además, si una curva es atípica para la clase verde, lo será aún más para la clase roja, por lo tanto esta también será clasificada como verde. Luego, la diferencia de dispersión provoca un sesgo a la hora de clasificar según este método, ya que clasificaremos todo elemento no central para su respectiva clase como verde. Esto se puede ver en la Figura 3.7 (b) que representa en el eje horizontal

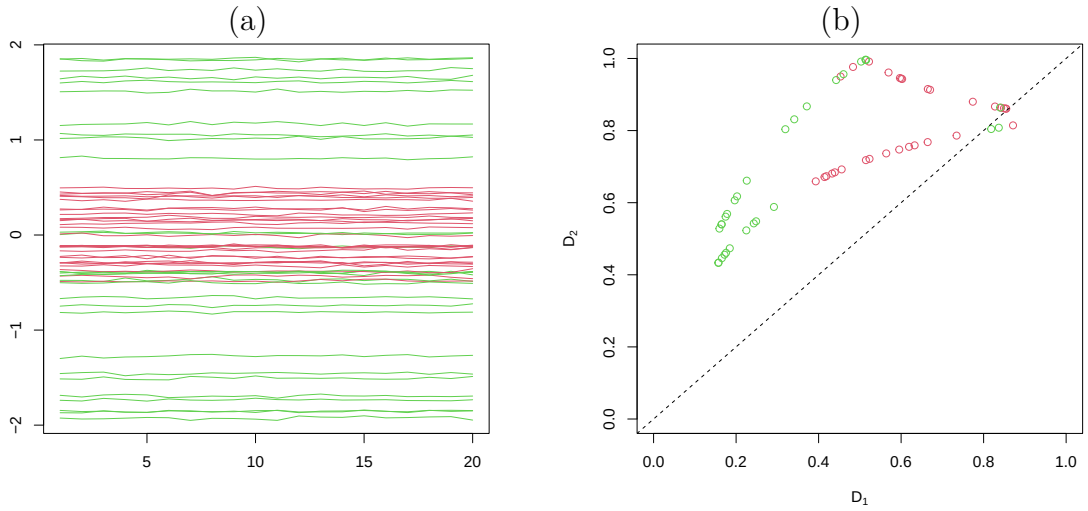


Figura 3.7: Conjunto generado según (3.7). (a) Gráfico de curvas, (b) la profundidad de cada curva según ambas clases, con la frontera de las regiones de clasificación dada por la regla de clasificación de máxima profundidad.

$D_1(X_{i,j}) := D(X_{i,j}; \mathbb{P}_{X_1})$), es decir, la profundidad de $X_{i,j}$ respecto del grupo $g_1$ y en el vertical $D_2(X_{i,j}) := D(X_{i,j}; \mathbb{P}_{X_2})$, o sea, la profundidad de $X_{i,j}$ respecto del grupo $g_2$. Este gráfico se denomina DD-plot. En el mismo la recta $D_1 = D_2$ sirve de frontera de clasificación ya que es natural clasificar a una nueva observación en el grupo donde esta tiene más profundidad. Sin embargo, en el ejemplo podemos observar que toda curva roja tendrá una profundidad alta para la clase verde, que incluso suele ser mayor que la profundidad respecto de su propia clase. Además este problema no lo tendrán las curvas verdes, ya que, excepto la curvas centrales, las curvas verdes serán vistas como atípicas para la clase roja y por ende tendrán una profundidad baja para esta clase.

Para solucionar este problema, Li et al. (2012) proponen no utilizar la recta de identidad en el DD-plot como frontera de las regiones de clasificación, puesto que hay situaciones como en el ejemplo anterior donde esto puede ser muy problemático. Por lo tanto, proponen buscar la función continua y creciente $f : [0, 1] \to \mathbb{R}$ con $f(0) = 0$ que mejor separe los datos. De esta manera, al ser $f$ continua con $f(0) = 0$, quedará el gráfico $DD$ separado en dos regiones siendo el gráfico de la función $f$ la frontera de las dos regiones de clasificación. Para resolver el problema de encontrar dicha función es conveniente elegir una familia de funciones, para luego buscar la función que minimice el error de clasificación utilizando métodos de validación cruzada o bootstrap para evitar sobreajuste. En dicho trabajo se utilizan polinomios crecientes de grado menor a $q$ con término independiente nulo, ya que este conjunto es lo suficientemente flexible para el problema, y además sólo tendremos que buscar $q$ parámetros.

Como alternativa al anterior clasificador, Hubert et al. (2016) proponen un método totalmente no paramétrico que se puede extender fuera de la clasificación binaria. Este método

considera el DD-plot, donde tendremos nuestras observaciones proyectadas a un espacio de dimensión igual a la cantidad de clases de forma $x \to (D(x;\mathbb{P}_{X_1}),\ldots,D(x;\mathbb{P}_{X_M}))^{\mathrm{T}} = (D_1(x),\ldots,D_M(x))^{\mathrm{T}}$. Más precisamente, para cada observación $X_{i,j}$ calculamos las profundidades $D_k(X_{i,j})$, obteniendo el vector de $\mathbb{R}^M$ definido como $\mathbf{D}_{i,j} = (D_1(X_{i,j}),\ldots,D_M(X_{i,j}))^{\mathrm{T}}$. Con estos vectores $\mathbf{D}_{i,j} \in \mathbb{R}^M$ podemos utilizar un método de clasificación multivariado como pueden ser los descriptos en la Sección 3.3. Además, como comentamos anteriormente respecto de las ventajas de usar la atipicidad en lugar de la profundidad para evitar empates, en el trabajo proponen utilizar este método en el gráfico análogo al DD-plot, tomando las atipicidades respecto de cada clase en lugar de las profundidades.

La principal ventaja de esta variante es que, a diferencia de los dos métodos nombrados anteriormente, no tiene problemas en el caso donde no estamos trabajando con clases unimodales. Para ejemplificar consideremos el siguiente ejemplo. Sean

$$X_{1,j}(t) \equiv \begin{cases} U_{1,j}^{(1)} & \text{si } Z_{1,j} = 0 \\ U_{1,j}^{(2)} & \text{en caso contrario} \end{cases} \tag{3.8}$$

y

$$X_{2,j}(t) \equiv \begin{cases} U_{2,j}^{(1)} & \text{si } Z_{2,j} = 0 \\ U_{2,j}^{(2)} & \text{en caso contrario} \end{cases} \tag{3.9}$$

donde $U_{1,j}^{(1)} \sim U(-1,-1/2)$ independientes, $U_{1,j}^{(2)} \sim U(1/2,1)$ independientes, $Z_{1,j} \sim \mathrm{Bi}(1,1/2)$ independientes entre sí y $U_{2,j}^{(1)} \sim U(-3,-1)$ independientes, $U_{2,j}^{(2)} \sim U(-1/2,1/2)$ independientes, $Z_{2,j} \sim \mathrm{Bi}(1,3/10)$ independientes entre sí. A partir de dichas variables aleatorias obtenemos los elementos aleatorios $X_{1,j}$ y $X_{2,j}$.

En la Figura 3.8 (a) se presentan los datos generados con $1 \leq j \leq 50$ para cada elemento aleatorio. En este ejemplo de datos generados, podemos ver claramente que tanto el borde de decisión dado por la función de identidad, como la frontera de decisión dada por el gráfico de cualquier función continua y creciente $f : [0,1] \to \mathbb{R}$ con $f(0) = 0$, tendrán problemas al clasificar a partir del DD-plot. Esto sucede por la existencia del conjunto de curvas no centrales que tiene la clase verde. Ninguno de los métodos anteriores tiene la flexibilidad suficiente para incorporarlo adecuadamente al método de clasificación. Por otra parte, podemos utilizar el método propuesto por Hubert et al. (2016), donde usamos métodos no paramétricos de clasificación como los presentados en la Sección 3.3.3 sobre las nuevas observaciones $\mathbf{D}_{i,j} \in \mathbb{R}^M$ que se obtienen gracias al DD-plot. Este último método no tendría dificultades en afrontar problemas de clasificación donde aparezcan distribuciones multimodales. Presentamos en la Figura 3.8 (b) la frontera de la regla de clasificación utilizando vecinos más cercanos.

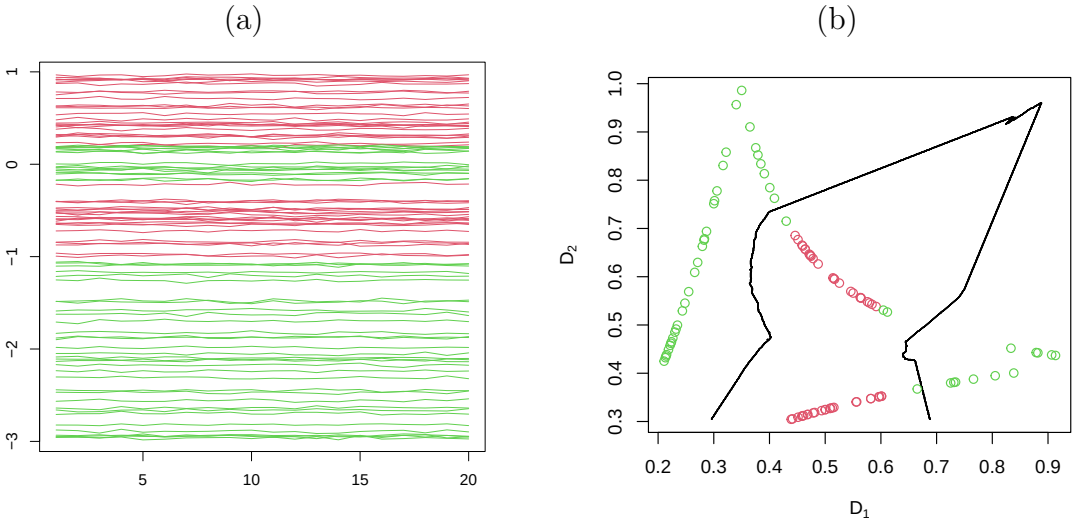


Figura 3.8: Conjunto de datos generados de acuerdo a (3.8) y (3.9). (a) Gráfico de las curvas obtenidas. (b) Profundidad de cada curva en cada población con la frontera de las regiones de clasificación (en negro) según el método de $k$ vecinos más cercanos con $k = 9$.

# Capítulo 4

# Profundidades y atipicidades

## 4.1 Introducción

En este capítulo presentaremos nociones de profundidad y atipicidad, y mostraremos varios ejemplos de ellas. Estas nociones serán de suma importancia, ya que a partir de ellas es posible construir una familia de clasificadores, como explicamos en la Sección 3.4.3.

El objetivo principal que motivó a Tukey (1975) cuando desarrolló la primera profundidad fue la de tener una herramienta de visualización de datos para cuando se trabaja en dimensiones mayores a 3. De esta manera surgió una forma de reducción de la dimensión que actualmente se utiliza no sólo en ese área, sino dentro de la estadística no paramétrica como método para detectar valores atípicos o dentro de la clasificación supervisada y no supervisada.

La profundidad surge como una forma de medir la centralidad de un punto respecto de una distribución. Parte de la idea de que, a mayor profundidad, más cerca se está del centro de dicha distribución, donde cómo se mide la cercanía y cuál es el centro serán dos cuestiones que dependerán de cada función de profundidad. A partir de esto, lograremos un orden de adentro hacia afuera para nuestros puntos dado por la profundidad de cada uno de estos, siendo nuestro centro el valor que maximiza la profundidad. Una propiedad deseable en la profundidad es que, a medida que nos alejamos de dicho centro, la profundidad se vaya tornando nula.

Comencemos con las propiedades que debería tener una profundidad en el caso multivariado. Sea $\mathcal{P}$ el espacio de las probabilidades sobre $\mathbb{R}^p$. Sea $D : \mathbb{R}^p \times \mathcal{P} \to \mathbb{R}$ una función de profundidad. Está claro que, si buscamos obtener un orden a partir de la profundidad, necesitaremos que la imagen de $D$ esté contenida en los reales positivos. Además, necesitamos que tenga al menos un valor donde se maximice la profundidad. Por lo tanto la imagen de $D$ va a tener que estar contenida en un intervalo acotado $[0, a]$, que usualmente se toma con $a = 1$.

A partir de lo previo, en el trabajo Serfling y Zuo (2000) se proponen cuatro condiciones para que una función sea una profundidad en el caso multivariado. Diremos que $D : \mathbb{R}^p \times \mathcal{P} \to [0, 1]$ es una profundidad si cumple:

1. *Invarianza por transformaciones afines.* La profundidad no debería depender del sistema de coordenadas, como tampoco de la escala de dichas coordenadas. Si buscamos que la profundidad muestre un orden entre los puntos respecto de alguna distribución, cualquier transformación afín debería preservar dicho orden. Es decir, si $\mathbb{P}_{\mathbf{x}}$ indica la medida de probabilidad asociada al vector aleatorio $\mathbf{x}$,

$$D(\mathbf{Au} + \mathbf{b}; \mathbb{P}_{\mathbf{Ax}+\mathbf{b}}) = D(\mathbf{u}; \mathbb{P}_{\mathbf{x}})$$

donde $\mathbf{A} \in \mathbb{R}^{p\times p}$ es una matriz inversible y $\mathbf{b}$ es un vector en $\mathbb{R}^p$.

2. *Maximalidad en el centro.* Si una distribución tiene un centro respecto de algún tipo de simetría, este debería ser el punto más profundo. Por ejemplo, si existe $\boldsymbol{\theta} \in \mathbb{R}^p$ tal que $\mathbf{x}$ cumple que $\mathbf{x} - \boldsymbol{\theta}$ tiene igual distribución que $-(\mathbf{x} - \boldsymbol{\theta})$, entonces $\boldsymbol{\theta}$ es el centro y debe tener profundidad máxima. Luego, si $\mathbb{P}_{\mathbf{x}}$ es una ley de probabilidad con un único centro de simetría $\boldsymbol{\theta}$ en $\mathbb{R}^p$, entonces $D(\boldsymbol{\theta}; \mathbb{P}_{\mathbf{x}}) = \sup_{\mathbf{u}\in\mathbb{R}^p} D(\mathbf{u}; \mathbb{P}_{\mathbf{x}})$.

3. *Monotonía respecto del punto más profundo.* A partir del punto más profundo $\boldsymbol{\theta}$, la profundidad debería descender de manera monótona a medida que el punto al que se le calcula la profundidad se aleja. Dicho de otra forma, la profundidad debe ser monótona decreciente sobre cualquier semi-recta que parta del punto más profundo $\boldsymbol{\theta}$. Por lo tanto, $D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) \leq D(\boldsymbol{\theta} + t(\mathbf{u} - \boldsymbol{\theta}); \mathbb{P}_{\mathbf{x}})$ para todo $t \in [0, 1]$ y $\mathbf{u} \in \mathbb{R}^p$.

4. *Desvanecimiento en el infinito.* La profundidad debería tender a cero a medida que nos alejamos del punto más profundo, que se puede asociar al hecho de que las distribuciones tienden a cero a medida que se tiende al infinito. Es decir que $D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) \to 0$ cuando $\|\mathbf{u}\| \to \infty$.

Como agregado a esta definición, en su trabajo separan las profundidades en cuatro categorías distintas a partir de su construcción. De esta manera, podremos clasificar cada profundidad dentro de las cuatro categorías presentadas. A lo largo del capítulo iremos mostrando diversos ejemplos de profundidades tanto en el caso univariado y multivariado, como el caso de nuestro interés que es el de los datos funcionales, que corresponden a esta clasificación.

1. *Profundidades de tipo A.* Sean $\mathbf{x}_1, \ldots, \mathbf{x}_r$ vectores aleatorios independientes e idénticamente distribuidos con medida de probabilidad común $\mathbb{P}_{\mathbf{x}}$. Dados $\mathbf{u}$, $\mathbf{u}_i \in \mathbb{R}^p$, $1 \leq i \leq r$, sea $h(\mathbf{u}; \mathbf{u}_1, \ldots, \mathbf{u}_r)$ una función acotada y positiva que mida de alguna manera la cercanía entre $\mathbf{u}$ y los puntos $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$. Luego, se define la profundidad de tipo A a partir de $h(\mathbf{u}; \mathbf{u}_1, \ldots, \mathbf{u}_r)$ como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = \mathbb{E}h(\mathbf{u}; \mathbf{x}_1, \ldots, \mathbf{x}_r)\,.$$

De esta forma, estamos observando la cercanía esperada entre $\mathbf{u}$ y $r$ vectores aleatorios $\mathbf{x}_1, \ldots, \mathbf{x}_r$ independientes e idénticamente distribuidos con ley de probabilidad $\mathbb{P}_{\mathbf{x}}$.

2. *Profundidades de tipo B.* Sea ahora $h(\mathbf{u}; \mathbf{u}_1, \ldots, \mathbf{u}_r)$ una función no acotada y positiva que mida de alguna manera la distancia entre $\mathbf{u}$ y los puntos $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$. De manera análoga a la profundidad de tipo A, se define la profundidad de tipo B a partir de $h(\mathbf{u}; \mathbf{u}_1, \ldots, \mathbf{u}_r)$ como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = (1 + \mathbb{E}h(\mathbf{u}; \mathbf{x}_1, \ldots, \mathbf{x}_r))^{-1}.$$

Entonces, ahora estaremos midiendo la distancia esperada entre $\mathbf{u}$ y $r$ vectores aleatorios $\mathbf{x}_1, \ldots, \mathbf{x}_r$ independientes e idénticamente distribuidos con ley de probabilidad $\mathbb{P}_{\mathbf{x}}$ en primer lugar. Luego calcularemos la inversa de $1 + \mathbb{E}h(\mathbf{u}; \mathbf{x}_1, \ldots, \mathbf{x}_r)$, para mantener el objetivo de la profundidad que es medir la centralidad de una observación.

3. *Profundidades de tipo C.* Sea $O(\mathbf{u}; \mathbb{P}_{\mathbf{x}})$ una medida de atipicidad de $\mathbf{u}$ respecto de la ley de probabilidad $\mathbb{P}_{\mathbf{x}}$. Algo natural que se puede observar es que la profundidad se ve directamente relacionada con la atipicidad de la siguiente forma: cuanta mayor profundidad tenga un punto, menor será su atipicidad. De esta forma, a partir de una se podrá construir la otra. Surgiendo la profundidad de tipo C como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = (1 + O(\mathbf{u}; \mathbb{P}_{\mathbf{x}}))^{-1}.$$

Es importante observar que, aunque las profundidades de tipo B y tipo C parecerían tener la misma forma, es conveniente separarlas en dos categorías distintas. Esto se debe a que utilizan conceptos distintos, pues la profundidad de tipo B parte de la distancia esperada de $\mathbf{u}$ a $r$ vectores aleatorios con medida de probabilidad $\mathbb{P}_{\mathbf{x}}$, mientras que la profundidad de tipo C utiliza la atipicidad de $\mathbf{u}$ respecto de $\mathbb{P}_{\mathbf{x}}$.

4. *Profundidades de tipo D.* Sea $\mathcal{C}$ un conjunto de subconjuntos cerrados de $\mathbb{R}^p$. A partir de $\mathcal{C}$ definimos la profundidad de $\mathbf{u}$ respecto de la probabilidad $\mathbb{P}_{\mathbf{x}}$ como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = \inf_{\{C \in \mathcal{C}: \mathbf{u} \in C\}} \mathbb{P}(\mathbf{x} \in C).$$

De esta manera, esta profundidad se puede ver como la menor densidad o probabilidad acumulada por un subconjunto $C \in \mathcal{C}$ con la condición de que $\mathbf{u}$ pertenezca a $C$.

## 4.2 Profundidades para el caso univariado

### 4.2.1 Profundidad del semiespacio univariada

La profundidad del semiespacio definida en Tukey (1975) fue la primera profundidad que se definió con el objetivo particular que tienen las profundidades, es decir, como herramienta de reducción de dimensión y visualización de datos. Sea $X$ una variable aleatoria con probabilidad $\mathbb{P}_X$. Luego, definimos la profundidad del semiespacio para un punto $x \in \mathbb{R}$ como

$$D(x; \mathbb{P}_X) = \min\{\mathbb{P}(x \leq X), \mathbb{P}(x \geq X)\},$$

donde el punto más profundo de $X$ será la mediana, con profundidad igual a $1/2$.

A través de esta definición se puede ver que es un ejemplo de una profundidad de tipo D, siendo el más clásico ejemplo de esta categoría. Esto sucede porque, si pensamos a $\mathcal{C}$ como el conjunto de las semirrectas cerradas reales,

$$\min\{\mathbb{P}(x \leq X), \mathbb{P}(x \geq X)\} = \inf_{\{C \in \mathcal{C}: x \in C\}} \mathbb{P}(X \in C)\,.$$

### 4.2.2 Profundidad simplicial univariada

La profundidad simplicial univariada definida por Liu (1990) es una de las profundidades clásicamente usadas junto con la profundidad del semiespacio. Dado un punto $x \in \mathbb{R}$ y una variable aleatoria $X$ con probabilidad $\mathbb{P}_X$, se define la profundidad simplicial univariada como

$$D(x; \mathbb{P}_X) = 2\mathbb{P}(x \leq X)\mathbb{P}(x \geq X)\,.$$

La intuición detrás de esta fórmula es que la profundidad de $x$ será la probabilidad de que este valor pertenezca al intervalo con extremos $X_1$ y $X_2$, con $X_1$ y $X_2$ dos variables aleatorias independientes con la misma distribución que $X$. A partir de esto, si $S[X_1, X_2]$ es el intervalo con extremos $X_1$ y $X_2$, podemos observar que si dichas variables aleatorias son continuas, entonces

$$\mathbb{P}(x \in S[X_1, X_2]) = 2\mathbb{P}(x \leq X)\mathbb{P}(x \geq X)\,. \tag{4.1}$$

Efectivamente, notemos que

$$\begin{aligned}
\mathbb{P}(x \in S[X_1, X_2]) &= \mathbb{P}(\min(X_1, X_2) \leq x \leq \max(X_1, X_2)) \\
&= \mathbb{P}\{(x \leq \max(X_1, X_2)) \cap (\min(X_1, X_2) \leq x)\} \\
&= \mathbb{P}(x \leq \max(X_1, X_2)) - \mathbb{P}(x \leq \min(X_1, X_2))\,.
\end{aligned}$$

Como $X_1$ y $X_2$ son independientes, tenemos que

$$\mathbb{P}(x \leq \max(X_1, X_2)) = 1 - \mathbb{P}(\max(X_1, X_2) \leq x) = 1 - [\mathbb{P}(X \leq x)]^2\,,$$

donde usamos que $\mathbb{P}(x \leq \max(X_1, X_2)) = \mathbb{P}(x \leq X_1,\, x \leq X_2))$. De igual forma, usando que $\mathbb{P}(x \leq \min(X_1, X_2) = \mathbb{P}(x \leq X_1,\, x \leq X_2)$ y la independencia, obtenemos que

$$\mathbb{P}(x \leq \min(X_1, X_2)) = [\mathbb{P}(x \leq X)]^2 = [1 - \mathbb{P}(X \leq x)]^2\,.$$

De las igualdades anteriores se deduce que

$$\begin{aligned}
\mathbb{P}(x \leq \max(X_1, X_2)) - \mathbb{P}(x \leq \min(X_1, X_2)) &= 1 - [\mathbb{P}(X \leq x)]^2 - [1 - \mathbb{P}(X \leq x)]^2 \\
&= 1 - [\mathbb{P}(X \leq x)]^2 - 1 + 2\mathbb{P}(X \leq x) - [\mathbb{P}(X \leq x)]^2 \\
&= 2\mathbb{P}(X \leq x)(1 - \mathbb{P}(X \leq x)) = 2\mathbb{P}(x \leq X)\mathbb{P}(x \geq X)\,,
\end{aligned}$$

lo que concluye la demostración de (4.1).

Notemos que la mediana será el punto más profundo según esta profundidad, y $D(\text{MED}(X); \mathbb{P}_X) = 1/2$. Además, esta profundidad es un ejemplo de una profundidad de tipo A, tomando $h(x; x_1, x_2) = \mathbb{1}_{\{x \in S[x_1, x_2]\}}$.

### 4.2.3 Profundidad de Mahalanobis univariada

En Mahalanobis (1936) se presenta la distancia de Mahalanobis con el objetivo de medir la distancia entre un punto y la media de distribución teniendo en cuenta la varianza que esta tiene. De esta manera, dada $X$ una variable aleatoria con media $\mu$ y varianza $\sigma^2$, para un punto $x \in \mathbb{R}$ se define la distancia de Mahalanobis como

$$O(x; \mathbb{P}_X) = \frac{|x - \mu|}{\sigma}\,.$$

Por lo tanto, al estar interesados en profundidades, definimos la profundidad de Mahalanobis como

$$D(x; \mathbb{P}_X) = [1 + O(x; \mathbb{P}_X)]^{-1}\,.$$

Es evidente que el punto más profundo en este caso es la media de la distribución. También se puede observar que este es un típico caso de profundidad de tipo C, ya que podemos entender la distancia de Mahalanobis como una medida de atipicidad.

### 4.2.4 Profundidad basada en posición y dispersión robustas

Esta profundidad está basada en los trabajos de Stahel (1981) y Donoho (1982) donde se define una distancia entre un punto y una distribución de manera análoga a la distancia de Mahalanobis. Se puede pensar esta medida como una versión robusta de la anterior, ya que se define esta nueva noción de atipicidad más robusta como

$$O(x; \mathbb{P}_X) = \frac{|x - \text{MED}(X)|}{\text{MAD}(X)}\,,$$

notando cómo se reemplaza la media por la mediana, y el desvío estándar por la desviación absoluta media.

Luego, definimos la profundidad como en el caso anterior

$$D(x; \mathbb{P}_X) = [1 + O(x; \mathbb{P}_X)]^{-1}\,.$$

De esta manera tenemos una nueva profundidad de tipo C, cuyo punto más profundo es la mediana de $X$.

### 4.2.5 Profundidad con asimetría ajustada

A partir de la profundidad de Mahalanobis univariada, en el trabajo de Brys et al. (2005) se propone una modificación para poder trabajar con distribuciones asimétricas. Para esto comenzaremos viendo una noción previa de asimetría investigada en Brys et al. (2004) llamada medcouple. Esta noción nos servirá para medir de manera robusta cuán asimétrica es una muestra de datos univariados.

Sean $x_1, \ldots, x_n$ observaciones independientes e idénticamente distribuidas de una variable aleatoria $X$. Sin pérdida de generalidad, asumamos que el conjunto $\{x_1, \ldots, x_n\}$ está formado por las observaciones ordenadas, es decir $x_1 = x^{(1)}, \ldots, x_n = x^{(n)}$, donde $x^{(1)}, \ldots x^{(n)}$ son los estadísticos de orden. Sea $x_m$ la mediana de dicha muestra. Definimos el medcouple como

$$MC = \text{MED}_{x_i \leq x_m \leq x_j} h(x_i, x_j),$$

donde el núcleo $h$ está dado por

$$h(x_i, x_j) = \frac{(x_j - x_m) - (x_m - x_i)}{x_j - x_i} \tag{4.2}$$

si $x_i < x_j$.

Para el caso donde $x_i = x_m = x_j$, definimos el núcleo de la siguiente manera. Sean $m_1 < \cdots < m_k$ los índices de las observaciones que son iguales a la mediana. Luego,

$$h(x_{m_i}, x_{m_j}) = \begin{cases} -1 & \text{si } i + j - 1 < k \\ 0 & \text{si } i + j - 1 = k \\ +1 & \text{si } i + j - 1 > k. \end{cases} \tag{4.3}$$

Veamos en detalle qué calcula este núcleo. Primero es importante observar que a causa del denominador en (4.2), el núcleo $h$ tendrá siempre valores entre -1 y 1. También se puede ver que, si $x_i$ y $x_j$ son distintos a la mediana, $h(x_i, x_j)$ es una medida estandarizada que refleja la diferencia entre la distancia de dichos valores a la mediana, donde si $x_i$ está más lejos de la mediana que $x_j$, $h(x_i, x_j)$ tendrá un valor más cercano a -1. Análogamente, si $x_i$ está más cerca de la mediana que $x_j$, $h(x_i, x_j)$ tendrá un valor más cercano a 1. En los casos donde uno, y solo uno, de estos valores es igual a la mediana, $h(x_m, x_j) = 1$, que expresa intuitivamente que $x_j$ está infinitamente alejado de la mediana. De igual manera tenemos el caso $h(x_i, x_m) = -1$. Luego, en estos casos, $h(x_i, x_j)$ tendrá tantos unos como elementos mayores a la mediana haya, multiplicado por la cantidad de elementos iguales a la mediana, y opuestamente tendrá tantos -1 como elementos menores a la mediana haya, multiplicando nuevamente por la cantidad de elementos iguales a la mediana. Luego, una muestra que tenga más valores mayores a la mediana que menores a esta, tendrá una mayor cantidad de 1 que de -1, y en el caso opuesto tendrá una menor cantidad.

Para comprender la definición dada por (4.3), notemos primero que, por la simetría de $h(x_{m_i}, x_{m_j})$, tendremos tanta cantidad de 1 como de -1, ya que $h(x_{m_i}, x_{m_j}) = -h(x_{m_{(k-i+1)}}, x_{m_{(k-j+1)}})$, pues

$$i + j - 1 < k \iff (k - i + 1) + (k - j + 1) - 1 > k\,.$$

Aunque estos valores en principio no cumplan ninguna función, ya que al final calcularemos la mediana y se cancelarán, tienen la función de que el núcleo $h$ esté bien definido, como también buenas propiedades dentro del análisis numérico para calcular eficientemente el

medcouple que se pueden encontrar en el trabajo de Brys et al. (2004). Por otra parte, si $k$ es par, al observar que tendremos $k$ pares $(x_{m_i}, x_{m_j})$ para los cuales $i+j-1=k$, estaremos agregando $k$ ceros. De igual manera, si $k$ es impar, estaremos agregando $k-1$ ceros. Esto provocará que el medcouple esté más cerca de cero si hay más valores iguales a la mediana, cosa que se puede pensar como que la distribución es menos asimétrica, ya que tiene muchas observaciones en su mediana.

A partir de esta definición del medcouple, Brys et al. (2005) proponen la medida de atipicidad con asimetría ajustada como

$$O(x;\mathbb{P}_X) = \begin{cases} \dfrac{x - \text{MED}(X)}{w_2(X) - \text{MED}(X)} & si\, x > \text{MED}(X) \\ \dfrac{\text{MED}(X) - x}{\text{MED}(X) - w_1(X)} & si\, x \leq \text{MED}(X)\,, \end{cases}$$

donde

$$w_1(X) = Q_1(X) - 1.5e^{-4MC(X)}IQR(X) \quad \text{y}$$
$$w_2(X) = Q_3(X) + 1.5e^{3MC(X)}IQR(X)\,,$$

siendo $Q_1$ y $Q_3$ el primer y tercer cuartil de $X$, $IQR(X) = Q_3(X) - Q_1(X)$ la distancia intercuartil de $X$ y $MC(X)$ el medcouple de $X$. Si $MC(X) < 0$, reemplazamos en la fórmula para el cálculo de la atipicidad $(x, X)$ por $(-x, -X)$. La gran diferencia entre esta atipicidad y la atipicidad presentada en la Sección 4.2.4 la encontramos en el denominador, donde en este caso estaremos utilizando una medida de dispersión basada en el boxplot, pero con un ajuste para los casos asimétricos. Además, en el caso donde $MC(X) = 0$, $w_1(X)$ y $w_2(X)$ corresponden a las expresiones para definir los bigotes dentro de un boxplot.

A partir de la noción de atipicidad anterior, como en los casos de la profundidad de Mahalanobis y la de la Sección 4.2.4, se define la profundidad con asimetría ajustada como

$$D(x;\mathbb{P}_X) = [1 + O(x;\mathbb{P}_X)]^{-1}\,,$$

siendo este un nuevo ejemplo de profundidad de tipo C.

## 4.3 Profundidades para el caso multivariado

### 4.3.1 Profundidad del semiespacio

La profundidad del semiespacio que definimos en la sección anterior puede ser extendida al caso multivariado teniendo que necesariamente cambiar el conjunto $\mathcal{C}$, que previamente era el de las semirrectas cerradas reales. Al trabajar en un espacio de dimensión finita mayor a uno tomaremos, en lugar de dichas semirrectas, la familia de los semiespacios cerrados, es decir,

$$\mathcal{C} := \left\{ \{\mathbf{u} : \mathbf{a}^{\mathrm{T}}\mathbf{u} \leq b\} : \mathbf{a} \in \mathbb{R}^p, b \in \mathbb{R} \right\}\,.$$

A partir de esta familia de conjuntos, podremos definir la profundidad del semiespacio como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = \inf_{\{C \in \mathcal{C}: \mathbf{u} \in C\}} \mathbb{P}(\mathbf{x} \in C)\,,$$

obteniendo nuevamente una profundidad de tipo D. Serfling y Zuo (2000) muestran que esta profundidad cumple con las cuatro condiciones que establecieron para que una función sea una profundidad.

Para la versión muestral de la profundidad del semiespacio tendremos que buscar una alternativa al cálculo de $\mathbb{P}(\mathbf{x} \in C) = \mathbb{P}(\mathbf{x} \in \{\mathbf{u} : \mathbf{a}^{\mathrm{T}}\mathbf{u} \leq b\})$. Para esto usaremos, en lugar de dicha probabilidad, la distribución empírica que da origen a la estimación dada por $\#\{i : \mathbf{a}^{\mathrm{T}}\mathbf{x}_i \leq \mathbf{a}^{\mathrm{T}}\mathbf{u}\}/n$. Esto surge de contar, para cada dirección $\mathbf{a}$, la fracción de elementos de la muestra que pertenecen al semiespacio cerrado dado por la dirección $\mathbf{a}$, notando también cómo $b$ es reemplazado por $\mathbf{a}^{\mathrm{T}}\mathbf{u}$, ya que estamos buscando el conjunto que minimiza la probabilidad. De esta manera,

$$\widehat{D}(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = D(\mathbf{u}; \mathbb{P}_n) = \inf_{\mathbf{a} \in \mathbb{R}^p} \frac{\#\{i : \mathbf{a}^{\mathrm{T}}\mathbf{x}_i \leq \mathbf{a}^{\mathrm{T}}\mathbf{u}\}}{n}\,,$$

donde $\mathbb{P}_n$ es la distribución empírica de $\mathbf{x}_1 \ldots \mathbf{x}_n$.

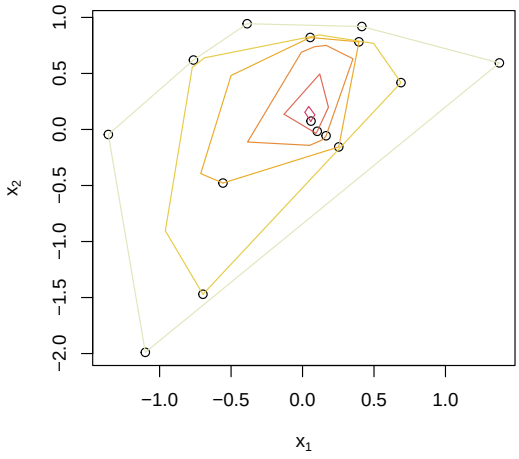


Figura 4.1: Gráfico con la profundidad del semiespacio calculada para un conjunto de puntos. Cada polígono delimita una región con profundidad constante, y a mayor intensidad de color hacia el rojo, mayor es la profundidad de dicha región.

Para entender la intuición detrás de esta profundidad podemos ver la Figura 4.1. En esta figura cada punto es un elemento de la muestra y cada polígono marca el borde de una región con profundidad constante según la profundidad del semiespacio, siendo la región convexa del centro la que alcanza la profundidad máxima y siendo cero la profundidad por fuera del polígono más abarcativo.

### 4.3.2 Profundidad simplicial

La profundidad presentada en el trabajo de Liu (1990) se extiende sin problemas al caso multivariado. Recordando primero que en el caso univariado definimos la profundidad como la probabilidad de que un punto $x$ esté en un intervalo cerrado con extremos dados por observaciones de la variable aleatoria $X$. Para el caso multivariado extenderemos la idea del intervalo con extremos dados por dos observaciones de $\mathbf{x}$ al símplex cerrado, o clausura convexa, dado por $p+1$ observaciones de $\mathbf{x}$, siendo $p$ la dimensión de nuestro espacio. De esta manera, para un punto $\mathbf{u} \in \mathbb{R}^p$ y un vector aleatorio $\mathbf{x}$, definimos la profundidad de $\mathbf{u}$ como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = \mathbb{P}(\mathbf{u} \in S[\mathbf{x}_1 \dots \mathbf{x}_{p+1}]) \,,$$

donde $S[\mathbf{x}_1 \dots \mathbf{x}_{p+1}]$ es el símplex cerrado formado por $p+1$ observaciones $\mathbf{x}_1 \dots \mathbf{x}_{p+1}$ del vector aleatorio $\mathbf{x}$, es decir,

$$S[\mathbf{x}_1 \dots \mathbf{x}_{p+1}] = \left\{ \sum_{i=1}^{p+1} \alpha_i \mathbf{x}_i \,\middle|\, \sum_{i=1}^{p+1} \alpha_i = 1, \alpha_i \geq 0 \right\} .$$

Luego, como en el caso univariado, tendremos una profundidad de tipo A, con $h(\mathbf{u}; \mathbf{x}_1, \dots, \mathbf{x}_{p+1}) = \mathbb{1}_{\{\mathbf{u} \in S[\mathbf{x}_1 \dots \mathbf{x}_{p+1}]\}}$. Por otra parte, Serfling y Zuo (2000) muestran que esta profundidad no cumple con las cuatro condiciones para que sea una función de profundidad. Esto se puede ver intuitivamente en el caso que, si estamos trabajando con una distribución multimodal, en la región de cada moda la profundidad será alta, mientras que por fuera de esta será baja, provocando que no se cumpla la monotonía respecto del punto más profundo.

Como ejemplo de esto veamos el caso univariado de una variable aleatoria $X$ con probabilidades puntuales dadas por

$$\mathbb{P}(X = -2) = \mathbb{P}(X = -1) = \mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \mathbb{P}(X = 2) = \frac{1}{5} \,.$$

Al ser una distribución simétrica, su centro está claro que es el 0, luego la profundidad simplicial debería ser decreciente a medida que nos alejamos de dicho valor. Pero

$$D\left(\frac{1}{2}; \mathbb{P}_X\right) = 2\mathbb{P}\left(X \geq \frac{1}{2}\right) \mathbb{P}\left(X \leq \frac{1}{2}\right) = 2 \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{12}{25} \,,$$

mientras que

$$D(1; \mathbb{P}_X) = 2\mathbb{P}(X \geq 1)\mathbb{P}(X \leq 1) = 2 \cdot \frac{4}{5} \cdot \frac{2}{5} = \frac{16}{25} \,,$$

contradiciendo la propiedad de monotonía respecto del punto más profundo.

En su versión muestral para un nuevo punto $\mathbf{u}$, usaremos las observaciones $\mathbf{x}_1 \dots \mathbf{x}_n$, con $n > p$, para calcular la fracción de símplex a los que $\mathbf{u}$ pertenece

$$\widehat{D}(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = D(\mathbf{u}; \mathbb{P}_n) = \binom{n}{p+1}^{-1} \sum_{\circledast} \mathbb{1}_{\mathbf{u} \in S[\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p+1}}]} \,,$$

donde $\circledast$ se mueve por todos los subconjuntos de $p+1$ elementos sin reposición de la muestra $\mathbf{x}_1 \dots \mathbf{x}_n$. De esta manera un punto con mayor profundidad será un punto que se encuentra en mayor cantidad de símplex, mostrando que intuitivamente se encuentra más inmerso en los datos.

### 4.3.3 Profundidad de Mahalanobis

La profundidad de Mahalanobis se puede extender sin problemas al caso muestral manteniendo su idea inicial. Esta es que, dado un vector aleatorio $\mathbf{x}$ con media $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$, se mire la distancia entre un punto $\mathbf{u}$ y la media $\boldsymbol{\mu}$ teniendo en cuenta la variabilidad de la distribución en esa dirección. Esto se asemeja al caso univariado donde se divide por la desviación estándar, aunque aquí tendremos que utilizar la matriz de covarianza. Luego, en direcciones donde la variabilidad del vector aleatorio sea alta, la distancia será dividida por un mayor valor que en direcciones donde la variabilidad sea baja. De esta manera, definimos la distancia de Mahalanobis como

$$O(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = \sqrt{(\mathbf{u}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{u}-\boldsymbol{\mu})}\,.$$

Luego, la profundidad de Mahalanobis se define como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = [1 + O(\mathbf{u}; \mathbb{P}_{\mathbf{x}})]^{-1}\,.$$

Para el caso muestral no tendremos inconvenientes ya que simplemente se puede reemplazar a $\boldsymbol{\mu}$ y a $\boldsymbol{\Sigma}$ por estimaciones hechas utilizando la muestra $\mathbf{x}_1, \dots, \mathbf{x}_n$. Estos estimadores pueden obtenerse como

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}} \quad \text{y} \quad \mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} (\mathbf{x}_i - \overline{\mathbf{x}})\,.$$

Entonces, para el caso muestral la distancia de Mahalanobis se define como

$$\widehat{O}(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = O(\mathbf{u}; \mathbb{P}_n) = \sqrt{(\mathbf{u}-\widehat{\boldsymbol{\mu}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{u}-\widehat{\boldsymbol{\mu}})}\,,$$

quedando la profundidad de Mahalanobis muestral como

$$\widehat{D}(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = D(\mathbf{u}; \mathbb{P}_n) = [1 + \widehat{O}(\mathbf{u}; \mathbb{P}_{\mathbf{x}})]^{-1}\,.$$

Teniendo en cuenta la sensibilidad de la media $\overline{\mathbf{x}}$ y varianza muestral $\mathbf{S}$ a observaciones atípicas, y con la finalidad de poder detectar dichos datos, se pueden utilizar en lugar de $\overline{\mathbf{x}}$ y $\mathbf{S}$ estimadores robustos de posición y dispersión como los S-estimadores, por ejemplo, que se pueden ver en Maronna et al. (2019).

### 4.3.4 Profundidad de proyección

La profundidad de proyección busca extender al caso multivariado las nociones de distancia o atipicidad de las Secciones 4.2.4, 4.2.5 y 4.2.3 utilizando distintas direcciones donde detectar la atipicidad de un dato. Para esto, notemos con $O^{(1)}(x; \mathbb{P}_X)$ la noción de distancia o atipicidad definida para el caso univariado. Luego, dada una dirección $\mathbf{a} \in \mathbb{R}^p$, con $\|\mathbf{a}\| = 1$, calcularemos la atipicidad de nuestra observación $\mathbf{u}$ al proyectarla en la dirección $\mathbf{a}$. Es decir que podremos calcular $O^{(1)}(\mathbf{a}^{\mathrm{T}}\mathbf{u}; \mathbb{P}_{\mathbf{a}^{\mathrm{T}}\mathbf{x}})$. A partir de esto, podremos definir para el caso multivariado la atipicidad dada por

$$O(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = \sup_{\|\mathbf{a}\|=1} O^{(1)}(\mathbf{a}^{\mathrm{T}}\mathbf{u}; \mathbb{P}_{\mathbf{a}^{\mathrm{T}}\mathbf{x}}) \,.$$

Observemos que esta definición busca la peor dirección para obtener la atipicidad de $\mathbf{u}$ como el valor de la atipicidad univariada de la proyección $\mathbf{a}^{\mathrm{T}}\mathbf{u}$ que maximiza $O^{(1)}(\mathbf{a}^{\mathrm{T}}\mathbf{u}; \mathbb{P}_{\mathbf{a}^{\mathrm{T}}\mathbf{x}})$. La profundidad de proyección asociada se define como

$$D(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = [1 + O(\mathbf{u}; \mathbb{P}_{\mathbf{x}})]^{-1} \,.$$

El problema de esta atipicidad es que la cantidad de direcciones es infinita, por eso, para el caso muestral, se aproxima al $\sup_{\|\mathbf{a}\|=1} O^{(1)}(\mathbf{a}^{\mathrm{T}}\mathbf{u}; \mathbb{P}_{\mathbf{a}^{\mathrm{T}}\mathbf{x}})$ tomando el máximo sobre un número acotado de direcciones ortonormales $\mathbf{a}_i$, con $\|\mathbf{a}_i\| = 1$, $1 \leq i \leq k$ adecuadamente elegidas, por ejemplo a partir del proceso de ortonormalización de Gram–Schmidt sobre una muestra de vectores generada a partir de un vector aleatorio $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. Por lo tanto, calcularemos $O_{\mathbf{a}_i} := \widehat{O}^{(1)}(\mathbf{a}_i^{\mathrm{T}}\mathbf{u}; \mathbb{P}_{\mathbf{a}_i^{\mathrm{T}}\mathbf{x}})$ para cada $1 \leq i \leq k$, quedándonos con el valor máximo de dichas atipicidades. De esta forma, se define el valor de la atipicidad muestral como

$$\widehat{O}(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = \max_{1 \leq i \leq k} O_{\mathbf{a}_i} \,,$$

y la profundidad muestral de proyección asociada

$$\widehat{D}(\mathbf{u}; \mathbb{P}_{\mathbf{x}}) = [1 + \widehat{O}(\mathbf{u}; \mathbb{P}_{\mathbf{x}})]^{-1} \,.$$

## 4.4 Profundidades para el caso funcional

Para el caso funcional tendremos que separarnos del caso multivariado por diversas razones. Para comenzar, recordando las condiciones para que una función sea una profundidad, ya vemos que nos encontramos con un problema al pedir que sea invariante por transformaciones afines. En el caso funcional, un cambio de coordenadas entre $t_i$ y $t_j$, con $t_i \neq t_j$, puede provocar un cambio de regularidad tanto dentro de nuestros elementos aleatorios como de nuestras observaciones, cosa que puede alterar el análisis si por ejemplo estamos trabajando con funciones derivables. Por esta razón, Nieto-Reyes y Battey (2016) así como Gijbels y

Nagy (2017), proponen nuevas condiciones desarrolladas específicamente para el caso funcional, aunque no entraremos en detalle al ser un área en desarrollo que todavía no tiene consenso, ya que ninguna propuesta logra solucionar de manera efectiva el problema.

Por otro lado, tendremos dificultades para extender al caso funcional varias de las nociones de profundidad previamente explicadas tanto por motivos teóricos como prácticos. Dentro de las dificultades teóricas, algo que se presenta es la imposibilidad de extender ciertas nociones que funcionan en el caso multivariado, por ejemplo, en el caso de la profundidad simplicial no podremos tomar símplex de medida positiva generados por $p+1$ puntos, con $p$ la dimensión de nuestro espacio, o en la profundidad de Mahalanobis no podremos calcular la inversa del operador de covarianza por ser este compacto. Por otro lado, tendremos problemas numéricos al trabajar en espacios de dimensión infinita, ya que pueden haber profundidades que, aunque puedan ser extendidas al caso funcional, como la profundidad del semiespacio o las profundidades de proyección, se vuelven inviables para el caso muestral por no ser computables, ya que se debe resolver un problema de optimización tan costoso que se vuelve imposible.

### 4.4.1 Método de la integral

La profundidad basada en el método de la integral o profundidad integrada fue definida en Fraiman y Muñiz (2001). Esta noción surge como una forma de calcular la profundidad media de una observación $x$ respecto de un elemento aleatorio $X$ a lo largo del intervalo $\mathcal{I}$ sobre el cual están definidos. Esto se puede definir a través de la integral

$$D(x;\mathbb{P}_X) = \int_{\mathcal{I}} D^{(1)}(x(t);\mathbb{P}_{X(t)})dt\,,$$

siendo $D^{(1)}(x(t);\mathbb{P}_{X(t)})$ una profundidad univariada, como cualquiera vista en la Sección 4.2.

Para el caso muestral no tendremos problemas en extender esta profundidad, ya que podemos tomar, para cada $t \in \mathcal{I}$, la versión muestral de la profundidad univariada $\widehat{D}^{(1)}(x(t);\mathbb{P}_{X(t)}) = D^{(1)}(x(t);\mathbb{P}_{n,X(t)})$ y calcular la profundidad de la observación $x$ a través de

$$\widehat{D}(x;\mathbb{P}_X) = \int_{\mathcal{I}} \widehat{D}^{(1)}(x(t);\mathbb{P}_{X(t)})dt$$

de ser posible el cálculo de la integral. En el caso donde sólo tengamos muestras sobre una grilla $t_1 < \cdots < t_p$, simplemente calcularemos

$$\widehat{D}(x;\mathbb{P}_X) = \frac{1}{p}\sum_{i=1}^{p} \widehat{D}^{(1)}(x(t_i);\mathbb{P}_{X(t_i)})\,.$$

### 4.4.2 Método de proyección aleatoria

El método de proyección aleatoria surge motivado por la búsqueda de extender la profundidad del semiespacio al caso funcional. Como mencionamos en el comienzo de la sección, esta profundidad no tiene problemas teóricos para dar el salto al caso funcional, ya que seguimos trabajando con un espacio de Hilbert donde contaremos con un producto interno. Luego, podremos definir nuevamente el conjunto

$$\mathcal{C} := \{\{x : \langle a, x\rangle \leq b\} : a \in \mathbb{H}, b \in \mathbb{R}\} \ ,$$

de manera tal que tendremos la profundidad del semiespacio como

$$D(x; \mathbb{P}_X) = \inf_{\{C \in \mathcal{C}: x \in C\}} \mathbb{P}(X \in C) \, .$$

Pero para el caso muestral tendremos serios inconvenientes a la hora de calcular la versión muestral dada por

$$\widehat{D}(x; \mathbb{P}_X) = \inf_{a \in \mathbb{H}} \frac{\#\{i : \langle a, X_i\rangle \leq \langle a, x\rangle\}}{n} \, ,$$

ya que este resulta ser un problema de optimización imposible por trabajar en espacios de dimensión infinita.

Para el caso funcional Cuesta-Albertos et al. (2006) propusieron un método para resolver este problema. La propuesta de estos autores busca calcular una nueva profundidad basándonos en una profundidad univariada, que en principio es la de semiespacios, pero que se puede extender sin problemas a cualquier otra. Para esto se elige un valor $k$ y se calculan al azar $k$ direcciones $a_i \in \mathbb{H}$, con $\|a_i\| = 1$, $1 \leq i \leq k$. Luego, para nuestra nueva observación $x$, se calcula su profundidad como el promedio de las profundidades univariadas de la proyección de $x$ sobre las direcciones $\{a_i\}_{i=1}^k$. Más precisamente, se calcula primero $\widehat{D}^{(1)}(\langle a_i, x\rangle; \mathbb{P}_{\langle a_i, X\rangle})$, con $1 \leq i \leq k$, donde $\widehat{D}^{(1)}$ es la profundidad del semiespacio univariada, recordando que en el caso univariado esta es de rápido cálculo, y luego se promedian estos valores.

Además, basados en este método de cálculo de profundidades para el caso funcional, Cuesta-Albertos et al. (2006) consideran un segundo procedimiento de proyección aleatoria diseñado específicamente para funciones diferenciables, ya que se busca extraer información de la derivada. De esta manera, dadas $k$ direcciones aleatorias $a_i \in \mathbb{H}$, se calcula en primer lugar, la profundidad bivariada de la proyección de $x$ en $\mathbb{R}^2$ dada por $x \to (\langle a_i, x\rangle, \langle a_i, x'\rangle)^{\mathrm{T}}$, donde $x'$ indica la derivada de $x$. Luego, se computa la profundidad como el promedio de $\left\{\widehat{D}^{(2)}\left((\langle a_i, x\rangle, \langle a_i, x'\rangle)^{\mathrm{T}}; \mathbb{P}_{(\langle a_i, X\rangle, \langle a_i, X'\rangle)^{\mathrm{T}}}\right)\right\}_{i=1}^k$, donde $\widehat{D}^{(2)}$ es una profundidad bivariada.

Como alternativa a los anteriores métodos, también se puede buscar aproximar el valor $\inf_{a \in \mathbb{H}}\{\#\{i : \langle a, X_i\rangle \leq \langle a, x\rangle\}/n\}$ de la siguiente manera. Primero, tomamos $k$ direcciones aleatorias ortonormales $a_i \in \mathbb{H}$, y luego, para cada una de ellas, calculamos $\widehat{D}_i = \#\{j : \langle a_i, X_j\rangle \leq \langle a_i, x\rangle\}/n$. Finalmente, tomamos el mínimo sobre los distintos valores $\widehat{D}_i$, que

será el valor más cercano al ínfimo, dando lugar a una aproximación de la profundidad del semiespacio para el caso funcional.

De igual manera se puede utilizar dicho método para el caso de estar trabajando con profundidades de tipo C, como por ejemplo la nociones de distancia o atipicidad de las Secciones 4.2.3, 4.2.4 y 4.2.5 . Primero, definimos la atipicidad en el caso funcional dada por

$$O(x;\mathbb{P}_X) = \sup_{\|a\|=1} O^{(1)}(\langle a,x\rangle;\mathbb{P}_{\langle a,X\rangle})\,,$$

donde $O^{(1)}(x;\mathbb{P}_X)$ es alguna atipicidad univariada como, por ejemplo, una dada en la Sección 4.2.4, con lo cual

$$O^{(1)}(\langle a,x\rangle;\mathbb{P}_{\langle a,X\rangle}) = \frac{|\langle a,x\rangle - \text{MED}(\langle a,X\rangle)|}{\text{MAD}(\langle a,X\rangle)}\,.$$

Luego, aproximamos $\sup_{\|a\|=1} O^{(1)}(\langle a,x\rangle;\mathbb{P}_{\langle a,X\rangle})$ como en el caso anterior, utilizando $k$ direcciones aleatorias $a_1,\ldots,a_k$ ortonormales entre sí, para buscar el máximo de $\{O^{(1)}(\langle a_i,x\rangle;\mathbb{P}_{\langle a_i,X\rangle})\}_{i=1}^k$, que notaremos como $\widehat{O}(x;\mathbb{P}_X)$. Por último, ya obtenida nuestra aproximación de la atipicidad de $x$ respecto de $X$, calcularemos la profundidad como

$$\widehat{D}(x;\mathbb{P}_X) = [1+\widehat{O}(x;\mathbb{P}_X)]^{-1}\,.$$

### 4.4.3 Profundidad h-modal

La profundidad h-modal, definida en Cuevas et al. (2007), puede entenderse como una forma de extender la noción de verosimilitud o densidad del análisis multivariado al caso funcional. Al no tener, en este último caso, una función de densidad para aproximar, vamos a buscar calcular la profundidad de una observación $x$ a partir de cuán rodeada está de otras observaciones. De esta forma, usaremos que estamos trabajando en un espacio métrico para medir la cercanía entre nuestra nueva observación y el elemento aleatorio, para luego procesar esta información con un núcleo $K$ positivo, acotado y decreciente. Por lo tanto, calcularemos la profundidad de $x$ respecto de un elemento aleatorio $X$ como

$$D_h(x;\mathbb{P}_X) = \mathbb{E}[K_h(d(x,X))]\,,$$

donde $K$ es un núcleo y $h$ es el parámetro de suavizado llamado ancho de banda, de forma que $K_h(x) = h^{-1}K(x/h)$. La versión muestral se puede calcular sin complicaciones al trabajar, en lugar de la distancia entre $x$ y $X$, con la distancia entre $x$ y cada observación $X_i$. Luego,

$$\widehat{D}_h(x;\mathbb{P}_X) = \frac{1}{n}\sum_{i=1}^n K_h(d(x,X_i))\,.$$

Para esta profundidad en particular será importante la elección de la métrica, donde podremos utilizar tanto las métricas de $L^p(\mathcal{I})$ como métricas que incluyan también a las derivadas en el caso de ser posible. La elección del ancho de banda también será importante para decidir qué tan local es la mirada de nuestro método, cosa que puede ser crucial al trabajar con pocas observaciones.

### 4.4.4 Profundidad de bandas

La profundidad de bandas, introducida en López-Pintado y Romo (2009), busca medir la profundidad a través de ver cuán sumergida está una observación $x$ entre el conjunto de curvas de la muestra. Esto lo haremos utilizando los gráficos de las curvas para delimitar una región que llamaremos banda. Para esto, comencemos por definir la banda generada por las curvas $x_1, \ldots, x_j$ como

$$V(x_1, \ldots, x_j) = \{(t, y) : t \in \mathcal{I}, \min_{1 \leq i \leq j} x_i(t) \leq y \leq \max_{1 \leq i \leq j} x_i(t)\}.$$

A partir de esta definición, dada una nueva observación $x$ junto con una muestra $\{X_i\}_{i=1}^n$ generada en la misma ley de probabilidad que $X$ y dado $j$ fijo, $2 \leq j \leq n$, buscaremos contar la proporción de bandas generadas por $j$ elementos de la muestra tales que el gráfico de $x$ está contenido en ellas, es decir

$$S_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} \mathbb{1}_{\{G(x) \subset V(X_{i_1}, \ldots, X_{i_j})\}},$$

donde $G(x)$ es el gráfico de $x$. De esta manera, un valor de $S_n^{(j)}(x)$ alto denota que la curva $x$ se encuentra más inmersa gráficamente en los datos, pudiendo asociarse con la profundidad simplicial del caso multivariado. Para calcular la profundidad de $x$ respecto de $X$, tendremos que elegir $J$, para luego calcular cada $S_n^{(j)}(x)$, $2 \leq j \leq J$ utilizando una muestra $\{X_i\}_{i=1}^n$ y por último obtener la profundidad como

$$D_J(x; \mathbb{P}_X) = \sum_{j=2}^{J} S_n^{(j)}(x).$$

A partir de la anterior definición de profundidad, López-Pintado y Romo (2009) proponen una profundidad llamada profundidad de bandas generalizada. En esta versión, en lugar de contar los casos donde la gráfica de $x$ está contenida en la banda, se ponderará la medida del conjunto donde la gráfica de $x$ pertenece a dicha banda. Sea $x$ una nueva observación y sea $\{X_i\}_{i=1}^n$ una muestra, definimos

$$A_n^{(j)}(x) = \{t \in \mathcal{I} : \min_{1 \leq i \leq j} X_i(t) \leq x(t) \leq \max_{1 \leq i \leq j} X_i(t)\}.$$

Luego $|A_n^{(j)}(x)|/|\mathcal{I}|$ nos dará la fracción de "tiempo" sobre la cual el gráfico de $x$ está contenido en la banda generada por la muestra. Como en el caso anterior, definimos ahora la media de estas fracciones para todas las combinaciones de $j$ elementos de nuestra muestra

$$GS_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} \frac{|A_n^{(j)}(x)|}{|\mathcal{I}|}.$$

Por lo tanto, para definir la profundidad generalizada de $x$ respecto de $X$ como antes, debemos elegir $J$, para luego definir la profundidad de $x$ como

$$D_J(x;\mathbb{P}_X) = \sum_{j=2}^{J} GS_n^{(j)}(x)\,.$$

Para la elección del parámetro $J$, López-Pintado y Romo (2009) recomiendan tomar $J = 3$ cuando se considera la profundidad de bandas y $J = 2$ para la profundidad de bandas generalizada.

### 4.4.5 Profundidad basada en la atipicidad direccional

La atipicidad direccional definida en Dai y Genton (2019), busca medir paralelamente dos nociones de atipicidad para una observación $x$. Por un lado, la atipicidad de magnitud, y, por el otro, la atipicidad de forma, algo de suma importancia en el caso funcional pero que no aparece explícitamente en ninguna de las profundidades anteriores. Para esto, primero trabajaremos con una noción de atipicidad univariada que indicaremos por $O^{(1)}$, como puede ser la distancia de Mahalanobis, la atipicidad basada en posición y dispersión robustas o de asimetría ajustada. A partir de esta, definimos la atipicidad direccional que contiene la atipicidad univariada multiplicada por la dirección a donde esta apunta el vector que va desde $Z(t)$ a $x(t)$, es decir

$$O(x(t);\mathbb{P}_{X(t)}) = O^{(1)}(x(t);\mathbb{P}_{X(t)}) \cdot \frac{x(t) - Z(t)}{\|x(t) - Z(t)\|} = O^{(1)}(x(t);\mathbb{P}_{X(t)}) \cdot v(t)\,,$$

con $Z(t)$ el valor de menor atipicidad, o valor más central de la distribución, definiendo en el caso particular en que $x(t) = Z(t)$, $O(x(t);\mathbb{P}_{X(t)}) = 0$. Para ejemplificar, si estamos trabajando con funciones cuyo gráfico está en $\mathbb{R}^2$, la dirección de atipicidad $v(t)$ será 1 si la observación está por encima de la curva más central, 0 si son iguales, y -1 si está por debajo.

Tomando la definición de atipicidad direccional, definimos la atipicidad direccional funcional como

$$FO(x;\mathbb{P}_X) = \int_{\mathcal{I}} \|O(x(t);\mathbb{P}_X(t))\|^2 dt\,.$$

Esta medida de atipicidad representa la atipicidad total de $x$, y la podemos pensar fuertemente asociada a la profundidad integrada de Fraiman y Muñiz (2001) definida anteriormente. La única diferencia en este caso es que incluimos el cuadrado de $O^{(1)}(x(t);\mathbb{P}_{X(t)})$.

También definimos la atipicidad direccional media como

$$MO(x;\mathbb{P}_X) = \int_{\mathcal{I}} O(x(t);\mathbb{P}_{X(t)}) dt\,. \tag{4.4}$$

En este caso podemos pensar que estamos calculando la posición relativa de $x$ respecto de la curva central $Z$ de la distribución $X$. Luego, $\|MO(x;\mathbb{P}_X)\|$ representa la atipicidad

de magnitud, ya que ignora la dirección, que tiene norma 1. Entonces, una observación con un valor alto de norma de atipicidad direccional media se puede pensar como que su propio "centro" está alejado del centro del elemento aleatorio $X$.

Por último, definimos la variación de la atipicidad direccional como

$$VO(x;\mathbb{P}_X) = \int_{\mathcal{I}} \|O(x(t);\mathbb{P}_{X(t)}) - MO(x;\mathbb{P}_X)\|^2 dt\,. \tag{4.5}$$

Esta atipicidad va a medir los cambios, tanto en norma como en dirección, de la atipicidad direccional a lo largo del intervalo. Se puede pensar que estamos estudiando la atipicidad del comportamiento de la observación $x$ en el caso en el que estuviera centrada. Por lo tanto, esta noción se asocia a la atipicidad de forma.

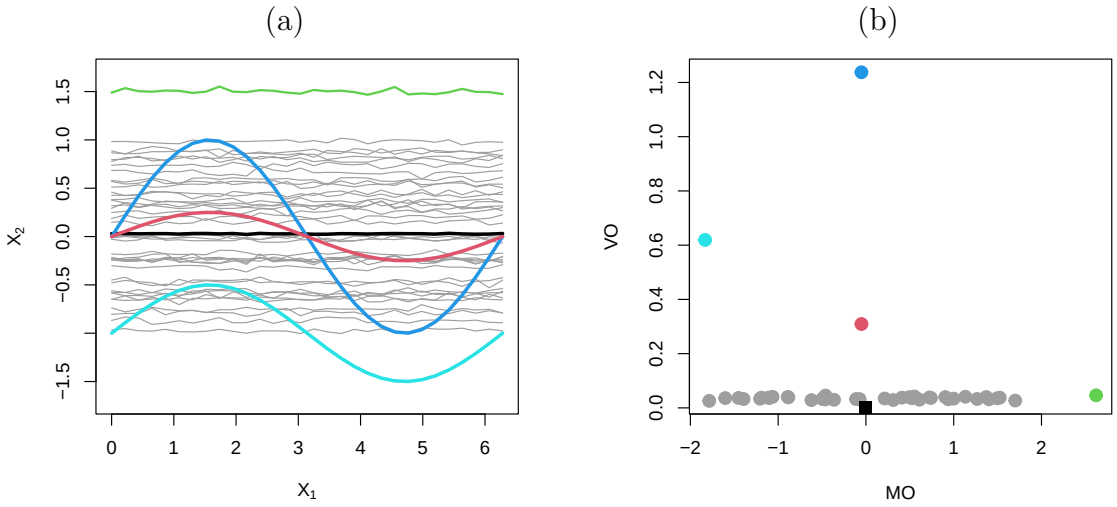


Figura 4.2: Ejemplo del cálculo de la atipicidad direccional para cuatro observaciones nuevas (curvas con color), siendo las curvas grises la muestra y la curva negra su centro: (a) Gráfico de las curvas. (b) La atipicidad direccional media y la variación de la atipicidad direccional para cada curva del gráfico anterior, respetando sus colores.

Para visualizar estos valores, veamos en la Figura 4.2 el comportamiento de estas atipicidades para cuatro observaciones nuevas, indicadas en rojo, azul, celeste y verde, respecto de una muestra de 40 observaciones en gris. Además, se indica en negro la curva central, es decir, la curva con mayor profundidad. Podemos notar como las curvas azul y roja, al estar centradas respecto de nuestra muestra, tendrán un bajo valor de atipicidad direccional media. En cambio, las curvas celeste y verde tendrán alto, en términos del valor absoluto, dicho valor, aunque se puede ver cómo, el tener la atipicidad de magnitud en direcciones opuestas, en el segundo gráfico estas direcciones se ven reflejadas con signos opuestos. Por otra parte, la variación de la atipicidad direccional de la curva verde, al tener una forma típica para la muestra, rondará el valor cero. No así el resto de nuestras nuevas observaciones, que tienen dicho valor más alto al tener una forma más ondulada.

Una de las principales propiedades de estas definiciones de atipicidad direccional es la igualdad que logra relacionar la atipicidad direccional total con las otras dos nociones de atipicidad a través del siguiente teorema.

**Teorema 4.4.1.** Sea $x \in L^2(\mathcal{I})$ y sea $X$ un elemento aleatorio en dicho espacio. Se cumple que

$$FO(x;\mathbb{P}_X) = \|MO(x;\mathbb{P}_X)\|^2 + VO(x;\mathbb{P}_X)\,.$$

*Demostración.* Comencemos desarrollando la atipicidad direccional total

$$\begin{aligned} FO(x;\mathbb{P}_X) &= \int_{\mathcal{I}} \|O(x(t);\mathbb{P}_{X(t)})\|^2 dt \\ &= \int_{\mathcal{I}} \|O(x(t);\mathbb{P}_{X(t)}) - MO(x;\mathbb{P}_X) + MO(x;\mathbb{P}_X)\|^2 dt \\ &= \int_{\mathcal{I}} \|O(x(t);\mathbb{P}_{X(t)}) - MO(x;\mathbb{P}_X)\|^2 dt + \|MO(x;\mathbb{P}_X)\|^2 \\ &= VO(x;\mathbb{P}_X) + \|MO(x;\mathbb{P}_X)\|^2, \end{aligned}$$

donde la tercera igualdad se debe a la ortogonalidad entre dichos valores, ya que

$$\begin{aligned} &\big\langle O(x(t);\mathbb{P}_{X(t)}) - MO(x;\mathbb{P}_X), MO(x;\mathbb{P}_X)\big\rangle \\ &\quad = \left\langle O(x(t);\mathbb{P}_{X(t)}) - \int_{\mathcal{I}} O(x(t);\mathbb{P}_{X(t)})dt, \int_{\mathcal{I}} O(x(t);\mathbb{P}_{X(t)})dt \right\rangle = 0\,, \end{aligned}$$

lo que completa la demostración. □

De esta manera logramos obtener una descomposición de la noción de atipicidad direccional total, análoga al método de la integral, en dos medidas de suma importancia en el caso funcional, la magnitud y la forma.

Finalmente, como en los casos de atipicidad anteriores, podremos definir las profundidades como

$$MD(x;\mathbb{P}_X) = [1 + MO(x;\mathbb{P}_X)]^{-1} \quad \text{y} \quad VD(x;\mathbb{P}_X) = [1 + VO(x;\mathbb{P}_X)]^{-1}\,.$$

### 4.4.6 Profundidad de reconocimiento de forma

La profundidad introducida en Nagy et al. (2017) se basa en la extensión de dos métodos de profundidad para el espacio funcional que permiten analizar la forma de las observaciones. Así como en el caso de la atipicidad direccional descripta en la Sección 4.4.5, las profundidades definidas en dicho trabajo utilizan una noción de profundidad univariada $D^{(1)}$.

A partir de esto, por un lado, utilizamos la noción de profundidad integral presentada en la Sección 4.4.1, es decir, calculamos la profundidad univariada media de la observación $x$ a

lo largo del intervalo $\mathcal{I}$ que indicaremos por,

$$FD(x;\mathbb{P}_X) = \int_{\mathcal{I}} D^{(1)}(x(t);\mathbb{P}_{X(t)})dt\,. \tag{4.6}$$

Por el otro lado, definimos la profundidad infimal donde, en lugar de calcular la profundidad univariada media de $x$, utilizamos el ínfimo de dicha profundidad univariada a lo largo del intervalo $\mathcal{I}$. De esta manera definimos la profundidad infimal como

$$ID(x;\mathbb{P}_X) = \inf_{t\in\mathcal{I}} D^{(1)}(x(t);\mathbb{P}_{X(t)})\,. \tag{4.7}$$

Tomando estas nociones de profundidad para el caso funcional, vamos a buscar una forma de encontrar una atipicidad de $J-$ésimo orden.

**Definición 4.4.1.** Sea $x \in C(\mathcal{I})$, con $C(\mathcal{I})$ el espacio de las funciones continuas sobre $\mathcal{I}$ y sea $X$ un elemento aleatorio en ese mismo espacio. Diremos que $x$ es una observación atípica de primer orden, si existe un $t \in \mathcal{I}$ tal que $x(t)$ es atípico respecto a la probabilidad $\mathbb{P}_{X(t)}$. Para $J \in \mathbb{N}_{>1}$, diremos que $x$ es una observación atípica de $J-$ésimo orden, si existe una colección de puntos $(t_1,\ldots,t_J) \in \mathcal{I}^J$ tales que $(x(t_1),\ldots,x(t_J))^{\mathrm{T}}$ es un valor atípico respecto de la probabilidad conjunta del vector $(x(t_1),\ldots,x(t_J))^{\mathrm{T}}$, que indicaremos por $\mathbb{P}_{(X(t_1),\ldots,X(t_J))}$, pero a su vez no existe una colección de $J-1$ elementos de $\mathcal{I}$ tales que $x$ sea una observación atípica de $(J-1)$-ésimo orden.

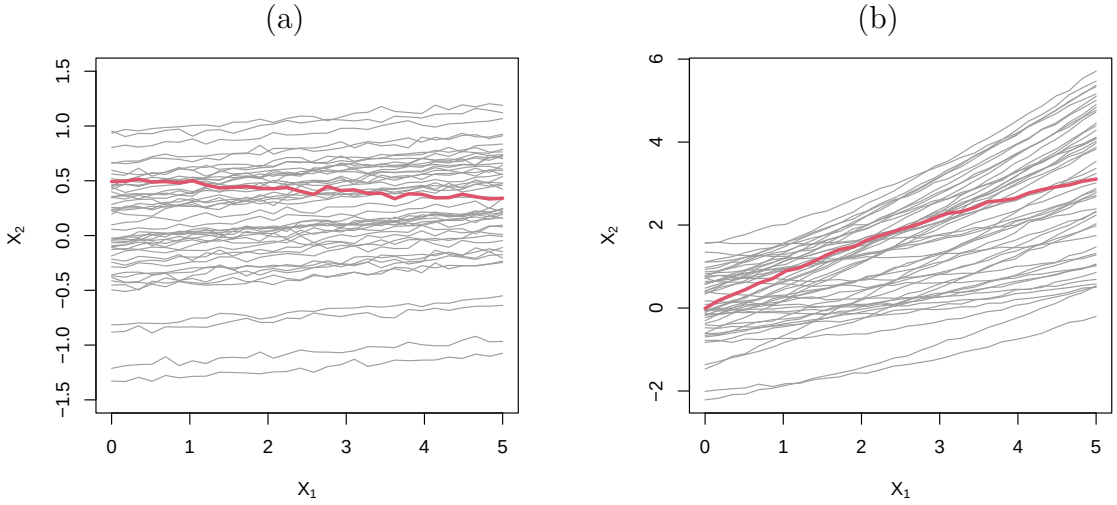


Figura 4.3: Ejemplos de valores atípicos de mayor orden que 1: (a) Observación atípica de segundo orden (en rojo) respecto de la muestra (en gris), (b) Observación atípica de tercer orden (en rojo) respecto de la muestra (en gris).

Para ejemplificar el tipo de problemas que representan estos valores atípicos de $J-$ésimo orden, consideremos la Figura 4.3. En el panel (a) podemos observar en gris una muestra de

rectas crecientes, aunque alteradas por ruido, y en rojo una nueva recta decreciente también afectada por el mismo ruido. Está claro que esta no es una observación atípica de primer orden ya que no existe ningún valor entre 0 y 5 tal que la curva roja sea atípica en dicho valor, mientras que si tomamos dos puntos $t_1$ y $t_2$, con $t_1$ cercano a 0 y $t_2$ cercano a 5, está claro que $(x(t_1), x(t_2))^{\mathrm{T}}$ es atípico, pues es el único caso donde $x(t_1) > x(t_2)$. Por lo tanto, podremos pensar la atipicidad de segundo orden como atipicidad respecto de la derivada o del crecimiento.

En el panel (b) de la Figura 4.3 vemos un caso de atipicidad de orden 3, con una muestra en gris de polinomios de grado 2, con el coeficiente principal positivo positivo y el resto de los coeficientes con signo variado, alterados igual que en la figura anterior por un ruido. En rojo tenemos una observación, también alterada con ruido, que corresponde a un polinomio de grado 2 pero con el coeficiente principal negativo. Este caso es claro que no tiene atipicidad de orden 1, ya que no se destaca en ningún punto del intervalo, y además tampoco es de orden 2, ya que en ningún momento crece de manera excepcional para la muestra. En cambio, si miramos la atipicidad de tercer orden, tomando $t_1$ cercano a 0 y $t_2$ y $t_3$ cercanos a 5, veremos que se destaca al ser la única curva donde $x(t_2)$ y $x(t_3)$ se parecen, pero que estos valores están alejados de $x(t_1)$, pues en la muestra si $X(t_2)$ y $X(t_3)$ están cercanos, entonces $X(t_1)$ también lo estará al ser una función convexa creciente. Luego, podremos pensar la atipicidad de tercer orden como atipicidad respecto de la convexidad o concavidad.

Motivados por el problema de detección de valores atípicos de orden mayor a uno, como las dos profundidades dadas en (4.6) y (4.7) son insuficientes para detectarlos, las extenderemos para que sean capaces de detectar dichas observaciones atípicas. Para el caso donde estemos trabajando con funciones diferenciables hasta el orden $J$ y para poder detectar cuando hay atipicidad de orden $J$, definimos

$$FD^{(J)}(x; \mathbb{P}_X) = \int_{\mathcal{I}} D^{(J)}((x(t), \ldots, x^{(J)}(t))^{\mathrm{T}}; \mathbb{P}_{(X(t),\ldots,X^{(J)}(t))}) dt \,,$$

y

$$ID^{(J)}(x; \mathbb{P}_X) = \inf_{t \in \mathcal{I}} D^{(J)}((x(t), \ldots, x^{(J)}(t))^{\mathrm{T}}; \mathbb{P}_{(X(t),\ldots,X^{(J)}(t))}) \,,$$

donde $D^{(J)}(\mathbf{x}_0, \mathbb{P}_{\mathbf{x}})$ es una profundidad multivariada en $\mathbb{R}^J$. El gran problema de este método es que deja de funcionar en el caso en que las observaciones no pertenezcan a $C^J(\mathcal{I})$, el espacio de funciones continuas sobre $\mathcal{I}$ continuamente diferenciables hasta orden $J$. Para solucionar este problema, Nagy et al. (2017) definen las siguientes nuevas profundidades, utilizando exactamente la definición previa de atipicidad de $J-$ésimo orden,

$$FD_J(x; \mathbb{P}_X) = \int_{\mathcal{I}} \cdots \int_{\mathcal{I}} D^{(J)}((x(t_1), \ldots, x(t_J))^{\mathrm{T}}; \mathbb{P}_{(X(t_1),\ldots,X(t_J))}) dt_1 \ldots dt_J \,,$$

y

$$ID_J(x; \mathbb{P}_X) = \inf_{t_1,\ldots,t_J \in \mathcal{I}} D^{(J)}((x(t_1), \ldots, x(t_J))^{\mathrm{T}}; \mathbb{P}_{(X(t_1),\ldots,X(t_J))}) \,.$$

Con estas definiciones habremos solucionado el problema de necesitar funciones diferenciables, pero es importante observar que, a medida que el parámetro $J$ crece, la complejidad

del método crecerá exponencialmente. Es por esto que es conveniente aproximar dichas profundidades. Por ejemplo, podemos limitar la cantidad de valores $t_i \in \mathcal{I}$ que usaremos para el armado de los posibles valores $(t_1, \ldots, t_J)^{\mathrm{T}}$ que calcularemos. Otra propuesta es tomar un parámetro a elección $S$, y definir $((t_{1,1}), \ldots, x(t_{1,J})), \ldots, ((t_{S,1}), \ldots, x(t_{S,J})) \in \mathcal{I}^J$, donde cada $((t_{m,1}), \ldots, x(t_{m,J}))^{\mathrm{T}}$ es una muestra tomada de manera aleatoria a partir de una distribución uniforme en $\mathcal{I}^J$, luego

$$FD_J^{(S)}(x; \mathbb{P}_X) = \frac{1}{S} \sum_{m=1}^{S} D((x(t_{m,1}), \ldots, x(t_{m,J}))^{\mathrm{T}}; \mathbb{P}_{(X(t_{m,1}),\ldots,X(t_{m,J}))}) \,, \tag{4.8}$$

y

$$ID_J^{(S)}(x; \mathbb{P}_X) = \inf_{m=1,\ldots,S} D((x(t_{m,1}), \ldots, x(t_{m,J}))^{\mathrm{T}}; \mathbb{P}_{(X(t_{m,1}),\ldots,X(t_{m,J}))}) \,.$$

De esta manera, Nagy et al. (2017) definen para el caso funcional la profundidad dada por $FD_J^{(S)}(x; \mathbb{P}_X)$ en (4.8) o la profundidad bidimensional dada por el par $(FD_J^{(S)}(x; \mathbb{P}_X), ID_J^{(S)}(x; \mathbb{P}_X))^{\mathrm{T}}$.

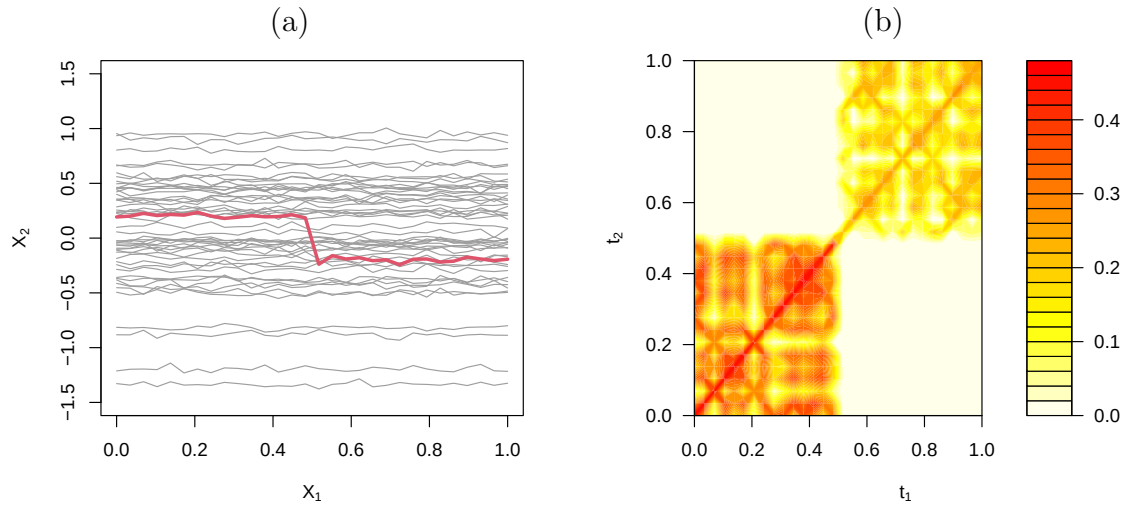


Figura 4.4: Ejemplo de la profundidad de reconocimiento de forma de orden 2: (a) Gráfico de las curvas (gris) con la nueva observación (rojo), (b) Profundidad bivariada $D^{(2)}((x(t_1), x(t_2))^{\mathrm{T}}; \mathbb{P}_{((X(t_1),X(t_2)))})$ de la nueva observación calculada para cada par $t_1$ y $t_2$ del intervalo $[0, 1]$.

La Figura 4.4 ilustra con un ejemplo el uso de la profundidad de segundo orden para una nueva observación. En el panel (a) presentamos en color gris la muestra de tamaño $n = 50$ que corresponde a rectas con distinta ordenada al origen, perturbadas por ruido, junto con una nueva observación en rojo. Vale la pena mencionar que dicha observación puede verse como una función no continua definida por partes, que en la primera mitad tiene un valor constante, y en la segunda mitad otro, siempre perturbada por ruido. En el panel (b), mostramos un gráfico de intensidad para la profundidad bivariada de esta nueva observación

en cada par $(t_1, t_2) \in [0, 1]^2$. Podemos observar que si $t_1$ y $t_2$ son ambos menores o mayores a la mitad del intervalo, entonces el par $(x(t_1), x(t_2))^{\mathrm{T}}$ no es atípico para la muestra, ya que presenta una profundidad mayor. En cambio, si tomamos $t_1$ y $t_2$ en distintas mitades, la profundidad de $(x(t_1), x(t_2))^{\mathrm{T}}$ se vuelve nula, ya que un cambio tan grande entre $x(t_1)$ y $x(t_2)$ es sumamente atípico.

# Capítulo 5

# Método propuesto

## 5.1 Introducción

En este capítulo utilizaremos las herramientas presentadas previamente en la tesis para desarrollar reglas de clasificación específicas para el caso funcional. Primero expliquemos cuál es la motivación para desarrollar un nuevo clasificador: por un lado, tenemos métodos de clasificación para datos funcionales que tienen el problema de no trabajar con la forma de las observaciones, que es el núcleo del análisis de datos funcionales, y lo que separa esta rama del análisis multivariado. Por otro lado, tenemos métodos que sí observan la forma de los datos, pero en su mayoría utilizan scores dados a través la descomposición dada por Karhunen-Loève o por distintos métodos de regularización como pueden ser los splines. Dichos métodos por lo tanto requieren de una noción de regularidad de los datos que no siempre es posible, ya que en el caso de trabajar con datos discontinuos o no diferenciables, la aproximación por splines tendrá un alto error.

Es por esto que para la propuesta de clasificación de esta tesis utilizaremos herramientas que puedan considerar la forma de nuestros datos sin por eso requerir nociones de regularidad. Estas serán dos nociones de profundidad o atipicidad que, además de distinguir la distancia convencional entre funciones, como la métrica de $L^2(\mathcal{I})$, pueden reconocer la forma de los datos funcionales sin requerir su regularidad, siendo estas la atipicidad direccional y la profundidad de reconocimiento de forma presentadas en las Secciones 4.4.5 y 4.4.6. A partir de estos métodos, utilizaremos la herramienta del DD-plot para reducir la dimensionalidad de nuestros datos. De esta forma, lograremos reducir la dimensión del espacio original que es infinito, sin por eso perder la información propia de los datos funcionales como lo es su forma. Finalmente, podremos clasificar nuestros datos representados en el DD-plot a través de distintas reglas de clasificación para el caso multivariado como las vistas en la Sección 3.3.

En la Sección 5.2 se describe nuestra propuesta, mientras que en la Sección 5.3 se presentan los resultados de un estudio de simulación y el análisis de un conjunto de datos

reales.

## 5.2 Propuesta de clasificador para datos funcionales

La propuesta de clasificación de esta tesis, que será estudiada numéricamente en la Sección 5.3, puede describirse como sigue. Sean $X_{i,j} \in L^2(\mathcal{I})$, $1 \leq j \leq n_i$, $1 \leq i \leq M$, observaciones independientes tales que $X_{i,j} \sim \mathbb{P}_i$, donde $\mathbb{P}_i$ es la ley de probabilidad del $i-$ésimo grupo. A la muestra $\{X_{i,j}\}_{\substack{1\leq j\leq n_i \\ 1\leq i\leq M}}$ se la llama muestra de entrenamiento. Sea $x_0$ una nueva observación que deseamos clasificar, es decir, que deseamos asignar a alguna de las $M$ clases.

Dado un grupo $g_\ell$, $1 \leq i \leq M$, indicaremos por $D(x, \mathbb{P}_\ell)$ la profundidad de reconocimiento de forma $FD_J^{(S)}$ presentada en la Sección 4.4.6 o la atipicidad direccional definida en la Sección 4.4.5. En este último caso, $D(x, \mathbb{P}_\ell)$ corresponderá al vector $(MO(x; \mathbb{P}_\ell), VO(x; \mathbb{P}_\ell))^{\mathrm{T}}$ dado en (4.4) y (4.5). En el caso de la profundidad de reconocimiento de forma utilizaremos solamente $FD_J^{(S)}$ y no $ID_J^{(S)}$ pues esta última produce gran cantidad de empates al asignarle profundidad nula a varios elementos de la muestra, lo que sugiere que los errores de clasificación empeorarán en este caso. Entonces, la profundidad $D(x, \mathbb{P}_\ell)$ en un caso será un valor real mientras que en el otro será un vector bidimensional.

El procedimiento de clasificación para una observación $x_0$ puede describirse como sigue:

a) Para $1 \leq \ell \leq M$, calcule la profundidad de cada observación $X_{i,j}$ respecto de la medida de probabilidad $\mathbb{P}_\ell$, o sea,
$$D_{i,j,\ell} = D(X_{i,j}; \mathbb{P}_\ell)^{\mathrm{T}},$$
donde en el caso unidimensional la traspuesta puede omitirse. Defina los vectores
$$\mathbf{D}_{i,j} = (D_{i,j,1}, \ldots, D_{i,j,M})^{\mathrm{T}}.$$

b) Consideremos una regla de clasificación en $\mathbb{R}^p$ con $p = M$ si se utiliza la profundidad de reconocimiento de forma, y $p = 2M$ si se utiliza la atipicidad direccional. Indiquemos por $\{\mathcal{G}_\ell\}_{\ell=1}^M$ las regiones de clasificación obtenidas en $\mathbb{R}^p$ a partir de las pseudo-observaciones $\mathbf{D}_{i,j}$, $1 \leq j \leq n_i$, $1 \leq i \leq M$.

c) Calcule $\mathbf{D}_0 = (D_{0,1}, \ldots, D_{0,M})^{\mathrm{T}}$ donde $D_{0,\ell} = D(x_0, \mathbb{P}_\ell)^{\mathrm{T}}$.

d) Asigne $x_0$ al grupo $i$-ésimo si $\mathbf{D}_0 \in \mathcal{G}_i$.

En el caso de tener varios puntos a clasificar $\{x_1 \ldots, x_m\}$ se repiten los pasos c) y d) tomando $x_0 = x_k$ para $k = 1, \ldots, m$.

En el estudio numérico de la Sección 5.3 llamaremos al conjunto $\{x_1 \ldots, x_m\}$ conjunto de testeo, pues para dichos elementos conoceremos la clase de pertenencia.

## 5.3 Estudio numérico

### 5.3.1 Reglas de clasificación consideradas

En esta sección compararemos el método propuesto con distintas alternativas ya existentes en la clasificación de datos funcionales, tanto en casos reales como simulados. Los cuatro métodos que usaremos en este estudio comparativo son: el método basado en núcleos utilizando la semi-métrica basada en componentes principales que se describe en la Sección 3.4.1, el método de vecinos más cercanos con la distancia de $L^2(\mathcal{I})$ descripto en la Sección 3.4.1 y los dos métodos mencionados previamente en la Sección 5.2, que utilizan el DD-plot junto con la atipicidad direccional y la profundidad de reconocimiento de forma para reducir la dimensión junto con un clasificador multivariado. Además, definiremos el criterio de comparación entre los distintos métodos.

Sea $\{x_1 \ldots, x_m\}$ el conjunto de testeo, es decir, $\{x_1 \ldots, x_m\} = \{x_{i,j}\}_{1\leq i\leq M}^{1\leq j\leq m_i}$, donde sabemos que la observación $x_{i,j}$ pertenece al grupo $i-$ésimo, $1 \leq j \leq m_i$, $1 \leq i \leq M$. Para simplificar la notación consideremos los pares $\{(x_1, G_1), \ldots, (x_m, G_m)\}$, donde $G_j = i$ si $x_j \in \{x_{i,1}, \ldots, x_{i,m_i}\}$. Dada una regla de clasificación $C$ con regiones de clasificación $\mathcal{G}_i$, $1 \leq i \leq M$, definimos $G_j^* = i$ si $x_j \in \mathcal{G}_i$, es decir, $G_j^* = C(x_j) = i$ indica la clase a la que se asigna la observación $j-$ésima.

El error de mala clasificación puede aproximarse mediante la expresión $\hat{e}$ dada por

$$\hat{e} = \sum_{i=1}^{M} \hat{e}_i \frac{m_i}{m}\,,$$

donde $\hat{e}_i$ es el porcentaje de observaciones de la población $i-$ésima mal clasificadas, o sea

$$\hat{e}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{1}(C(x_{i,j}) \neq i)\,.$$

Por lo tanto,

$$\hat{e} = \frac{1}{m} \sum_{i=1}^{M} \sum_{j=1}^{m_i} \mathbb{1}(C(x_{i,j}) \neq i) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(G_j^* \neq G_j)\,. \tag{5.1}$$

A continuación se describen los parámetros elegidos para cada método de clasificación.

Como primer método ya existente para el estudio comparativo se tomó el método de vecinos más cercanos con la norma de $L^2(\mathcal{I})$ descripto en la Sección 3.4.1. Se eligió el parámetro $k$ del clasificador que minimizara el error de mala clasificación para 10 iteraciones iniciales e independientes al resto en un estudio preliminar en cada ejemplo. Se eligió este método por ser el más simple dentro de los clasificadores para datos funcionales, que además logra buenos resultados y es un método estándar en cualquier estudio comparativo. En las tablas y gráficos este método se indicará como KNN.

El siguiente método usado es el método basado en núcleos. Para dicho procedimiento, se utiliza la semi-métrica basada en componentes principales funcionales tal como se describe en la Sección 3.4.1. El núcleo elegido fue $K(s) = \mathbb{1}_{[0,1]}(s)(1-s)$, mientras que la cantidad de direcciones principales $q$ y la ventana $h$ se eligieron minimizando el error de mala clasificación en 10 iteraciones iniciales e independientes al resto para cada modelo. Este clasificador se eligió para comparar la reducción de la dimensión utilizando componentes principales y la reducción de la dimensión utilizando el DD-plot, observando que ambos métodos tienen en cuenta la forma de los datos. En las tablas y gráficos este método se indicará como NFPC.

Se compararon estos dos procedimientos con nuestra propuesta tomando dos profundidades. Para la profundidad de reconocimiento de forma se usó la versión integral solamente, ya que, como se mencionó, la versión que utiliza el mínimo asigna profundidad nula a varios elementos de la muestra. También se eligió como parámetro $J = 2$, basándonos en que en ambos estudios comparativos hay atipicidad de orden 2, aunque esto podría cambiar en otra situación. Respecto a la profundidad multivariada que utiliza el método, se tomó la profundidad del semiespacio descripta en la Sección 4.3.1.

Respecto a la atipicidad direccional, dependiendo del conjunto de datos se eligió utilizar o no la atipicidad direccional media y utilizar o no la variación de la atipicidad direccional, tomando la decisión en base a los errores de mala clasificación obtenidos en 10 iteraciones iniciales e independientes al resto para cada ejemplo, eligiendo el procedimiento que dio como resultado el menor error.

Para la elección de la regla de clasificación multivariada, nos limitamos a considerar algunas de las reglas descriptas en la Sección 3.3. En particular, tomamos la regla de clasificación cuadrática, la de máquinas del vector soporte y el método de vecinos más cercanos. Para la elección de la regla de clasificación, fijada la noción de profundidad o atipicidad, efectuamos 10 replicaciones iniciales e independientes al resto en un estudio preliminar en ambos estudios comparativos. De esta manera, observamos el comportamiento del procedimiento descripto en la Sección 5.2 utilizando las tres posibles reglas de clasificación. Luego, elegimos la regla que minimizaba el error de mala clasificación en dicho estudio preliminar. En ambos estudios comparativos y para ambas nociones de profundidad y atipicidad, dicha regla de clasificación resultó ser la regla de discriminación cuadrática. En las tablas y gráficos indicaremos a estos dos procedimientos como RQDA y DQDA cuando se utilizan la profundidad de reconocimiento de forma y la atipicidad direccional, respectivamente.

### 5.3.2 Estudio de Montecarlo

Consideremos el mismo esquema de simulación considerado en Ferraty y Vieu (2003). El modelo considerado estudia el comportamiento de la regla de clasificación para observaciones provenientes de 3 clases distintas, donde cada curva estará definida en el intervalo $[1, 21]$ en una grilla con paso constante igual a 0.2, es decir que tendremos la grilla

$(t = 1, 1.2, 1.4, ..., 21)$. Las observaciones se generan con la misma distribución que $X$, donde

$$\begin{aligned} X(t) &= uh_1(t) + (1-u)h_2(t) + \epsilon(t) \quad \text{para la clase 1,} \\ X(t) &= uh_1(t) + (1-u)h_3(t) + \epsilon(t) \quad \text{para la clase 2,} \\ X(t) &= uh_2(t) + (1-u)h_3(t) + \epsilon(t) \quad \text{para la clase 3,} \end{aligned} \tag{5.2}$$

siendo $u$ una variable aleatoria uniforme en $(0, 1)$, $\epsilon(t)$ normales estándar independientes para distintos tiempos. Las funciones $h_i$, $1 \leq i \leq 3$, son traslaciones de las ondas triangulares dadas por

$$h_1(t) = \max(6 - |t - 11|, 0), \qquad h_2(t) = h_1(t-4) \quad \text{y} \quad h_3(t) = h_1(t+4)\,.$$

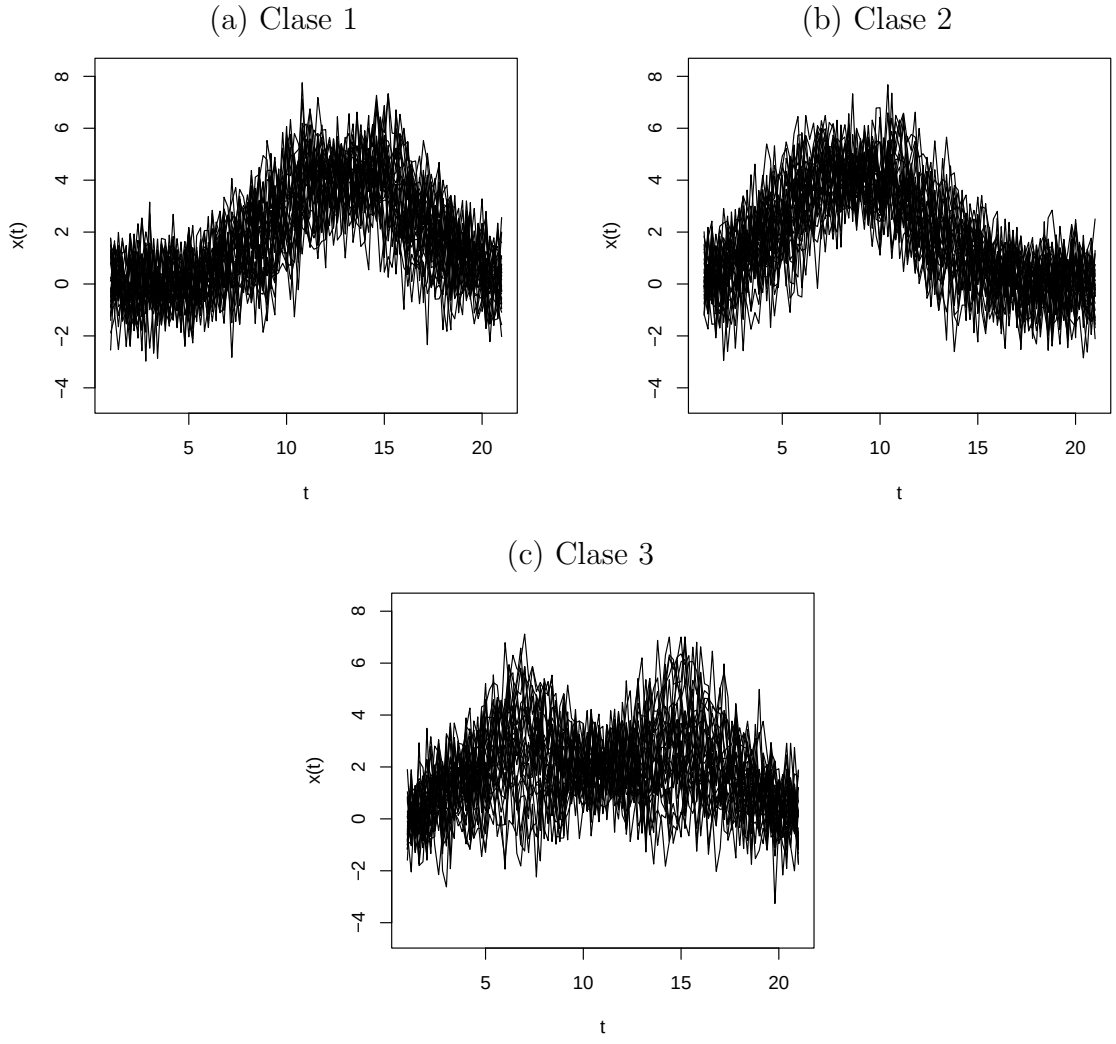


Figura 5.1: Gráficos correspondientes a una muestra generada de acuerdo al Modelo (5.2).

Se realizaron $NR = 500$ replicaciones generando en cada replicación muestras $X_{i,j}$, $1 \leq j \leq n_i$, $1 \leq i \leq 3$ según el modelo (5.2) con $n_i = 200$ para $i = 1, 2, 3$ y muestras de testeo de

tamaño $m_i = 100$ en cada población como se describe en la Sección 5.3.2. En la Figura 5.1 se pueden visualizar una muestra de 30 elementos generada para cada clase. Para resumir el comportamiento de cada uno de los procedimientos, en cada replicación se obtiene el error $\hat{e}$ de mala clasificación dado en (5.1), siendo $\hat{e}_s$ el correspondiente a la replicación $s$. La Tabla 5.1 reporta los promedios obtenidos para cada regla de clasificación en la columna que se indica con $\hat{e}$, es decir, $\hat{e} = \sum_{s=1}^{NR} \hat{e}_s/NR$, la tercer columna, denotada $\text{SD}_{\hat{e}}$, indica el desvío estándar de $\hat{e}_1, \ldots, \hat{e}_{NR}$, mientras que en la última columna se reporta el tiempo de cómputo total de las 500 replicaciones realizadas en una computadora con 16gb de Ram y procesador Intel i7 11gen.

En la simulación, según el estudio preliminar de 10 iteraciones independientes al resto, los parámetros utilizados fueron los siguientes. Para el clasificador de vecinos más cercanos utilizando la distancia de $L^2(\mathcal{I})$ el parámetro utilizado fue $k = 41$. Para el clasificador basado en núcleos con la semi métrica de componentes principales funcionales se utilizaron dos componentes principales, es decir $q = 2$, y el parámetro de ancho de banda $h = 4$. Como se mencionó antes, tanto para el método que utiliza la profundidad basada en el reconocimiento de forma como para el que utiliza la atipicidad direccional, el procedimiento de clasificación multivariado basado en la regla de disciminante cuadrático resultó ser el mejor clasificador sobre las 10 replicaciones consideradas, aunque sin mucha diferencia con la clasificación mediante máquinas del vector soporte. Para el método que utiliza la atipicidad direccional, no se utilizó la atipicidad direccional media, es decir que en el procedimiento descripto en la Sección 5.2 en lugar de usar $D(x, \mathbb{P}_X)$ correspondiente al vector $(MO(x; \mathbb{P}_X), VO(x; \mathbb{P}_X))^{\mathrm{T}}$, se utilizó $D(x, \mathbb{P}_X) = VO(x; \mathbb{P}_X)$, ya que de esta forma resultaba con menor error de mala clasificación sobre las 10 replicaciones consideradas. En las tablas y gráficos este método se indicará como VDQDA.

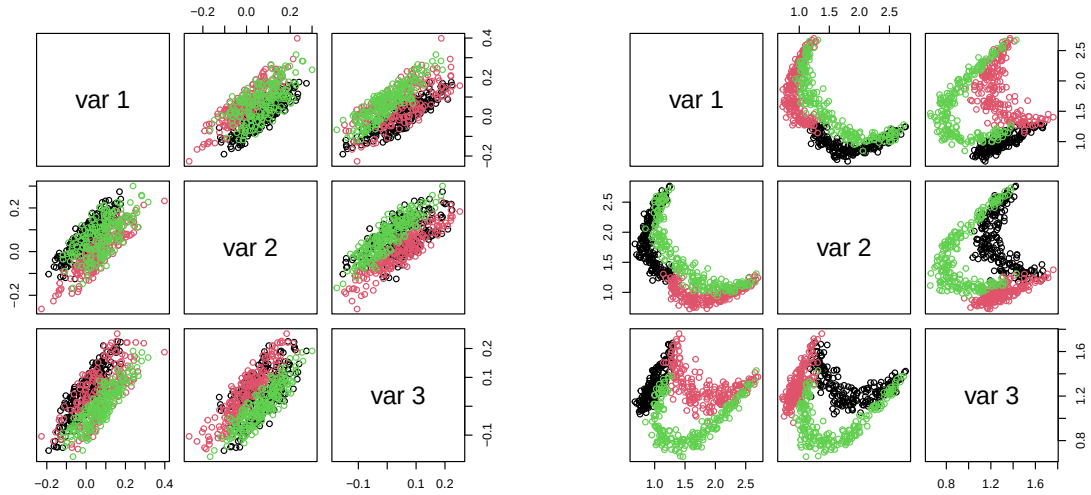


Figura 5.2: Pairplot de la atipicidad direccional respecto de cada una de las 3 clases, con cada clase en distinto color. (a) Atipicidad direccional media. (b) Variación de la atipicidad direccional.

Para mostrar los problemas que ocasiona usar el vector $(MO(x;\mathbb{P}_X), VO(x;\mathbb{P}_X))^{\mathrm{T}}$, la Figura 5.2 presenta los gráficos con 6 variaciones de los pares $(D_{i,j,\ell_1}, D_{i,j,\ell_2})$, con $1 \leq j \leq n$, $1 \leq i \leq 3$, $1 \leq \ell_1 \leq 3$, $1 \leq \ell_2 \leq 3$, con $\ell_1 \neq \ell_2$ para una muestra. En ambos gráficos, el texto "var $\ell$" representa a la atipicidad en el grupo $\ell$, $1 \leq \ell \leq 3$. Luego, podemos notar cómo la variación de la atipicidad direccional logra diferenciar mejor los datos al observar la diferencia de formas, mientras que la atipicidad direccional media parecería agregar ruido.

| Clasificador | $\hat{e}$ | $\mathrm{SD}_{\hat{e}}$ | Tiempo |
|---|---|---|---|
| KNN | 0.0915 | 0.0162 | 5min |
| NFPC | 0.1580 | 0.0277 | 8seg |
| RQDA | 0.0987 | 0.0142 | 40hs |
| VDQDA | 0.0761 | 0.0126 | 33seg |

Tabla 5.1: Errores de mala clasificación para los distintos métodos considerados.

A partir de los resultados en la Tabla 5.1, podemos notar cómo el error de mala clasificación $\hat{e}$ para el método basado en núcleos es mayor que para el resto de los métodos. Este incremento en el error de mala clasificación puede atribuirse a la dificultad de reconocer las autofunciones principales del tercer grupo y a la falta de regularidad de las observaciones, provocando un mayor promedio del error de mala clasificación y un mayor desvío que puede atribuirse a la variabilidad provocada por tener que estimar las direcciones principales y no contar con las verdaderas direcciones. Por otra parte, el método basado en la atipicidad direccional donde sólo se utiliza la variación de la atipicidad direccional junto con la regla de clasificación dada por la función discriminante cuadrática es la que obtiene mejores resultados, es decir, los menores errores promedio de mala clasificación. Esto puede atribuirse al hecho de que la mayor diferencia entre las clases se debe a la forma de las curvas, cosa que este método logra diferenciar bien. Es importante notar además la diferencia de tiempo de cómputo entre los distintos métodos del total de las 500 iteraciones. El método que utiliza la profundidad de reconocimiento de forma debe calcular reiteradas veces (en este caso $101^2$) la profundidad del semiespacio para cada elemento del grupo de testeo, por lo que el costo computacional del cálculo de dicha profundidad en dos dimensiones se vuelve no despreciable al multiplicarse por tal magnitud, cosa que se ve proyectado al tiempo de cómputo de dicha profundidad funcional sin disminuir el error de mala clasificación. Cabe mencionar que el método de vecinos más cercanos da resultados comparables a este último procedimiento en menor tiempo de cómputo.

En la Figura 5.3 se presenta el boxplot de los errores de mala clasificación obtenidos para cada método. Teniendo en cuenta que dichos errores $\hat{e}$ son cantidades no negativas y que se espera que tengan distribución asimétrica, utilizamos el boxplot ajustado definido en Hubert y Vandervieren (2008). De esta manera confirmamos lo comentado respecto a los resultados de la Tabla 5.1, ya que observamos que el método que utiliza la variación de la atipicidad direccional junto con la función discriminante cuadrática logra destacar entre los métodos, mientras que el método de vecinos más cercanos y el método que utiliza la profundidad de

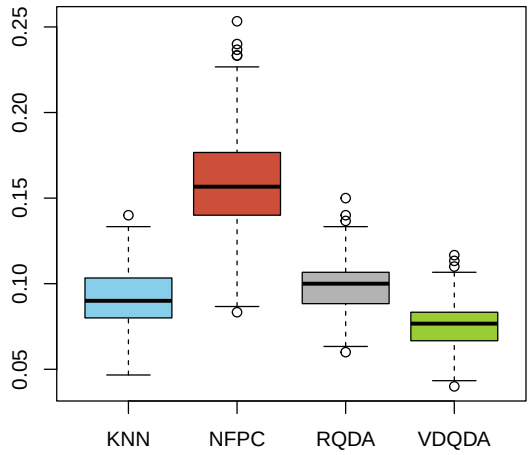


Figura 5.3: Boxplots ajustados de $\hat{e}_1, \ldots, \hat{e}_{NR}$ para cada método de clasificación.

reconocimiento junto con la función discriminante cuadrática obtienen resultados similares en una gran cantidad de replicaciones, aunque los mejores resultados del método de vecinos parecerían ser mejores que el de profundidad de reconocimiento de forma. Por el otro lado, con este gráfico podemos confirmar el bajo rendimiento a la hora de clasificar correctamente del método basado en núcleos junto con componentes principales en comparación a los otros métodos, ya que la caja del boxplot se encuentra por encima del bigote superior de cualquiera de los otros 3 métodos.

### 5.3.3 Conjunto de datos de fonemas

El conjunto de datos de fonemas, disponible en el repositorio del libro de Ferraty y Vieu `https://www.math.univ-toulouse.fr/staph/npfda/npfda-datasets.html`, consiste en observaciones correspondientes a 5 fonemas del idioma inglés: "sh" de "she", "dcl" de "dark", "iy" la vocal de "she", "aa" la vocal de "dark" y "ao" la primera vocal de "water". Para cada fonema, se registraron 500 observaciones que corresponden al logaritmo del periodograma medidas en 150 frecuencias. En la Figura 5.4 se presenta un gráfico de dichos datos funcionales donde en el eje horizontal se encuentra la frecuencia y en el vertical $x(t)$, el logaritmo del periodograma en la frecuencia $t$.

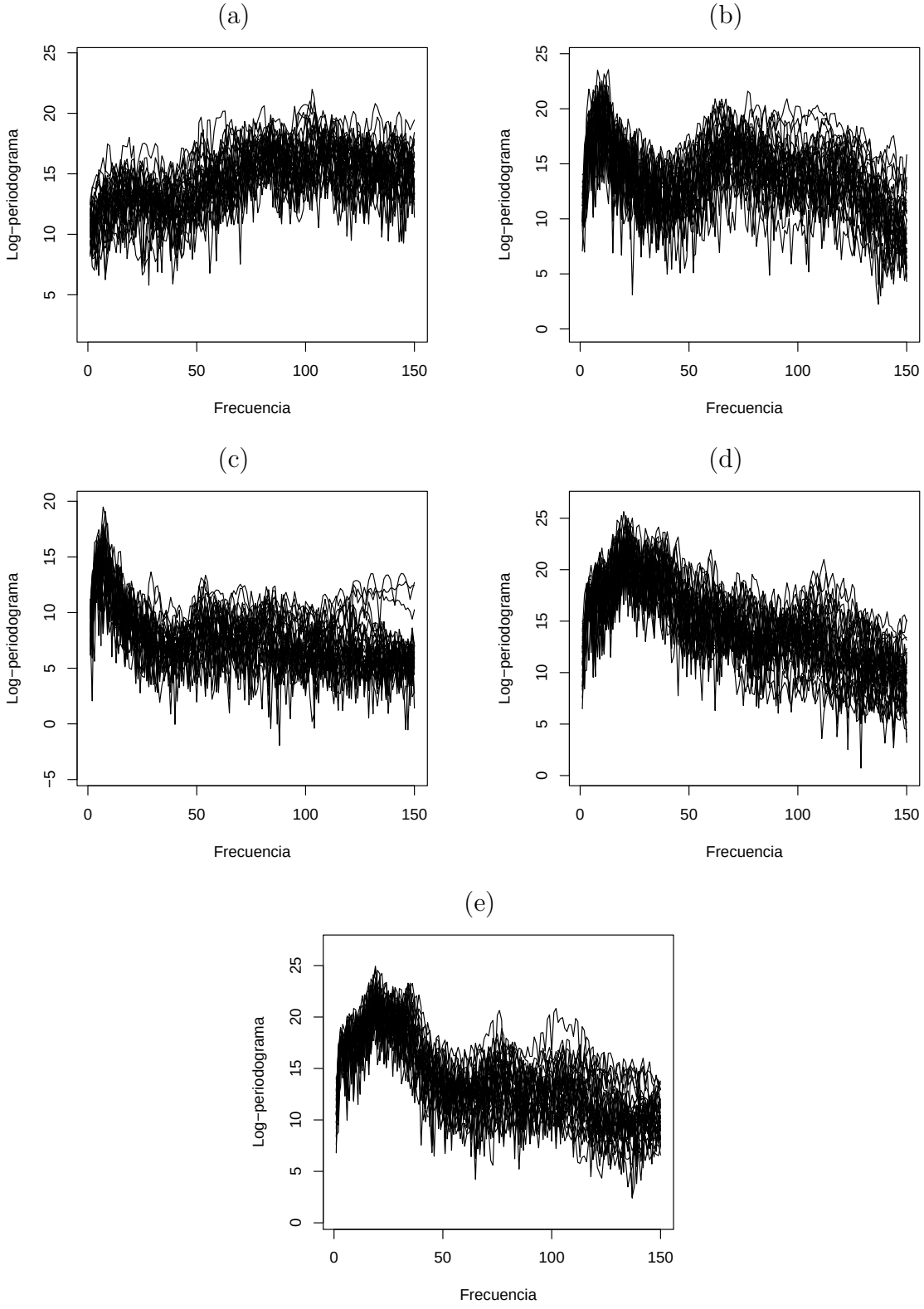


Figura 5.4: Datos de log-periodogramas de 5 fonemas distintos: (a) “sh”, (b) “iy”, (c) “dcl”, (d) “aa”, (e) “ao”.

Para el conjunto de entrenamiento usamos $n_i = 150$ observaciones de cada clase, mientras que el conjunto de testeo tiene $m_i = 250$ elementos de cada clase. Igual que en el estudio de simulación reportado en la Sección 5.3.2, se realizaron 500 replicaciones eligiendo en cada replicación las observaciones a asignar a la muestra de entrenamiento y de testeo al azar dentro de cada grupo $\ell$, $\ell = 1, \ldots, 5$. Si indicamos por $\hat{e}_s$ el error de mala clasificación en la replicación $s$, la Tabla 5.2 reporta el promedio sobre las replicaciones $\hat{e} = \sum_{s=1}^{500} \hat{e}_s/500$ como estimación del error de mala clasificación de cada método. Asimismo se reporta el desvío estándar $\text{SD}_{\hat{e}}$ y el tiempo de cómputo total en una computadora con 16gb de Ram y un procesador Intel i7 11gen.

Para cada uno de los cuatro métodos, sus parámetros fueron elegidos para minimizar el error medio de mala clasificación sobre 10 iteraciones realizadas en forma independiente a las 500 replicaciones posteriores. Como resultado de este estudio preliminar, se eligió $k = 27$ para el procedimiento de clasificación basado en vecinos más cercanos y $q = 9$ y $h = 25$ para el método basado en núcleos y componentes principales. Por otra parte, al utilizar los métodos de clasificación basados en la profundidad de reconocimiento de forma o la atipicidad direccional, el método multivariado basado en la regla discriminante cuadrática dio origen a los mejores resultados y, por ello, se la utilizó en las 500 iteraciones. Finalmente, para la profundidad de reconocimiento de forma, basado en este estudio preliminar, se eligió la profundidad de semiespacio y $J = 2$.

| Clasificador | $\hat{e}$ | $\text{SD}_{\hat{e}}$ | Tiempo |
|---|---|---|---|
| KNN | 0.0986 | 0.0066 | 19min |
| NFPC | 0.1406 | 0.0098 | 4min |
| RQDA | 0.1829 | 0.0100 | 120hs |
| DQDA | 0.0866 | 0.0058 | 4min |

Tabla 5.2: Errores de mala clasificación para el conjunto de datos de fonemas.

Los resultados obtenidos se reportan en la Tabla 5.2. Por otra parte, la Figura 5.5 muestra el boxplot ajustado de los errores de mala clasificación obtenidos. Recordemos que se indica con DQDA al procedimiento descripto en la Sección 5.2 que combina la atipicidad direccional, utilizando el vector $(MO(x; \mathbb{P}_X), VO(x; \mathbb{P}_X))^{\mathrm{T}}$, junto con la regla de clasificación cuadrática. Dichos resultados reflejan que el clasificador basado en la profundidad de reconocimiento de forma no logra tener un bajo error de mala clasificación duplicando el basado en la atipicidad direccional. Esto puede suceder ya que la profundidad del semiespacio que utiliza es una profundidad que apunta a establecer un orden de centralidad de los valores de la muestra. Más precisamente, curvas "fuera" de la muestra tendrán asignada una profundidad nula sin poder obtener más información, ya que sólo se obtiene un "ranking" de centralidad respecto de la muestra. Este hecho se vuelve particularmente problemático al trabajar con varias poblaciones. Observemos que, como en el estudio de simulación, el costo computacional del método resulta mucho más elevado que el de los otros sin reducir el error de mala clasificación, haciéndolo no competitivo. Este costo computacional se podría reducir

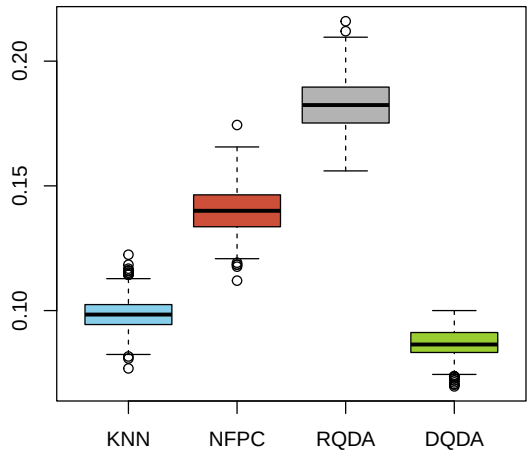


Figura 5.5: Boxplots ajustados de $\hat{e}_1, \ldots, \hat{e}_{NR}$ para cada método para el conjunto de datos de fonemas.

a través de la aproximación del cálculo de la profundidad usando la profundidad aproximada $FD_J^{(S)}(x; \mathbb{P}_X)$ definida en (4.8). Sin embargo, en base a los errores de clasificación obtenidos, no realizaremos en esta tesis dicha comparación que queda como trabajo futuro. En cuanto al método basado en núcleos junto con componentes principales, otra vez obtenemos un pobre rendimiento en comparación con el método de vecinos más cercanos y el método que utiliza la atipicidad direccional junto con la función discriminante cuadrática. Este último se vuelve a destacar al lograr reconocer de manera eficaz la diferencia de magnitud y forma entre las poblaciones, obteniendo los mejores resultados.

### 5.3.4 Conclusión

La propuesta de combinar el DD-plot con la atipicidad direccional para la reducción de la dimensión, junto con un clasificador multivariado genera un clasificador para datos funcionales que no sólo es competente al compararse con otros métodos existentes respecto al error de mala clasificación, sino que también lo logra con bajo costo computacional. Tomar la atipicidad en lugar de la profundidad integral descripta en la Sección 4.4.1 resulta un acierto para el problema de clasificación de datos funcionales, ya que de esta manera conseguimos la flexibilidad de utilizar distintas nociones de atipicidad univariada que puedan amoldarse a los requerimientos de nuestros datos, ya que podemos utilizar versiones más o menos robustas o incluso preparadas para datos asimétricos. También es importante mencionar que la versatilidad para elegir la regla de clasificación multivariada nos va a permitir trabajar en situaciones donde tengamos poblaciones con distintas modas.

## 5.4 Apéndice: Código

### 5.4.1 Clasificador de datos funcionales utilizando la atipicidad direccional

```
library(mrfDepth) # AO
library(e1071) # svm
library(MASS) # qda
library(class) # knn

dirOut <- function(test, train, outly.method, L.norm, depth, abs) {

  nTrain <- dim(train)[1]
  nTest <- dim(test)[1]
  p <- dim(train)[2]

  ### Univariate method for outlyingness calculation
  if (outly.method=="Mahalanobis") {
    Zcenter <- apply(train,2,mean)
    Zsd <- apply(train,2,sd)
    dirOutMatrix <- (test-rep(Zcenter,each=nTest))/rep(Zsd,each=nTest)
  }
  else if (outly.method=="SDO") {
    Zcenter <- apply(train,2,median)
    Zmad <- apply(train,2,mad)
    dirOutMatrix <- (test-rep(Zcenter,each=nTest))/rep(Zmad,each=nTest)
  }
  else if (outly.method=="AO"){
    Zcenter <- apply(train,2,median)
    adjOutlCol <- apply(test,2,function(columna) {adjOutl(columna[1:nTrain],columna[(↩
        nTrain+1):nTest])})
    dirOutMatrix <- sapply(1:p,function(ncol) {c(adjOutlCol[[ncol]]$outlyingnessX,↩
        adjOutlCol[[ncol]]$outlyingnessZ)})
    dirOutMatrix <- dirOutMatrix*sign(test-rep(Zcenter,each=nTest))
  }


  ### False for outlyingness, True for depth
  if (depth==TRUE) {
    dirOutMatrix <- 1/(1+abs(dirOutMatrix))
  }

  out_avr <- apply(dirOutMatrix, 1, mean)

  ### L norm to calculate variation
  if (L.norm=="Inf") {
    out_var <- apply(matrix(apply(dirOutMatrix, 2, (function(columna) abs(columna-out_avr)↩
        )),nTest),1,max)
  }
  else {
    out_var <- (apply(matrix(apply(dirOutMatrix, 2, (function(columna) (abs(columna-out_↩
        avr))**L.norm)),nTest),1,mean))**(1/L.norm)
  }

  ### True for the absolute value of out_avr
  if (abs==TRUE) {
    dirOutMatrix <- abs(dirOutMatrix)
  }
  out_avr <- apply(dirOutMatrix, 1, mean)
```

```
  M <- cbind(out_avr,out_var)
  colnames(M) <- c("out_avr","out_var")
  return(M)
}


dirOutClass <- function(test, train, target, outly.method="Mahalanobis", L.norm=2, abs=↩
    FALSE, depth=FALSE, classifier="qda", classify=TRUE, out.avr=TRUE, out.var=TRUE) {

  train <- as.matrix(train)
  test <- as.matrix(test)
  nTrain <- dim(train)[1]
  nTest <- dim(test)[1]
  trainTestBind <- rbind(train,test)

  if (classify==FALSE) {

    ### If classify==FALSE, only calculate the Directional Outlyingness
    predictions <- NA
    M <- dirOut(trainTestBind,train, outly.method=outly.method, L.norm=L.norm, depth=depth↩
        , abs=abs)
    out_avrTrain <- M[1:nTrain,1]
    out_avrTest <- M[(nTrain+1):(nTrain+nTest),1]
    out_varTrain <- M[1:nTrain,2]
    out_varTest <- M[(nTrain+1):(nTrain+nTest),2]
  }

  else {

    ### Dimensionality reduction using Directional Outlyingness
    levels <- levels(target)
    out_avrTrain <- matrix(NA, nTrain, length(levels))
    out_avrTest <- matrix(NA, nTest, length(levels))
    out_varTrain <- matrix(NA, nTrain, length(levels))
    out_varTest <- matrix(NA, nTest, length(levels))

    for (i in 1:length(levels)) {
      M <- dirOut(trainTestBind,train[target==levels[i],], outly.method=outly.method, L.↩
          norm=L.norm, depth=depth, abs=abs)
      out_avrTrain[1:nTrain,i] <- M[1:nTrain,1]
      out_avrTest[1:nTest,i] <- M[(nTrain+1):(nTrain+nTest),1]
      out_varTrain[1:nTrain,i] <- M[1:nTrain,2]
      out_varTest[1:nTest,i] <- M[(nTrain+1):(nTrain+nTest),2]
    }

    ### Classification
    ### If out.avr==FALSE or out.avr==FALSE, the method won't use those values
    if (classifier=="svm") {
      if (out.avr==FALSE) {
        model <- svm(out_varTrain,target,kernel="sigmoid")
        predictions <- predict(model, out_varTest)
      }
      else if (out.var==FALSE) {
        model <- svm(out_avrTrain,target,kernel="sigmoid")
        predictions <- predict(model, out_avrTest)
      }
      else {
        model <- svm(cbind(out_avrTrain,out_varTrain),target,kernel="sigmoid")
        predictions <- predict(model, cbind(out_avrTest,out_varTest))
      }
    }

    else if (classifier=="qda") {
      if (out.avr==FALSE) {
        model <- qda(out_varTrain,target)
```

```
        predictions <- predict(model, out_varTest)$class
      }
      else if (out.var==FALSE) {
        model <- qda(out_avrTrain,target)
        predictions <- predict(model, out_avrTest)$class
      }
      else {
        model <- qda(cbind(out_avrTrain,out_varTrain),target)
        predictions <- predict(model, cbind(out_avrTest,out_varTest))$class
      }
    }

    else if (classifier=="knn") {
      if (out.avr==FALSE) {
        predictions <- knn(out_varTrain,out_varTest,target,k=floor(sqrt(nTrain)))
      }
      if (out.var==FALSE) {
        predictions <- knn(out_avrTrain,out_avrTest,target,k=floor(sqrt(nTrain)))
      }
      else {
        predictions <- knn(cbind(out_avrTrain,out_varTrain), cbind(out_avrTest,out_varTest↩
            ),target,k=floor(sqrt(nTrain)-22))
      }
    }
  }

  return(list(predictions = predictions,
              out_avrTrain = out_avrTrain, out_avrTest = out_avrTest,
              out_varTrain = out_varTrain, out_varTest = out_varTest))
}
```

### 5.4.2 Clasificador de datos funcionales utilizando la profundidad de reconocimiento de forma

```
library(e1071) # svm
library(MASS) # qda
library(class) # knn
library(depth.fd)# shape.fd.analysis


shapeDepthClass <- function(test, train, target, order=2, depth="Halfspace",
approx=0, classifier="qda", classify=TRUE) {

  train <- as.matrix(train)
  test <- as.matrix(test)
  nTrain <- dim(train)[1]
  nTest <- dim(test)[1]
  trainTestBind <- rbind(train,test)

  ### If classify==FALSE, only calculate the Shape Depth
  if (classify==FALSE) {
    predictions <- NA
    depths <- rep(NA,(nTrain+nTest))

    for (i in 1:(nTrain+nTest)) {

      ### Multivariate method for depth calculation
      if (depth=="Halfspace") {
        depths[i] <- shape.fd.analysis(t(as.matrix(trainTestBind[i,])), train,
```

```
        order = order,approx = approx, plot = FALSE)$Half_FD
      }
      else if (depth=="Simplicial") {
        depths[i] <- shape.fd.analysis(t(as.matrix(trainTestBind[i,])), train,
        order = order,approx = approx, plot = FALSE)$Simpl_FD
      }
    }
    depthTrain <- depths[1:nTrain]
    depthTest <- depths[(nTrain+1):(nTrain+nTest)]
  }

  else {

    ### Dimensionality reduction using Shape Depth
    levels <- levels(target)
    depthTrain <- matrix(NA, nTrain, length(levels))
    depthTest <- matrix(NA, nTest, length(levels))

    for (i in 1:length(levels)) {
      depths <- rep(NA,nTrain+nTest)
      for (j in 1:(nTrain+nTest)) {
        if (depth=="Halfspace") {
          depths[j] <- shape.fd.analysis(t(as.matrix(trainTestBind[j,])), train[target==↩
              levels[i],],
          order = order, approx = approx, plot = FALSE)$Half_FD
        }
        else if (depth=="Simplicial") {
          depths[j] <- shape.fd.analysis(t(as.matrix(trainTestBind[j,])), train[target==↩
              levels[i],],
          order = order, approx = approx, plot = FALSE)$Simpl_FD
        }
      }
      depthTrain[,i] <- depths[1:nTrain]
      depthTest[,i] <- depths[(nTrain+1):(nTrain+nTest)]
    }
    ### Classification
    if (classifier=="svm") {
      model <- svm(depthTrain,target)
      predictions <- predict(model,depthTest)
    }

    else if (classifier=="qda") {
      model <- qda(depthTrain,target)
      predictions <- predict(model, depthTest)$class
    }

    else if (classifier=="knn") {
      predictions <- knn(depthTrain,depthTest,target,k=floor(sqrt(nTrain)))
    }
  }

  return(list(predictions = predictions,
              depthTrain = depthTrain, depthTest = depthTest))
}
```

# Bibliografía


Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Lecture Notes in Statistics, Springer, New York.

Brys, G., Hubert, M., y Rousseeuw, P. (2005). A robustification of independent component analysis. *Journal of Chemometrics*, 19:364–375.

Brys, G., Hubert, M., y Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13:996–1017.

Chang, C., Chen, Y., y Ogden, R. (2014). Functional data classification: a wavelet approach. *Computational Statistics*, 29:1497–1513.

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

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

Dai, W. y Genton, M. G. (2019). Directional outlyingness for multivariate functional data. *Computational Statistics and Data Analysis*, 131:50–65.

Debnath, L. y Mikusinski, P. (2005). *Introduction to Hilbert Spaces with Applications*. Elsevier Science.

Delaigle, A. y Hall, P. (2010). Defining probability density for a distribution of random functions. *Annals of Statistics*, 38:1171–1193.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*. Harvard University. Qualifying paper.

Efron, B. y Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, USA.

Ferraty, F. y Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis*, 44:161–173.

Fraiman, R. y Muñiz, G. (2001). Trimmed means for functional data. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 10:419–440.

Giambartolomei, G. (2015). *The Karhunen-Loève theorem*. Facoltà di scienze matematiche, fisiche e naturali, Università di Bologna. Master thesis.

Gijbels, I. y Nagy, S. (2017). On a general definition of depth for functional data. *Statistical Science*, 32:630–639.

Hastie, T., Tibshirani, R., y Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Horvath, L. y Kokoska, P. (2012). *Inference for Functional Data with Applications*. Springer, New York.

Hubert, M., Rousseeuw, P., y Segaert, P. (2016). Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, 11:445–466.

Hubert, M. y Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52:5186–5201.

Leng, X. y Müller, H.-G. (2005). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22:68–76.

Li, J., Cuesta-Albertos, J., y Liu, R. (2012). Dd-classifier: Nonparametric classification procedure based on dd-plot. *Journal of The American Statistical Association*, 107:737–753.

Liu, R. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, 18:405–414.

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

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences*, 2:49–55.

Maronna, R., Martin, D., Yohai, V. J., y Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. Wiley, New York.

Nagy, S., Gijbels, I., y Hlubinka, D. (2017). Depth-based recognition of shape outlying functions. *Journal of Computational and Graphical Statistics*, 26:883–893.

Nieto-Reyes, A. y Battey, H. (2016). A Topologically Valid Definition of Depth for Functional Data. *Statistical Science*, 31:61–79.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33:1065–1076.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27:832–837.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Seber, G. A. F. (1984). *Multivariate Observations*. Wiley, New York.

Serfling, R. y Zuo, Y. (2000). General notions of statistical depth function. *Annals of Statistics*, 28:461–482.

Stahel, W. A. (1981). *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. ETH Zürich. Ph.D. dissertation.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Canad. Math. Congress, Montreal*, 523–531.

Wang, J.-L., Chiou, J.-M., y Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.