UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Detección de datos atípicos para datos funcionales asimétricos

Pablo Vena

Directora: Dra. Graciela Boente Boente

Marzo de 2014

Detección de datos atípicos para datos funcionales asimétricos

En todo análisis, la detección de datos atípicos es un paso importante aún cuando se usen estimadores robustos. En particular, la distancia de Mahalanobis robustificada (Rousseeuw y van Zomeren, 1990) es una medida natural si uno se concentra en distribuciones elípticas multivariadas. Por otra parte, Hubert y Van der Veeken (2008) propusieron un método de detección de datos atípicos que no necesita la hipótesis de simetría y no depende de la inspección visual de los datos. Este método es una generalización de la medida de atipicidad de Stahel–Donoho que asigna a cada observación una medida de atipicidad obtenida por *projection pursuit* y que utiliza solamente medidas robustas univariadas de posición y escala. Para permitir asimetría en los datos, Hubert y Van der Veeken (2008) realizan un ajuste de esta medida usando una medida de asimetría univariada. Para datos funcionales, no es posible extender la distancia de Mahalanobis robustificada debido a que los operadores de covarianza son compactos. Sun y Genton (2011) dan una extensión del boxplot univariado al caso de datos funcionales que detecta bien datos atípicos provenientes de distribuciones simétricas. En esta tesis, se da una extensión de este boxplot y de las medidas de asimetría multivariadas dadas por Hubert y Van der Veeken (2008) de modo de detectar datos atípicos sin necesidad de suponer simetría y que se adapta bien a una clase amplia de distribuciones funcionales asimétricas.

Palabras Claves: Datos Funcionales; Outliers; Robustez.

Indice

Capítulo 1

Introducción

La detección de datos atípicos es un paso importante en cualquier análisis aún cuando se usen estimadores robustos. Los métodos utilizados en el caso univariado y multivariado, en general, necesitan hipótesis adicionales sobre la distribución de los datos, usualmente desconocida. Un supuesto habitual es la simetría. Por ejemplo, si nos restringimos a distribuciones elípticas multivariadas, la distancia de Mahalanobis robustificada (Rousseeuw y van Zomeren, 1990) es una medida natural de atipicidad. Para subsanar esta limitación, Hubert y Van der Veeken (2008) propusieron un método de detección de datos atípicos que no necesita la hipótesis de simetría y tampoco depende de la inspección visual de los datos. El método se inspira en los estimadores robustos de posición y escala de Stahel–Donoho que pondera cada observación con una medida de atipicidad obtenida por *projection pursuit* y que utiliza solamente medidas robustas univariadas de posición y escala. Para permitir asimetría en los datos, Hubert y Van der Veeken (2008) ajustan esta atipicidad con un estimador robusto de asimetría univariada.

La consideración de datos funcionales como objeto de estudio surge como extensión natural del análisis de datos multivariados. Las observaciones en distintas disciplinas como la medicina, meteorología o economía, entre otras, que se representan más adecuadamente como una función que como un vector de dimensión grande, son inherentemente funcionales. En los últimos años, el análisis de datos funcionales ha extendido numerosas técnicas multivariadas como el análisis de componentes principales, análisis de la varianza y métodos de regresión (ver Ramsay y Silverman, 2005) y ha abordado el problema de la detección de datos atípicos. No es posible extender en forma directa la distancia de Mahalanobis robustificada debido a que los operadores de covarianza son compactos. Una propuesta para detectar observaciones atí‘picas se encuentra en Febrero *et al.* (2007) para la clasificación de mediciones de gases contaminantes según días laborales o no (comparación de medias funcionales entre grupos). Los autores introducen una noción de *profundidad funcional* para ordenar las observaciones basada en la integración de medidas de profundidad

univariadas. Con esta idea de orden, extienden la definiciones de estimadores robustos de posición y escala como la media, $\widehat{\mu}_{\mathrm{TM},\alpha}$, y desvío estándar, $\widehat{\sigma}_{\mathrm{TSD},\alpha}$, α-podados funcionales y definen una medida de atipicidad con la que clasifican datos atípicos algún valor de corte a definir. La medida de atipicidad se define como

$$O_\alpha(x_i(t)) = \left\| \frac{x_i(t) - \widehat{\mu}_{\mathrm{TM},\alpha}(t)}{\widehat{\sigma}_{\mathrm{TSD},\alpha}(t)} \right\|$$

donde

$$\begin{aligned} \widehat{\mu}_{\mathrm{TM},\alpha(t)} &= \frac{1}{n-[\alpha n]} \sum_{i=1}^{n-[\alpha n]} x_{[i]}(t)\,, \\ \widehat{\sigma}_{\mathrm{TSD},\alpha}(t) &= \left(\frac{1}{n-[\alpha n]} \sum_{i=1}^{nn-[\alpha n]} \left(x_{[i]}(t) - \widehat{\mu}_{\mathrm{TM},\alpha}(t)\right)^2 \right)^{\frac{1}{2}}, \end{aligned}$$

con $x_{[i]}(t)$ las observaciones ordenadas de acuerdo a su profundidad, es decir, $x_{[1]}(t)$ es la curva más profunda (más central) y $x_{[n]}(t)$ la más exterior.

Por otro lado, Sun y Genton (2010) dan una extensión del boxplot univariado, basada en la *profundidad de banda* definida por López-Pintado y Romo (2009), al caso de datos funcionales que detecta bien datos atípicos provenientes de distribuciones simétricas pero, nuevamente, no detecta apropiadamente la falta de simetría.

En esta tesis, se da una extensión de este boxplot y de las medidas de atipicidad multivariadas dadas por Hubert y Van der Veeken (2008) de modo de detectar datos atípicos sin necesidad de suponer simetría y que se adapta bien a una clase amplia de distribuciones funcionales asimétricas. La tesis se estructura de la siguiente manera:

En el **Capítulo 2**, se introduce, en la Sección 2.1, el boxplot como herramienta de visualización una distribución univariada unimodal y clasificación de las observaciones de la muestra entre típicas y atípicas. El desempeño del criterio de clasificación es pobre para distribuciones que no sean normales, en particular, para distribuciones como las que son objeto de esta tesis con falta de simetría. Para tener en cuenta este factor, en la Sección 2.2, se introduce el *medcouple*, una medida robusta de asimetría para muestras univariadas propuesta en Brys *et al.* (2004), con la que se modifican los bigotes del boxplot dando lugar al *boxplot ajustado* presentado en la Sección 2.3 desarrollado en Hubert y Vandervieren (2007).

En el **Capítulo 3**, se extiende la familia de distribuciones normales a una familia paramétrica que incorpora un parámetro de forma que regula la asimetría. Se trata el caso univariado y su extensión al caso multivariado que permiten luego definir procesos asimétricos en el contexto funcional. Estos procesos servirán de base para generar observaciones en el estudio de simulación.

En el **Capítulo 4**, se describe la propuesta multivariada de Hubert y Van der Veeken (2008) para la detección de datos atípicos por el método de proyecciones univariadas inspirada en la medida de atipicidad de Stahel-Donoho.

En el **Capítulo 5**, se define la profundidad de banda para datos funcionales de López-Pintado y Romo (2009) que permite ordenar una muestra de datos funcionales y definir estimadores de orden. Luego, en la Sección 5.2, se describe la extensión de Sun y Genton (2010) del boxplot univariado a una muestra de datos funcionales. El final del capítulo, Sección 5.3, se dedica a la definición de dos procesos asimétricos introducidos uno en Zhang y El-Shaarawe (2010) y el otro en Staicu *et al.* (2010), para el contexto funcional. Introducimos además una familia particular de procesos asimétricos, los procesos Gaussianos asimétrico funcional, que generalizan las distribuciones normales asimétricas multivariadas y que pueden verse como un caso particular de los considerados en Staicu *et al.* (2010).

En el **Capítulo 6** ofrecemos tres propuestas de detección de outliers para el caso funcional de datos asimétricos. Dos se basan en modifaciones del boxplot de Genton y la otra en la extensión del método de proyecciones descripto en Hubert y Van der Veeken (2008) para el caso multivariado.

Finalmente, en el **Capítulo 7**, se presentan los resultados de un estudio de simulación con el objetivo de analizar el comportamiento de las propuestas de detección. Consideramos varios modelos de datos funcionales asimétricos, con diferentes tipos de asimetría y proporciones de contaminación. Las conclusiones se dan en la **Sección 7.3** y todos los códigos utilizados para las simulaciones, escritos en R, se presentan en el **Apéndice**.

Capítulo 2

Boxplot y Boxplot ajustado

2.1 Boxplot

El *diagrama de cajas* o *boxplot* es un método gráfico para representar la distribución de una muestra univariada unimodal. Sintetiza información sobre la posición, dispersión y forma de la distribución a través de cinco estadísticos descriptivos: la mediana, el primer y tercer cuartil y la mínima y máxima observación regular. El boxplot provee además un criterio para clasificar dentro de las observaciones a aquellas que sean **potenciales** *datos atípicos*. Indistintamente hablaremos de datos atípicos o, su denominación en inglés, *outliers*.

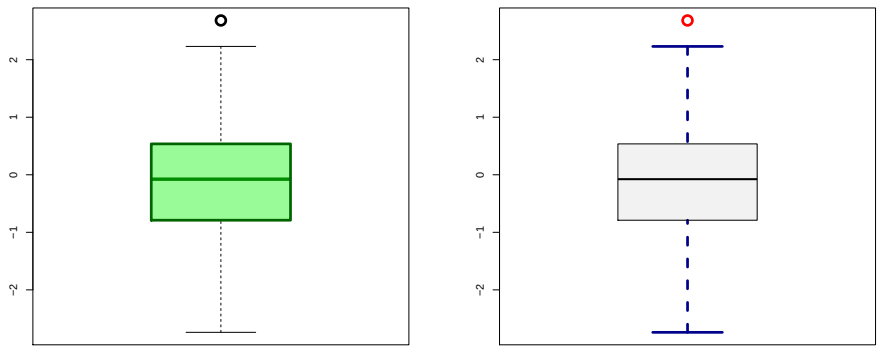
Desde su introducción por Tukey (1977), el boxplot se ha convertido en una de las herramientas más difundidas en el análisis exploratorio de datos. Es particularmente útil para comparar distribuciones en distintas muestras sin recurrir a supuestos sobre las distribuciones involucradas. En este sentido podemos decir que es una herramienta *no paramétrica*.

Dado un conjunto de datos univariado y unimodal, $X_n = \{x_1, x_2, \ldots, x_n\}$, el boxplot se construye dibujando una línea a la altura de la **mediana**, Q_2, una **caja** limitada por el primer, Q_1, y tercer, Q_3, cuartil y **bigotes**[1], w_1 y w_2, desde los bordes de la caja hasta las observaciones más alejadas (de la mediana) que no superen los límites del intervalo:

$$\Big[Q_1 - 1.5\, IQR\,,\, Q_3 + 1.5\, IQR\Big]. \tag{2.1}$$

Si todos los datos cayeran dentro del intervalo, es decir que fueran todos regulares, los bigotes corresponderían al máximo y mínimo de la muestra. Aquellos datos que excedieran el intervalo, se marcarían y clasificarían como potenciales datos atípicos. Los bigotes

[1] Un nombre alternativo del boxplot es *box-and-whisker plot* o box-and-whisker diagram, o sea diagrama de cajas y bigotes.



(a) Mediana y caja. (b) Bigotes y outliers

Figura 2.1: Elementos constitutivos del *boxplot*: caja (en verde en a)) y bigotes (en azul en b)). Los valores que superan los bigotes se marcan como posibles datos atípicos (círculos).

resultan ser las observaciones más extremas dentro del intervalo. En la Figura 2.1 se ilustran, en color, las componentes del boxplot.

Otro ejemplo se presenta en la Figura 2.2 que muestra los boxplots correspondientes a cinco muestras de experimentos de medición de la velocidad de la luz[2]. El primer y tercer boxplot presentan potenciales datos atípicos marcados en el gráfico con un círculo.

Si bien no hay una definición precisa de dato atípico, los entendemos como aquellos que se desvían lo suficiente del resto de las observaciones para despertar la sospecha de haber sido generados por otro mecanismo. Más concretamente, supondremos que los atípicos son aquellos datos que provienen de otra población, es decir que siguen otra distribución.

La presencia de outliers podría enmascarar las características de la distribución de los datos regulares. Para minimizar su efecto, el boxplot se basa en medidas **robustas**, que no son sensibles a los outliers. De aquí que su construcción dependa de los cuartiles de la muestra. La posición se estima con la **mediana** mientras que para la dispersión se considera la **distancia intercuartil** dada por $IQR = Q_3 - Q_1$, donde Q_1 y Q_3 son el primer y tercer cuartil respectivamente. La longitud de la caja, la IQR, es una medida robusta de la escala de la muestra.

La regla definida por el intervalo dado en (2.1) para clasificar posibles datos atípicos depende del factor 1.5. Este valor es arbitrario pero queda justificado en términos de la

[2]Los datos fueron tomados del dataset de R: *morley*. Corresponden a un experimento realizado por el físico Albert Michelson en 1879 para determinar la velocidad de la luz.

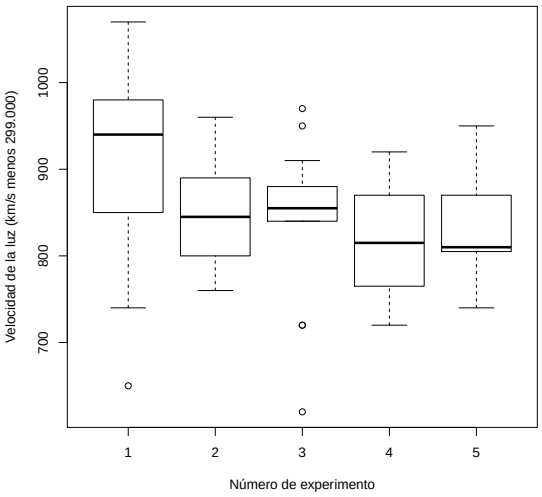


Figura 2.2: Cada diagrama se corresponde con una muestra del experimento realizado por Michelson para determinar la velocidad de la luz.

distribución normal. En caso de ser conocida la distribución de los datos observados sería posible cuantificar el alejamiento de los potenciales datos atípicos, su distancia al centro de la muestra, y establecer un valor de corte. Si las observaciones se distribuyeran como una normal estándar, $\mathcal{N}(0,1)$, los límites asintóticos del intervalo serían $L_1 = Q_1 - 1.5 \cdot IQR = -2.698$ y $L_2 = Q_3 + 1.5 \cdot IQR = 2.698$ de manera que la probabilidad de pertenecer al intervalo sería alta: 99.3%. Por el contrario, la probabilidad de que un dato regular sea detectado, erróneamente, como atípico es baja: 0.7%. Si el factor fuera, digamos, 2 en vez de 1.5, la probabilidad de que un dato sea declarado atípico sería aun menor: 0.07%. Esta regla, basada en el factor 1.5, se comporta bien frente a muestras de datos distribuídos normalmente. La Figura 2.3 muestra la relación entre el boxplot de una muestra con distribución normal estándar y los cuartiles de la distribución subyacente.

Los datos clasificados como potenciales datos atípicos no necesariamente son outliers reales, de aquí la insistencia con en hablar de **potenciales**. Veamos cómo se comporta el boxplot frente a una distribución de Cauchy estándar. Esta distribución también es simétrica como la normal pero tiene colas pesadas. Informalmente, no es tan poco probable caer en las colas de la distribución como en el caso normal. Dada una muestra generada con esta distribución, construimos su boxplot (Figura 2.4 b)) donde se observa que varios puntos son clasificados como outliers mientras que la muestra no fue contaminada con datos atípicos. Por el contrario, los datos generados de una distribución con colas livianas difícilmente excedan los límites del intervalo dado en (2.1).

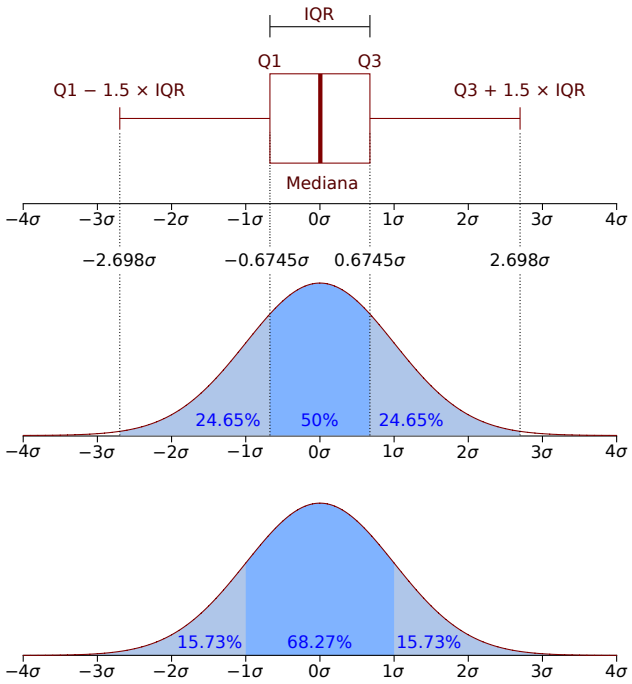


Figura 2.3: Relación entre las partes del boxplot y la función de densidad en el caso de una distribución $\mathcal{N}(0,1)$.

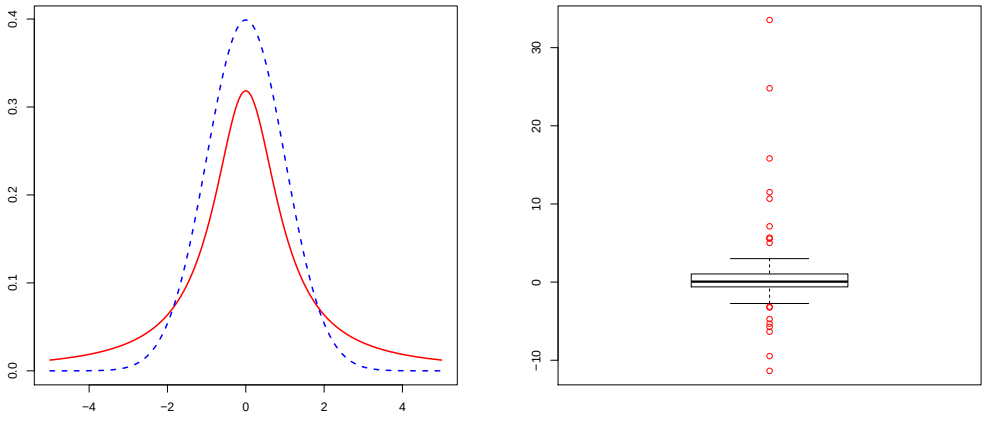
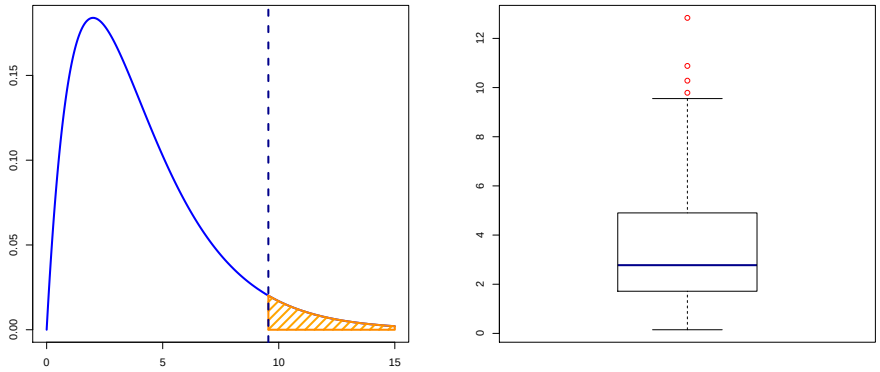


Figura 2.4: a) Gráficos de la Densidad Normal (azul) y Cauchy (rojo). b) Boxplot proveniente de una muestra de $n = 100$ datos con distribución Cauchy $\mathcal{C}(0,1)$.

Otro caso en el que la regla de detección no es satisfactoria corresponde a las distribuciones asimétricas. Si tomamos una muestra generada a partir de una distribución χ^2_2, la probabilidad de que un dato, generado con esa distribución, exceda el límite superior es aproximadamente 7.56%. Esto se ve en el boxplot de la Figura 2.5 donde nuevamente se clasifican de manera incorrecta varios puntos de una muestra de $n = 100$ observaciones. Nótese además que la mediana está *descentrada* respecto de la caja.



(a) Función de densidad de una distribución χ^2_2.

(b) El boxplot detecta varios outliers sobre el lado asimétrico de la densidad.

Figura 2.5: Densidad χ^2_2 y Boxplot correspondiente a una muestra de $n = 100$ datos. En una distribución asimétrica hacia la derecha la probabilidad de que un dato regular exceda el límite del intervalo de clasificación (línea punteada) es alta (área rayada).

El boxplot nos informa, además de la posición y la escala, la presencia de colas pesadas o asimetría a través de las distancias entre sus partes (caja y bigotes) y la clasificación excesiva de datos atípicos. Sin embargo esa aparente ventaja conspira contra la capacidad de detección de datos atípicos por parte del boxplot. Es necesario entonces introducir una modificación en el boxplot para contemplar estas características.

En este trabajo, nos ocuparemos de distribuciones para datos funcionales con asimetría. Por esta razón, haremos primero una descripción de la medida robusta de asimetría definida por Brys *et al.* (2003) para una muestra univariada que permitió a Hubert y Vandervieren (2007) dar una modificación en la construcción del boxplot y un nuevo criterio de detección de outliers en ese contexto.

2.2 Medidas robustas de asimetría

Una distribución de probabilidad se dice simétrica si y solo si existe un valor x_0 tal que

$$f(x_0 - \delta) = f(x_0 + \delta) \quad \forall\, \delta \in \mathbb{R}$$

donde f es la función de densidad si la distribución es continua o la función de probabilidad puntual si es discreta. Ejemplos de distribuciones simétricas son la distribución normal, la Cauchy y más generalmente, la distribución de Student.

Sin embargo, en el caso univariado, existe también una gran variedad de distribuciones asimétricas como la distribución Beta, la Gamma y en particular la Chi–cuadrado. En el Capítulo 3 describiremos una clase de distribuciones asimétricas amplia que permite su generalización al caso multivariado y funcional. Es, por lo tanto, importante en base a una muestra proveniente de una distribución desconocida determinar si la distribución subyacente es simétrica o no lo es.

Clásicamente, la asimetría de un conjunto de datos univariado, $X_n = \{x_1, x_2, \ldots, x_n\}$, generado a partir de una distribución continua se mide con el coeficiente de asimetría, b_1, dado por:

$$b_1(X_n) = \frac{m_3(X_n)}{m_2(X_n)^{3/2}}\,,$$

donde m_2 y m_3 denotan el segundo y tercer momento empírico de los datos, es decir, $m_k(X_n) = (1/n)\sum_1^n x_i^k$. El valor de b_1 se ve fuertemente afectado por datos atípicos. Un solo *outlier* en la cola de una distribución simétrica puede hacer b_1 negativo o aumentar su valor complicando su interpretación. En la Figura 2.2 vemos el boxplot de una muestra simétrica (normal estándar) con un claro outlier. Para el conjunto completo de datos, $b_1 = 1.69$, pero si quitamos el dato atípico, b_1 cae a 0.31.

La robustificación de estas medidas se hace, como en el boxplot, con los cuartiles de la distribución. En esa línea, Bowley (1920) propuso la *quartile skewness* o asimetría de cuartiles definida como:

$$QS = \frac{(Q_{0.75} - Q_{0.5}) - (Q_{0.5} - Q_{0.25})}{Q_{0.75} - Q_{0.25}}\,,$$

donde Q_i es el cuartil i-ésimo de X_n, esto es $Q_1 = F^{-1}(0.25)$, $Q_2 = F^{-1}(0.5)$ y $Q_3 = F^{-1}(0.75)$.

Es fácil ver que para cualquier distribución simétrica esta medida computa cero. El denominador $Q_3 - Q_1$ escala el coeficiente de manera que $QS \in [-1; 1]$.

El coeficiente definido por Bowley fue generalizado por Hinkley (1975) quien introdujo, para $\alpha \in [0; 0.5]$, el coeficiente $SK(\alpha)$ dado por

$$SK(\alpha) = \frac{\big(F^{-1}(1-\alpha) - Q_{0.5}\big) - \big(Q_{0.5} - F^{-1}(\alpha)\big)}{F^{-1}(1-\alpha) - F^{-1}(\alpha)}\,.$$

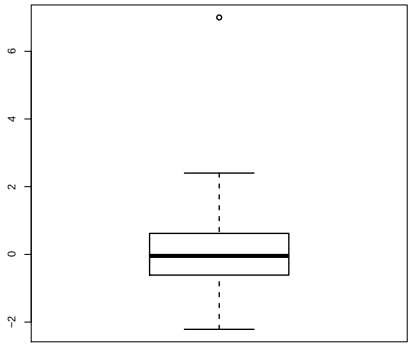


Figura 2.6: Boxplot de una muestra normal estándar con un outlier. El outlier tiene un gran impacto en el cómputo del coeficiente b_1.

Si $\alpha = 0.25$ se recupera el coeficiente QS como caso particular. Otra medida que volveremos a mencionar, el *octile skewness*, corresponde a $\alpha = 0.125$ y está dado por

$$OS = \frac{(Q_{0.875} - Q_{0.5}) - (Q_{0.5} - Q_{0.125})}{Q_{0.875} - Q_{0.125}}$$

La definición de OS usa más información de las colas de la distribución de manera tal que será más apropiado para detectar asimetría en los datos. Como contrapartida, lo hace más sensible a la presencia de datos atípicos. Por el contrario, QS es más resistente a outliers pero será menos propenso a detectar pequeñas asimetrías. Ambas medidas detectan la asimetría relevando las diferentes distancias entre algunos cuantiles de la muestra y la mediana. Un compromiso entre ambas características, buena detección y resistencia frente a outliers, se obtiene comparando la posición de todas las observaciones a uno y otro lado de la mediana, lo que nos lleva a la definición del *Medcouple* introducida por Brys *et al.* (2003).

2.2.1 Medcouple

Consideremos una *muestra* de n observaciones $X_n = \{x_1, \ldots, x_n\}$ provenientes de una distribución continua univariada F. Por conveniencia, supongamos que la muestra está ordenada: $x_1 \leq x_2 \leq \cdots \leq x_n$. Sea m_n la mediana de X_n definida como:

$$m_n = \begin{cases} \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{si } n \text{ es par} \\ x_{\frac{(n+1)}{2}} & \text{si } n \text{ es impar} \end{cases}.$$

En Brys *et al.* (2003), se define el medcouple, MC_n, como

$$MC_n = \underset{x_i \leq m_n \leq x_j}{\text{mediana}}\, h(x_i, x_j),$$

donde la función h se define para los valores $x_i \neq x_j$ como

$$h(x_i, x_j) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i}.$$

De esta forma, $h(x_i, x_j)$ provee una medida normalizada de la distancia de las observaciones que se encuentran a un lado y otro de la mediana de la muestra. Para los casos en que haya observaciones iguales a la mediana, $x_i = x_j = m_n$, se procede como sigue. Sean $m_1 < \cdots < m_k$ los índices de las observaciones iguales a la mediana, o sea, $x_{m_l} = m_n$ para todo $l = 1, \ldots, k$, entonces,

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

Miremos estos casos particulares. Cuando la mediana m_n coincide con un solo elemento de la muestra tenemos que $h(m_n, x_j) = +1$ para todo $x_j > m_n$, lo cual expresa que x_j se encuentra infinitamente más alejado de la mediana que m_n. Análogamente, $h(x_i, m_n) = -1$ para todo $x_i < m_n$. En este caso, la cantidad de elementos de la muestra estrictamente mayor a la mediana es la misma que los menores, luego se agregan tantos $+1$ como -1. De forma que el medcouple no se ve afectado por estos valores extremos.

En el caso de que más de un elemento de la muestra sea igual a la mediana podría ocurrir, por ejemplo, que haya más valores estrictamente mayores a la mediana que menores, así se incluirían más $+1$ que -1. Notemos que se incluirían tantos ceros como elementos coincidentes con la mediana haya (basta contar los índices). Esto atrae al medcouple hacia cero, que se corresponde con la idea, intuitiva, de que varios valores coincidentes con la mediana disminuyen la asimetría de una distribución.

Podemos considerar también la versión poblacional del medcouple, $MC(F)$, para una distribución continua F. Sea $m_F = F^{-1}(0.5)$ la mediana poblacional de F, entonces la definición de $MC(F)$ se sigue de manera análoga al caso muestral tratado antes:

$$MC(F) = \underset{x_1 \leq m_F \leq x_2}{\text{mediana}}\, h(x_1, x_2)$$

con x_1, x_2 independientes $x_i \sim F$.

La función h es la misma que antes si reemplazamos la mediana muestral m_n por m_F. A través de la función indicadora $\mathbb{I}_A(u) = 1$ si $u \in A$ y cero en otro caso, podemos escribir

$$\mathbb{P}(h \leq u) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbb{I}_{(-\infty, u]}(h(x_1, x_2))\, dF(x_1) dF(x_2).$$

Por simetría respecto de m_F, la integral se reescribe como

$$H_F(u) = 4 \int_{m_F}^{+\infty} \int_{-\infty}^{m_F} \mathbb{I}_{(-\infty,u]}(h(x_1, x_2))\, dF(x_1) dF(x_2), \tag{2.2}$$

y obtenemos una formulación más compacta

$$MC(F) = H_F^{-1}(0.5)\ .$$

Observemos por un lado que el dominio de H_F es $[-1, 1]$ y que las condiciones

$$h(x_1, x_2) \leq u, \qquad x_1 \leq m_F \leq x_2$$

son equivalentes a

$$x_1 \leq \frac{x_2(u-1) + 2m_F}{u+1}\ ,\ x_2 \geq m_F.$$

Esto permite simplificar la expresión dada en (2.2)

$$H_F(u) = 4 \int_{m_F}^{+\infty} F\left(\frac{x_2(u-1) + 2m_F}{u+1}\right) dF(x_2)\ .$$

La expresión anterior se usa para probar que la función de influencia de $MC(F)$ está acotada sobre la cual nos referiremos más adelante.

2.2.2 Algunas propiedades del Medcouple

Una primera ventaja del medcouple sobre el coeficiente clásico b_1 es que el $MC(F)$ puede ser calculado para distribuciones sin momentos finitos. Además, una medida de asimetría tiene que verificar algunas propiedades naturales que enunciamos a continuación para el medcouple. Sea X una variable aleatoria con una distribución continua F_X.

Teorema 2.1. *MC es invariante por traslaciones y cambios de escala:*

$$MC(F_{aX+b}) = MC(F_X)$$

para cualquier $a > 0$ y $b \in \mathbb{R}$.

Teorema 2.2. *Si invertimos una distribución, MC también se invierte, es decir, $MC(F_{-X}) = -MC(F_X)$.*

Teorema 2.3. *Si F es simétrica entonces $MC(F) = 0$.*

Las demostraciones de los Teoremas 2.1, 2.2 y 2.3 son inmediatas a partir de la definición de $MC(F)$.

El Teorema 2.4 muestra que MC respeta cierta noción de orden entre distribuciones.

Definición 2.1. Sean F y G distribuciones continuas con soporte en un intervalo. Decimos que G es al menos tan asimétrica hacia la derecha como F y lo indicaremos como $F <_c G$, si $G^{-1}(F(x))$ es convexa en el soporte de F.

Teorema 2.4. *Si $F <_c G$ entonces $MC(F) \leq MC(G)$.*

Demostración. Sin pérdida de generalidad podemos suponer que $F^{-1}(0.5) = G^{-1}(0.5) = 0$. Tenemos que ver que $H_F \leq H_G$ con

$$H_F(u) = 4 \int_0^{+\infty} \int_{-\infty}^{0} I(h(x_1, x_2) \leq u) dF(x_1) dF(x_2)$$
$$H_G(u) = 4 \int_0^{+\infty} \int_{-\infty}^{0} I(h(y_1, y_2) \leq u) dG(y_1) dG(y_2) \ .$$

Como F y G tienen soporte en un intervalo, su función cuantil es estríctamente monótona. Entonces, para cualquier par (x_1, x_2) con $x_1 \leq 0 \leq x_2$ podemos encontrar un único par (y_1, y_2) con $y_1 \leq 0 \leq y_2$ tal que

$$x_1 = F^{-1}(p) \quad x_2 = F^{-1}(q) \quad y_1 = G^{-1}(p) \quad y_1 = G^{-1}(q) \ ,$$

con $p \in \left[0, \frac{1}{2}\right]$ y $q \in \left[\frac{1}{2}, 1\right]$. Basta con mostrar que

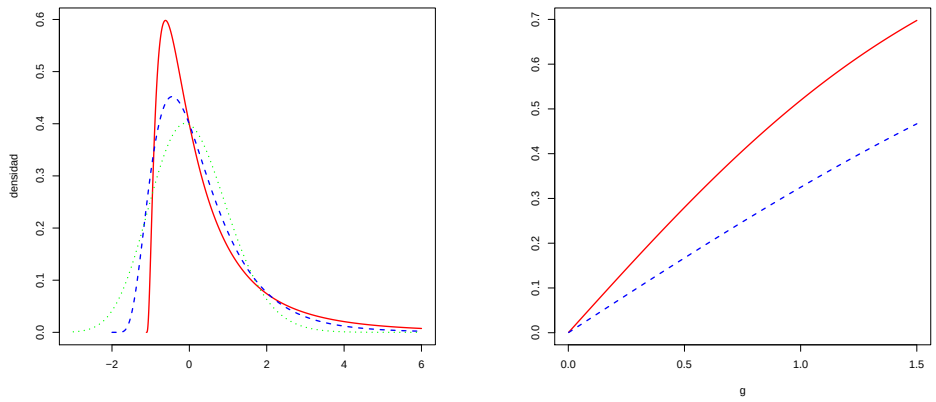
$$\frac{F^{-1}(q) + F^{-1}(p)}{F^{-1}(q) - F^{-1}(p)} \leq \frac{G^{-1}(q) + G^{-1}(p)}{G^{-1}(q) - G^{-1}(p)} \ .$$

Groeneveld y Meeden (1984) probaron que esta desigualdad se satisface si $F <_c G$ y $p + q = 1$. La misma prueba se extiende para $p + q < 1$. □

Una familia de distribuciones que verifica este ordenamiento es la formada por son aquellas de la forma $Y_g = (\exp(gZ) - 1)/g$ con $g \in \mathbb{R}$ y Z una distribución gaussiana. Para $g = 0$, definimos $Y_0 = Z$ que resulta simétrica. A su vez, $F_{Y_{-g}}(x) = 1 - F_{Y_g}(-x)$, luego basta considerar sólo las asimétricas hacia la derecha, o sea, cuando $g > 0$. Para esta familia de distribuciones se verifica que si $g_1 \leq g_2$ entonces $G_{g_1} \leq_c G_{g_2}$.

Con esta familia paramétrica podemos comparar los valores de las tres medidas de asimetría, QS,OS y MC. La Figura 2.7a) muestra el gráfico de las densidades correspondientes a varias distribuciones G_g con g variando de 0 a 1.5.

La Figura 2.7b) muestra el gráfico de OS y QS contra el parámetro g. En el trabajo de Brys *et al.* (2004) puede consultarse un gráfico comparativo también para MC que la ubica entre los gráficos de OS y QS. Por un lado, vemos que respetan la monotonía dentro de la familia de distribuciones. Por otra parte, observamos que si bien no estiman la misma cantidad, todos reflejan el grado de asimetría de los datos.



(a) Densidad de las distribuciones de Y_g para $g = 0.1$ (línea punteada), $g = 0.5$ (línea rayada) y $g = 0.9$ (línea completa).

(b) Relación monótona entre el parámetro g y QS (en azul) y OS (en rojo) para las distribuciones G_g.

Figura 2.7: Densidad de las distribuciones de Y_g y relación con las medidas de asimetría.

2.2.3 Robustez del Medcouple

Punto ruptura

El punto de ruptura de un estimador T_n dada una muestra $X_n = \{x_1, \ldots, x_n\}$ mide cuántas observaciones entre $x_1, \ldots, x_n$ deben ser modificadas para que el estimador pierda significado. En el caso de un estimador de posición univariado, por ejemplo, el punto de ruptura indica la proporción de observaciones que el estimador tolera sin que el valor absoluto del estimador se vuelva arbitrariamente grande.

Dado que $MC(F) \in [-1, 1]$, el punto de ruptura de una muestra queda definido como:

$$\epsilon_n^*(MC_n; X_n) = \min \left\{ \frac{m}{n} \quad \text{tales que} \quad \sup_{X_n'} \left| MC_n(X_n') \right| = 1 \right\},$$

donde el conjunto X_n' se obtiene reemplazando m observaciones de X_n por valores arbitrarios.

Puede demostrarse que el punto de ruptura de OS_n es 12.5% y QS_n es 25%. Probemos ahora que MC_n resiste también hasta un 25% de outliers en los datos.

Teorema 2.5. *Sea X_n una muestra tal que no hay elementos iguales. Entonces,*

$$\frac{1}{n}\left(\left\lceil \frac{n}{4} \right\rceil - 1\right) \leq \epsilon_n^*(MC_n; X_n) \leq \frac{1}{n}\left(\left\lceil \frac{n}{4} \right\rceil + 1\right) \tag{2.3}$$

donde $\lceil x \rceil$ indica la parte entera del número real x.

Demostración. Primero vamos a probar que $\epsilon_n^* \leq (\lceil n/4 \rceil + 1)/n$. Como MC_n es invariante por traslaciones, podemos suponer sin pérdida de generalidad que $m_n(X_n) = 0$. Por simetría, también podemos suponer que $MC(X_n) \geq 0$. Sea B tal que $MC(X_n) < B < 1$. Debemos mostrar ahora que podemos construir una muestra X_n' reemplazando $\lceil n/4 \rceil + 1$ datos de X_n tal que $MC(X_n') > B$. Para esto, sumemos una constante $k > 2 \max |x_i|/(1 - B)$ a los $\lceil n/4 \rceil + 1 = n - \lceil 3n/4 \rceil + 1$ valores más grandes de la muestra original X_n, es decir, definimos

$$x_i' = \begin{cases} x_i & i = 1, \ldots, \lceil \frac{3n}{4} \rceil - 1 \\ x_i + k & i = \lceil \frac{3n}{4} \rceil, \ldots, n \end{cases} .$$

Luego, $m_n(X_n') = m_n(X_n)$ y para todos los $x_i \leq m_n$ resulta

$$h(x_i, x_j)' = \begin{cases} h(x_i, x_j) & j = 1, \ldots, \lceil \frac{3n}{4} \rceil - 1 \\ \frac{x_j + x_i + k}{x_j - x_i + k} & j = \lceil \frac{3n}{4} \rceil, \ldots, n \end{cases} .$$

Por otro lado,

$$\frac{x_j + x_i + k}{x_j - x_i + k} > B \Leftrightarrow k > \frac{x_j(B-1) - x_i(B+1)}{1 - B} . \tag{2.4}$$

Entonces, si $x_i < x_j$ obtenemos $h(x_i, x_j') > B$ para cada $j \geq \lceil 3n/4 \rceil$. Como $i \leq \lceil n/2 \rceil$ al menos $\lceil n/2 \rceil (\lceil (n/4) + 1)$ de los $h(x_i, x_j')$ son mayores que B.

Como X_n no tiene elementos iguales, tampoco los tiene X_n'. Por lo tanto, para n par el medcouple está definido como la mediana sobre $(n/2)^2$ números mientras que para n impar, la mediana se toma sobre $(n + 1/2)^2$ números. De donde, $MC(X_n')$ será mayor que B porque al menos $\lceil n^2/8 \rceil + 1$ para n par, respectivamente $\lceil (n+1)^2/8 \rceil + 1$ para n impar, de los $h(x_i, x_j')$ son mayores que B.

Probemos la otra desigualdad: $\epsilon_n^* \geq (\lceil n/4 \rceil - 1)/n$. Reemplacemos ahora $k < \lceil n/4 \rceil - 1$ puntos de la muestra por valores arbitrarios x_i'. Mostraremos que el medcouple de la muestra contaminada sigue dependiendo de los datos originales y que consecuentemente su valor absoluto es menor a 1. Llamemos a la mediana de la nueva muestra m_n, a a la cantidad de valores originales en la muestra que están a la izquierda de m_n y b los que están a la derecha. Es claro que $a + b \geq \lceil 3n/4 \rceil + 2$. Más aún, si n es par, entonces

$$\frac{n}{4} + 1 \leq \min\{a, b\} \quad \text{y} \quad \max\{a, b\} \leq \frac{n}{2}$$

mientras que para los n impares

$$\frac{n+1}{4} + 1 \leq \min\{a, b\} \quad \text{y} \quad \max\{a, b\} \leq \frac{n+1}{2} .$$

El número de expresiones no contaminadas $h(x_i, x_j)$ es $ab \geq a\left(\lceil 3n/4 \rceil + 2 - a\right)$. Esta cota inferior es estrictamente mayor que $\left[(n^2/4+1)/2\right]$ para n par y $\left[\left((n+1)^2/4+1\right)/2\right]$ para n impar, entonces el medcouple se obtiene como el promedio de uno o dos de estos núcleos no contaminados. □

Función de influencia

Sea $T : \mathcal{P} \to \mathbb{R}$ un funcional sobre el espacio de la probabilidades $\mathcal{P}$ y F_n la distribución empírica asociada a una muestra $x_1, \ldots, x_n$. La función de influencia de un estimador $T_n = T(F_n)$ sobre cierta distribución F mide el efecto sobre T al agregar un pequeño porcentaje de datos en el punto x. Si Δ_x es la masa puntual de x, luego la función de influencia se define como:

$$IF(x, T, F) = \lim_{\epsilon \searrow 0} \frac{T((1-\epsilon)F + \epsilon\Delta_x) - T(F)}{\epsilon} .$$

Teorema 2.6. *Sea F una distribución absolutamente continua con densidad f tal que $MC_F \neq -1$,$f(m_F) \neq 0$ y $H'_F(MC_F) \neq 0$, entonces*

$$\begin{aligned} IF(x, MC, F) =& \frac{1}{H'_F(MC_F)} \Bigg[1 - 4F(g_1(x))\mathbb{I}_{(m_F, +\infty)}(x) - 4(F(g_2(x)) - 0.5)\mathbb{I}_{(-\infty, m_F)}(x) \\ & + signo(x - m_F)\left(1 - \frac{4}{f(m_F)(MC_F + 1)} \int_{m_F}^{+\infty} f(g_1(w))dF(w)\right) \Bigg] . \end{aligned}$$

La demostración de este resultado puede verse en Brys *et al.* (2004).

A partir de este teorema se deduce que la función de influencia del medcouple es acotada en comparación con la de la medida clásica de asimetría b_1. Las funciones de influencia de QS y OS son acotadas también.

2.2.4 Un algoritmo eficiente para el cálculo del Medcouple

Una primera aproximación en el desarrollo de un algoritmo para el cálculo del medcouple es la evaluación de la función núcleo $h(x_i, x_j)$ para cada $x_i \leq m_n \leq x_j$. Este enfoque demanda $O(n^2)$ operaciones, lo cual se torna lento para conjuntos de datos demasiado grandes. Brys *et al.* (2004) introdujeron un algoritmo con menor tiempo de cómputo, para ser más preciso, el algoritmo calcula el medcouple con velocidad $O(n \log n)$. Por completitud describimos a continuación el procedimiento.

Supongamos que $X_n = \{x_1, \ldots, x_n\}$ es la muestra observada de una distribución continua univariada. El pseudo–código del algoritmo es el siguiente:

1. Ordenar la muestra de menor a mayor. Hay algoritmos eficientes como el Quicksort o Mergesort que lo hacen en tiempo $O(n \log n)$.

2. Para facilitar la notación y por estabilidad numérica, transformamos los datos restando la mediana m_n a X_n. Esto puede hacerse porque MC_n es invariante por traslaciones. Sea $Z_n = X_n - m_n$ la muestra corrida, luego la función núcleo se reduce a:
$$h(z_i, z_j) = \frac{z_j + z_i}{z_j - z_i}\,.$$
Sean Z^- y Z^+ definidos como:
$$Z^- = \{z_i^- = z_k \in Z_n; z_k \leq 0\} \qquad Z^+ = \{z_j^- = z_l \in Z_n; z_l \geq 0\}\,.$$
Tanto Z^- como Z^+ permanecen ordenados. Sea p (resp. q) el cardinal de Z^- (resp. Z^+).

3. Supongamos primero que no hay observaciones que coincidan con la mediana. Consideremos luego la siguiente matriz de $q \times p$ para cada $i = 1, \ldots, p$ y $j = 1, \ldots, q$ contiene $h(z_i^-, z_j^+)$ en la columna i y la fila j
$$\begin{pmatrix} h(z_1^-, z_1^+) & \ldots & h(z_p^-, z_j^+) \\ \vdots & \ddots & \vdots \\ h(z_1^-, z_q^+) & \ldots & h(z_p^-, z_q^+) \end{pmatrix}\,.$$
Usando la definición de h y el orden de Z^- y Z^+ se verifica que
$$h(z_i^-, z_j^+) \geq h(z_{i+1}^-, z_j^+)\,,$$
para cada $i = 1, \ldots, p-1$ y $j = 1, \ldots, q$ y
$$h(z_i^-, z_j^+) \geq h(z_i^-, z_{j+1}^+)\,,$$
para cada $i = 1, \ldots, p$ y $j = 1, \ldots, q-1$. Luego se obtiene el siguiente esquema:
$$\begin{pmatrix} h(z_1^-, z_1^+) & \overset{\geq}{\rightarrow} & h(z_p^-, z_j^+) \\ \downarrow \geq & \searrow^{\geq} & \downarrow \geq \\ h(z_1^-, z_q^+) & \overset{\geq}{\rightarrow} & h(z_p^-, z_q^+) \end{pmatrix}\,.$$
Notemos que no es necesario calcular todos los valores de la tabla que sería $O(n^2)$ sino solo los que específicamente se usarán en el paso 4 del algoritmo.

4. Aplique el algoritmo de Johnson y Mizoguchi (1978), que encuentra en tiempo $O(n \log n)$ el $k-$ésimo estadístico de orden en una tabla $(x_i + y_j)_{i,j}$ con vectores ordenados $\mathbf{x} = (x_1, \ldots, x_n)$ $\mathbf{y} = (y_1, \ldots, y_n)$. Esencialmente, solamente se usa la monotonía en las filas, columnas y diagonales de la tabla y esta condición se cumple en la tabla construída en el paso 3, luego encontramos la mediana en tiempo $O(n \log n)$.

Como conclusión general podemos decir que MC_n combina la sensibilidad de OS_n con las robustez de QS_n frente a datos atípicos. Sumado lo anterior al algoritmo eficiente para su cálculo, de complejidad $O(n \log n)$, estas razones convierten al medcouple en una interesante medida robusta de asimetría.

2.3 Boxplot ajustado

El boxplot clásico tratado en la Sección 2.1 provee un criterio de detección de outliers que se ajusta al caso de datos distribuídos de manera normal. Si la distribución no fuera simétrica, la regla del factor 1.5 no toma en cuenta la asimetría alterando el porcentaje de detección de outliers. En la Figura 2.8 se ve cómo el bigote superior del boxplot (línea punteada) no acompaña a la cola derecha de la distribución.

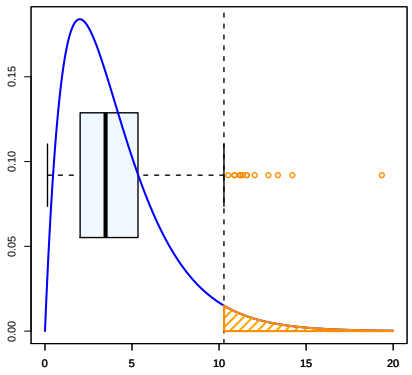


Figura 2.8: La asimetría de la distribución no se contempla en la regla de detección del boxplot clásico. Varios datos regulares son clasificados como datos atípicos en la muestra de tamaño 500 generada con una distribución χ^2_2.

Para obtener un criterio que se adapte al caso asimétrico es necesario incluir una medida de asimetría de la muestra y con ella modificar la definición de los bigotes del boxplot. Hubert y Vandervieren (2008) abordan este problema introduciendo la asimetría de la

muestra en la definición de los bigotes del boxplot clásico a través del medcouple presentado en la Sección 2.2.

Se definen primero dos funciones $h_l(MC)$ y $h_u(MC)$ las cuales deben verificar que $h_l(0) = h_u(0) = 1.5$ para recuperar la regla original en el caso de tratar con una distribución simétrica ($MC = 0$). De esta manera el nuevo intervalo queda:

$$\left[Q_1 - h_l(MC)\ IQR\,;\, Q_3 + h_u(MC)\ IQR\right].$$

Observemos que como el medcouple es invariante por traslaciones y cambios de escala, el intervalo resulta equivariante. En Hubert y Vandervieren (2008), se estudian tres modelos diferentes para las funciones h_l y h_u

a) lineal

$$h_l(MC) = 1.5 + aMC$$
$$h_u(MC) = 1.5 + bMC$$

b) cuadrático

$$h_l(MC) = 1.5 + a_1MC + a_2MC^2$$
$$h_u(MC) = 1.5 + b_1MC + b_2MC^2$$

c) exponencial

$$h_l(MC) = 1.5e^{aMC}$$
$$h_u(MC) = 1.5e^{bMC}$$

con $a, a_1, a_2, b, b_1, b_2 \in \mathbb{R}$. Para determinar los valores de las constantes estos autores simulan muestras libres de outliers generadas de distribuciones asimétricas y ajustan los valores de las constantes para que el porcentaje de outliers clasificados por los bigotes del boxplot sea 0.7% (que coincide con la regla de detección para el boxplot clásico para la distribución normal). Se utilizaron 12605 distribuciones de las familias Γ, χ^2, F, $Pareto$ y G_g cuyos parámetros se eligieron para que el medcouple no fuera mayor que 0.6. Los detalles del análisis pueden consultarse en Hubert y Vandervieren (2008). Los autores concluyen que el modelo que mejor se comporta bajo distribuciones que no son extremadamente asimétricas es el exponencial y fijan los parámetros como $a = -4$ y $b = 3$. Entonces, para $MC \geq 0$, todas las observaciones fuera del intervalo

$$\mathcal{I} = \left[Q_1 - 1.5e^{-4MC}IQR; Q_3 + 1.5e^{3MC}IQR\right],$$

serán marcadas como potenciales outliers, mientras que cuando $MC \leq 0$ el intervalo correspondiente es

$$\mathcal{I} = \left[Q_1 - 1.5e^{-3MC} IQR; Q_3 + 1.5e^{4MC} IQR \right],$$

y las observaciones fuera es $\mathcal{I}$ serán marcadas como potenciales outliers. Observemos que bajo una distribución simétrica, $MC = 0$, el intervalo definido se reduce al caso del boxplot clásico.

En el siguiente ejemplo consideramos las mediciones de la velocidad del viento (en millas por hora) para 156 días consecutivos provistas en el dataset *airquality* del paquete *datasets* de *R*. Como $MC = 0.048$ es decir, el conjunto presenta baja asimetría, ambas versiones del boxplot son similares como se observa en la Figura 2.9. Notemos que el bajo valor del MC modifica ligeramente los límites de clasificación pero no afecta a los bigotes ya que son las observaciones más extremas no atípicas. Podemos, por lo tanto, considerar el boxplot ajustado como una generalización del boxplot clásico para distribuciones asimétricas.

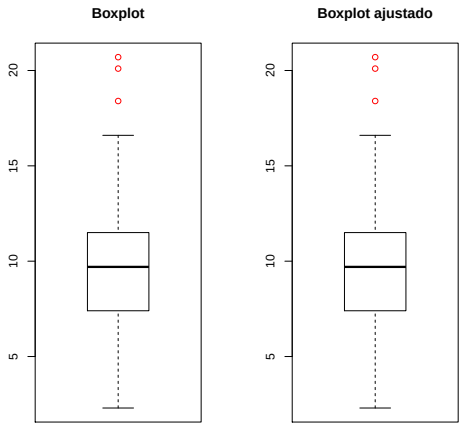
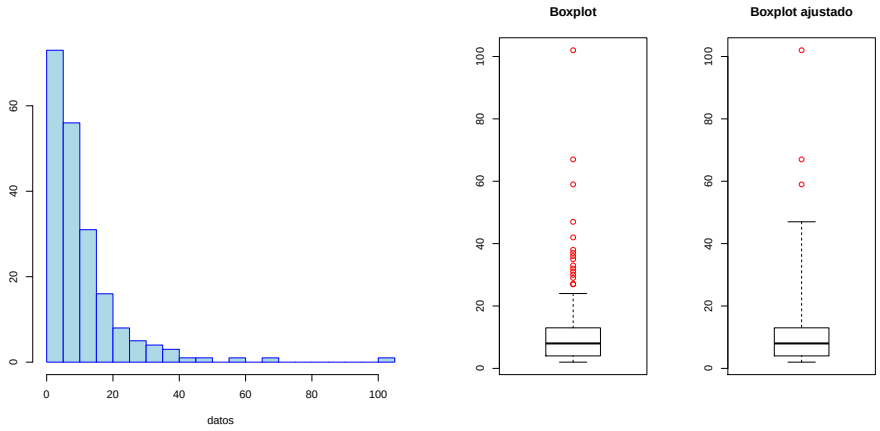


Figura 2.9: Boxplot clásico y ajustado para el conjunto *airquality*.

Ilustremos el desempeño del boxplot ajustado en un caso asimétrico. El siguiente ejemplo corresponde a los datos de 201 pacientes, ingresados en el Hospital Universitario de Lausanne en el año 2000. Los datos están disponibles en el dataset *LOS* ("length of stay") del paquete *robustbase*. Un objetivo natural sería estimar y predecir el consumo total de recursos de este grupo de pacientes. Para este propósito, podemos enfocarnos en la variable "duración de la estadía", *LOS*, que es un indicador fácilmente medible de la actividad del hospital. Un estimador natural del valor esperado de la variables es el promedio. Sin embargo, la distribución subyacente de la variable *LOS* tiene dos características que

vuelven cuestionable el uso de este estimador. Primero la distribución es asimétrica hacia la derecha, como se ve en el histograma de la Figura 2.10(a). Además, tres observaciones están claramente separadas de la mayoría y podrían considerarse como outliers de la muestra. El boxplot ajustado se muestra en la Figura 2.10(b) que permite compararlo con el boxplot clásico. Queda claro cómo el boxplot clásico clasifica por exceso los datos atípicos mientras que el ajustado considera la asimetría y los bigotes se ajustan para clasificar sólo los tres datos observados en el histograma. Tomando el promedio de las observaciones dentro de los bigotes del boxplot ajustado se obtiene una mejor estimación del tiempo esperado de estadía en el hospital.



(a) Histograma de la variables *LOS* para 201 datos.

(b) Boxplot clásico (izq.) y ajustado (der.) para la misma muestra.

Figura 2.10: El boxplot ajustado del conjunto *LOS* modifica los bigotes incorporando la asimetría de la muestra. Detecta los tres datos atípicos que aparecen en el histograma.

La construcción del boxplot ajustado está implementada en la rutina *boxadj* del paquete *robustbase* desarrollado por Vandervieren, disponible en el repositorio de paquetes de R[3].

Los resultados sobre datos simulados y reales sugieren que el uso del boxplot ajustado para distribuciones asimétricas alcanza mejor distinción entre observaciones regulares y atípicas. Esto convierte al boxplot ajustado en una herramienta interesante y rápida para la detección de outliers sin recurrir a supuestos adicionales sobre la distribución de los datos.

[3]http://cran.r-project.org/web/packages/robustbase/index.html

Capítulo 3

Distribución normal asimétrica

En muchas aplicaciones, se sabe que los datos observados no siguen una distribución simétrica. Para su modelado es necesario contar con familias de distribuciones que tengan cierta versatilidad para adaptarse a ellos. Idealmente, desearíamos tener disponibles familias reguladas por pocos parámetros con alta flexibilidad de forma (asimetría, kurtosis y, en el caso multivariado, estructura de dependencia) que a su vez sean sencillas en su tratamiento matemático y gocen de buenas propiedades (e.g. cerradas bajo marginalización). Más aun, en el contexto multivariado el número de familias paramétricas es notablemente menor que en el caso univariado. En un trabajo fundamental, Azzalini (1985) propone un método que genera nuevas familias de distribuciones a partir de una *perturbación* de densidades simétricas.

Lema 3.1. *Sea f_0 una función de densidad simétrica alrededor de 0 y G una distribución absolutamente continua tal que $g = G'$ es simétrica alrededor de 0. Entonces, la función $f(z) = 2G(\lambda z)f_0(z)$, $z \in \mathbb{R}$ es una función de densidad para cualquier $\lambda \in \mathbb{R}$.*

Demostraremos una versión ligeramente más general que se enuncia en Azzalini (2005).

Lema 3.2. *Sea f_0 una función de densidad simétrica alrededor de 0 y G una distribución absolutamente continua tal que $g = G'$ es simétrica alrededor de 0. Entonces, la función*

$$f(z) = 2G(w(z))f_0(z) \qquad z \in \mathbb{R} \tag{3.1}$$

es una función de densidad para cualquier función impar $w : \mathbb{R} \to \mathbb{R}$.

Demostración. Sean $Y \sim f_0$ y $X \sim g$ variables aleatorias independientes. Por hipótesis, como w es impar, tanto $w(Y)$ como $X - w(Y)$ tienen distribuciones simétricas respecto de 0, de donde

$$\frac{1}{2} = \mathbb{P}(X - w(Y) \leq 0) \quad = \quad \mathbb{E}_Y\left[\mathbb{P}(X - w(Y) \leq 0|Y)\right] = \mathbb{E}G(w(Y)) = \int_{\mathbb{R}} G(w(z))f_0(z)dz$$

lo que concluye la demostración. □

La Figura 3.1 ilustra el funcionamiento del Lema 3.2 tomando G como la función de distribución de una normal estándar y como densidad f_0 la densidad de una distribución de Student con dos grados de libertad, $\mathcal{T}_2$. Se eligieron como funciones w las funciones $w(x) = x$ y $w(x) = x^3 - 2x^2$. Se parte de una densidad *base* f_0 y se modifica su forma con el factor $G(w(z))$ resultando en una nueva densidad. La libertad de elección de la función de base y los ingredientes de la perturbación, G y w, permiten generar un gran número de nuevas familias de distribuciones. Un ejemplo inmediato y de suma interés que trataremos en el resto del capítulo es la generalización de la distribución normal. Este mismo método puede aplicarse a la distribución de Student, de Cauchy o, en el caso multivariado, a cualquier distribución elíptica (Azzalini, 2005).

Un resultado interesante que se conecta con el Lema 3.2 es el siguiente.

Proposición 3.1. *Si $X \sim g$ y $Y \sim f_0$ son variables independientes y $Z = Y|(X \leq w(Y))$, entonces $Z \sim f$ donde f está dada por (3.1).*

Demostración. Llamemos $W = \mathbb{I}_{X \leq w(Y)}$ y calculemos la densidad de Z. Tenemos que

$$f_{Y|W}(y|W=1) = \frac{\mathbb{P}(W=1|Y=y)f_Y(y)}{\mathbb{P}(W=1)} = \frac{\mathbb{P}(X \leq w(y))f_0(y)}{1/2} = 2G(w(y))\ f_0(y)$$

□

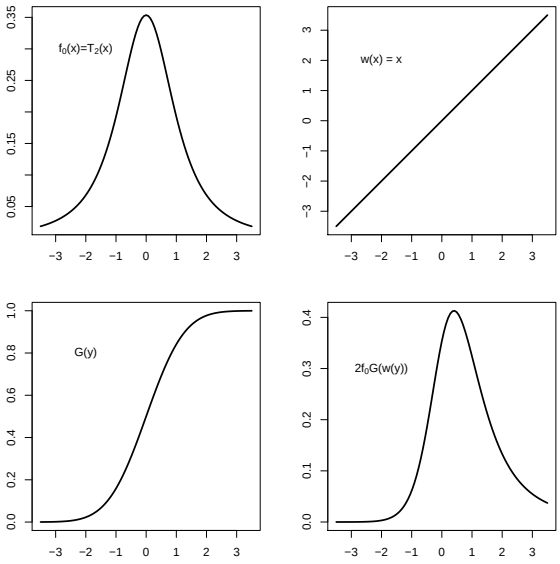
Este resultado da una representación estocástica de la nueva distribución que permite generar números aleatorios. Una variante más eficiente desde el punto de vista práctico que aprovecha la simetría de las densidades es definir

$$Z = \begin{cases} Y & \text{si } X \leq w(Y) \\ -Y & \text{si no}\,, \end{cases} \tag{3.2}$$

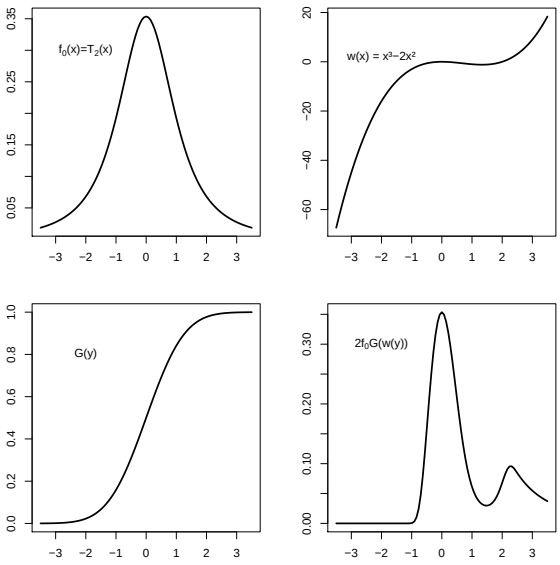
evitando así rechazar resultados.

El trabajo de Azzalini (1985) disparó un interés en la generación de nuevas familias de distribuciones que continúa hasta hoy. En Azzalini (2005) se presentan extensiones de familias elípticas y formulaciones semiparamétricas superando el interés solo de incorporar asimetría en las distribuciones.

En lo que sigue centraremos nuestra atención en la extensión de la familia normal para incorporar un parámetro de asimetría, describiremos el caso univariado, la extensión propuesta por Azzalini y Dalla Valle (1996) al caso multivariado y adaptaciones al contexto funcional que utilizaremos para nuestras simulaciones del Capítulo 7.



(a) $w(x) = x$



(b) $w(x) = x^3 - 2x^2$

Figura 3.1: Generación de distribuciones por el método de perturbación tomando como densidad f_0 la densidad de una distribución de Student con dos grados de libertad, $\mathcal{T}_2$. En ambos casos, $G = \Phi$, la distribución normal estándar.

3.1 Normal asimétrica univariada

Si en el Lema 3.2 tomamos $f_0 = \phi$ y $G = \Phi$, la densidad y la función de distribución de una normal estándar $\mathcal{N}(0,1)$, respectivamente y $w(x) = \lambda x$, $\lambda \in \mathbb{R}$, obtenemos la densidad

$$\phi(z;\lambda) = 2\phi(z)\Phi(\lambda z) \qquad z \in \mathbb{R} \tag{3.3}$$

que se denomina *normal asimétrica* ($\mathcal{SN}$) con parámetro de forma λ y se nota $Z \sim \mathcal{SN}(\lambda)$.

Por otra parte, para permitir cambios de posición y escala, dada $Z \sim \mathcal{SN}(\lambda)$, diremos que $Y \sim \mathcal{SN}(\xi, \omega^2, \lambda)$ si $Y = \xi + \omega Z$ donde $\xi \in \mathbb{R}$, $\omega \in \mathbb{R}^+$.

El parámetro λ regula la asimetría y varía en $(-\infty, \infty)$. En la Figura 3.2 se muestra la forma de la densidad para algunos valores del parámetro. Se muestran solamente los gráficos correspondientes a valores positivos de λ dado que si $Z \sim \mathcal{SN}(\lambda)$ entonces $-Z \sim \mathcal{SN}(-\lambda)$. Observemos además que esta familia de densidades incluye a la normal estándar cuando $\lambda = 0$.

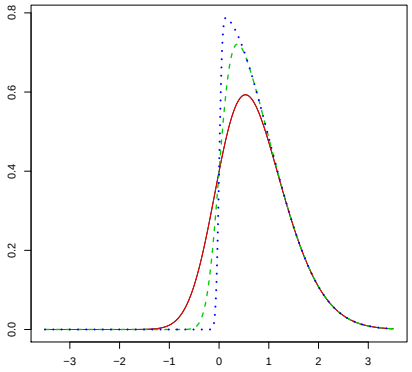


Figura 3.2: Función de densidad de $\mathcal{SN}(\lambda)$ para valores de $\lambda = 2$ (rojo), 5 (verde) y 20 (azul).

Una característica que hace interesante el estudio de esta familia de distribuciones es la cantidad de buenas propiedades y las expresiones simples de esperanza y varianza que posee. Enunciamos a modo de ejemplo algunas propiedades que pueden encontrarse en Azzalini (2005) donde también se enuncian otras propiedades sobre estas familias.

1. Sea $Z \sim \mathcal{SN}(\lambda)$ entonces

   (a) $\mathbb{E}(Z) = \delta\sqrt{2/\pi}$.

(b) $\text{VAR}(Z) = 1 - (2/\pi)\delta^2$, donde $\delta(\lambda) = \lambda/\sqrt{1+\lambda^2}$. Definamos además $\lambda(\delta) = \delta/\sqrt{1-\delta^2}$.

(c) La definición (3.2) permite mostrar fácilmente que $Z^2 \sim \chi^2_1$.

2. Sean $X \sim \mathcal{N}(0,1)$ y $Z_n \sim \mathcal{SN}(\lambda_n)$ donde $\lambda_n \to +\infty$. Luego, $Z_n \xrightarrow{D} |X|$.

   En forma análoga, se obtiene que si $Z_n \sim \mathcal{SN}(\lambda_n)$ donde $\lambda_n \to -\infty$ entonces, $Z_n \xrightarrow{D} -|X|$

Una representación alternativa a la dada en la Proposición 3.1 se basa en la distribución condicional de una normal bivariada. Indiquemos con $\mathbf{0}_p \in \mathbb{R}^p$ y $\mathbf{1}_p \in \mathbb{R}^p$ a los vectores con todas su componentes iguales a 0 y 1 respectivamente, y por $\mathbf{I}_p \in \mathbb{R}^{p\times p}$ a la matriz identidad.

Proposición 3.2. *Sea $(X,Y)^{\mathrm{T}}$ un vector normal bivariado $\mathcal{N}_2(\mathbf{0}_2, \delta\mathbf{I}_2 + (1-\delta)\mathbf{1}_2\mathbf{1}_2^{\mathrm{T}})$, es decir, $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ y la correlación entre X e Y es δ. Entonces, la distribución condicional de Y dado $X \geq 0$ es $\mathcal{SN}(\lambda(\delta))$ donde $\lambda(\delta) = \delta/\sqrt{1-\delta^2}$.*

Demostración. Como $(X,Y)^{\mathrm{T}} \sim \mathcal{N}_2(\mathbf{0}_2, \delta\mathbf{I}_2 + (1-\delta)\mathbf{1}_2\mathbf{1}_2^{\mathrm{T}})$ tenemos que la distribución condicional de X dado $Y = y$, $X|Y = y$, es $\mathcal{N}(\delta y, 1-\delta^2)$. Por lo tanto, si $W = (\delta Y - X)/\sqrt{1-\delta^2}$ tenemos que $W|Y = y \sim \mathcal{N}(0,1)$, de donde $W \sim \mathcal{N}(0,1)$ y W es independiente de Y. Por lo tanto, si tomamos $w(u) = \lambda u$ con $\lambda = \lambda(\delta) = \delta/\sqrt{1-\delta^2}$ por la Proposición 3.1 obtenemos que $Y|(W < \lambda Y)$ tiene densidad dada por (3.1) que por la elección hecha de $w(u)$ se reduce a (3.3).

La demostración se concluye observando que, como $\lambda = \delta/\sqrt{1-\delta^2}$, la condición $W < \lambda Y$ es equivalente a $-X < 0$, es decir, $X > 0$ como queríamos. □

La siguiente representación de una variable $\mathcal{SN}(\lambda)$ se debe a Henze (1986) y es útil para dar extensiones al caso multivariado y funcional.

Proposición 3.3. *Si Y_0 y Y_1 son variables $N(0,1)$ independientes y $\delta \in (-1,1)$ entonces $Z = \delta|Y_0| + \sqrt{(1-\delta^2)}Y_1$ es $\mathcal{SN}(\lambda(\delta))$ con $\lambda = \delta/\sqrt{1-\delta^2}$.*

Demostración. Tenemos que

$$F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{E}\left[\mathbb{P}\left(Z \leq z \middle| |Y_0|\right)\right] = \int_0^\infty \mathbb{P}\left(\delta|Y_0| + \sqrt{(1-\delta^2)}Y_1 \leq z \middle| |Y_0| = y\right) 2\phi(y)dy$$
$$= 2\int_0^\infty \Phi\left(\frac{z-\delta y}{\sqrt{1-\delta^2}}\right)\phi(y)dy\ .$$

Por lo tanto, si llamamos $b = \sqrt{1-\delta^2}$, la función de densidad de Z queda

$$f(z) = F_Z'(z) = 2\int_0^\infty \frac{1}{b}\phi\left(\frac{z-\delta y}{\sqrt{1-\delta^2}}\right)\phi(y)dy\ .$$

Usando que

$$\exp\left\{-\frac{1}{2}z^2\right\}\exp\left\{-\frac{\delta^2}{2(1-\delta^2)}z^2\right\} = \exp\left\{-\frac{1}{2(1-\delta^2)}z^2\right\}$$

y una expresión análoga para y, se obtiene que

$$\phi\left(\frac{z-\delta y}{\sqrt{1-\delta^2}}\right)\phi(y) = \phi\left(\frac{y-\delta z}{\sqrt{1-\delta^2}}\right)\phi(z)\,,$$

de donde

$$\begin{aligned} f(z) &= 2\phi(z)\int_0^\infty \frac{1}{b}\phi\left(\frac{y-\delta z}{\sqrt{1-\delta^2}}\right)dy = 2\phi(z)\int_0^\infty \phi\left(u-\frac{\delta z}{\sqrt{1-\delta^2}}\right)du \\ &= 2\phi(z)\int_{-\frac{\delta z}{\sqrt{1-\delta^2}}}^\infty \phi(u)\,du = 2\phi(z)\left(1-\Phi\left(-\frac{\delta}{\sqrt{1-\delta^2}}z\right)\right) = 2\phi(z)\Phi\left(\frac{\delta}{\sqrt{1-\delta^2}}z\right) \end{aligned}$$

lo que muestra que $Z \sim \mathcal{SN}\left(\delta/\sqrt{1-\delta^2}\right)$. □

Teniendo en cuenta que no es fácil el cálculo de $MC(F_Z)$ cuando $Z \sim \mathcal{SN}(\lambda)$ se realizó una simulación para obtener el rango de variación de $MC(F_Z)$. Se realizaron 50 replicaciones y en cada replicación se generaron n observaciones independientes $Z_1, \ldots, Z_n$ tales que $Z_i \sim \mathcal{SN}(\lambda)$ donde $\lambda = \delta/\sqrt{1-\delta^2}$. Para cada muestra, se calculó el medcouple y luego se promedió sobre las 50 replicaciones. Se tomó $n = 100000$ y 50 valores de δ en el intervalo $[0,1]$. Como $MC(F_{-Z}) = -MC(F_Z)$ basta considerar el intervalo $[0,1]$ ya que si llamamos $Z_\delta = \delta|Y_0| + \sqrt{(1-\delta^2)}Y_1$ tenemos que $-Z_\delta = -\delta|X_0| - \sqrt{(1-\delta^2)}Y_1 \sim Z_{-\delta}$ pues $Y_1 \sim -Y_1$. La Figura 3.3 muestra como varía el medcouple con δ, en particular, se obtiene que cuando $\delta \in [-1,1]$ $MC(F_Z)$ varía entre -0.2 y 0.2.

3.2 Extensión al caso multivariado

Azzalini y Dalla Valle (1996) proponen una extensión al caso multivariado en la que un vector multivariado $\mathbf{Z}$ tenga cada componente normal asimétrica. Es natural entonces definir la distribución conjunta de $\mathbf{Z}$ como $\mathcal{SN}$ multivariada.

Definición 3.1. Consideremos un vector aleatorio normal de dimensión k, $\mathbf{X} = (X_1, \ldots, X_k)^{\mathrm{T}}$, con marginales estandarizadas pero correlacionadas, independiente de $X_0 \sim N(0,1)$, es decir,

$$\begin{pmatrix} Y_0 \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}_{k+1}\left(\mathbf{0}_{k+1}, \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{\Psi} \end{pmatrix}\right)\,,$$

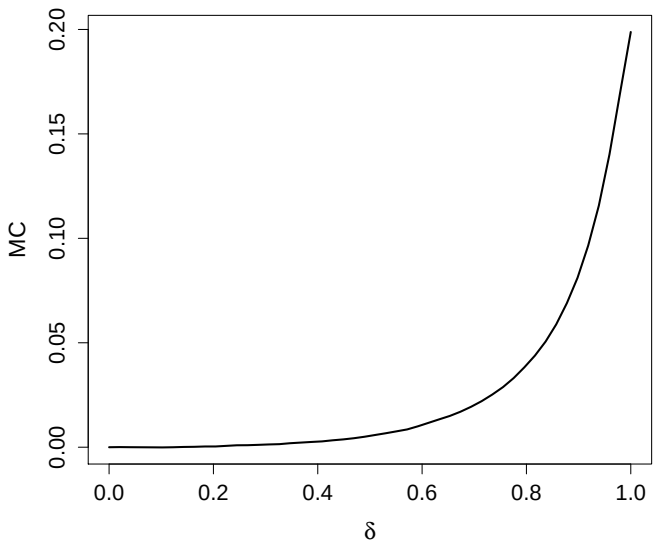


Figura 3.3: Medcouple de la distribución $\mathcal{SN}(\lambda)$ con $\lambda = \delta/\sqrt{1-\delta^2}$ en función de δ.

donde $\boldsymbol{\Psi} \in \mathbb{R}^{k\times k}$ es una matriz de correlación. Para $\delta_j \in (-1,1), j = 1,\ldots,k$ se definen

$$Z_j = \delta_j |X_0| + \sqrt{(1-\delta_j^2)} X_j \qquad j = 1,\ldots,k\,. \tag{3.4}$$

Por la Proposición 3.3, se tiene $Z_j \sim \mathcal{SN}\left(\lambda(\delta_j)\right)$. El cálculo de la función de densidad f_k de la distribución conjunta de $\mathbf{Z} = (Z_1,\ldots,Z_k)^{\mathrm{T}}$ puede consultarse en Azzalini y Dalla Valle (1996), quienes deducen que

$$f_k(\mathbf{z}) = 2\phi_k(\mathbf{z};\boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}) \qquad\qquad \mathbf{z} \in \mathbb{R}^k\,, \tag{3.5}$$

donde

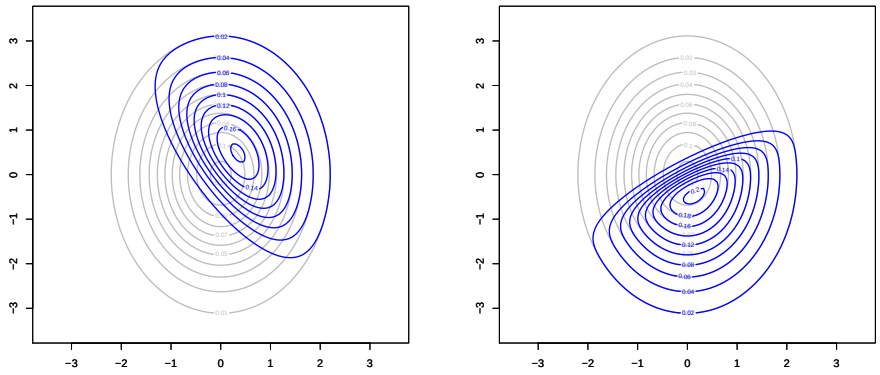
$$\begin{aligned}
\boldsymbol{\alpha}^{\mathrm{T}} &= \frac{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\Delta}^{-1}}{\sqrt{1+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}}} \\
\boldsymbol{\Delta} &= \operatorname{diag}\left(\sqrt{1-\delta_1^2},\ldots,\sqrt{1-\delta_k^2}\right) \\
\boldsymbol{\lambda} &= (\lambda(\delta_1),\ldots,\lambda(\delta_k))^{\mathrm{T}} \\
\boldsymbol{\Omega} &= \boldsymbol{\Delta}(\boldsymbol{\Psi}+\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}})\boldsymbol{\Delta}
\end{aligned}$$

y $\phi_k(z;\boldsymbol{\Omega})$ denota la función de densidad de una distribución normal multivariada con marginales estandarizadas y matriz de correlación $\boldsymbol{\Omega}$, es decir, $\phi_k(z;\boldsymbol{\Omega})$ es la densidad de $\mathbf{W}$ donde $\mathbf{W} \sim \mathcal{N}_k\left(\mathbf{0}_k, \boldsymbol{\Omega}\right)$.

Diremos que el vector $\mathbf{Z}$ con función de densidad dada por (3.5) es un vector normal multivariado asimétrico, con vector de parámetros de forma $\boldsymbol{\lambda}$ y parámetro de dependencia y se indicará $\boldsymbol{\Psi}$, $\mathbf{Z} \sim \mathcal{SN}_k(\boldsymbol{\lambda}, \boldsymbol{\Psi})$.

Una familia más amplia de distribuciones se obtiene por transformaciones afines de $\mathbf{Z}$, es decir, tomando $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$ con $\boldsymbol{\mu} \in \mathbb{R}^k$ y $\mathbf{A} \in \mathbb{R}^{k \times k}$.

De esta manera se extiende al caso multivariado la distribución normal asimétrica de manera de que las marginales pertenezcan a la misma familia (cerrada por marginalización). La Figura 3.4 muestra las curvas de nivel de un vector $\mathbf{Z} \sim \mathcal{SN}_2(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ para $\boldsymbol{\Psi} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.



(a) $\boldsymbol{\lambda} = (2, 2)$ (b) $\boldsymbol{\lambda} = (2, -5)$

Figura 3.4: Curvas de nivel para $\mathbf{Z} \sim \mathcal{SN}_2(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ (en azul) superpuestas sobre las curvas de nivel de $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ (en gris).

Capítulo 4

Detección de datos atípicos

En el contexto univariado, las herramientas de visualización de datos como el boxplot o el boxplot ajustado sirven para definir criterios de detección de datos atípicos. Para el caso multivariado, una práctica usual es estimar primero la posición y dispersión de los datos a través de un estimador robusto. Algunas opciones son el estimador de Stahel-Donoho, el MCD-estimador (Minimum Covariance Determinant), MM-estimador y los S-estimadores. Son afínmente equivariantes, permiten transformaciones afines de los datos, traslaciones, rotaciones y cambios de escala, y tienen un alto punto de ruptura para resistir hasta un 50% de outliers.

Luego se calcula para cada observación multivariada, $\mathbf{x} \in \mathbb{R}^p$, su distancia (robusta) de Mahalanobis

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \widehat{\mu})^T \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{x} - \widehat{\mu})}.$$

donde $\widehat{\mu}$ y $\widehat{\mathbf{\Sigma}}$ son las estimaciones robustas del centro y dispersión ya obtenidas. Por último se define un valor de corte basado en la distribución de estas distancias como umbral de clasificación. Para conocer esa distribución, es necesario hacer supuestos adicionales. Es común suponer que los datos han sido generados a partir de una distribución elíptica, entre las cuales la normal multivariada es una de las más populares. Consecuentemente estos métodos de detección de outliers no funcionarán apropiadamente cuando los datos no sean simétricos.

Una opción ante la falta de simetría es aplicar una transformación a los datos para obtener una muestra simétrica. Un ejemplo común es la transformación de Box y Cox. Sin embargo, además de no ser invariante por transformaciones afines, la transformación de Box y Cox se basa en una estimación de máxima verosimilitud y luego no es resistente a la presencia de outliers.

Hubert y Van der Veeken (2008) proponen un método automático de detección de

datos atípicos para datos multivariados que presenten asimetría. El método se inspira en el estimador de Stahel-Donoho. Este estimador se basa en una medida de atipicidad de las observaciones, que esencialmente se obtiene proyectando la muestra en varias direcciones univariadas y calculando el centro y escala, de manera robusta, en cada dirección.

El primer paso del método propuesto es ajustar la atipicidad de Stahel-Donoho para que contemple asimetría, lo que nos lleva a definir la atipicidad ajustada. Este paso utiliza el boxplot ajustado para datos asimétricos tratado en la Sección 2.3 del Capítulo 2. Como segundo paso, una observación se declara como atípica si su AO es demasiado grande. En tanto la distribución de las AO es en general desconocida aplicamos nuevamente la regla de detección provista por el boxplox ajustado. Trataremos primero el caso univariado y luego la extensión multivariada por proyecciones unidimensionales.

4.1 Caso univariado

Sea $X_n = \{x_1, x_2, \ldots, x_n\}$ un conjunto de datos univariado (continuo, unimodal), la atipicidad de Stahel y Donoho (SDO) de una observación x_i se define como:

$$SDO_i = SDO^{(1)}(x_i, X_n) = \frac{x_i - \text{mediana}(X_n)}{\text{MAD}(X_n)}$$

donde $\text{mediana}(X_n)$ es la mediana muestral y $\text{MAD}(X_n) = b{\cdot}\text{mediana}_{1\leq i\leq n}\,|x_i - \text{mediana}(X_n)|$ es la mediana de los desvíos absolutos. La constante $b = 1.483$ se elige para hacer al MAD asintóticamente insesgado cuando la muestra es normal. Este estimador es una medida robusta de la dispersión de la muestra que lo vuelve más útil ante la presencia de datos atípicos que, por ejemplo, el desvío estándar.

El valor de atipicidad de un dato mide su lejanía respecto del centro de la muestra. La definición de SDO no considera si el dato es menor o mayor que la mediana, sólo su distancia. Nuevamente, esta medida presupone simetría en los datos. Cuando la distribución es asimétrica, de la misma forma que en el boxplot ajustado, es necesario usar una escala distinta para los datos que se encuentran a uno y otro lado de la mediana. Luego, se define la *atipicidad ajustada* (AO) como:

$$AO_i = AO^{(1)}(x_i, X_n) = \begin{cases} \dfrac{x_i - \text{mediana}(X_n)}{w_2 - \text{mediana}(X_n)} & x_i \geq \text{mediana}(X_n) \\ \\ \dfrac{x_i - \text{mediana}(X_n)}{\text{mediana}(X_n) - w_1} & x_i \leq \text{mediana}(X_n) \end{cases} \tag{4.1}$$

donde w_1 y w_2 son los bigotes inferior y superior, respectivamente, del boxplot ajustado aplicado a la muestra X_n. Notemos que la medida AO_i se reduce a SDO_i bajo distribuciones simétricas.

La atipicidad ajustada se ilustra en la Figura 4.1. La observación x_1 tiene $AO_1 = d_1/s_1 = (\text{mediana}(X_n) - x_1)/(\text{mediana}(X_n) - w_1)$ mientras que para x_2 se tiene $AO_2 = d_2/s_2 = (x_2 - \text{mediana}(X_n))/(w_2 - \text{mediana}(X_n))$. Entonces, aunque x_1 y x_2 estén a la misma distancia de la mediana, x_1 tiene un valor mayor de atipicidad, porque la escala del lado menor a la mediana es menor que la escala del lado mayor. Una propiedad importante de ambas medidas de atipicidad es que son invariantes por cambios de posición o de escala en la muestra.

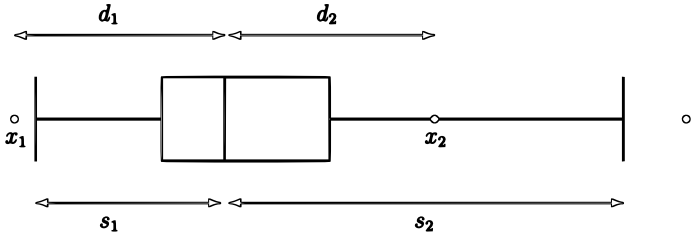


Figura 4.1: Ilustración de la atipicidad ajustada

En lo que refiere a robustez, la *AO* está basada es medidas robustas de posición, escala y asimetría (se incluye en los bigotes del boxplot ajustado). Teóricamente, puede alcanzarse una resistencia hasta el 25% de outliers aunque en la práctica se ha observado un importante sesgo en el medcouple cuando la contaminación supera el 10% de los datos. Más aún, puede mostrarse que la función de influencia de *AO* es acotada, ver Hubert y Van der Veeken (2008).

4.2 Caso multivariado

Hubert y Van der Veeken (2008) proponen una extensión al caso multivariado a través de proyecciones unidimensionales del procedimiento de detección de datos atípicos univariado para datos asimétricos.

Consideremos ahora muestra p-dimensional $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ con $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. La medida de atipicidad de Stahel-Donoho para $\mathbf{x}_i$ se define como

$$SDO_i = SDO(\mathbf{x}_i, \mathbf{X}_n) = \sup_{\mathbf{a} \in \mathbb{R}^p} SDO^{(1)}(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i, \mathbf{X}_n\mathbf{a}) \; . \tag{4.2}$$

La expresión dada en (4.2) puede interpretarse de la siguiente manera: para cada dirección univariada $\mathbf{a} \in \mathbb{R}^p$ se considera la distancia estandarizada de la proyección $\mathbf{a}^{\mathrm{T}}\mathbf{x}_i$ de la observación $\mathbf{x}_i$ al centro (robusto) de las proyecciones de toda la muestra. Si se obtuviera un valor de $SDO(\mathbf{x}_i, \mathbf{X})$ grande, entonces existiría una dirección en la cual la proyección de $\mathbf{x}_i$ estaría lejos del grueso de las otras proyecciones. Así uno podría sospechar que la observación $\mathbf{x}_i$ es atípica.

Como ya se comentó, la definición de la atipicidad SD no tiene en consideración una posible asimetría y, entonces, es adecuada solamente para datos con simetría elíptica. Para permitir asimetría en los datos, definimos análogamente la atipicidad ajustada de una observación multivariada $\mathbf{x}_i$ como

$$AO_i = AO(\mathbf{x}_i, \mathbf{X}_n) = \sup_{\mathbf{a}\in\mathbb{R}^p} AO^{(1)}(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i, \mathbf{X}_n\mathbf{a}),$$

donde $AO^{(1)}$ está definido en (4.1). Observemos que como $AO^{(1)}(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i, \mathbf{X}_n\mathbf{a})$ es invariante por cambios de escala podemos tomar supremo sobre los vectores $\mathbf{a} \in \mathbb{R}^p$ tales que $\|\mathbf{a}\| = 1$. En la práctica, no es posible proyectar en *todas* las direcciones univariadas $\mathbf{a} \in \mathbb{R}^p$, $\|\mathbf{a}\| = 1$. Debemos restringirnos a un conjunto finito m de direcciones aleatorias. En Hubert y Van der Veeken (2008) se sugiere considerar $m = 250p$. Esta elección muestra en las simulaciones un buen balance entre eficiencia y tiempo de cómputo. Las direcciones se obtienen como la dirección perpendicular al subespacio generado por p observaciones elegidas al azar de la muestra. Como la AO es invariante por transformaciones afines de los datos, en la implementación se elige $\|\mathbf{a}\| = 1$ aunque no es necesario por la invariancia de escala.

Una vez calculadas las AO de cada observación, podemos usar esta información para clasificar outliers. Salvo en el caso de la distribución normal, para la cual la AO^2 (o SDO^2) se distribuye asintóticamente como χ^2_p, la distribución de AO es en general desconocida pero típicamente asimétrica hacia la derecha dado que está acotada por 0. Por esta razón, Hubert y Van der Veeken (2008) recomiendan tomar el boxplot ajustado para la muestra formada por los valores de AO y declarar que una observación multivariada es atípica si su AO_i excede el bigote superior. Más precisamente, como el medcouple, MC, de la muestra $\{AO_i\}_{1\le i\le n}$ es positivo, el valor de corte es igual a $v_{\text{corte}} = Q_3 + 1.5\exp(3MC)IQR$ donde Q_3 es el tercer cuartil de las AO_i y similarmente para IQR y MC.

Una característica importante de este método es que tanto la construcción del boxplot ajustado como de la atipicidad ajustada no presuponen ninguna distribución asimétrica subyacente (solo unimodalidad), luego es un enfoque *distribution-free*.

El concepto de *robustez frente a outliers* puede resultar ambiguo en el contexto de distribuciones asimétricas. Supongamos que la mayoría de las observaciones son generadas de una distribución simétrica y que un grupo más chico (como máximo el 25%) es atípico. Cuando los outliers estás ubicados lejos de los datos regulares, un estimador robusto de asimetría debería ser capaz de detectar la simetría del grupo principal. Un detector de outliers basado en tal estimador robusto de asimetría, combinado con estimadores robustos de posición y escala es capaz de marcar las observaciones atípicas.

Cuando la misma metodología se utiliza con estimadores no robustos, los valores de atipicidad se verían afectatos por los outliers (aumentando el valor de la asimetría e inflando la escala) de manera que el grupo de outliers quedaría enmascarado. Esta diferencia

entre un enfoque robusto y no robusto también aplica cuando el grupo mayoritario tiene una distribución asimétrica. En tal situación, los outliers podrían dar la impresión de que la distribución entera es altamente asimétrica, cuando esto podría no ser cierto para la mayoría. Si por otro lado no hubiera outliers y la distribución fuera efectivamente asimétrica, un estimador robusto de la asimetría sería capaz de detectarla.

Sin embargo, cuando los outliers no están lejos de la cola de la distribución principal, la distinción entre puntos regulares y atípicos es difusa. Sobre este punto, Hubert y Van der Veeken (2008) comentan que ningún estimador (robusto o no) puede ser capaz de hacer una correcta clasificación. Si se presume que la asimetría es causada por los outliers, y que el grupo principal tiene distribución simétrica, se aconseja comparar los valores AO con SDO (ambas atipicidades). Si las conclusiones son muy diferentes queda a cargo del analista decidir si la simetría es del grupo principal o no.

Capítulo 5

Datos funcionales

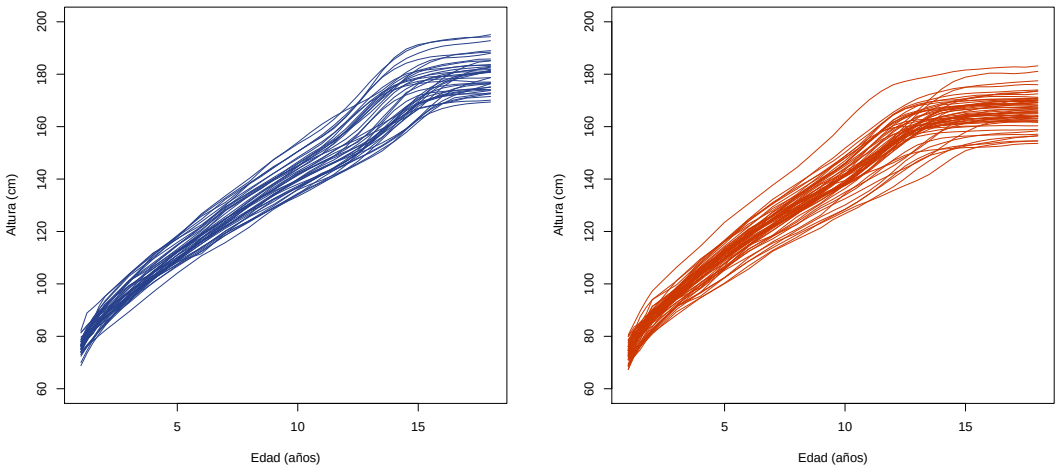
Los datos funcionales son una generalización natural de los datos multivariados de dimensión finita a dimensión infinita. Varias razones justifican esta generalización. Primero, en distintas áreas de investigación, como medicina, meteorología, biología o economía, entre otras, los procesos que generan los datos son por naturaleza funciones estocásticas. Son mediciones de una o varias magnitudes a lo largo del tiempo, curvas temporales, o sobre una región espacial, superficies espaciales, u otro conjunto continuo. Algunos ejemplos prácticos son las curvas de crecimiento[1] de la Figura 5.1 o las curvas anuales de temperatura diaria de distintas estaciones climáticas[2] que muestra la Figura 5.2.

Actualmente es posible, con el desarrollo de equipamiento cada vez más sofisticado y preciso, adquirir grandes cantidades de datos, usualmente llamados de **alta frecuencia**. Esto facilita su representación como funciones a través de métodos de interpolación y suavizado. Segundo, numerosos problemas se abordan mejor si las observaciones son consideradas como funciones. Por ejemplo, si cada curva de una muestra se observa en distintos puntos, un análisis multivariado, en general, no sería válido. Lo mismo ocurre si la cantidad de puntos en los que se observan los datos es superior al tamaño de la muestra, la matriz de covarianza muestral asociada es degenerada. Resulta más conveniente utilizar la estructura de los datos como provenientes de una función continua que se observa con error.

Una gran cantidad de técnicas multivariadas como el análisis de componentes principales, análisis de la varianza y métodos de regresión ya fueron extendidos al contexto funcional (ver Ramsay y Silverman, 2005). Una tarea fundamental en el análisis de datos funcionales es dotar de un **orden** natural a una muestra de curvas y poder definir los estadísticos de orden, mediana, mediana podada, etc. Una herramienta razonable es la de

[1] Los datos fueron extraidos del *dataset growth* del paquete *fda* de R.

[2] Las mediciones corresponden al dataset *CanadianWeather* disponible en el paquete *fda* de R.



(a) Curvas para 39 varones.

(b) Curvas para 54 mujeres.

Figura 5.1: Las curvas de crecimiento son las alturas medidas en 31 edades no distribuidas uniformemente entre 1 y 18 años.

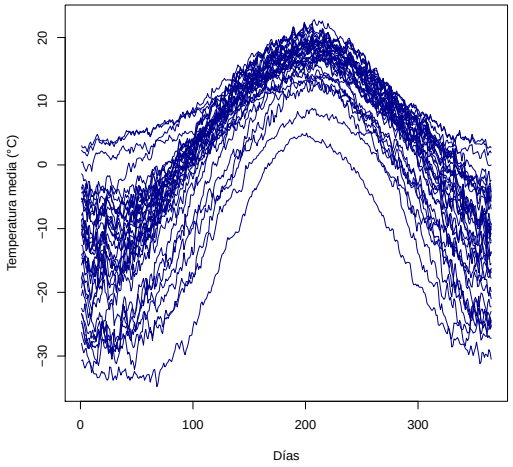


Figura 5.2: Temperaturas diarias medida en 35 estaciones climáticas de Canadá promediadas entre 1960 y 1994.

profundidad estadística. La profundidad de los datos se introduce para medir la *centralidad* o la *perifericidad* de una observación dentro de un conjunto de datos o bajo alguna distribución subyacente.

Presentaremos una medida de profundidad para observaciones funcionales introducida por López-Pintado y Romo (2009). Basados en esta medida, se ordena la muestra, lo que permitió a Sun y Genton (2010) extender la construcción del boxplot al caso funcional, a través de la generalización de estimadores robustos, mediana y distancia intercuartil.

5.1 Profundidad de banda para datos funcionales

La noción de profundidad introducida en López-Pintado y Romo (2009) se basa en la representación gráfica de los datos y de las bandas que determinan en el plano. El gráfico de una función x es el subconjunto del plano $G(x) = \{(t, x(t)) : t \in \mathcal{I}\}$. Dadas $x_1(t), \ldots, x_n(t) \in C(\mathcal{I})$, con $\mathcal{I}$ compacto[3], la banda determinada en $\mathbb{R}^2$ determinada por k de tales curvas, $x_{i_1}, \ldots, x_{i_k}$, es

$$\begin{aligned} V(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) &= \{(t, y) : t \in I, \min_{r=1,\ldots,k} x_{i_r}(t) \leq y \leq \max_{r=1,\ldots,k} x_{i_r}(t)\} \\ &= \{(t, y) : t \in I, y = \alpha_t \min_{r=1,\ldots,k} x_{i_r}(t) + (1 - \alpha_t) \max_{r=1,\ldots,k} x_{i_r}(t), \alpha_t \in [0, 1]\} \end{aligned}.$$

La Figura 5.3(a) muestra en rojo con sombreado azul la banda $V(x_1, x_2)$ delimitada por dos curvas; el gráfico de las funciones x_i se incluye en la banda en gris. La Figura 5.3(b) presenta la banda dada por tres curvas $V(x_1, x_2, x_3)$.

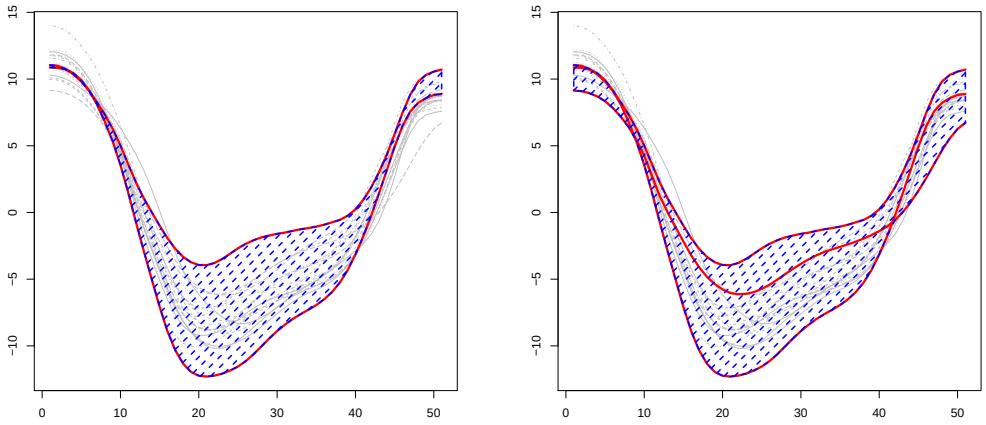
La profundidad de una curva en la muestra se medirá en términos de cuántas bandas la contienen. Para cualquier función $x \in \{x_1, \ldots, x_n\}$, calculamos para $j \geq 2$

$$BD_n^{(j)}(x) = \binom{n}{k}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \cdots \leq i_j \leq n} \mathbb{I}\left\{G(x) \subset V(x_{i_1}, x_{i_2}, \ldots, x_{i_j})\right\}$$

donde $\mathbb{I}(A)$ es el indicador de la condición A, o sea, $\mathbb{I}(A)$ vale 1 si se verifca A y 0 si no.

El numerador de esta expresión cuenta la cantidad de bandas $V(x_{i_1}, \ldots, x_{i_j})$ determinadas por j curvas diferentes $x_{i_1}, \ldots, x_{i_j}$ que contienen el gráfico de x. El denominador representa la cantidad de bandas que pueden formarse con j curvas elegidas de las n que conforman la muestra. Luego, el valor de $BD_n^{(j)}(x)$ es la proporción de bandas formadas por j curvas que contienen el gráfico de la curva x.

[3] Aunque las siguientes definiciones pueden ser extendidas a observaciones más generales nos restringiremos a funciones del espacio $C(\mathcal{I})$ de funciones continuas sobre un intervalo compacto $\mathcal{I}$.



(a) Banda formada por dos curvas

(b) Banda formada por dos curvas

Figura 5.3: Ejemplos de bandas delimitadas por dos y tres curvas.

Definición 5.1. Dadas las funciones $x_1, \ldots, x_n$, se define la *profundidad de banda* de una función x como

$$BD_{n,J}(x) = \sum_{j=2}^{J} BD_n^{(j)}(x) \,, \qquad J \geq 2 \,.$$

A partir de la definición muestral es directa la versión poblacional de la profundidad de banda. Si $X_1, X_2, \ldots, X_n$ son copias independientes del proceso estocástico X que genera las observaciones $x_1, x_2, \ldots, x_n$, las correspondientes versiones poblacionales son

$$BD^{(j)}(x) = \mathbb{P}\left\{G(x) \subset V(X_{i_1}, X_{i_2}, \ldots, X_{i_j})\right\}$$
$$BD_J(x) = \sum_{j=2}^{J} BD^{(j)}(x) = \sum_{j=2}^{J} \mathbb{P}\left\{G(x) \subset V(X_{i_1}, X_{i_2}, \ldots, X_{i_j})\right\} \,.$$

Con esta noción de profundidad, dada una muestra, definimos su mediana muestral como la observación $\widehat{m}_{n,J}$ con mayor valor de profundidad:

$$\widehat{m}_{n,J} = \underset{x \in \{x_1, \ldots, x_n\}}{\operatorname{argmax}} \; BD_{n,J}(x) \,.$$

Análogamente, la mediana poblacional será aquella función $m_J \in C(I)$ que maximice $BD_J(\cdot)$. Si no fueran únicas, la mediana será el promedio de las curvas que maximicen la profundidad.

En principio, la definición propuesta para la profundidad de banda depende del parámetro J, es decir de la cantidad máxima curvas que generan una banda. Cabe preguntarse si existe alguna dependencia. En tal caso el punto de mayor profundidad, la mediana, podría variar con J. En López-Pintado y Romo (2009), se simulan curvas de un proceso gaussiano en $C[0,1]$ con media $f(t) = 4t$ y covarianza $\gamma(s,t) = \exp\left\{-|t-s|^2\right\}$ y se estudia para cada J la distancia entre la mediana de profundidad $\widehat{m}(J)$ y el valor real $f(t) = 4t$ aproximando $\mathbb{E}I_{\widehat{m}}(J)$ mediante el promedio sobre replicaciones de $I_{\widehat{m}}(J) = \int_0^1 (\widehat{m}_{n,J}(t) - f(t))^2 dt$. López-Pintado y Romo (2009) muestran que este error se minimiza para $J = 3$ y que permanece constante para $J \geq 3$, es decir que para $J = 3, 4, 5, \ldots$ la curva más profunda es la misma. Luego, recomiendan trabajar con BD_3.

5.1.1 Propiedades de la profundidad de banda funcional

Una noción de profundidad debería satisfacer naturalmente ciertas propiedades. Por ejemplo, si consideramos datos univariados, la profundidad de banda produce la misma noción de mediana. A continuación, listamos las propiedades más relevantes de la profundidad de banda para funciones cuyas pruebas pueden encontrarse en López-Pintado y Romo (2009).

Sea X un proceso en $C(\mathcal{I})$ con una distribución ajustada $\mathbb{P}$, es decir, $\mathbb{P}(\|X\|_\infty \geq M) \to 0$ cuando $M \to \infty$.

Proposición 5.1 (Invariancia)**.** *Sea $T(x) = ax + b$, donde x,a y b son funciones continuas en $\mathcal{I}$, con $a(t) \neq 0$ para cada $t \in \mathcal{I}$. Entonces $S_J(x, \mathbb{P}) = S_J(ax + b, \mathbb{P}_{aX+b})$.*

Proposición 5.2 (Convergencia a cero)**.** *Cuando $M \to \infty$:*

$$\sup_{\|x\|_\infty \geq M} BD_J \to 0 \quad y \quad \sup_{\|x\|_\infty \geq M} BD_{n,J}(x) \xrightarrow{a.s.} 0\,.$$

Proposición 5.3 (Consistencia fuerte)**.** *$BD_{n,J}(x)$ es un estimador fuertemente consistente de $BD_J(x)$, i.e., $BD_{n,J}(x) \xrightarrow{a.s.} BD_J(x)$, cuando $n \to \infty$.*

Proposición 5.4 (Simetría)**.** *Si la variable aleatoria X sobre $C(\mathcal{I})$ es simétrica entonces la distribución de $m_{n,J}$ es también simétrica.*

Proposición 5.5 (Continuidad)**.** *BD_J es una función semicontinua superior. Más aún, si la distribución de probabilidad $\mathbb{P}$ en $C(\mathcal{I})$ tiene distribuciones marginales absolutamente continuas, entonces BD_J es un funcional continuo en $C(I)$.*

Teorema 5.1. *Sea $\mathbb{P}$ una distribución de probabilidad en $C(I)$ con distribuciones marginales absolutamente continuas. Entonces*

1. *$BD_{n,J}(\cdot)$ es uniformemente consistente sobre cualquier conjunto equicontinuo E:*

$$\sup_{x \in E} |BD_{n,J}(x) - BD_J(x)| \xrightarrow{a.s.} 0\,,$$

cuando $n \to \infty$.

2. *Si* $BD_J(\cdot)$ *se maximiza únicamente en* $m \in E$ *y* m_n *es una sucesión de funciones en* E *con* $BD_{n,J}(m_n) = \sup_{x \in E} BD_{n,J}(x)$ *entonces* $m_n \xrightarrow{a.s.} m$, *cuando* $n \to \infty$.

Por ejemplo, el conjunto $Lip_{\alpha,A}(\mathcal{I}) = \{x : I \to \mathbb{R}, |x(t_1) - x(t_2)| \leq A|t_1 - t_2|^\alpha, t_1, t_2 \in \mathcal{I}\}$ es equicontinuo y verifica las condiciones del teorema anterior. Así, $BD_{n,J}$ converge uniformemente a BD_J sobre $Lip_{\alpha,A}(\mathcal{I})$. En particular, la profundidad de banda es uniformemente consistente sobre funciones Lipschitz.

5.1.2 Profundidad de banda generalizada

Cuando las curvas son muy irregulares, pocas bandas contendrán por completo una curva. Varias curvas de la muestra tendrán el mismo valor de profundidad, lo cual resulta en un ordenamiento pobre de la muestra con muchos empates. Por ejemplo, la banda definida por dos curvas ($J = 2$) que se corten en un punto, con probabilidad 1 no contendrá ninguna otra curva y no contribuirá al valor de la profundidad.

En este sentido, la Definición 5.1 es restrictiva. Esto proviene de usar la función indicadora (vale 0 o 1). Una definición alternativa, más flexible, es medir el conjunto donde la función queda contenida en la correspondiente banda. Para cualquier función x en $x_1, \ldots, x_n$, sea para $j \geq 2$

$$A_{i_1,\ldots,i_j}(x) = \left\{ t \in I : \min_{r=i_1,\ldots,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,\ldots,i_j} x_r(t) \right\},$$

el conjunto de puntos del intervalo $\mathcal{I}$ donde la función x está dentro de la banda determinada por las observaciones $x_{i_1}, \ldots, x_{i_j}$. Si λ es la medida de Lebesgue en $\mathcal{I}$, $\lambda_r(A_j(x)) = \lambda(A_j(x))/\lambda(\mathcal{I})$ es la "proporción de tiempo" que x está en la banda.

Definamos entonces:

$$MBD_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \cdots \leq i_j \leq n} \lambda_r \left(A(x; x_{i_1}, x_{i_2}, \ldots, x_{i_j}) \right), \; j \geq 2 \qquad (5.1)$$

Efectivamente, (5.1) es una generalización de $BD_n^{(j)}(x)$ ya que si x está siempre dentro de la banda, $\lambda_r(A_j(x)) = 1$ y extiende la definición anterior.

Definición 5.2. Dado $J \geq 2$, para funciones $x_1, \ldots, x_n$, la profundidad de banda generalizada de una curva x es

$$MBD_{n,J}(x) = \sum_{j=2}^{J} MBD_n^{(j)}(x).$$

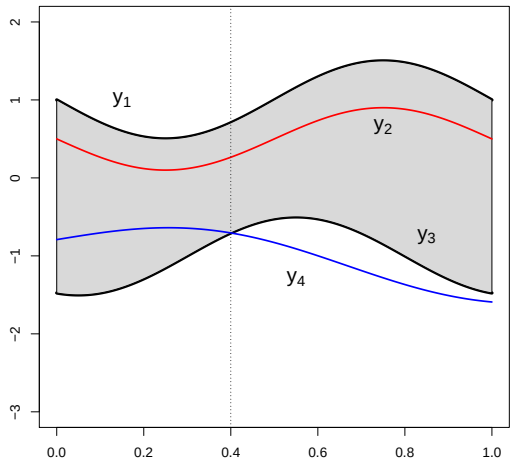


Figura 5.4: Un ejemplo del cálculo de BD y MBD: la región pintada es la banda delimitada por las curvas $y_1(t)$ y $y_3(t)$. La curva $y_2(t)$ pertenece completamente a la banda mientras que $y_4(t)$ lo hace parcialmente.

Si $X_1, X_2, \ldots, X_n$ son copias independientes de un proceso X que proveen las observaciones $x_1, \ldots, x_n$, la versión poblacional de las cantidades anteriores está dada por

$$\begin{aligned} MBD^j(x) &= \mathbb{E}\left(\lambda_r\left(A(x; X_{i_1}, X_{i_2}, \ldots, X_{i_j})\right)\right), \quad \text{si} \quad j \geq 2 \\ MBD_J(x) &= \sum_{j=2}^{J} GS_n^{(j)}(x), \quad \text{si} \quad J \geq 2 \end{aligned}$$

Nuevamente, las simulaciones dadas en López-Pintado y Romo (2009), muestran que el orden inducido por esta definición es estable cuando aumenta J. Para evitar problemas computacionales, se elige $J = 2$, y por simplicidad, escribimos directamente BD y MBD.

Como la profundidad de banda modificada toma en cuenta la proporción de tiempo que una curva permanece dentro de una banda es más conveniente para obtener la curva más representativa en términos de magnitud. La profundidad de banda depende más de la forma de las curvas dando más empates. Puede usarse para obtener la curva más representativa en términos de forma. Entonces, la atipicidad de un dato, en el caso funcional, puede ser de dos tipos: de magnitud o forma. Los *outliers de magnitud* están lejos de la media y los (*outliers de forma*) tienen un patrón diferente de las otras curvas.

La Figura 5.4 muestra un ejemplo sencillo con $n = 4$ curvas sobre cómo se calculan BD y MBD. Para $J = 2$, hay 6 posibles bandas limitadas por 2 curvas. Por ejemplo, el

área pintada en la Figura 5.4 es la banda limitada por $y_1(t)$ y $y_3(t)$. Se ve que la curva y_2 pertenece totalmente a la banda, mientras que la curva $y_4(t)$ lo hace parcialmente. Si una curva cae en el borde de una banda se considera en ella. Luego $BD(y_2) = 5/6 = 0.83$ mientras que solo la banda delimitada por $y_3(t)$ y $y_4(t)$ no contiene completamente a la curva $y_2(t)$ y $BD(y_4) = 3/6 = 0.5$ dado que es la única completamente contenida en las bandas limitadas por sí misma y otra curva. Similarmente, podemos calcular $BD(y_1) = 0.5$ y $BD(y_3) = 0.5$. Para calcular MBD, notemos que la curva $y_2(t)$ esta siempre contenida en las cinco bandas, entonces $MBD(y_2) = 0.83$, el mismo valor que BD. En contraste, la curva $y_4(t)$ solo pertenece a la banda pintada el 40% del tiempo, entonces $MBD(y_4) = (3 + 0.4 + 0.4)/6 = 0.63$ por definición. Para las otras dos curvas, $MBD(y_1) = 0.5$ y $MBD(y_3) = 0.7$.

Una diferencia entre ambas profundidades de banda es su comportamiento ante curvas que abandonan el centro de la muestra en un intervalo corto, o sea permanecen en el interior de la muestra casi todo el tiempo pero tomando valores extremos en subintervalos pequeños: el valor de MBD sigue siendo grande mientras que BD es sensible a este tipo de contaminación y pasa a tomar valores pequeños. Es decir que BD es resistente a los outliers de forma mientras que la profundidad MBD es robusta cuando los datos están contaminados en magnitud.

Hoy en día es posible recolectar datos de manera masiva generando conjuntos de datos de gran tamaño. Se vuelve necesario contar con un algoritmo eficiente para que el costo computacional no limite la aplicación de la profundidad de banda. En Sun *et al.* (2012) se presenta un método rápido y exacto para ordenar un millón de curvas.

La profundidad de banda y su versión numérica permiten ordenar muestras y extender los métodos univariados basados en los estadísticos de orden al caso funcional. En la próxima sección extenderemos la construcción del boxplot.

5.2 Boxplot funcional

La noción de profundidad presentada en la sección anterior nos permite ordenar curvas definiendo estadísticos de orden y extender el boxplot a conjuntos de datos funcionales.

Supongamos que cada observación es una función real $x_i(t), i = 1, \ldots, n, t \in \mathcal{I}$, donde $\mathcal{I}$ es un intervalo en $\mathbb{R}$. Sea $x_{[i]}(t)$ la observación asociada al i-ésimo mayor valor de profundidad. Luego $x_{[1]}(t), \ldots, x_{[n]}(t)$ son los estadísticos de orden con $x_{[1]}(t)$ siendo la curva más profunda (más central), y $x_{[n]}(t)$ la más exterior. Los estadísticos de orden inducidos por la medida de profundidad comienzan en la curva muestral más central y se alejan hacia afuera en todas las direcciones.

En el boxplot clásico, la caja representa el 50% de los datos. En el boxplot funcional

buscamos el 50% de las observaciones más profundas:

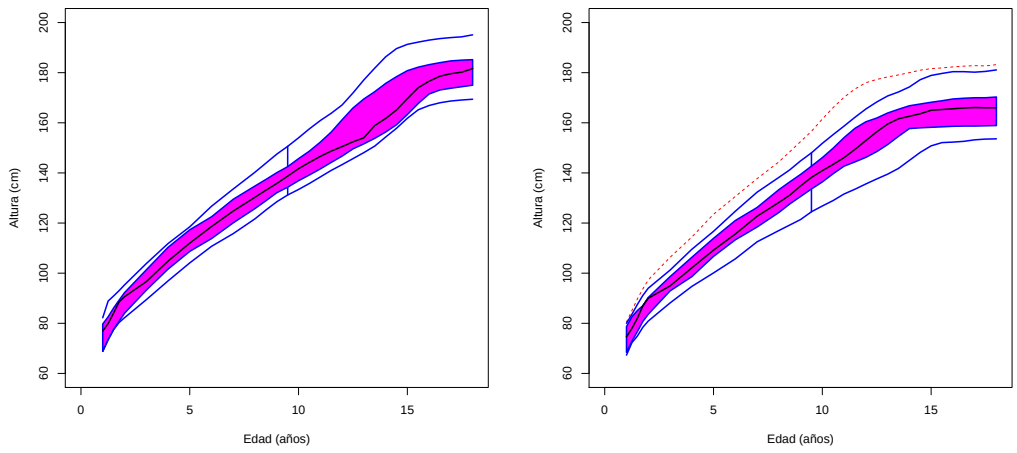
$$C_{0,5} = \{(t, x(t)) : \min_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}(t) \leq x(t) \leq \max_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}\}$$

donde $\lceil \frac{n}{2} \rceil$ es el menor entero mayor a $\frac{n}{2}$.

Esta región central es análoga a la determinada por la distancia intercuartil y da una medida robusta de la dispersión del 50% de las curvas más centrales. La curva $x_{[1]}(t)$ es la que indica la mediana, o sea la curva más central, la de mayor profundidad. La mediana funcional también es un estadístico robusto para medir centralidad. La idea de regiones centrales puede ampliarse para definir la región del 25% y 75%

Los bigotes del boxplot son las líneas verticales del gráfico que se extienden desde la caja hasta la última observación que no es atípica. Necesitamos entonces una regla de detección de outliers. Nuevamente, extendemos la regla del boxplot clásico, inflar la caja 1.5 veces al caso funcional. Definimos la *región exterior* inflando la región central 1.5 veces su tamaño. Cualquier curva fuera de estos límites se clasifica como un potencial outlier.

Es inmediato ver que si las funciones fueran constantes el boxplot funcional se reduce al boxplot univariado.



(a) Boxplot funcional (varones).

(b) Boxplot funcional (mujeres) con un outlier (rojo).

Figura 5.5: Elementos de un boxplot funcional. En negro se dibuja la mediana, la zona pintada es la región central. El dato atípico se representa en rojo.

La Figura 5.5 ilustra la construcción del boxplot funcional para los datos de la Figura 5.1b).

Una aproximación intuitiva sugeriría considerar boxplots puntuales, que no trata a cada curva como una observación. Obviamente este enfoque pierde la información de las formas de las curvas.

En general, la mediana funcional no tiene por qué ser equivalente a las medianas puntuales. Sólo si todos los puntos de la curva de la mediana funcional fueran simultáneamente cuantiles 50%, lo cual es raro, más ante curvas irregulares. Más aun, contectar las medianas puntuales como la más representativa no es adecuado porque no corresponde a una curva de la muestra o a un promedio de curvas de los datos.

Al igual que en el boxplot univariado, la región central, los bigotes y la mediana pueden revelar información útil sobre la muestra mirando su posición, tamaño y forma. Más aun, los espacios entre las diferentes partes de la caja indican el grado de asimetría en los datos. Sin embargo, como en el caso univariado, cuando las trayectorias son asimétricas, usualmente muchos puntos exceden los límites del boxplot y son erróneamente declarados como outliers. En la Sección 5.3 introduciremos algunas familias de distribuciones asimétricas para datos funcionales que permitirán en el Capítulo 7 evaluar el comportamiento del boxplot funcional. En particular, como en el caso univariado, si el boxplot funcional tendrá capacidad para no detectar como atípicos datos que corresponden a observaciones de la distribución.

5.3 Distribuciones asimétricas en el caso funcional

5.3.1 Procesos gaussianos asimétricos

A partir de la extensión al caso multivariado dada en la Sección 3.2, Zhang y El-Shaarawe (2010) proponen una adaptación al caso funcional. Sean $X_0(s)$ y $X(s)$ dos procesos gaussianos estacionarios con marginales estandarizadas. Ambos procesos son independientes y pueden tener diferentes funciones de covarianza. Definamos

$$Z(s) = \delta|X_0(s)| + \sqrt{(1-\delta^2)}X(s)\ . \tag{5.2}$$

El proceso $Z(s)$ es estrictamente estacionario con marginales normales asimétricas en virtud de la Proposición 3.3.

Un inconveniente en esta definición es que, aunque $Z(s)$ tenga distribución normal asimétrica, la distribución finito dimensional $\mathbf{Z_s} = (Z(s_1), \ldots, Z(s_m))^{\mathrm{T}}$ con $\mathbf{s} = (s_1, \ldots, s_m)^{\mathrm{T}}$ no coincide con la distribución normal asimétrica multivariada vista en la Sección 3.2. La razón es que el proceso $X_0(s)$ varía con s.

Para que cualquier distribución finito dimensional de un proceso definido por (5.2) fuera normal asimétrica multivariada, $X_0(s)$ no dependería de s. Es decir, (5.2) devendría en

$$Z(s) = \delta|X_0| + \sqrt{(1-\delta^2)}X(s)$$

donde X_0 es una variable normal estándar independiente del proceso $X(s)$. La definición del proceso $Z(s)$ cuando $X_0(s) = X_0$ tiene algunos inconvenientes. Por un lado, su comportamiento queda esencialmente reducido al de un proceso gaussiano con media $\delta|X_0|$ ya que a diferencia de la definición dada en (3.4) donde la ponderación entre el módulo de la variable normal y cada una de las componentes del vector normal dependía de la coordenada, en este caso, el factor δ en $Z(s) = \delta|X_0| + \sqrt{(1-\delta^2)}X(s)$ no depende de s. Por otro, el grado de asimetría se mezcla con la correlación del proceso lo que lo hace poco apto si se estudian procesos en todo $\mathbb{R}$ para los cuales la correlación debe decrecer con la distancia entre tiempos. Concretamente, para cualesquiera s_1 y s_2, el coeficiente de correlación entre $Z(s_1)$ y $Z(s_2)$ está dado por

$$\textsc{corr}(Z(s_1), Z(s_2)) = \frac{\delta^2(1-2/\pi) + (1-\delta^2)\textsc{corr}(X(s_1), X(s_2))}{\delta^2(1-2/\pi) + (1-\delta^2)}$$

que es cercano a 1 si $Z(s)$ es extremadamente asimétrica, es decir, si $\delta \approx 1$), sin importar la separación entre s_1 y s_2. Por lo tanto, estos procesos son pocos aptos cuando consideramos procesos estacionarios en $\mathbb{R}$ o en $\mathbb{Z}$.

Para resolver este problema, Zhang y El-Shaarawe (2010) definen el siguiente proceso estacionario que generaliza (5.2), al que nos referiremos como *proceso gaussiano asimétrico* (SGP)

$$Z(s) = m(s) + \sigma_1|X_1(s)| + \sigma_2 X_2(s) + \sigma_0\epsilon(s) \tag{5.3}$$

donde $\sigma_0 \geq 0$, $\sigma_2 \geq 0$ y $\sigma_1 \in \mathbb{R}$, $m(s)$ es una función no aleatoria, $X_i(s)$, $i = 1, 2$ son procesos estacionarios gaussianos con marginales estandarizadas y covarianzas $\textsc{Cov}(X_i(s), X_i(t)) = \rho_i(|t-s|)$ y $\epsilon(s)$ es un ruido blanco gaussiano de media 0 y varianza 1. Los tres procesos $X_1(s)$, $X_2(s)$ y $\epsilon(s)$ son independientes. Vale la pena mencionar que si $\sigma_0 \neq 0$ el operador de covarianza de $Y(s)$ no resultará compacto.

Proposición 5.6. *Sea $Z(s)$ un proceso definido por (5.3) entonces, para cada s,*

$$V(s) = \frac{Z(s) - m(s)}{\sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}} \sim \mathcal{SN}(\lambda)\,,$$

donde $\lambda = \sigma_1/\sqrt{\sigma_0^2 + \sigma_2^2}$. Por lo tanto,

$$\begin{aligned} \mathbb{E}Z(s) &= m(s) + \sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}\mathbb{E}V(s) = m(s) + \sigma_1\sqrt{\frac{2}{\pi}} \\ \textsc{Var}(Z(s)) &= \sigma_0^2 + \sigma_2^2 + \sigma_1^2\left(1 - \frac{2}{\pi}\right)\,. \end{aligned}$$

Demostración. Primero observemos que $\sigma_2 X_2(s)+\sigma_0\epsilon(s) \sim \mathcal{N}(0, \sigma_0^2+\sigma_2^2)$. La Proposición 3.3, con $Y_0 = X_1(s)$, $Y_1 = (\sigma_2 X_2(s) + \sigma_0\epsilon(s))/\sqrt{\sigma_0^2+\sigma_2^2}$ y $\delta = \sigma_1/\sqrt{\sigma_0^2+\sigma_1^2+\sigma_2^2}$ implica que la densidad de $(Z(s) - m(s))/\sqrt{\sigma_0^2+\sigma_1^2+\sigma_2^2}$ está dada por $2\phi(z)\Phi(\lambda z)$ con $\lambda = \delta/\sqrt{1-\delta^2} = \sigma_1/\sqrt{\sigma_0^2+\sigma_2^2}$. Por lo tanto,

$$V(s) = \frac{Z(s)-m(s)}{\sqrt{\sigma_0^2+\sigma_1^2+\sigma_2^2}} \sim \mathcal{SN}(\lambda)\,.$$

Usando que $V(s) \sim \mathcal{SN}(\lambda)$ obtenemos que

$$\mathbb{E}V(s) = \delta\sqrt{\frac{2}{\pi}} \qquad\qquad \textsc{Var}(V(s)) = 1 - \frac{2}{\pi}\delta^2\,,$$

con $\delta = \sigma_1/\sqrt{\sigma_0^2+\sigma_1^2+\sigma_2^2}$, de donde usando que $Z(s) = V(s)\sqrt{\sigma_0^2+\sigma_1^2+\sigma_2^2} + m(s)$ se deduce que

$$\begin{aligned}
\mathbb{E}Z(s) &= m(s) + \sqrt{\sigma_0^2+\sigma_1^2+\sigma_2^2}\mathbb{E}V(s) = m(s) + \sigma_1\sqrt{\frac{2}{\pi}} \\
\textsc{Var}Z(s) &= (\sigma_0^2+\sigma_1^2+\sigma_2^2)\textsc{Var}V(s) = (\sigma_0^2+\sigma_1^2+\sigma_2^2)\left\{1 - \frac{2}{\pi}\frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2+\sigma_2^2}\right\}
\end{aligned}$$

lo que concluye la demostración. □

Observemos que, por la condición de estacionaridad pedida, la Proposición 5.6 implica que la distribución marginal no depende de s y por lo tanto, el grado de asimetría es el mismo para todo s.

Por otra parte, para obtener la función de covarianza del proceso $Y(s)$ necesitamos el siguiente Lema cuya demostración puede verse en Zhang y El-Shaarawe (2010).

Lema 5.1. *Si dos variables aleatorias X e Y tienen distribución conjunta normal con marginales estandarizadas y coeficiente de correlación $\rho \geq 0$, es decir, si $(X,Y)^{\mathrm{T}} \sim \mathcal{N}_2(\mathbf{0}_2, \rho\mathbf{I}_2 + (1-\rho)\mathbf{1}_2\mathbf{1}_2^{\mathrm{T}})$ entonces*

$$\textsc{Cov}(|X|,|Y|) = \frac{2}{\pi}\left(\sqrt{1-\rho^2} + \rho\arcsin(\rho) - 1\right)\,.$$

Sean $\rho_i(|t-s|) = \textsc{Cov}\left(X_i(s), X_i(t)\right)$, para $i = 1, 2$, el Lema 5.1 permite deducir que $\textsc{Cov}\left(Z(s), Z(t)\right) = C(|t-s|)$ donde la función C está dada por

$$C(h) = \frac{2\sigma_1^2}{\pi}\left(\sqrt{1-\rho_1(h)^2} + \rho_1(h)\arcsin(\rho_1(h)) - 1\right) + \sigma_2^2\rho_2(h) + \sigma_0^2\mathbb{I}_{h=0} \tag{5.4}$$

Observemos que $\textsc{Var}Z(s) = C(0)$ y usando que $\rho_i(0) = 1$ y $\arcsin(1) = \pi/2$ de (5.4) obtenemos la expresión dada en la Proposición 5.6.

En Minozzo y Ferracuti (2012) se discuten más propiedades de estos procesos.

Observación 5.1. En el caso de datos funcionales es usual suponer que las observaciones corresponden a un proceso que es un elemento aleatorio de $L^2(\mathcal{I})$ donde $\mathcal{I} \subset \mathbb{R}$ es un intervalo acotado. Por esta razón, los problemas mencionados por Zhang y El-Shaarawe (2010) para el proceso $Z(s) = \delta|X_0| + \sqrt{(1-\delta^2)}X(s)$ no es esencial.

Para el proceso $Z(s) = \delta|X_0| + \sqrt{(1-\delta^2)}X(s)$, tenemos que dados $\alpha_1, \ldots, \alpha_k \in L^2(\mathcal{I})$ el vector definido por $\mathbf{Y} = (Y_1, \ldots, Y_k)^{\mathrm{T}} = (\langle Z, \alpha_1\rangle, \ldots, \langle Z, \alpha_k\rangle)^{\mathrm{T}}$ es tal que

$$Y_j = \delta|X_0| \int_{\mathcal{I}} \alpha_j(t)dt + \sqrt{(1-\delta^2)}X_j \,,$$

donde $X_j = \langle X, \alpha_j\rangle \sim N(0, \langle \alpha_j, \Gamma_X \alpha_j\rangle)$ con Γ_X es el operador de covarianza de X. Por lo tanto, si $\langle \alpha_j, \Gamma_X \alpha_j\rangle \neq 0$, $Y_j = \nu_j W_j$ donde $\nu_j^2 = \delta^2\langle \alpha_j, 1\rangle^2 + (1-\delta^2)\langle \alpha_j, \Gamma_X \alpha_j\rangle$,

$$W_j = a_j|X_0| + \sqrt{1-a_j^2}\left(\frac{X_j}{\sqrt{\langle \alpha_j, \Gamma_X \alpha_j\rangle}}\right) \qquad a_j = \delta\frac{\langle \alpha_j, 1\rangle}{\nu_j} \,.$$

Sea $V_j = X_j/\sqrt{\langle \alpha_j, \Gamma_X \alpha_j\rangle}$. Luego, $\mathbf{V} = (V_1, \ldots, V_k)^{\mathrm{T}}$ es un vector normal $k-$variado con marginales estandarizadas, independiente de $X_0 \sim N(0,1)$, o sea, $\mathbf{V} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{\Psi})$.

Luego, $\mathbf{W} = (W_1, \ldots, W_k)^{\mathrm{T}} \sim \mathcal{SN}_k(\boldsymbol{\lambda}, \mathbf{\Psi})$ con $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^{\mathrm{T}}$ y $\lambda_j = \lambda(a_j) = a_j/\sqrt{1-a_j^2}$ lo que implica que $\mathbf{Y} = \mathrm{diag}(\nu_1, \ldots, \nu_k)\mathbf{W}$ es una transformación lineal de vector normal multivariado asimétrico.

Esta propiedad no se preserva si consideramos el proceso $Z(s) = m(s) + \sigma_1|X_1(s)| + \sigma_2 X_2(s)$ ya que el proceso Gaussiano $X_1(s)$ varía con s.

Sin embargo, se preserva si consideramos el proceso, que llamaremos proceso Gaussiano asimétrico funcional (FSG), definido por

$$Z(s) = \delta(s)|X_0| + \sqrt{(1-\delta^2(s))}X(s) \tag{5.5}$$

donde $\delta : \mathcal{I} \to [-1,1]$, $\delta \in L^2(\mathcal{I})$ y X_0 es una variable normal estándar independiente del proceso gaussiano $X(s)$, ya que basta tomar

$$\alpha_j^\star = \sqrt{1-\delta^2}\,\alpha_j \qquad \nu_j^2 = \langle \alpha_j, \delta\rangle^2 + \langle \alpha_j^\star, \Gamma_X \alpha_j^\star\rangle$$
$$W_j = a_j|X_0| + \sqrt{1-a_j^2}\left(\frac{X_j}{\sqrt{\langle \alpha_j^\star, \Gamma_X \alpha_j^\star\rangle}}\right) \qquad a_j = \frac{\langle \alpha_j, \delta\rangle}{\nu_j} \,.$$

El siguiente resultado cuya demostración es inmediata a partir de las propiedades vistas en la Sección 3.1 del Capítulo 3 permite obtener la covarianza del proceso definido por (5.5).

Proposición 5.7. *Sea $X(s)$ un proceso gaussiano estacionario con marginales estandarizadas, X_0 una variable normal estándar independiente de $X(s)$ y $\delta : \mathcal{I} \to \mathbb{R}$. Sea $Z(s) = \delta(s)|X_0| + \sqrt{(1-\delta^2(s))}X(s)$, entonces, para cada s, $Z(s) \sim \mathcal{SN}(\lambda(s))$ con*

$$\lambda(s) = \frac{\delta(s)}{\sqrt{1-\delta(s)^2}}\,.$$

Por lo tanto,

$$\begin{aligned}
\mathbb{E}Z(s) &= \delta(s)\sqrt{\frac{2}{\pi}} \\
\textsc{Var}Z(s) &= 1 - \frac{2}{\pi}\delta^2(s) \\
\textsc{Cov}\left(Z(s), Z(t)\right) &= \delta(s)\delta(t)\left(1 - \frac{2}{\pi}\right) + \sqrt{1-\delta^2(s)}\sqrt{1-\delta^2(t)}\,\rho(s,t)\,,
\end{aligned}$$

donde $\rho(s,t) = \textsc{Cov}\left(X(s), X(t)\right)$. Por lo tanto, si $\delta \in L^2(\mathcal{I})$ con $\mathcal{I}$ un intervalo acotado, el proceso $Z(s)$ cumple que $\mathbb{E}\|Z\|^2 < \infty$.

5.3.2 Modelo de cuantiles inducidos

Una familia de procesos asimétricos fue introducida por Staicu *et al.* (2010) y se define como

$$Y(t) = \mu(t) + \sigma(t)G^{-1}\left(W(t), \boldsymbol{\alpha}(t)\right)\,, \tag{5.6}$$

donde $\mu(t)$ es una función no aleatoria que corresponde a la media del proceso, $\sigma(t)$ es el desvío estándar del proceso y $W_i(t)$ es un proceso tal que $W(t) \sim \mathcal{U}(0,1)$ para cada t. Por ejemplo, si $V(t)$ es un proceso Gaussiano con media 0 y función de covarianza $\gamma_V(t,s)$ podemos tomar

$$W(t) = \Phi\left(\frac{V(t)}{\sqrt{\gamma_V(t,t)}}\right)\,,$$

o más generalmente, podemos definir

$$W(t) = F_{X,t}\left(X(t)\right)\,,$$

donde $F_{X,t}$ es la función de distribución del proceso $X(t)$ tal que para cada t, $X(t)$ tiene densidad.

Por otra parte, $G^{-1}\left(\cdot, \boldsymbol{\alpha}\right)$ indica la inversa de la función de distribución $G\left(\cdot, \boldsymbol{\alpha}\right)$. La función $G\left(\cdot, \boldsymbol{\alpha}\right)$ es una familia de distribuciones paramétricas con media 0, varianza 1 y

parámetro de forma $\boldsymbol{\alpha}$. Por ejemplo, $\boldsymbol{\alpha}$ puede ser el parámetro λ de las distribución normal asimétrica. En este caso, si $Z \sim \mathcal{SN}(\lambda)$ y $\delta = \lambda/\sqrt{1+\lambda^2}$, $G(\cdot, \boldsymbol{\alpha})$ es la distribución de

$$U = \frac{Z - \delta\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}}\,.$$

Otras elecciones para $\boldsymbol{\alpha}$ son el parámetro de forma bidimensional de una distribución $\mathcal{T}$ asimétrica definido en Azzalini y Capitanio (2003).

El modelo (5.6) se conoce como el modelo se denomina *modelo funcional de cuantiles inducidos* ya que dicho modelo implica que el cuantil p de $Y(t)$, indicado por $Q_p(t)$, cumple

$$Q_p(t) = \mu(t) + \sigma(t) G_t^{-1}(p), \qquad 0 < p < 1\,,$$

donde $G_t^{-1} = G(\cdot, \boldsymbol{\alpha}(t))$.

Observación 5.2. El *modelo funcional de cuantiles inducidos* incluye los *procesos gaussianos asimétrico* definidos por (5.3). Efectivamente, si $Z \sim \mathcal{SN}(\lambda)$ y $\delta = \lambda/\sqrt{1+\lambda^2}$, llamemos G_λ a la distribución de

$$U = \frac{Z - \delta\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}}\,.$$

Recordemos que si $Z(t)$ es un proceso definido por (5.3) entonces, para cada t,

$$V(t) = \frac{Z(t) - m(t)}{\sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}} \sim \mathcal{SN}(\lambda)\,,$$

con $\lambda = \sigma_1/\sqrt{\sigma_0^2 + \sigma_2^2}$. Por lo tanto, si $\delta = \sigma_1/\sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}$

$$U(t) = \frac{V(t) - \delta\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}} \sim G_\lambda\,.$$

Definamos $W(t) = G_\lambda(U(t))$ y tomemos en (5.6), $\boldsymbol{\alpha}(t) \equiv \lambda$, $G(\cdot, \boldsymbol{\alpha}(t)) = G_\lambda(\cdot)$ y

$$\sigma^2(t) \equiv \left(1 - \frac{2}{\pi}\,\delta^2\right)\left(\sigma_0^2 + \sigma_1^2 + \sigma_2^2\right)\,.$$

Entonces, el proceso $Y(t)$ definido en (5.6) cumple

$$\begin{aligned} Y(t) &= \mu(t) + \sigma(t) G^{-1}\left(W(t), \boldsymbol{\alpha}(t)\right) = \mu(t) + \sigma(t) U(t) \\ &= \mu(t) + \sigma(t) \frac{V(t) - \delta\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}} \\ &= \mu(t) - \sigma(t) \left\{ \frac{\delta\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}} + \frac{m(t)}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}\,\sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}} \right\} + \sigma(t) \frac{Z(t)}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}\,\sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}} \\ &= \mu(t) - \delta\sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}\,\sqrt{\frac{2}{\pi}} - m(t) + Z(t) \end{aligned}$$

con lo que eligiendo $\mu(t) = \delta\sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}\,\sqrt{\frac{2}{\pi}} + m(t)$ obtenemos que $Y(t) = Z(t)$.

En forma análoga, el proceso definido en (5.5) también está contenido en el proceso dado en (5.6). Si $Z(t) = \mu(t) + \delta(s)|X_0| + \sqrt{(1 - \delta^2(t))}X(t)$, para cada t, $V(t) = Z(t) - \mu(t) \sim \mathcal{SN}(\lambda(t))$ con $\lambda(t) = \delta(t)/\sqrt{1 - \delta^2(t)}$. Sea

$$U(t) = \frac{V(t) - \sqrt{\frac{2}{\pi}\delta(t)}}{\sqrt{1 - \frac{2}{\pi}\delta^2(t)}}$$

y $G_{\lambda(t)}$ su distribución.

Definamos $W(t) = G_{\lambda(t)}\left(U(t)\right)$ y tomemos $\boldsymbol{\alpha}(t) = \lambda(t)$, $G\left(\cdot, \boldsymbol{\alpha}(t)\right) = G_{\lambda(t)}(\cdot)$ y $\sigma^2(t) \equiv 1 - (2/\pi)\delta^2(t)$. Entonces, el proceso $Y(t)$ definido en (5.6) cumple

$$\begin{aligned} Y(t) &= \mu(t) + \sigma(t)\, G^{-1}\left(W(t), \boldsymbol{\alpha}(t)\right) = \mu(t) + \sigma(t) U(t) \\ &= \mu(t) + \sigma(t) \frac{V(t) - \delta(t)\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\,\delta^2(t)}} \\ &= \mu(t) - \sigma(t) \left\{ \frac{\delta(t)\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\,\delta^2(t)}} + \frac{m(t)}{\sqrt{1 - \frac{2}{\pi}\,\delta^2}} \right\} + \sigma(t) \frac{Z(t)}{\sqrt{1 - \frac{2}{\pi}\,\delta^2(t)}} \\ &= \mu(t) - \delta(t)\sqrt{\frac{2}{\pi}} - m(t) + Z(t) \end{aligned}$$

con lo que eligiendo $\mu(t) = \delta(t)\sqrt{\frac{2}{\pi}} + m(t)$ obtenemos que $Y(t) = Z(t)$.

El proceso definido por (5.6) permite tener procesos no estacionarios y en particular, procesos cuya asimetría varía con t mediante la elección de $\boldsymbol{\alpha}(t)$.

Capítulo 6

Propuestas de detección

El boxplot funcional propuesto por Sun y Genton (2011) clasifica potenciales datos atípicos de manera análoga al boxplot univariado visto en la Sección 2.1 del Capítulo 2. Para ser más precisos, sean $x_1, \ldots, x_n$ independientes, $x_i : \mathcal{I} \to \mathbb{R}$. Indiquemos por $x_{[1]}(t), \ldots, x_{[n]}(t)$ los estadísticos de orden respecto de la profundidad de banda con $x_{[1]}(t)$ siendo la curva más profunda (más central), y $x_{[n]}(t)$ la más exterior. Indicaremos por $m(t)$ a la mediana funcional $m_n(t)$ que es igual a $x_{[1]}(t)$ si no hay empates o el promedio de las curvas con mayor profundidad de banda si hay empates. Sea $Q_1(t) = \min_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}(t)$, $Q_3(t) = \max_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}$ y $D(t) = Q_3(t) - Q_1(t)$. Estas medidas corresponden al primer y tercer cuartil y a la distancia intercuartil en el caso univariado. El boxplot funcional dado por Sun y Genton (2011) dilata la región central

$$C_{0,5} = \{(t, x(t)) : Q_1(t) \leq x(t) \leq Q_3(t)\}$$

mediante un factor de 1.5, es decir, considera la región

$$\mathcal{R} = \{(t, x(t)) : Q_1(t) - 1.5\, D(t) \leq x(t) \leq Q_3(t) + 1.5\, D(t)\}\,, \tag{6.1}$$

declarando como atípica una curva x que no yace completamente en la proyección de $\mathcal{R}$, es decir, tal que para algún $t \in \mathcal{I}$, $x(t) \notin [Q_1(t) - 1.5\, D(t), Q_3(t) + 1.5\, D(t)]$.

De la misma forma que en el caso univariado, frente a datos asimétricos, los bigotes del diagrama no acompañan la asimetría clasificando de manera incorrecta observaciones regulares. En el caso univariado se resuelve este inconveniente incorporando una medida robusta de asimetría, el medcouple, en la definición de los bigotes. Para el caso multivariado, se proyecta la muestra en direcciones univariadas y se trabaja con la muestra de atipicidades ajustadas. A partir de estas ideas, modificación de los bigotes y proyección univariada, se ofrecen tres propuestas para la detección de datos atípicos en el caso funcional.

6.1 Semi-distancia intercuartil

Por analogía con el bagplot introducido por Rousseeuw *et al.* (1999), proponemos modificar el factor 1.5 dilatando la región en forma asimétrica según la distancia a la mediana. Más precisamente, dada la muestra $x_1(t), \ldots, x_n(t)$, sea $m_n(t)$ la mediana funcional y $Q_1(t) = \min_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}(t)$, $Q_3(t) = \max_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}$ los límites inferior y superior del boxplot funcional, respectivamente de la caja del boxplot funcional definido por Sun y Genton (2011).

Consideramos la semi-distancia intercuartil superior $SIQR_{\mathrm{S}}(t) = Q_3(t) - m_n(t)$ e inferior $SIQR_{\mathrm{I}}(t) = m_n(t) - Q_1(t)$. Se considera entonces la región

$$\mathcal{R} = \{(t, x(t)) : Q_1(t) - 3\, SIQR_{\mathrm{I}}(t)) \leq x(t) \leq Q_3(t) + 3\, SIQR_{\mathrm{S}}(t)\}\,,$$

y se decide que $x(t)$ es un potencial outlier si para algún $t \in \mathcal{I}$,

$$x(t) \notin [Q_1(t) - 3\, SIQR_{\mathrm{I}}(t)), Q_3(t) + 3\, SIQR_{\mathrm{S}}(t))]\,.$$

En Hubert y Vandervieren (2008) se estudia la performance de la semi-distancia intercuartil para datos univariados y se muestra que en muchos casos presenta inconvenientes, siendo superada por el boxplot ajustado. Por esta razón, no se presentan los resultados correspondientes a esta medida en el Capítulo 7.

6.2 Boxplot con corrección mediante el medcouple

Una forma directa de incorporar el medcouple es en cada coordenada. Podríamos construir para cada $t \in \mathcal{I}$ el boxplot ajustado visto en la Sección 2.3 del Capítulo 2. Así se perdería por completo el enfoque funcional como se discutió en el caso del boxplot funcional. Por ejemplo, la mediana que se obtiene de unir las medianas puntuales no tiene por qué coincidir con la mediana funcional, además de no ser un elemento de la muestra.

Como se describió al inicio del Capítulo, sean $x_{[1]}(t), \ldots, x_{[n]}(t)$ son los estadísticos de orden con $x_{[1]}(t)$ siendo la curva más profunda (más central), y $x_{[n]}(t)$ la más exterior. Llamemos $m_n(t)$ a la mediana, $Q_1(t) = \min_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}(t)$, $Q_3(t) = \max_{r=1,\ldots,\lceil \frac{n}{2} \rceil} x_{[r]}$ y $D(t) = Q_3(t) - Q_1(t)$.

Centremos los datos respecto de la mediana funcional y luego, para cada $t \in \mathcal{I}$, calculemos el medcouple. Para ser más precisos, sea para cada t, $MC(t)$ indica el medcouple de los datos centrados respecto de la mediana funcional $m_n(t)$. Definimos

$$MC(t) = \underset{\widetilde{x}_i(t) \leq 0 \leq \widetilde{x}_j(t)}{\text{mediana}} \widetilde{h}(\widetilde{x}_i(t), \widetilde{x}_j(t)),$$

donde $\widetilde{x}_i(t) = x_i(t) - m(t)$ y la función $\widetilde{h}$ se define para los valores $u \neq v$ como

$$\widetilde{h}(u,v) = \frac{u+v}{u-v}\,.$$

Para tener en cuenta la posible asimetría de las curvas adaptaremos la región $\mathcal{R}$ definida en (6.1) mediante el medcouple $MC(t)$. Definamos entonces

$$\mathcal{R} = \{(t,x(t)) : Q_1(t) - 1.5\,D(t)e^{\alpha_1(t)MC(t)} \leq x(t) \leq Q_3(t) + 1.5\,D(t)e^{\alpha_2(t)MC(t)}\,,$$

donde $(\alpha_1(t),\alpha_2(t)) = (-4,3)$ si $MC(t) \geq 0$ y $(\alpha_1(t),\alpha_2(t)) = (-3,4)$ si $MC(t) \leq 0$. Luego, una curva x_i será atípica si no yace completamente en $\mathcal{R}$, es decir, si para algún $t \in \mathcal{I}$, $x(t) \notin \left[Q_1(t) - 1.5\,D(t)e^{\alpha_1(t)MC(t)}, Q_3(t) + 1.5\,D(t)e^{\alpha_2(t)MC(t)}\right]$.

6.3 Detección por proyecciones

A partir del trabajo de Hubert y Van der Veeken (2008) es posible emular el método de proyecciones en el caso funcional. Dadas $x_1,\ldots,x_n \in L^2(\mathcal{I})$ observaciones de una muestra aleatoria, sea v una dirección sobre la que vamos a proyectar. Indiquemos por $x_i^v = \langle x_i, v\rangle$ la proyección de la observación x_i sobre la dirección v con el producto interno usual de $L^2(\mathcal{I})$ y por X^v el conjunto de observaciones proyectadas, o sea, $X^v = \{x_1^v,\ldots,x_n^v\}$. Definimos la atipicidad ajustada de la observación funcional x como el supremo sobre las atipicidades ajustadas de todas las proyecciones univariadas:

$$AO(x) = \sup_{v \in L^2(\mathcal{I})} AO^{(1)}(x^v, X^v)$$

donde $AO^{(1)}$ está definido en (4.1). El supraíndice indica que la muestra es univariada. Recordemos que la AO es invariante por cambios de escala, luego podemos restringirnos a direcciones de norma 1.

En la práctica se debe usar un conjunto finito de direcciones en las que se proyectar los datos. Aquí surgen dos alternativas:

a) Usar las mismas observaciones centradas con la mediana funcional $m_n(t)$ para generar las direcciones de proyección. Si existe i_0 tal que $x_{i_0} = m_n(t)$ (es decir, si no hay varias curvas con profundidas máxima empatada) luego las direcciones son

$$d_i = \frac{x_i - m_n}{\|x_i - m_n\|} \quad i = 1,\ldots,n \quad i \neq i_0\,.$$

b) Generar al azar una cantidad M de observaciones simulando, por ejemplo, un proceso gausiano.

Una vez calculada la atipicidad $AO(x_i)$ para cada observación de la muestra se procede de manera idéntica al caso multivariado. Se construye el boxplot ajustado para la muestra univariada de atipiciades $\{AO^2(x_i)\}_{1\leq i\leq n}$ y se clasifican los outliers como aquellas observaciones funcionales cuya atipicidad ajustada exceda el límite superior del boxplot ajustado.

Capítulo 7

Estudio de Monte Carlo

En este capítulo investigamos a través de un estudio de simulación el comportamiento de las propuestas para detectar outliers para datos funcionales asimétricos. En todos los casos, se consideraron muestras de tamaño $n = 100$ y $n = 500$ de observaciones en $L^2(\mathcal{I})$, $\mathcal{I} = [0, 1]$, y se realizaron $NR = 1000$ replicaciones.

Al utilizar el método descripto en la Sección 6.3 del Capítulo 6, generamos las direcciones de proyección además de las muestras de datos funcionales con algún tipo asimetría. Luego, cada observación funcional se proyecta sobre una de las direcciones generadas digamos α_j, se calcula la atipicidad ajustada de todos los datos funcionales proyectados sobre α_j que llamaremos $AO_j(x_i)$ y posteriormente, para cada observacion x_i se toma el máximo sobre las direcciones, es decir, se calcula $\max_j AO_j(x_i)$, definiendo de esta forma una aproximación para $AO(x_i)$. Se aplica a la muestra de atipicidades ajustadas $AO(x_i)$ la regla de clasificación del boxplot ajustado. Los valores clasificados como datos atípicos por el boxplot serán las observaciones que declaremos como outliers.

Las direcciones de proyección se obtuvieron por dos métodos que llamaremos *al azar* y *muestral* que corresponden a lo descripto en la Sección 6.3. En el caso de generar direcciones al azar, se generaron $M = 10n$ direcciones normalizadas como procesos gausianos de media 0 y operador de covarianza con núcleo de covarianza dado por $\gamma(s, t) = \sigma \exp\left(-(t-s)^2/\left(2\nu^2\right)\right)$, donde $\sigma = \nu = 1$. Con este núcleo, las trayectorias resultan continuas.

Para el método muestral, se calculó la mediana de la muestra, $m_n(t)$, a través de la profundidad de banda modificada MBD. Sea i_0 tal que $x_{i_0} = m_n$. Se definieron las direcciones normalizadas como

$$d_i = \frac{x_i - m_n}{\|x_i - m_n\|} \qquad i \neq i_0\,,$$

es decir, una dirección menos que el tamaño de la muestra cuando una trayectoria coincide

con la mediana.

Todas las rutinas fueron implementadas en R y el código correspondiente se encuentra en el Apéndice 8.

7.1 Condiciones de la simulación

Para la generación de las observaciones se usaron dos métodos. El primero que llamaremos de rango finito consiste en armar combinaciones lineales finitas de elementos de una base del espacio $L^2(\mathcal{I})$ con coeficientes aleatorios distribuídos como una normal asimétrica multivariada. El segundo es el proceso gausiano asimétrico de Zhang y El-Shaarawe (2010) dado por (5.3) y la modificación definida en (5.5) como proceso Gaussiano asimétrico funcional (FSG). En todos los casos, las observaciones se generaron sobre una grilla de $G = 200$ puntos equiespaciados.

- **Modelo 1: Caso de rango finito**. Las observaciones $Z_1, \ldots, Z_n$ de rango finito se generaron con una base ortonormal de *Fourier* utilizando los primeros 5 términos de la base

$$\{1\,,\ \sqrt{2}\sin(2\pi t)\,,\ \sqrt{2}\cos(2\pi t)\,,\ \sqrt{2}\sin(4\pi t)\,,\ \sqrt{2}\cos(4\pi t)\} = \{\delta_j(t)\}_{j=1}^5$$

  y los coeficientes $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_5)^{\mathrm{T}}$ según una distribución normal multivariada asimétrica con parámetro de forma $\boldsymbol{\lambda} = (10, 10, 4, 4, 4)$ y matriz de correlación $\mathbf{I}_5$ definida en la Sección 3.2 del Capítulo 3. Es decir, Z_i tiene la misma distribución que el proceso definido como $Z(t) = \sum_{j=1}^5 \xi_j \delta_j$.

  La contaminación con datos atípicos se hizo reemplazando un porcentaje de las observaciones por datos normales multivariados de poca escala y distinta media de la del proceso, de manera tal que las colas distribuciones de los coeficientes no se solapen (demasiado).

  Recordemos que $\mathbb{E}(\xi_j) = (\sqrt{2/\pi})\ \lambda_j/\sqrt{1+\lambda_j^2}$, de donde, $\mathbb{E}\boldsymbol{\xi} = (0.7939, 0.7939, 0.7741, 0.7741)^{\mathrm{T}}$. Por esta razón, se escogió una distribución $\mathcal{N}(-k\,\mathbf{1}_p, \mathbf{I}_p/20)$, es decir, los coeficientes quedan definidos ahora como $\boldsymbol{\xi}^{(c)} = (\xi_1^{(c)}, \ldots, \xi_5^{(c)})^{\mathrm{T}}$ donde $\boldsymbol{\xi}^{(c)} = (1-U)\,\boldsymbol{\xi} + U\mathbf{V}$ con $U \sim Bi(1, \epsilon)$, $\mathbf{V} \sim \mathcal{N}(-k\,\mathbf{1}_p, \mathbf{I}_p/20)$ con $k = 1$ y 2. Por lo tanto, el proceso correspondiente queda definido como $Z^{(c)}(t) = \sum_{j=1}^5 \xi_j^{(c)} \delta_j$ y la observaciones contaminadas $Z_1^{(c)}, \ldots, Z_n^{(c)}$ son independientes y tales que $Z_i^{(c)} \sim Z^{(c)}$. Los porcentajes usados fueron $\epsilon = 0.01$ y $\epsilon = 0.10$. Un ejemplo de las observaciones generadas puede verse en la Figura 7.1.

- **Modelo 2: Caso del modelo SGP definido por (5.3)**. Se generaron observaciones $Z_1, \ldots, Z_n$ con la misma distribución que $Z(t) = m(t) + \delta|X_0(t)| + \sqrt{1-\delta^2}X_1(t)$

donde $X_0(t)$ y $X_1(t)$ son dos procesos gaussianos de media 0 y núcleo de covarianza dado por $\gamma(s,t) = \exp\left(-(t-s)^2/2\right)$. Consideramos los valores de asimetría $\delta = 0, 0.5, 1$ y $m(t) = 4t$.

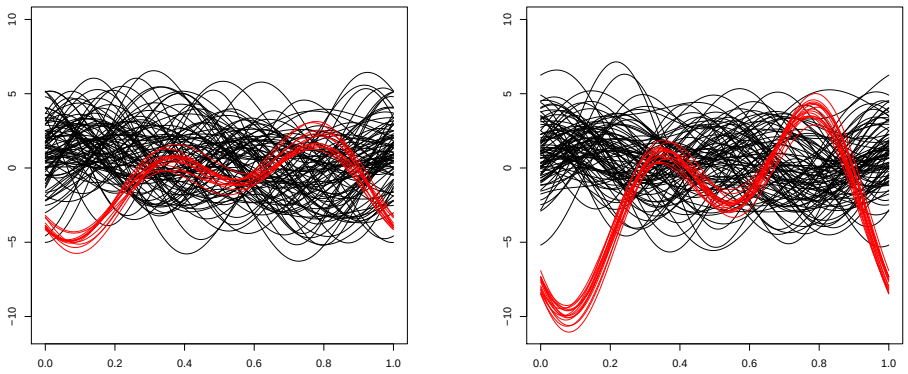
Indiquemos por $\mu(t) = \mathbb{E}Z(t) = m(t) + \delta\sqrt{2/\pi}$. Para generar outliers en el caso del SGP, se defineel proceso contaminado $Z^{(c)}(t) = (1-U)\,Z(t) + U\,V(t)$ donde $U \sim Bi(1,\epsilon)$ y $V(t)$ proceso Gaussiano con media $\mu_V(t)$ desplazada según un parámetro variable k como $\mu_V(t) = \mu(t) - k$. Como en el **Modelo 1**, $Z_1^{(c)}, \ldots, Z_n^{(c)}$ son independientes y tales que $Z_i^{(c)} \sim Z^{(c)}$. Elegimos $k = 1, 2, 3, 4$. El núcleo de covarianza de V está dado por $\gamma_V(s,t) = (1/20)\gamma(s,t)$. Es decir, reemplazamos una proporción ϵ de observaciones por datos generados por un proceso simétrico desplazado y concentrado alrededor de su media. El signo menos en el desplazamiento ubica los outliers *del otro lado* de la asimetría (ver Figura 7.4). Los porcentajes de generación de outliers fueron $\epsilon = 0.01$ y $\epsilon = 0.10$.

- **Modelo 3: Caso del modelo FSG definido por (5.5)**. Se generaron observaciones $Z(t) = \delta(t)|X_0| + \sqrt{1-\delta(t)^2}X_1(t)$ a partir de una variable aleatoria $X_0 \sim \mathcal{N}(0,1)$, un proceso gaussiano $X_1(t)$ de media 0 y operador de covarianza con núcleo dado por $\gamma(s,t) = \exp\left(-(t-s)^2/2\right)$ y las siguientes funciones de asimetría $\delta(s)$:
  - $\delta_1(t) = (\sin(2\pi t) + 1)/2$, $\lambda_1(t) = \delta_1(t)/\sqrt{1-\delta_1^2(t)}$,
  - $\delta_2(t) = \frac{\lambda_2(t)}{\sqrt{1+\lambda_2(t)^2}}$, $\lambda_2(t) = 5t^2 - 19t + 5$,
  - $\delta_3(t) = \frac{\lambda_3(t)}{\sqrt{1+\lambda_3(t)^2}}$, $\lambda_3(t) = -10\sin(2\pi t)$,

  En la Figura 7.3 se muestran los gráficos de las asimetrías consideradas que fueron escogidas de manera similar a las elegidas para simulaciones del trabajo de Staicu *et al.* (2011). Observemos que δ_1 es siempre positiva mientras que δ_2 y δ_3 recorren todo el intervalo $[-1,1]$. Las observaciones contaminadas se generaron del mismo modo que en el ítem anterior.

- **Modelo 4**. En este caso el proceso se generó con marginales χ_1^2 como en los casos anteriores se generaron observaciones $Z_1, \ldots, Z_n$ independientes $Z_i \sim Z$ donde $Z(t) = 4t + X_1^2(t)$ con $X_1(t)$ un proceso gaussiano de media 0 y operador de covarianza con núcleo dado por $\gamma(s,t) = \exp\left(-(t-s)^2/2\right)$. Observemos que en este caso, $\mu(t) = \mathbb{E}Z(t) = 4t + 1$, por lo que, como en el **Modelo 2**, las observaciones contaminadas $Z_1^{(c)}, \ldots, Z_n^{(c)}$ son independientes y tales que $Z_i^{(c)} \sim Z^{(c)}$ donde $Z^{(c)}(t) = (1-U)\,Z(t) + U\,V(t)$ donde $U \sim Bi(1,\epsilon)$ y $V(t)$ proceso Gaussiano con media $\mu_V(t) = 4t + 1 - k$, donde $k = 1, 2$.

En la Figura 7.4 se ilustra la generación de las observaciones para cada tipo de asimetría para los **Modelo 2 y 3** mientras que la Figura 7.2 muestra un ejemplo de observaciones correspondientes al **Modelo 4** con $\epsilon = 0.10$ y desplazamiento $k = 2$.



(a) Ejemplo de $n = 100$ observaciones de rango finito con 10% de contaminación generada como $\mathcal{N}(-\mathbf{1}_p, \mathbf{I}_p/20)$.

(b) Ejemplo de $n = 100$ observaciones de rango finito con 10% de contaminación generada como $\mathcal{N}(-\mathbf{2}_p, \mathbf{I}_p/20)$.

Figura 7.1: Ejemplo de observaciones correspondientes al **Modelo 1** de rango finito con 10% de contaminación ($n = 100$).

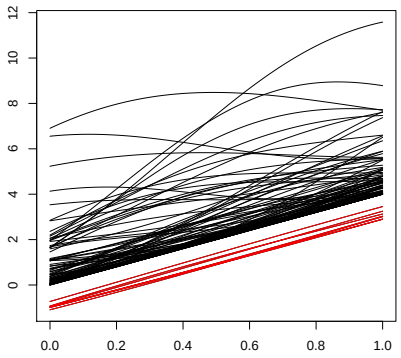
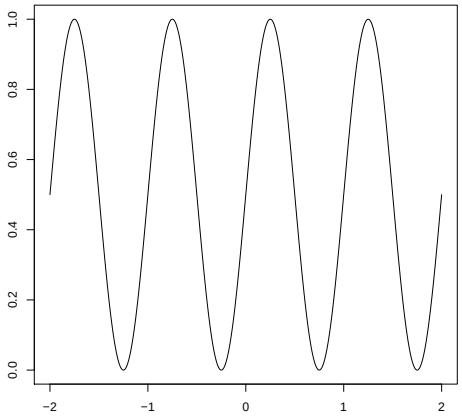
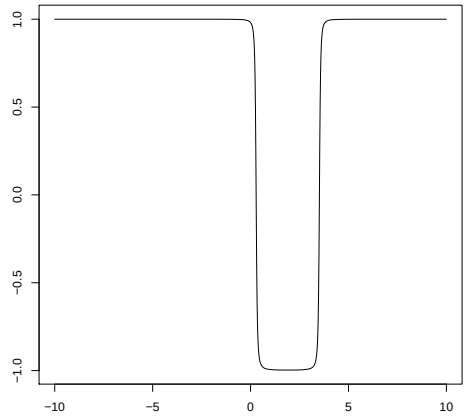


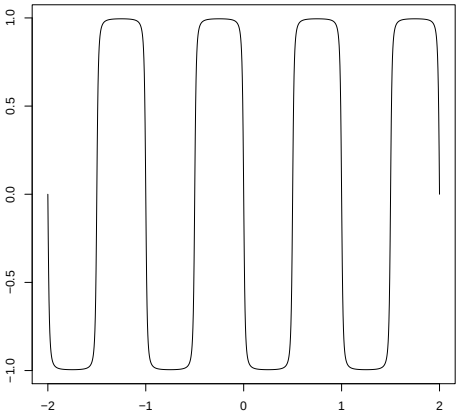
Figura 7.2: Muestra correspondiente la generación de $n = 100$ observaciones para el **Modelo 4** con $\epsilon = 0.10$ y desplazamiento $k = 2$.



(a) $\delta_1(t) = (\sin(2\pi t) + 1)/2$

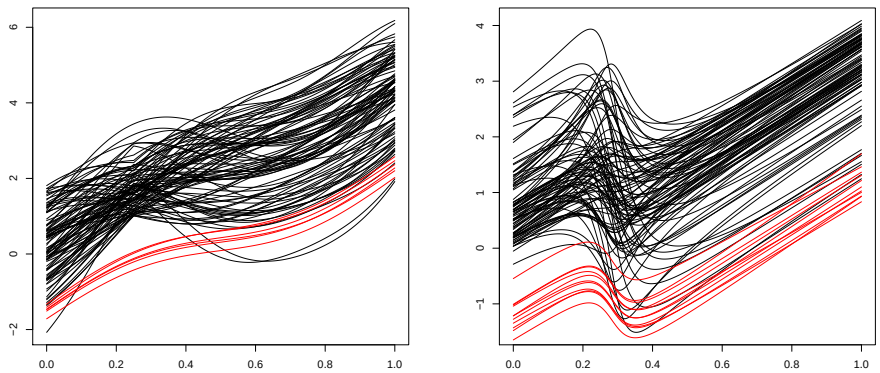


(b) $\delta_2(t) = \frac{\lambda_2(t)}{\sqrt{1+\lambda_2(t)^2}}; \lambda_2(t) = 5t^2 - 19t + 5$

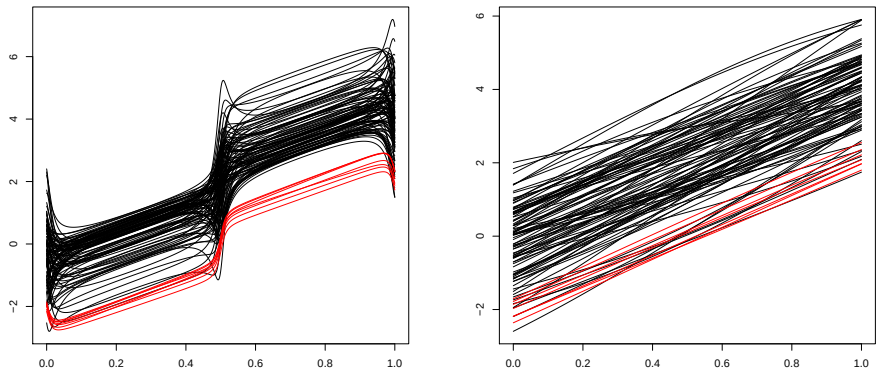


(c) $\delta_3(t) = \frac{\lambda_3(t)}{\sqrt{1+\lambda_3(t)^2}}$, $\lambda_3(t) = -10\sin(2\pi t)$

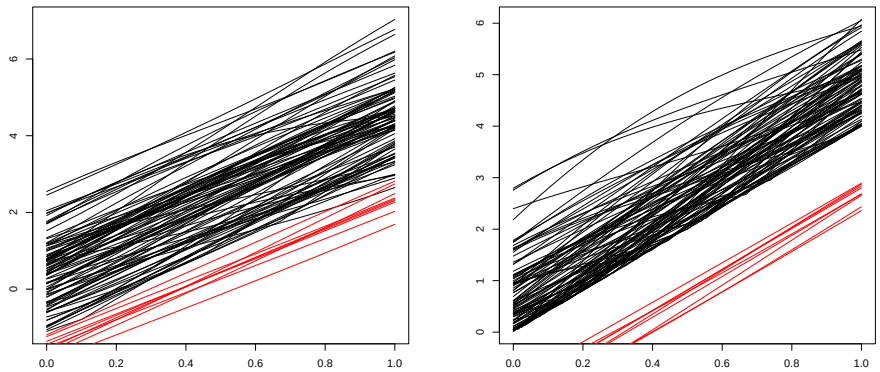
Figura 7.3: Funciones de asimetría para la generación de las observaciones del modelo definido por (5.3).



(a) $\delta_1(t)$ (b) $\delta_2(t)$



(c) $\delta_3(t)$ (d) $\delta = 0$



(e) $\delta = 0.5$ (f) $\delta = 1$

Figura 7.4: Muestras correspondientes a la generación de $n = 100$ observaciones para los modelos definidos por (5.3) y (5.5) con 10% de contaminación y desplazamiento $k = 2$.

7.2 Resultados

Las Tablas 7.1 a 7.12 resumen los resultados de la simulación para los modelos descriptos anteriormente y los distintos tipos de métodos de detección. Para detectar las observaciones atípicas, se consideraron el boxplot funcional de Sun y Genton (2011) y las propuestas de detección descriptas en las Secciones 6.2 y 6.3. Los resultados reportados en la Tablas corresponden al promedio sobre las 1000 replicaciones realizadas de la proporción de datos atípicos detectados como posibles outliers que indicaremos PO y de la proporción de observaciones no atípicas no detectadas como posibles outliers PB. En un marco ideal, en el que la regla de detección no comete error ambas medidas deben valer 1. La proporción PO corresponde a la sensibilidad mientras que PB a la especificidad del criterio de detección. Por otra parte, en las Tablas indicamos por AO, $Fbox$ y por $Fbox_{\text{CORR}}$ los resultados correspondientes al método que se basa en proyecciones descripto en la Sección 6.3, por el boxplot funcional descripto en la Sección 5.2 y por la modificación mediante el medcouple dada en la Sección 6.2, respectivamente.

Algunos comentarios sobre las tablas:

- Indicamos por $C_{\epsilon;k}$ a los resultados que corresponde a una proporción de outliers ϵ y un desplazamiento k como se indica en cada modelo.
- Para los datos sin contaminación solo consideró el caso $k = 0$ porque no hay outliers que desplazar.
- Los valores NaN corresponden a las proporciones PO de las corridas sin contaminación ya que se calcula dividiendo por la cantidad de outliers generados ($\epsilon \cdot n = 0$).
- La elección del método para la generación de las direcciones aleatorias no influye en las medidas de detección basadas en el boxplot funcional, es decir, para $Fbox$ y $Fbox_{\text{CORR}}$. Por esta razón, en la Tablas correspondientes a los métodos *al azar* y *muestral* para los **Modelos 2** y **3**, las columnas $Fbox$ (respectivamente, $Fbox_{\text{CORR}}$) arrojan los mismos datos.

	AO				$Fbox$		$Fbox_{\text{CORR}}$	
	azar		**muestral**					
n = 100	PO	PB	PO	PB	PO	PB	PO	PB
$C_{0;1}$	NaN	0.986	NaN	0.994	NaN	1.000	NaN	0.993
$C_{0.01;1}$	0.850	0.987	0.331	0.995	0.002	1.000	0.404	0.993
$C_{0.1;1}$	0.185	0.998	0.151	0.996	0.000	1.000	0.018	0.989
$C_{0.01;2}$	0.998	0.987	0.937	0.995	0.838	1.000	0.988	0.993
$C_{0.1;2}$	0.727	0.998	0.786	0.998	0.666	1.000	0.696	0.987

Tabla 7.1: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 100$ observaciones del **Modelo 1** de rango finito y para AO se tomaron direcciones al azar y con el método muestral.

	AO				$Fbox$		$Fbox_{\text{CORR}}$	
	azar		**muestral**					
n = 500	PO	PB	PO	PB	PO	PB	PO	PB
$C_{0;1}$	NaN	0.993	NaN	0.993	NaN	1.000	NaN	1.000
$C_{0.01;1}$	1.000	0.994	0.969	0.994	0.000	1.000	0.003	1.000
$C_{0.1;1}$	0.259	1.000	0.270	1.000	0.000	1.000	0.000	1.000
$C_{0.01;2}$	1.000	0.994	1.000	0.994	0.150	1.000	0.879	1.000
$C_{0.1;2}$	0.996	1.000	0.986	1.000	0.058	1.000	0.072	1.000

Tabla 7.2: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 500$ observaciones del **Modelo 1** de rango finito y para AO se tomaron direcciones al azar y con el método muestral.

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
azar	$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$	
n = 100	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.984	NaN	0.983	NaN	0.956	NaN	0.990	NaN	0.992	NaN	0.993	NaN	0.957	NaN	0.960	NaN	0.995
$C_{0.01;1}$	0.000	0.983	0.002	0.984	0.574	0.958	0.000	0.991	0.000	0.992	0.000	0.992	0.006	0.961	0.014	0.962	0.308	0.995
$C_{0.01;2}$	0.027	0.984	0.072	0.985	0.997	0.955	0.021	0.992	0.056	0.992	0.333	0.992	0.282	0.961	0.386	0.962	0.999	0.995
$C_{0.01;3}$	0.324	0.985	0.489	0.986	1.000	0.954	0.714	0.992	0.861	0.992	0.999	0.992	0.776	0.960	0.884	0.961	1.000	0.995
$C_{0.01;4}$	0.735	0.985	0.855	0.986	1.000	0.954	0.998	0.992	1.000	0.992	1.000	0.992	0.973	0.960	0.996	0.961	1.000	0.995
$C_{0.1;1}$	0.000	0.981	0.000	0.979	0.150	0.985	0.000	0.994	0.000	0.995	0.000	0.986	0.002	0.967	0.003	0.970	0.242	0.990
$C_{0.1;2}$	0.005	0.965	0.004	0.969	0.819	0.983	0.001	0.996	0.008	0.996	0.320	0.986	0.049	0.963	0.079	0.961	0.967	0.974
$C_{0.1;3}$	0.036	0.974	0.061	0.980	0.960	0.980	0.286	0.996	0.490	0.996	0.999	0.986	0.276	0.947	0.377	0.944	1.000	0.965
$C_{0.1;4}$	0.148	0.986	0.220	0.990	0.986	0.978	0.918	0.996	0.977	0.996	1.000	0.986	0.578	0.938	0.705	0.936	1.000	0.962

Tabla 7.3: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 100$ observaciones del **Modelo 2** con asimetrías constantes y para AO se tomaron direcciones al azar.

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
muestral	$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$	
n = 100	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.987	NaN	0.986	NaN	0.964	NaN	0.990	NaN	0.992	NaN	0.993	NaN	0.957	NaN	0.960	NaN	0.995
$C_{0.01;1}$	0.000	0.986	0.000	0.986	0.597	0.966	0.000	0.991	0.000	0.992	0.000	0.992	0.006	0.961	0.014	0.962	0.308	0.995
$C_{0.01;2}$	0.038	0.988	0.087	0.988	1.000	0.964	0.021	0.992	0.056	0.992	0.333	0.992	0.282	0.961	0.386	0.962	0.999	0.995
$C_{0.01;3}$	0.345	0.988	0.549	0.988	1.000	0.964	0.714	0.992	0.861	0.992	0.999	0.992	0.776	0.960	0.884	0.961	1.000	0.995
$C_{0.01;4}$	0.743	0.988	0.865	0.988	1.000	0.964	0.998	0.992	1.000	0.992	1.000	0.992	0.973	0.960	0.996	0.961	1.000	0.995
$C_{0.1;1}$	0.000	0.982	0.000	0.982	0.202	0.986	0.000	0.994	0.000	0.995	0.000	0.986	0.002	0.967	0.003	0.970	0.242	0.990
$C_{0.1;2}$	0.005	0.976	0.003	0.978	0.909	0.983	0.001	0.996	0.008	0.996	0.320	0.986	0.049	0.963	0.079	0.961	0.967	0.974
$C_{0.1;3}$	0.044	0.984	0.078	0.986	0.982	0.978	0.286	0.996	0.490	0.996	0.999	0.986	0.276	0.947	0.377	0.944	1.000	0.965
$C_{0.1;4}$	0.209	0.990	0.290	0.991	0.995	0.977	0.918	0.996	0.977	0.996	1.000	0.986	0.578	0.938	0.705	0.936	1.000	0.962

Tabla 7.4: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 100$ del **Modelo 2** con asimetrías constantes y para AO se tomaron direcciones con el método muestral.

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
azar	$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$	
n = 500	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.994	NaN	0.994	NaN	0.975	NaN	0.997	NaN	0.997	NaN	0.998	NaN	0.991	NaN	0.993	NaN	1.000
$C_{0.01;1}$	0.000	0.994	0.000	0.994	0.686	0.971	0.000	0.997	0.000	0.998	0.000	0.998	0.000	0.992	0.000	0.993	0.044	1.000
$C_{0.01;2}$	0.000	0.995	0.003	0.995	1.000	0.975	0.000	0.997	0.001	0.998	0.011	0.998	0.016	0.992	0.042	0.993	0.986	1.000
$C_{0.01;3}$	0.216	0.995	0.471	0.995	1.000	0.976	0.456	0.997	0.698	0.998	0.970	0.998	0.600	0.992	0.754	0.993	1.000	1.000
$C_{0.01;4}$	0.914	0.995	0.980	0.995	1.000	0.976	1.000	0.997	1.000	0.998	1.000	0.998	0.990	0.992	0.998	0.993	1.000	1.000
$C_{0.1;1}$	0.000	0.993	0.000	0.993	0.029	0.995	0.000	0.998	0.000	0.998	0.000	0.995	0.000	0.994	0.000	0.996	0.027	0.999
$C_{0.1;2}$	0.000	0.976	0.000	0.979	0.991	0.986	0.000	0.999	0.000	0.999	0.032	0.994	0.000	0.989	0.000	0.989	0.895	0.997
$C_{0.1;3}$	0.000	0.985	0.000	0.989	1.000	0.979	0.034	0.999	0.134	0.999	0.991	0.994	0.010	0.980	0.025	0.980	1.000	0.996
$C_{0.1;4}$	0.009	0.995	0.035	0.998	1.000	0.979	0.925	0.999	0.984	0.999	1.000	0.994	0.251	0.975	0.446	0.977	1.000	0.996

Tabla 7.5: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 500$ observaciones del **Modelo 2** con asimetrías constantes y para AO se tomaron direcciones al azar.

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
muestral	$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$		$\delta = 0$		$\delta = 0.5$		$\delta = 1$	
n = 500	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.995	NaN	0.995	NaN	0.981	NaN	0.997	NaN	0.997	NaN	0.998	NaN	0.991	NaN	0.993	NaN	1.000
$C_{0.01;1}$	0.000	0.995	0.000	0.995	0.685	0.975	0.000	0.997	0.000	0.998	0.000	0.998	0.000	0.992	0.000	0.993	0.044	1.000
$C_{0.01;2}$	0.001	0.996	0.004	0.995	1.000	0.980	0.000	0.997	0.001	0.998	0.011	0.998	0.016	0.992	0.042	0.993	0.986	1.000
$C_{0.01;3}$	0.211	0.996	0.467	0.996	1.000	0.981	0.456	0.997	0.698	0.998	0.970	0.998	0.600	0.992	0.754	0.993	1.000	1.000
$C_{0.01;4}$	0.884	0.996	0.974	0.996	1.000	0.982	1.000	0.997	1.000	0.998	1.000	0.998	0.990	0.992	0.998	0.993	1.000	1.000
$C_{0.1;1}$	0.000	0.993	0.000	0.992	0.037	0.995	0.000	0.998	0.000	0.998	0.000	0.995	0.000	0.994	0.000	0.996	0.027	0.999
$C_{0.1;2}$	0.000	0.981	0.000	0.986	0.995	0.986	0.000	0.999	0.000	0.999	0.032	0.994	0.000	0.989	0.000	0.989	0.895	0.997
$C_{0.1;3}$	0.000	0.993	0.000	0.995	1.000	0.980	0.034	0.999	0.134	0.999	0.991	0.994	0.010	0.980	0.025	0.980	1.000	0.996
$C_{0.1;4}$	0.009	0.998	0.039	0.999	1.000	0.980	0.925	0.999	0.984	0.999	1.000	0.994	0.251	0.975	0.446	0.977	1.000	0.996

Tabla 7.6: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 500$ observaciones del **Modelo 2** con asimetrías constantes y para AO se tomaron direcciones con el método muestral.

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
azar	$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$	
n = 100	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.984	NaN	0.981	NaN	0.979	NaN	0.996	NaN	0.992	NaN	0.989	NaN	0.963	NaN	0.979	NaN	0.980
$C_{0.01;1}$	0.007	0.982	0.450	0.981	0.822	0.980	0.000	0.996	0.000	0.992	0.004	0.989	0.056	0.966	0.057	0.981	0.303	0.982
$C_{0.01;2}$	0.374	0.985	0.970	0.982	0.999	0.980	0.021	0.996	0.535	0.992	0.687	0.989	0.886	0.966	0.806	0.980	0.979	0.981
$C_{0.01;3}$	0.899	0.985	0.998	0.981	1.000	0.980	0.849	0.996	0.997	0.992	1.000	0.989	0.999	0.965	0.994	0.980	1.000	0.981
$C_{0.01;4}$	0.988	0.985	1.000	0.981	1.000	0.980	1.000	0.996	1.000	0.992	1.000	0.989	1.000	0.965	1.000	0.980	1.000	0.981
$C_{0.1;1}$	0.001	0.976	0.120	0.981	0.237	0.990	0.000	0.997	0.000	0.997	0.001	0.994	0.018	0.973	0.012	0.986	0.150	0.989
$C_{0.1;2}$	0.032	0.982	0.536	0.994	0.775	0.996	0.006	0.997	0.232	0.997	0.391	0.994	0.522	0.959	0.302	0.966	0.752	0.973
$C_{0.1;3}$	0.245	0.991	0.836	0.996	0.947	0.996	0.587	0.997	0.960	0.997	0.996	0.994	0.907	0.943	0.750	0.957	0.949	0.964
$C_{0.1;4}$	0.507	0.995	0.946	0.996	0.991	0.996	0.997	0.997	1.000	0.997	1.000	0.994	0.982	0.934	0.938	0.954	0.993	0.962

Tabla 7.7: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos de detección propuestos. Se generaron $n = 100$ observaciones del **Modelo 3** con las asimetrías funcionales δ_1 δ_2 y δ_3 y para AO se eligieron direcciones al azar.

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
muestral	$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$	
n = 100	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.979	NaN	0.982	NaN	0.979	NaN	0.996	NaN	0.992	NaN	0.989	NaN	0.963	NaN	0.979	NaN	0.980
$C_{0.01;1}$	0.015	0.978	0.051	0.982	0.502	0.982	0.000	0.996	0.000	0.992	0.004	0.989	0.056	0.966	0.057	0.981	0.303	0.982
$C_{0.01;2}$	0.588	0.981	0.666	0.984	0.992	0.982	0.021	0.996	0.535	0.992	0.687	0.989	0.886	0.966	0.806	0.980	0.979	0.981
$C_{0.01;3}$	0.962	0.981	0.976	0.984	1.000	0.982	0.849	0.996	0.997	0.992	1.000	0.989	0.999	0.965	0.994	0.980	1.000	0.981
$C_{0.01;4}$	0.993	0.980	0.997	0.983	1.000	0.982	1.000	0.996	1.000	0.992	1.000	0.989	1.000	0.965	1.000	0.980	1.000	0.981
$C_{0.1;1}$	0.003	0.972	0.008	0.971	0.148	0.985	0.000	0.997	0.000	0.997	0.001	0.994	0.018	0.973	0.012	0.986	0.150	0.989
$C_{0.1;2}$	0.116	0.984	0.150	0.984	0.728	0.994	0.006	0.997	0.232	0.997	0.391	0.994	0.522	0.959	0.302	0.966	0.752	0.973
$C_{0.1;3}$	0.493	0.992	0.502	0.993	0.925	0.995	0.587	0.997	0.960	0.997	0.996	0.994	0.907	0.943	0.750	0.957	0.949	0.964
$C_{0.1;4}$	0.735	0.994	0.761	0.995	0.977	0.995	0.997	0.997	1.000	0.997	1.000	0.994	0.982	0.934	0.938	0.954	0.993	0.962

Tabla 7.8: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 100$ del **Modelo 3** con las asimetrías funcionales δ_1 δ_2 y δ_3 y para AO se eligieron direcciones con el método muestral.

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
azar	$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$	
n = 500	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.995	NaN	0.995	NaN	0.994	NaN	0.999	NaN	0.995	NaN	0.993	NaN	0.996	NaN	0.998	NaN	0.999
$C_{0.01;1}$	0.000	0.994	0.828	0.996	0.878	0.995	0.000	0.999	0.000	0.996	0.000	0.994	0.000	0.996	0.000	0.998	0.152	0.999
$C_{0.01;2}$	0.344	0.995	1.000	0.996	1.000	0.995	0.000	0.999	0.340	0.996	0.489	0.994	0.549	0.996	0.647	0.998	0.993	0.999
$C_{0.01;3}$	0.994	0.995	1.000	0.996	1.000	0.995	0.583	0.999	1.000	0.996	1.000	0.994	1.000	0.996	0.999	0.998	1.000	0.999
$C_{0.01;4}$	1.000	0.995	1.000	0.996	1.000	0.995	1.000	0.999	1.000	0.996	1.000	0.994	1.000	0.996	1.000	0.998	1.000	0.999
$C_{0.1;1}$	0.000	0.988	0.273	0.995	0.071	0.997	0.000	0.999	0.000	0.999	0.000	0.997	0.000	0.998	0.000	0.999	0.039	1.000
$C_{0.1;2}$	0.000	0.990	0.902	0.999	0.972	0.999	0.000	0.999	0.046	0.999	0.110	0.997	0.104	0.991	0.053	0.992	0.719	0.997
$C_{0.1;3}$	0.061	0.999	0.997	0.999	1.000	0.999	0.106	0.999	0.978	0.999	0.997	0.997	0.820	0.985	0.584	0.990	0.996	0.996
$C_{0.1;4}$	0.470	0.999	1.000	0.999	1.000	0.999	0.990	0.999	1.000	0.999	1.000	0.997	0.994	0.982	0.947	0.990	1.000	0.996

Tabla 7.9: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 500$ observaciones del **Modelo 3** con las asimetrías funcionales δ_1 δ_2 y δ_3 y para AO se eligieron direcciones al azar

	AO						$Fbox$						$Fbox_{\text{CORR}}$					
muestral	$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$		$\delta=\delta_1$		$\delta=\delta_2$		$\delta=\delta_3$	
n = 500	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.991	NaN	0.995	NaN	0.995	NaN	0.999	NaN	0.995	NaN	0.993	NaN	0.996	NaN	0.998	NaN	0.999
$C_{0.01;1}$	0.004	0.992	0.023	0.995	0.631	0.995	0.000	0.999	0.000	0.996	0.000	0.994	0.000	0.996	0.000	0.998	0.152	0.999
$C_{0.01;2}$	0.869	0.992	0.787	0.996	1.000	0.995	0.000	0.999	0.340	0.996	0.489	0.994	0.549	0.996	0.647	0.998	0.993	0.999
$C_{0.01;3}$	1.000	0.992	1.000	0.996	1.000	0.995	0.583	0.999	1.000	0.996	1.000	0.994	1.000	0.996	0.999	0.998	1.000	0.999
$C_{0.01;4}$	1.000	0.992	1.000	0.996	1.000	0.995	1.000	0.999	1.000	0.996	1.000	0.994	1.000	0.996	1.000	0.998	1.000	0.999
$C_{0.1;1}$	0.000	0.987	0.003	0.983	0.024	0.995	0.000	0.999	0.000	0.999	0.000	0.997	0.000	0.998	0.000	0.999	0.039	1.000
$C_{0.1;2}$	0.023	0.995	0.023	0.991	0.883	0.999	0.000	0.999	0.046	0.999	0.110	0.997	0.104	0.991	0.053	0.992	0.719	0.997
$C_{0.1;3}$	0.582	0.999	0.373	0.999	0.999	0.999	0.106	0.999	0.978	0.999	0.997	0.997	0.820	0.985	0.584	0.990	0.996	0.996
$C_{0.1;4}$	0.931	0.999	0.847	0.999	1.000	0.999	0.990	0.999	1.000	0.999	1.000	0.997	0.994	0.982	0.947	0.990	1.000	0.996

Tabla 7.10: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 500$ observaciones del **Modelo 3** con las asimetrías funcionales δ_1 δ_2 y δ_3 y para AO se eligieron direcciones con el método muestral.

	AO				$Fbox$		$Fbox_{\text{CORR}}$	
	azar		**muestral**					
n = 100	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.973	NaN	0.957	NaN	0.931	NaN	0.993
$C_{0.01;1}$	0.004	0.974	0.013	0.960	0.000	0.928	0.242	0.993
$C_{0.01;2}$	0.203	0.977	0.462	0.961	0.003	0.927	1.000	0.992
$C_{0.01;3}$	0.681	0.977	0.868	0.959	0.480	0.927	1.000	0.992
$C_{0.01;4}$	0.939	0.976	0.989	0.959	0.995	0.927	1.000	0.992
$C_{0.1;1}$	0.000	0.977	0.001	0.972	0.000	0.905	0.281	0.992
$C_{0.1;2}$	0.048	0.995	0.207	0.990	0.037	0.893	0.999	0.978
$C_{0.1;3}$	0.577	0.993	0.871	0.984	0.770	0.893	1.000	0.971
$C_{0.1;4}$	0.900	0.992	0.980	0.981	1.000	0.893	1.000	0.966

Tabla 7.11: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 100$ observaciones del **Modelo 4** y para AO se tomaron direcciones al azar y por el método muestral.

	AO				$Fbox$		$Fbox_{\text{CORR}}$	
	azar		**muestral**					
n = 100	PO	PB	PO	PB	PO	PB	PO	PB
C_0	NaN	0.988	NaN	0.981	NaN	0.949	NaN	0.999
$C_{0.01;1}$	0.000	0.988	0.000	0.982	0.000	0.947	0.115	0.999
$C_{0.01;2}$	0.016	0.990	0.088	0.982	0.000	0.947	0.999	0.999
$C_{0.01;3}$	0.575	0.989	0.816	0.981	0.090	0.947	1.000	0.999
$C_{0.01;4}$	0.988	0.988	0.998	0.980	0.984	0.947	1.000	0.999
$C_{0.1;1}$	0.000	0.981	0.000	0.982	0.000	0.927	0.188	0.999
$C_{0.1;2}$	0.003	0.997	0.013	0.996	0.000	0.917	0.999	0.994
$C_{0.1;3}$	0.904	0.994	0.968	0.992	0.579	0.917	1.000	0.992
$C_{0.1;4}$	1.000	0.992	1.000	0.989	1.000	0.917	1.000	0.991

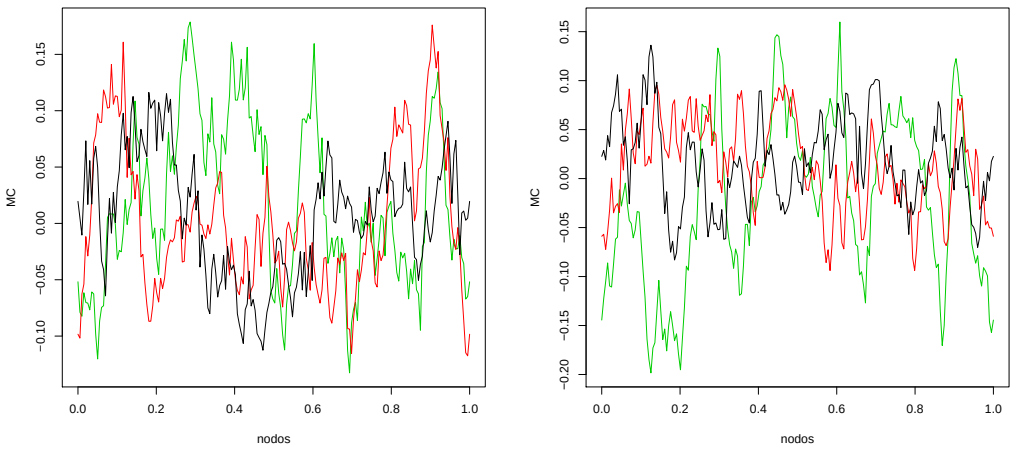
Tabla 7.12: Promedio sobre las replicaciones de la proporción de outliers detectados PO y de observaciones regulares no detectadas como atípicas PB para los tres métodos. Se generaron $n = 500$ observaciones del **Modelo 4** y para AO se tomaron direcciones al azar y por el método muestral.

7.3 Conclusiones

A partir de los resultados se ve claramente que un factor determinante para la clasificación de datos atípicos es la diferencia entre la media de las observaciones y de los outliers

generados. Cuanto mayor sea la separación, los outliers se "confunden" menos con las observaciones y aumenta la proporción de su detección (PO). Para los tres métodos evaluados en todas las situaciones, se confirma el mismo fenómeno discutido en el caso multivariado al final del del Capítulo 4 respecto de la propuesta de Hubert y Van der Veeken (2008).

En el **Modelo 1** el método de detección por proyecciones, AO, obtiene mejores resultados para ambos tipos de direcciones, al azar y muestral que los procedimientos basados en boxplots funcionales y esto puede atribuírse a la estructura intríinsecamente finito–dimensional de las observaciones. El boxplot de Genton, $Fbox$, tiene un nivel bajo de detección superado ligeramente por su versión corregida, $Fbox_{\text{CORR}}$. Para valores altos de contaminación, $\epsilon = 0.1$ el valor de MC aumenta inflando de más la región exterior del boxplot, por lo que la mayoría de los outliers generados caen dentro de los límites de clasificación. La Figura 7.5 muestra el valor de corrección de MC como función de t para una muestra de $n = 500$ observaciones del **Modelo 1** de rango finito. La línea negra corresponde a los datos sin contaminar y mientras que las líneas roja y verde corresponden a la muestra con un porcentaje del 1% y 10% de contaminación y $k = 2$, respectivamente. Para el desplazamiento $k = 2$ el rango de valores de la corrección por MC aumenta con la contaminación. Esta sobre-corrección enmascara los outliers.



(a) $k = 1$ (b) $k = 2$

Figura 7.5: Corrección por MC en cada coordenada para una muestra de $n = 500$ observaciones de rango finito con $\epsilon = 0\%$ (negro), $\epsilon = 0.01\%$ (rojo), $\epsilon = 0.1\%$ (verde).

Para los **Modelos 2 y 3** también es esperable que $Fbox$ no funcione bien ante datos asimétricos. La buena detección registrada para los valores más altos del corrimiento ($k = 3$ y $k = 4$) se debe a que los outliers generados quedan efectivamente separados del grueso

de los datos. Notemos que el parámetro de separación está vinculado con la asimetría, es claro para las asimetrías constantes que no ofrecen la misma separación entre observaciones y outliers para el valor fijo $k = 2$ (Figura 7.4).

En las separaciones intermedias, tanto el método AO como el boxplot corregido, $Fbox_{\text{CORR}}$, igualan o superan la performance del boxplot funcional de Sun y Genton (2011). Sólo en el caso simétrico, $\delta = 0$, domina sobre los otros, como es de esperar.

Con las asimetrías funcionales la performance es distinta para δ_1, por un lado, y δ_2, δ_3 por otro. Recordemos que, como diferencia, en el primer caso la función que da la asimetría es positiva mientras que las otras dos tienen como imagen el intervalo $[-1, 1]$. Las observaciones generadas tienen, coordenada a coordenada, distintos valores de asimetría que incluso cambian de signo. Recordemos que la Figura 3.3 mostraba la relación entre el valor de asimetría δ y el medcouple MC asociado para una distribución $\mathcal{SN}(\lambda)$ con $\lambda = \delta/\sqrt{1-\delta^2}$. Los casos más extremos de asimetría para cada coordenada, o sea, aquellos para los cuales $|\delta| = 1$ el valor absoluto del medcouple no supera el valor 0.2. En esta situación, otra vez el valor de MC tiene un amplio rango de variación.

Para la función de asimetría δ_1, $Fbox_{\text{CORR}}$ es mejor que el método AO pero para las funciones δ_2 y δ_3 ambos se comportan similarmente bien. Lo mismo sucede respecto del método elegido para generar las direcciones en el c"omputo de AO, para la función δ_1 la generación muestral da mejores resultados mientras que para δ_2 y δ_3 las direcciones al azar ofrecen mejor detección.

Para las asimetrías constantes correspondientes al **Modelo 2** (modelo SGP definido por 5.3), $Fbox_{\text{CORR}}$ en general se comporta bien. El método AO detecta mejor para $\delta = 1$ que para $\delta = 0$ y $\delta = 0.5$ donde el funcional de medcouple toma los valores 0 y 0.0054, respectivamente. Para datos más asimétricos, $\delta = 1$ el método AO iguala al boxplot funcional corregido $Fbox_{\text{CORR}}$ a quien supera en su capacidad de detección cuando $\epsilon = 0.01$ y $k = 1$.

Además de la separación entre las distribuciones de las observaciones y outliers, otro factor importante es el nivel de contaminación. En todas las tablas, la detección de outliers decae al aumentar su proporción. Recordemos que existen limitaciones tanto para el medcouple como para el boxplot ajustado. El medcouple, si bien tiene un punto de ruptura teórico de 25%, comienza a tener un sesgo para contaminaciones más altas. En Brys *et al.* (2004) los autores dan cuenta de este problema para contaminaciones del 15%. La definición de MC involucra la resta de una observación contra todo otro grupo de observaciones. Para cada outlier se introducen diferencias espúreas en la muestra que se multiplican con una alta proporción de outliers. Las altas contaminaciones afectan la robustez del medcouple.

Por último, el **Modelo 4** es el de asimetríia más fuerte. En este caso, todas las marginales del proceso $Z(t)-4t$ tienen distribución χ^2_1 para quien el funcional de medcouple

toma el valor 0.5048, Recordemos que, para definir los bigotes del boxplot ajustado, el límite de asimetría con el que trabajaron Hubert y Vandervieren (2008) fue $MC = 0.6$. El método $Fbox_{\text{CORR}}$ tiene buen poder de detección para ambos tamaños de muestras y niveles de contaminación. El método AO lo sigue con un mejor desempeño con las direcciones muestrales. Nuevamente, la separación entre los outliers y las observaciones determinan la bondad de la detección.

En líneas generales, la simulación no arroja un método uniformemente superador. Sí confirma la necesidad de mejorar el método de clasificación del boxplot de Sun y Genton (2011) y ofrece como opciones tanto al método por proyecciones, AO, como la modificación mediante el medcouple, $Fbox_{\text{CORR}}$ como caminos viables. En principio, la generación muestral de direcciones parece beneficiar la detección de los datos atípicos por sobre la generación al azar. Además, la generación muestral es computacionalmente menos costosa que su contrapartida al azar. Según el tipo de asimetría ambos métodos pueden detectar outliers cuya distribución se solape con el grueso de los datos. Es decir, los métodos son sensibles frente a outliers generados cerca del proceso original. Por último, ante las asimetrías más fuertes es $Fbox_{\text{CORR}}$ el método que parece predominar.

A la luz de estos resultados corresponde estudiar en detalle la dependencia con la distancia al centro de los outliers y los distintos tipos de asimetría. A su vez, se deben estudiar valores de contaminación menores y analizar la robustez del MC en este contexto y cuando la contaminación se produce en el mismo sentido de la asimetría. Por otro lado, queda pendiente determinar si los parámetros fijados en la Sección 2.3 para la corrección de los bigotes del boxplot ajustado son adecuados para el contexto funcional. Los próximos pasos podrían ir en este sentido.

Capítulo 8

Apéndice

8.1 Códigos

8.1.1 funsimulacion.R

```
#####################################################################
# RUTINA para n.rep replicaciones de 'generacion y deteccion' #
#####################################################################

source('replicacion.R')

corrida = function(n.samples,nodos,tipo.obs,tipo.dir,skew,out.ratio,out.shift){
#--------------------------------------------------------------------
# INPUT:
#   n.samples: cantidad de observaciones a generar
#   nodos: particion del intervalo
#   tipo.obs: metodo para generar las observaciones
#   tipo.dir: metodo para generar las direcciones
#   skew: grado de asimetria para el proceso (funcion o constante)
#   out.ratio: proporcion de outliers
#   out.shift: diferencia entre las medias de las obs y los out
#--------------------------------------------------------------------

  n.rep = 1000 #numero de replicaciones
  diregaus = NULL
  for(irep in 1:n.rep){
    print(irep)
    set.seed(999 + irep)
    r = replicacion(irep,n.samples,nodos,tipo.obs,tipo.dir,skew,out.ratio,out.↩
        shift,FALSE,diregaus=NULL)
    diregaus = r #reusa las direcciones al azar para cada corrida
  }
}
```

8.1.2 replicacion.R

```
#############################################################
# RUTINA para UNA replicacion de generacion y deteccion #
#############################################################
source('deteccion.R')

replicacion = function(irep,n.samples,nodos,tipo.obs,tipo.dire,skew,out.ratio,out.↩
    shift,plt,diregaus=NULL){

  # Parametros de las observaciones
  if(tipo.obs=='func'){
    media = function(x) 4*x # media del proceso
    obs = list( n = n.samples, # cantidad de observaciones
          # parametros para generar las observaciones
          media = media, # media del proceso
          cov = list(
            type = c("cuadratica","cuadratica"), # tipo de covarianza
            scale = c(1,1), # escala de la covarianza
            nu = c(1,1) # parametro de la covarianza
            ),
          wgn = FALSE, # termino white noise
          skew = skew, # parametro de asimetria en [0,1]
          # parametros para generar outliers
          out = list(
            ratio = out.ratio,
            shift = out.shift,
            cov = list(
              type = c("cuadratica","cuadratica"), #operador
              scale = c(1/20,1/20), #escala
              nu = c(1,1) #parametro de la covarianza
            )
          ),
          plt = FALSE
        )
  }

  if(tipo.obs=='nofunc'){
    n.term = c(5,0,FALSE)
    obs = list( n = n.samples, # cantidad de observaciones
          n.term = n.term, # cantidad de terminos de la parte finita, infinita y ↩
              browniana
          basis.tipo = "fourier", # base
          skew = skew,
          out = list(ratio=out.ratio,shift=out.shift),
          plt = FALSE)
  }

  ## Parametros de las direcciones
  if(tipo.dire=='sample'){dire = list(metodo = 'MBD',plt = FALSE)} #tipo de ↩
      profundidad de banda: 'MBD','BD2','Both'
  if(tipo.dire=='gauss') #direcciones de un proceso gausiano
    {cov = list(type = "cuadratica",scale = 2,nu = 1)
    dire = list(n = n.samples*10,diregaus = diregaus,cov = cov,plt = FALSE)}
  if(tipo.dire=='basis'){dire = list(tipo="fourier",n=11,plt = FALSE)} #las ↩
      direcciones se generan en alguna base fija
```

```
  #### Deteccion ###
  det = deteccion(tipo.obs,obs,tipo.dire,dire,nodos,plt)
  out.gener = det$out.gener
  out.detect = det$out.detect
  out.genton = det$out.genton
  out.genton.asim = det$out.genton.asim

  ### Performance ###
  nuestros = performance(n.samples,out.gener,out.detect)
  genton = performance(n.samples,out.gener,out.genton)
  genton.asim = performance(n.samples,out.gener,out.genton.asim)

  #### Guardo todo ####
  res.nuestros = c(irep,length(out.gener),length(out.detect),unlist(nuestros))
  res.genton = c(irep,length(out.gener),length(out.genton),unlist(genton))
  res.genton.asim = c(irep,length(out.gener),length(out.genton.asim),unlist(genton↩
      .asim))

  if(tipo.obs=="func"){
    covobs = paste(substr(obs$cov$type[1],1,2),substr(obs$cov$type[2],1,2),sep="")
    covout = paste(substr(obs$out$cov$type[1],1,2),substr(obs$out$cov$type[2],1,2)↩
        ,sep="")

    nombre.nuestro = paste("salida_tipo.obs",tipo.obs,"_n_",n.samples,"_asim_",↩
        skew$tipo,"_dirProj_",tipo.dire,"_propout_",obs$out$ratio,"_shiftOUT_",out↩
        .shift,"TipoCOV_",covobs,"COVOUT_",covout,".txt",sep="")
    nombre.genton = paste("genton_tipo.obs",tipo.obs,"_n_",n.samples,"_asim_",skew↩
        $tipo,"_dirProj_",tipo.dire,"_propout_",obs$out$ratio,"_shiftOUT_",out.↩
        shift,"TipoCOV_",covobs,"COVOUT_",covout,".txt",sep="")
    nombre.genton.asim = paste("gentonasim_tipo.obs",tipo.obs,"_n_",n.samples,"_↩
        asim_",skew$tipo,"_dirProj_",tipo.dire,"_propout_",obs$out$ratio,"_↩
        shiftOUT_",out.shift,"TipoCOV_",covobs,"COVOUT_",covout,".txt",sep="")
  }

  if(tipo.obs=="nofunc"){
    nombre.nuestro = paste("salida_tipo.obs",tipo.obs,"_n_",n.samples,"_asim_",↩
        skew$tipo,"_dirProj_",tipo.dire,"_propout_",out.ratio,".txt",sep="")
    nombre.genton = paste("genton_tipo.obs",tipo.obs,"_n_",n.samples,"_asim_",skew↩
        $tipo,"_dirProj_",tipo.dire,"_propout_",out.ratio,".txt",sep="")
    nombre.genton.asim = paste("gentonasim_tipo.obs",tipo.obs,"_n_",n.samples,"_↩
        asim_",skew$tipo,"_dirProj_",tipo.dire,"_propout_",out.ratio,".txt",sep=""↩
        )
  }

  if(!file.exists("resultados")){dir.create("resultados")}
  write(res.nuestros,file=paste("resultados/",nombre.nuestro,sep=""),ncol=length(↩
      res.nuestros),append=T)
  write(res.genton,file=paste("resultados/",nombre.genton,sep=""),ncol=length(res.↩
      genton),append=T)
  write(res.genton.asim,file=paste("resultados/",nombre.genton.asim,sep=""),ncol=↩
      length(res.genton.asim),append=T)

  return(det$diregaus)
}

performance = function(totales,generados,detectados){
  nodetectados = setdiff(1:totales,detectados)
```

```
  tp = intersect(generados,detectados)#true positives
  fp = setdiff(detectados,tp)#false positives
  fn = setdiff(generados,tp)#false negatives
  tn = setdiff(nodetectados,fn)#true negatives
  sen = length(tp)/(length(tp) + length(fn)) #sensibilidad
  esp = length(tn)/(length(tn) + length(fp)) #especificidad
  prop.detect = length(tp)/length(generados) #proporcion detectada
  return(list(fp=length(fp),fn=length(fn),tp=length(tp),tn=length(tn),sen=sen,esp=esp,prop.detect=prop.detect))
}
```

8.1.3 deteccion.R

```
# Dependencias
source("auxiliares.R")
source("generacion.R")
source("direccion.R")
source("atipicidad.R")
source("ffbplot.R")
source("ffbplot_adj.R")

deteccion = function(tipo.obs,obs,tipo.dire,dire,nodos,plt=FALSE){
  ## Genero las observaciones ##
  if(tipo.obs=='nofunc'){
    aux = obs.nofunc(obs$n,nodos,obs$basis.tipo,obs$n.term,obs$out,obs$skew,obs$plt)
    observ = aux$values
    out.gener = aux$out.index
  }
  if(tipo.obs=='func'){
    if(is.function(obs$skew$valor)){obs$out$media = function(x) obs$media(x) + obs$skew$valor(x)*sqrt(2/pi) - obs$out$shift}
    else{obs$out$media = function(x) obs$media(x) + obs$skew$valor *sqrt(2/pi) - obs$out$shift}
    aux = obs.func(obs$n,nodos,obs$media,obs$cov,obs$wgn,obs$skew,obs$out,obs$plt)
    observ = aux$values
    out.gener = aux$out.index
  }

  ## Genero las direcciones ##
  switch(tipo.dire,
    sample = {direc = dir.sample(observ,nodos,dire$metodo,dire$plt)}, #las direcciones se generan de la muestra
    gauss = {
        if(is.null(dire$diregaus)){direc = dir.gauss(dire$n,nodos,dire$cov$type,dire$cov$scale,dire$cov$nu,dire$plt)}
        else{direc = dire$diregaus}
        }, #direcciones de un proceso gausiano
    basis = {direc = dir.basis(dire$tipo,dire$n,nodos,dire$plt)} #las direcciones se generan en alguna base fija
  )

  ## Proyecto las observaciones en las direcciones
  p = L2.product(observ,direc,nodos) #cada fila es una direccion
```

```
  ## Calculo en cada proyeccion la atipicidad ajustada funcional y declaro ↩
      outliers
  atip = atipicidad(p,4,3)

  ## Calculo los outliers de Genton
  f = fbplot(observ,main = "Boxplot de Genton",method="MBD",plot=FALSE)
  ## Calculo los outliers de Genton modificado
  f.asim = fbplot.adj(observ,main = "Boxplot de Genton asimetrico",method="MBD",↩
      plot=FALSE)

  # Outliers generados, detectados por el metodo, por Genton y Genton modificado
  out.detect = which(atip$out==TRUE) # outliers
  out.genton = f$outpoint
  out.genton.asim = f.asim$outpoint

  ###### Graficos (opcionales) ##############
  if(plt)
  {
    # Cierro todas las ventanitas
    graphics.off()
    #
    par(mfrow=c(1,2),oma = c(0,0,3,0),new)
    b=boxplot(atip$adjout, main = "Boxplot")
    a=adjbox(atip$adjout, main = "Boxplot ajustado")
    mtext("Atipicidad ajustada", outer = TRUE, cex = 1.5)
    dev.new()
    #
    par(mfrow=c(2,2))
    matplot(nodos,observ,type="l",main=paste("Observaciones",tipo.obs),col="gray",↩
        xlab = "t")
    for(j in out.gener){lines(nodos,observ[,j],main=paste("Observaciones",tipo.obs↩
        ),col=3)}
    matplot(nodos,observ[,-out.gener],type="l",main="Observaciones sin contaminar"↩
        ,xlab = "t")
    matplot(nodos,observ[,out.gener],type="l",main = "Outliers generados",xlab = "↩
        t")
    matplot(nodos,observ[,out.detect],type="l",main = "Outliers detectados",xlab =↩
         "t")
    dev.new()
    #
    #plot(atip$adjout,type="l",ylim=c(0,1.2*atip$cutoff),xlab="Observaciones",ylab↩
        ="Atipicidad ajustada")
    plot(atip$adjout,type="l",ylim=c(-0.1,0.1),xlab="Observaciones",ylab="↩
        Atipicidad ajustada")
    hist(atip$adjout)
    abline(h=atip$cutoff); abline(h=atip$Qalph.adjout[2])
    for(j in out.gener){points(j,atip$adjout[j],col="red")}
  }

  return(list(out.gener = out.gener,out.detect = out.detect,out.genton = out.↩
      genton,out.genton.asim = out.genton.asim,diregaus = direc))
}
```

8.1.4 atipicidad.R

```
require("robustbase")

atipicidad = function(datos,clower=3,cupper=4,alpha.cutoff = 0.75, coef = 1.5)
#-------------------------------------------------------------------------
# Skewness-Adjusted Outlyingness: rutina que calcula la atipicidad ajustada
# por asimetria de un conjunto de observaciones (ya proyectadas)
# Calcula su boxplot ajustado y clasifica datos at picos
#-------------------------------------------------------------------------
# INPUT:
#   clower,cupper: coeficientes del boxplot ajustado (limites de la caja)
#   datos: matriz pxn de n observaciones proyectadas en p direcciones
#   alpha.cutoff: parametro que define el ancho de la caja del boxplot
#   coef: coeficiente que hincha la caja del boxplot
# OUTPUT:
#   adjout: vector de atipicidades
#   MCadjout: medcouple de la muestra de adjout
#   Qalph.adjout: cuantiles para el boxplot ajustado de adjout
#   cutoff: valor de corte para clasificar outliers
#   out: outliers (observaciones cuya atipicidad supera cutoff)
#-------------------------------------------------------------------------
{
  n = ncol(datos)
  p = nrow(datos)
  stopifnot(n >= 1, p >= 1, is.numeric(datos))

  # Calculo la mediana y centro los datos
  med = apply(datos, MARGIN = 1, median)
  datos = datos - rep(med, n)

  # Calculo el medcouple
  tmc = apply(datos, MARGIN = 1, mc)

  # Caja
  Q3 = apply(datos, MARGIN = 1, quantile, 0.75)
  Q1 = apply(datos, MARGIN = 1, quantile, 0.25)
  IQR = Q3-Q1

  # Defino los limites de la caja
  tup = Q3 + coef*IQR*exp( cupper*tmc*(tmc >= 0) + clower*tmc*(tmc < 0))
  tlo = Q1 - coef*IQR*exp(-clower*tmc*(tmc >= 0) - cupper*tmc*(tmc < 0))

  # Defino los bigotes (observacion tipica mas extrema)
  datosup = datoslo = datos
  datosup[!(datos < rep(tup, n))] = -Inf
  datoslo[!(datos > rep(tlo, n))] =  Inf
  tup =  apply(datosup, 1, max) # =  max{ Y[i,] ; Y[i,] < tup[i] }
  tlo = -apply(datoslo, 1, min) # = -min{ Y[i,] ; Y[i,] > tlo[i] }

  datospos = (datos >= 0) # TRUE para los datos positivos

  # Calculo la atipicidad ajustada en cada direccion
  aux = datospos*matrix(rep(tup,n),p,n)+(1-datospos)*matrix(rep(tlo,n),p,n)
  aux = replace(aux,which(abs(aux)<=10^(-50)),10^(-50)*sign(aux))
  datos = abs(datos)/aux

  adjout = apply(datos, 2, function(x) max(x[is.finite(x)]))
  adjout = adjout^2
```

```
  Qadj = quantile(adjout, probs = c(1 - alpha.cutoff, alpha.cutoff))
  mcadjout = mc(adjout)#[is.finite(adjout)])

  if(0>mcadjout){print("MCAdjOut es negativa: ALARMA")}
  cutoff = Qadj[2] + coef*(Qadj[2] - Qadj[1])*(if(mcadjout>0){exp(3*mcadjout)}else↩
      {1})

  #Devuelvo todo
  return(list(adjout=adjout,MCadjout=mcadjout,Qalph.adjout=Qadj,cutoff=cutoff,out↩
      =(cutoff < adjout)))
}
```

8.1.5 generacion.R

```
#-------------------------------------------------------------------------------
# Rutinas para generar datos funcionales (con o sin outliers) con asimetria
#-------------------------------------------------------------------------------

# Dependencias
require(fda) # libreria para las bases
require(sn) # libreria para la normal asimetrica
source("gausiano.R")

obs.func = function(n.samples,nodes,media,cov,wgn,skew,out,plt=FALSE)
#-------------------------------------------------------------------------------
# Caso funcional (propuesta de Zhang):
# observaciones asimetricas CON outliers gausianos
#-------------------------------------------------------------------------------
# INPUT:
#   n.samples: cantidad de observaciones a generar
#   nodes: particion del intervalo
#   media: 'media' del proceso (funcion)
#   cov: una lista con c(type,scale,nu) (string, vector R2, vector R2)
#   wgn: booleano para sumar un ruido blanco
#   skew: grado de asimetria para el proceso (funcion o constante)
#   out: lista con las cosas de los outliers
#   plt: si TRUE hace un plot las observaciones
# OUTPUT:
#   values: matriz con las observaciones generadas como columnas
#-------------------------------------------------------------------------------
{
  values = matrix(0,length(nodes),n.samples) # Inicializo

  out.pos = as.logical(rbinom(n.samples,1,out$ratio)) # Indices de contaminacion
  out.index = which(out.pos==TRUE)
  lout = length(out.index)
  mu = function(x){rep(0,length(x))} # Media de los GP

  wgn.values = 0
  if(wgn!=0){wgn.values = gausiano(n.samples,nodes,mu,"wn",NULL,wgn)} # White ↩
      noise

  g2 = gausiano(n.samples-lout,nodes,mu,cov$type[2],cov$scale[2],cov$nu[2])
```

```
  if(is.function(skew$valor)){
    g1 = rnorm(n.samples - lout)
    sigma = skew$valor(nodes)
    values[,!out.pos] = replicate(n.samples-lout,media(nodes)) + sigma %*% t(abs(↩
        g1)) + sqrt(1-sigma^2) * g2 + wgn.values
  }
  else{
    g1 = gausiano(n.samples - lout,nodes,mu,cov$type[1],cov$scale[1],cov$nu[1])
    sigma = c(skew$valor,sqrt(1-skew$valor^2))
    m = replicate(n.samples-lout,media(nodes))
    values[,!out.pos] = m + sigma[1] * abs(g1) + sigma[2] * g2 + wgn.values
  }

  #Contamino la muestra
  if(lout!=0){
    values[,out.pos] = replicate(lout,out$media(nodes)) + gausiano(lout,nodes,mu,↩
        out$cov$type[2],out$cov$scale[2],out$cov$nu[2])
  }

  # Grafico
  if(plt){
    matplot(nodes,values[,!out.pos],type="l",col="black",main="Observaciones")
    if(lout!=0){matlines(nodes,values[,out.pos],col="red")}
  }

  return(list(values=values,out.index=out.index))
}

obs.nofunc = function(n.samples,nodes,basis.tipo,n.term,out,skew,plt=FALSE)
#-------------------------------------------------------------------------
# Caso de dimensi n finita, infinita o componente browniana
#-------------------------------------------------------------------------
# INPUT:
#   n.samples: cantidad de observaciones a generar
#   nodes: partici n del intervalo
#   basis.tipo: string define el tipo de base ('fourier', 'bsplines', 'polinomios↩
    ')
#   n.term: vector con la cantidad de terminos de la parte finita, infinita y
#       browniana
#   out.ratio: probabilidad de contaminar una observacion
#   skew: vector de asimetria para la normal asimetrica multivariada
#   plt: si TRUE hace un plot las observaciones
# OUT
#   values: matriz con las observaciones generadas como columnas
#-------------------------------------------------------------------------
{
  # Inicializo
  obs.fin = obs.inf = obs.brw = values = matrix(0,length(nodes),n.samples)

  # Genero una discretizacion en los nodos de una base ON
  rangeval = c(nodes[1],nodes[length(nodes)])
  basisobj = base(basis.tipo,rangeval,n.term[1] + n.term[2])
  basiseval = getbasismatrix(nodes,basisobj)
  basisONeval = L2.gramm.schmidt(basiseval,nodes)

  # Genero coeficientes
  coefs.fin = sn.coef(n.samples,skew$valor)
```

```
  # Armo la parte finita
  obs.fin = basisONeval[,1:n.term[1]] %*% coefs.fin

  # Armo la parte "infinita" (truncada hasta n.term[2])
  if(n.term[2] > 0){
    lambda = sapply(n.term[1]+1:n.term[2],FUN=function(x) 10^(-x));
    print(paste("La varianza del termino infinito es: ",sum(lambda)))
    coefs.inf = t(sapply(lambda, function(x) rnorm(n.samples,0,x)))
    obs.inf = basisONeval[,n.term[1]+1:n.term[2]] %*% coefs.inf
  }

  # Armo la observacion sin outliers
  obs = obs.fin + obs.inf

  # Genero los outliers (y los marco)
  out.pos = rbinom(n.samples,1,out$ratio)
  out.cant = sum(out.pos)
  out.values = 0
  if(out.cant!=0){
  out.coef = matrix(rnorm(n.term[1]*out.cant,0,1/sqrt(20)),n.term[1],out.cant)
  out.values = basiseval[,1:n.term[1]]%*%matrix(-out$shift,n.term[1],out.cant)+↩
      basiseval[,1:n.term[1]]%*% out.coef
  }
  # Sumo un browniano si term.brw = TRUE
  if(n.term[3]){out.values = out.values + browniano(out.cant,nodes)}

  values[,out.pos==1] = out.values
  values[,out.pos!=1] = obs[,out.pos!=1]

  # Grafico
  if(plt){matplot(nodes,values,type="l",main="Observaciones")}

  return(list(values=values,out.index=which(out.pos==1),coefs=coefs.fin))
}
```

8.1.6 direccion.R

```
#-------------------------------------------------------------------------
# Rutinas para generar direcciones de proyeccion
#-------------------------------------------------------------------------

require('MASS') # paquete para generar normales multivariadas 'mvrnorm'
source('ffbplot.R')# paquete con el codigo del boxplot funcional (depth)
source('generacion.R') # paquete para la funcion que genera la base

dir.basis = function(tipo="fourier",n.basis=5,nodes=seq(0,1,0.001),plt)
#-------------------------------------------------------------------------
# Genera direcciones ortonormales a partir de una base dada
#-------------------------------------------------------------------------
# INPUT:
#   n.basis: cantidad de elementos generados de la base
#   tipo: tipo de base ('fourier','bsplines','polinomios')
#   nodos: particion del intervalo
#   plt: si TRUE hace un plot las direcciones
```

```
# OUTPUT:
#   values: matriz con las direcciones generadas como columnas
#-------------------------------------------------------------------------------
{
  # Evaluo la base en los nodos
  rangeval = c(nodes[1],nodes[length(nodes)])
  basisobj = base(tipo,rangeval,n.basis)
  basiseval = getbasismatrix(nodes,basisobj)
  values = L2.gramm.schmidt(basiseval,nodes)

  # Grafico todas las direcciones
  if(plt){matplot(nodes,values,type="l",main=paste("Direcciones muestrales"))}

  return(values)
}

dir.gauss = function(n.samples,nodes,cov.type,cov.scale,cov.nu,plt = FALSE)
#-------------------------------------------------------------------------------
# Genera direcciones normalizadas de un proceso gausiano con
# media cero y una funcion covarianza dada.
#-------------------------------------------------------------------------------
# INPUT:
#   n.samples: cantidad de direccciones a generar.
#   nodes: particion del intervalo
#   cov.type: tipo de funcion de covarianza (string)
#   cov.scale: parametros de escala para la covarianza
#   plt: si TRUE hace un plot las observaciones
# OUTPUT:
#   values: matriz con las direcciones normalizadas como columnas
#-------------------------------------------------------------------------------
{
  #fijo la semilla
  #set.seed(semilla)
  # Defino la media y genero las direcciones
  media = function(x) 0*x
  values = gausiano(n.samples,nodes,media,cov.type,cov.scale,cov.nu,plt)

  # Normalizo
  normas = L2.norm(values,nodes)
  values = t(apply(values,1,function(x) x/normas))
  return(values)
}

dir.sample = function(obs,nodes,metodo='MBD',plt = FALSE)
#-------------------------------------------------------------------------------
# Genera direcciones normalizadas a partir de la mediana de la muestra
#-------------------------------------------------------------------------------
# INPUT:
#   obs: observaciones de la muestra
#   nodes: particion del intervalo
#   metodo: metodo del Band-Depth ('BD2','MBD','Both')
#   depth: si NULL calcula las profundides, si no, las usa.
#   plt: si TRUE hace un plot las direcciones
# OUTPUT:
#   values: matriz con las direcciones generadas como columnas
#-------------------------------------------------------------------------------
{
  switch(metodo,
```

```
    BD2 = {depth=fBD2(obs)},
    MBD = {depth=fMBD(obs)},
    Both = {depth=round(fBD2(obs),4)*10000+fMBD(obs)}
  )

  # Armo las direcciones (resto la mediana a cada observacion)
  med = which.max(depth)
  medavg = apply(matrix(obs[,med],ncol=length(med),nrow=nrow(obs)),1,mean) # puede↩
      haber dos medianas

  values = obs[,-med] - rep(medavg,ncol(obs)-length(med))

  # Normalizo
  normas = L2.norm(values,nodes)
  values = t(apply(values,1,function(x) x/normas))

  # Grafico todas las direcciones
  if (plt){matplot(values, type="l",main=paste("Direcciones muestrales"))}
  return(values)
}
```

8.1.7 ffbplot.R

```
#combination
combinat=function(n,p){
        if (n<p){combinat=0}
        else {combinat=exp(lfactorial(n)-(lfactorial(p)+lfactorial(n-p)))}
}

#BD2
fBD2=function(data){
  p=dim(data)[1]
  n=dim(data)[2]
  rmat=apply(data,1,rank)
  down=apply(rmat,1,min)-1
  up=n-apply(rmat,1,max)
  (up*down+n-1)/combinat(n,2)
}

#MBD
fMBD=function(data){
  p=dim(data)[1]
  n=dim(data)[2]
  rmat=apply(data,1,rank)
  down=rmat-1
  up=n-rmat
  (rowSums(up*down)/p+n-1)/combinat(n,2)
}

#function boxplot
#fit: p by n functional data matrix, n is the number of curves
#method: BD2, MBD
fbplot=function(fit,x=NULL,method='MBD',depth=NULL,plot=TRUE,prob=0.5,color=6,↩
    outliercol=2,
```

```
        barcol=4,fullout=FALSE, factor=1.5,xlim=c(1,nrow(fit)),ylim=c(min(fit)-.5*↩
            diff(range(fit)),max(fit)+.5*diff(range(fit))),...){

  tp=dim(fit)[1]
  n=dim(fit)[2]
  if (length(x)==0) {x=1:tp}
    #compute band depth
    if (length(depth)==0){
    if (method=='BD2') {depth=fBD2(fit)}
    else if (method=='MBD') {depth=fMBD(fit)}
    else if (method=='Both') {depth=round(fBD2(fit),4)*10000+fMBD(fit)}
    }

  dp_s=sort(depth,decreasing=T)
  index=order(depth,decreasing=T)
  med=depth==max(depth)
  medavg=matrix(fit[,med],ncol=sum(med),nrow=tp)
  y=apply(medavg,1,mean)

  if (plot) {
  plot(x,y,lty=1,lwd=2,col=1,type='l',xlim,ylim,...)
  }
  for (pp in 1:length(prob)){
    m=ceiling(n*prob[pp])#at least 50%
    center=fit[,index[1:m]]
    out=fit[,index[(m+1):n]]
    inf=apply(center,1,min)
    sup=apply(center,1,max)

    if (prob[pp]==0.5){ #check outliers
      dist=factor*(sup-inf)
      upper=sup+dist
      lower=inf-dist
      outly=(fit<=lower)+(fit>=upper)
      outcol=colSums(outly)
      remove=(outcol>0)
      #outlier column
      colum=1:n
      outpoint=colum[remove==1]
      out=fit[,remove]
      woout=fit
      good=woout[,(remove==0),drop=FALSE]
      maxcurve=apply(good,1,max)
      mincurve=apply(good,1,min)
      if (sum(outly)>0){
        if (plot) {
        matlines(x,out,lty=2,col=outliercol,type='l',...)
        }
      }
      barval=(x[1]+x[tp])/2
      bar=which(sort(c(x,barval))==barval)[1]
      if (plot) {
      lines(c(x[bar],x[bar]),c(maxcurve[bar],sup[bar]),col=barcol,lwd=2)
        lines(c(x[bar],x[bar]),c(mincurve[bar],inf[bar]),col=barcol,lwd=2)
      }
    }
    xx=c(x,x[order(x,decreasing=T)])
    supinv=sup[order(x,decreasing=T)]
```

```
    yy=c(inf,supinv)
    if (plot) {
    if (prob[pp]==0.5) {polygon(xx,yy,col=color[pp],border=barcol,lwd=2)}
    else {polygon(xx,yy,col=color[pp],border=NA)}
    }
  }
  if (plot) {
  lines(x,fit[,index[1]],lty=1,lwd=2,col=1,type='l')
  lines(x,maxcurve,col=barcol,lwd=2)
  lines(x,mincurve,col=barcol,lwd=2)
  if (fullout) {
    if (sum(outly)>0){
        if (plot) {
        matlines(x,out,lty=2,col=outliercol,type='l',...)
        }
      }
    }
  }
  return(list(depth=depth,outpoint=outpoint,medcurve=which(med)))
}
```

8.1.8 ffbplot_adj.R

```
#combination
combinat=function(n,p){
        if (n<p){combinat=0}
        else {combinat=exp(lfactorial(n)-(lfactorial(p)+lfactorial(n-p)))}
}

#BD2
fBD2=function(data){
  p=dim(data)[1]
  n=dim(data)[2]
  rmat=apply(data,1,rank)
  down=apply(rmat,1,min)-1
  up=n-apply(rmat,1,max)
  (up*down+n-1)/combinat(n,2)

}

#MBD
fMBD=function(data){
  p=dim(data)[1]
  n=dim(data)[2]
  rmat=apply(data,1,rank)
  down=rmat-1
  up=n-rmat
  (rowSums(up*down)/p+n-1)/combinat(n,2)
}

#function boxplot
#fit: p by n functional data matrix, n is the number of curves
#method: BD2, MBD
#prob es un numero no un vector
```

```
fbplot.adj=function(fit,x=NULL,method='MBD',depth=NULL,plot=TRUE,prob=0.5,color=6,↩
    outliercol=2,
        barcol=4,fullout=FALSE, clower=4,cupper=3,xlim=c(1,nrow(fit)),ylim=c(min(↩
            fit)-.5*diff(range(fit)),max(fit)+.5*diff(range(fit))),...){

  tp=dim(fit)[1]
  n=dim(fit)[2]
  if (length(x)==0) {x=1:tp}
  #compute band depth
  if (length(depth)==0){
    if (method=='BD2') {depth=fBD2(fit)}
    else if (method=='MBD') {depth=fMBD(fit)}
    else if (method=='Both') {depth=round(fBD2(fit),4)*10000+fMBD(fit)}
    }

  dp_s=sort(depth,decreasing=T)
  index=order(depth,decreasing=T)
  med=depth==max(depth)
  medavg=matrix(fit[,med],ncol=sum(med),nrow=tp)
  y=apply(medavg,1,mean)

  if (plot) {
  plot(x,y,lty=1,lwd=2,col=1,type='l',xlim,ylim,...)
  }

    m=ceiling(n*prob)#at least 50%
    center=fit[,index[1:m]]
    out=fit[,index[(m+1):n]]
    inf=apply(center,1,min)
    sup=apply(center,1,max)

    if (prob==0.5){ #check outliers
      distup=1.5*(sup-inf)
      distlow=1.5*(sup-inf)

      datos = sweep(fit, 1,y)

      # Calculo el medcouple
      tmc = apply(datos, MARGIN = 1, mc)
      upper=sup+distup*exp( cupper*tmc*(tmc >= 0) + clower*tmc*(tmc < 0))
      lower=inf-distlow*exp(-clower*tmc*(tmc >= 0) - cupper*tmc*(tmc < 0))
      outly=(fit<=lower)+(fit>=upper)
      outcol=colSums(outly)
      remove=(outcol>0)
      #outlier column
      colum=1:n
      outpoint=colum[remove==1]
      out=fit[,remove]
      woout=fit
      good=woout[,(remove==0),drop=FALSE]
      maxcurve=apply(good,1,max)
      mincurve=apply(good,1,min)
      if (sum(outly)>0){
        if (plot) {
        matlines(x,out,lty=2,col=outliercol,type='l',...)
        }
      }
```

```
      barval=(x[1]+x[tp])/2
      bar=which(sort(c(x,barval))==barval)[1]
      if (plot) {
      lines(c(x[bar],x[bar]),c(maxcurve[bar],sup[bar]),col=barcol,lwd=2)
        lines(c(x[bar],x[bar]),c(mincurve[bar],inf[bar]),col=barcol,lwd=2)
      }

    xx=c(x,x[order(x,decreasing=T)])
    supinv=sup[order(x,decreasing=T)]
    yy=c(inf,supinv)
    if (plot) {
    if (prob==0.5) {polygon(xx,yy,col=color[1],border=barcol,lwd=2)}
    else {polygon(xx,yy,col=color[1],border=NA)}
    }
  }
  if (plot) {
  lines(x,fit[,index[1]],lty=1,lwd=2,col=1,type='l')
  lines(x,maxcurve,col=barcol,lwd=2)
  lines(x,mincurve,col=barcol,lwd=2)
  if (fullout) {
    if (sum(outly)>0){
        if (plot) {
        matlines(x,out,lty=2,col=outliercol,type='l',...)
        }
      }
    }
  }
  return(list(depth=depth,outpoint=outpoint,medcurve=which(med)))
}
```

8.1.9 gausiano.R

```
# ------------------------------------------------------------------
# Rutina para generar trayectorias de un proceso gausiano
# de media y operador de covarianza dados.
# ------------------------------------------------------------------

require('MASS') # paquete para generar normales multivariadas 'mvrnorm'

gausiano = function(n.samples,nodes=seq(0,1,0.001),media,cov.type,cov.scale,cov.nu↩
    ,plt = FALSE)
# ------------------------------------------------------------------
# INPUT:
#   n.samples: cantidad de observaciones a generar
#   nodes: particion del intervalo
#   media: media del proceso (funcion)
#   cov.type: tipo de funcion de covarianza (string)
#   cov.scale: parametros de escala para la covarianza
#   plt: si TRUE hace un plot las trayectorias
# OUTPUT:
#   values: matriz con las observaciones generadas como columnas
# ------------------------------------------------------------------
{
  # Menu de covarianzas
```

```
  switch(cov.type,
    cuadratica = {covfun=function(s,t) cov.scale*exp(-abs(t-s)^2/(2*cov.nu^2))}, #↩
        exp. cuadratica
    ornstein = {covfun=function(s,t) cov.scale*exp(-abs(s-t)/cov.nu)}, # Ornstein
    minimo = {covfun=function(s,t) cov.scale*pmin(s,t)}, #minim
    wn = {covfun=function(s,t) (s==t)*cov.scale} #ruido blanco de intensidad cov.↩
        scale
  )

  # Evaluo el operador de covarianza
  mu = media(nodes)
  covmat = outer(nodes,nodes,covfun)

  # Genero las observaciones
  if(n.samples==1){values = mvrnorm(n.samples,mu,covmat)}
  else{values = t(mvrnorm(n.samples,mu,covmat))}

  # Grafico todas las observaciones
  if (plt){matplot(nodes,values,type="l",main=paste("Proceso gausiano"))}
  return(values)
}

browniano = function(n.samples=1,nodes=seq(0,1,0.001), plt = FALSE)
#-------------------------------------------------------------
# Rutina que genera observaciones de un movimiento browniano
#-------------------------------------------------------------
# INPUT:
#   n.samples: cantidad de observaciones
#   nodes: particion del intervalo
#   plt: si TRUE hace un plot las observaciones
# OUTPUT:
#   values: matriz con las observaciones (como columnas)
#-------------------------------------------------------------
{
  jumps = sapply(diff(nodes),function(x) rnorm(n.samples,0,sqrt(x)))
  if(n.samples==1){
    values = cumsum(c(0,jumps))
  }else{
    values = apply((cbind(0,jumps)),1,cumsum)
  }

  if(plt){matplot(nodes,values,type="l",
    main = "Proceso browniano en [0,1]",
    xlab = "t"
    )
  }
  return(values)
}
```

8.1.10 auxiliares.R

```
############### Rutinas auxiliares ######################################

# Coeficientes con distribucion skew-normal
```

```
sn.coef = function(n.samples,shape){
  loc = rep(0,length(shape))
  Omega = diag(length(shape))
  return(t(rmsn(n.samples,loc,Omega,shape)))
}

# Producto interno real en L2
L2.product = function(f,g,nodos){
# f,g: matrices con las funciones (evaluadas en los nodos) como columnas
# deben tener el mismo tama o
  if(is.matrix(f) & is.matrix(g)){
    aux = rep(diff(nodos),ncol(g))
    t((t(f[-1,])%*%(g[-1,]*aux)))
  }
  else{(sum(f[-1]*g[-1]*diff(nodos)))}
}

# Norma en L2
L2.norm = function(datos,nodos){
  #datos: vector o matriz con la/s funcion/es (evaluada/s en los nodos) como ↩
      columnas
  #nodo: particion del intervalo donde se evaluan las funciones
  if(is.matrix(datos)){sqrt(colSums(apply(datos[-1,]*datos[-1,],2,function(x) x*↩
      diff(nodos))))}
  else{sqrt(sum(datos[-1]*datos[-1]*diff(nodos)))}
}

# Rutina que genera una base de L2 (no necesariamente ortonormalizada)
base = function(tipo="fourier",rangeval,n.basis){
  switch(tipo,
    fourier = create.fourier.basis(rangeval,n.basis), #n.basis tiene que ser impar
    bsplines = create.bspline.basis(rangeval,n.basis), #n.basis tiene que ser ↩
        mayor a 4
    polinomios = create.polynomial.basis(rangeval,n.basis)
    )
}

# Rutina de ortonormalizacion por Gramm-Schmidt para funciones en L2
L2.gramm.schmidt = function(basis, nodos){
  #basis: matriz donde cada columna es una funcion evaluada
  #nodos: particion del intervalo donde se evaluan las funciones
  baseON = matrix(rep(0, nrow(basis)*ncol(basis)), ncol = ncol(basis))
  baseON[,1] = basis[,1] / L2.norm(basis[,1],nodos)
  if (ncol(basis) == 1)
  {
    return(baseON)
  }
  for (k in 2:ncol(basis))
  {
    u = basis[,k]
    for (j in 1:(k-1))
    {
      u = u - L2.product(basis[,k], baseON[,j], nodos)* baseON[,j]
    }
    baseON[,k] <- u / L2.norm(u,nodos)
  }
  return(baseON)
}
```

Bibliografía

[1] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statististics*, **12**, 171-178.

[2] Azzalini, A.(2005).

[3] Azzalini, A. y Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistics Society*, Series B, **65**, 367-389.

[4] Azzalini, A. y Della Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* 1996, **83**, 715-726.

[5] Bowley, A.L. (1920). *Elements of statistics*, Charles Scribners Sons.

[6] Brys, G., Hubert, M. y Rousseeuw, P.J. (2005). A robustification of independent component analysis. *Journal of Chemometrics*,**19**, 364-375.

[7] Brys, G., Hubert, M. y Struyf, A. (2003). A comparison of some new measures of skewness. En: *Developments in Robust Statistics. International Conference on Robust Statistics 2001*. Heidelberg: Physica-Verlag, 98-113.

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

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

[10] Febrero, M., Galeano, P. and González-Manteiga, W. (2007). Functional analysis of NOx levels: location and scale estimation and outlier detection. *Computational Statistics*, **22** (3), 411-427.

[11] Groeneveld, R. A. y Meeden, G. (1984). Measuring Skewness and Kurtosis. *The Statistician*, **33**, 391-399.

[12] Henze, N. (1986). A probabilistic representation of the "skew-normal"distribution. *Scandinavian Journal of Statististics*, **13**, 271-275.

[13] Hinkley, D.V , (1975). On power transformations to symmetry. *Biometrika*, textbf62, 101111.

[14] Hubert, M., Rousseeuw, P.J. y Verdonck, T. (2009). Robust PCA for skewed data and its outlier map. *Computational Statistics and Data Analysis*, **53**, 2264-2274.

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

[16] Hubert, M. y Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics*,**22**, 235-246.

[17] Johnson, D. B. y Mizoguchi, T. (1978). Selecting the $K-$th Element in $X+Y$ and $X_1+X_2+\cdots+X_m$. *SIAM Journal of Computing*, **1**, 147-153.

[18] López-Pintado, S. y Romo, J. (2009). On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*, **104**, 718–734.

[19] Minozzo, M. y Ferracuti, L. (2012). On the existence of some skew-normal stationary processes. *Chilean Journal of Statistics*, **3**, 157-170.

[20] Ramsay, J.O. y Silverman, B.W. (1997). *Functional Data Analysis*. Springer Verlag, New York.

[21] Rousseeuw, P.J., Ruts, I. y Tukey, J.W. (1999). The bagplot: a bivariate boxplot. *American Statistician*, **53**, 382-387.

[22] Rousseeuw, P.J., y van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85**, 633-639.

[23] Staicu, A.M., Crainiceanu, C.M., Reich, D.S. y Ruppert, D. (2012). Modelling functional data with spatially heterogeneous shape characteristics, *Biometrics*, **68**, 331-343.

[24] Stahel, W. A. (1981) *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. PhD Thesis, ETH Zürich.

[25] Sun, Y. y Genton, M. G. (2011). Functional Boxplots. *Journal of Computational and Graphical Statistics*, **20**, 316-334.

[26] Sun, Y., Genton, M. G. y Nychka, D. (2012). Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked?. *Stat*, **1**, 68-74.

[27] Tukey, J. W. (1970). *Exploratory Data Analysis, Preliminary Edition*, Vol. 1, Ch. 5, Reading, MA: Addison-Wesley.

[28] Tukey, J. W. (1977). *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

[29] Zhang, H. y El-Shaarawi, A. (2010). On spatial skew-Gaussian processes and applications. *Environmetrics*, 21, 33-47.