Anales de la Academia Nacional de
Ciencias Exactas, Físicas y Naturales
Argentina
Tomo 72, 2021

INFERENCIA ROBUSTA: UN TRAYECTO DE LO FINITO A LO INFINITO-DIMENSIONAL*

Graciela Boente

Departamento de Matemáticas e Instituto de Cálculo, Universidad de Buenos Aires y CONICET. Email: gboente@dm.uba.ar

*Trabajo presentado por G. Boente en oportunidad de su incorporación como Académica Titular de la ANCEFN (31 de julio de 2020) (https://www.youtube.com/watch?v=GMDiAU-t520)

Palabras clave
Análisis de datos funcionales
Datos atípicos
Regresión no paramétrica
Robustez
Suavizado

Resumen El avance de las nuevas tecnologías ha hecho necesario desarrollar procedimientos estadísticos para estimar funciones o para analizar datos que corresponden a realizaciones de un proceso estocástico. Muchos de los procedimientos utilizados se basan en las mismas ideas que el estimador de mínimos cuadrados en el modelo de regresión lineal siendo por lo tanto muy sensibles a la presencia de un pequeño porcentaje de datos anómalos. En este trabajo, se presentan algunos de los avances obtenidos para definir métodos de inferencia confiables cuando la muestra puede contener datos atípicos tanto para modelos de regresión noparamétrica y semiparamétrica como para el análisis de datos funcionales.

Keywords
Functional data analysis
Nonparametric regression
Outliers
Robustness
Smoothing

Abstract **Robust inference: a path from the finite to the infinite-dimensional setting.** The development of new technologies clarified the need of developing new statistical procedures to estimate functions or to analyse data that correspond to realizations of a stochastic process. Many of the standard procedures used are based on the same ideas as the least squares estimator in the linear regression model, being therefore very sensitive to the presence of a small percentage of atypical data. In this paper, we present some of the advances obtained to define reliable inference methods when the sample can contain atypical data both for nonparametric and semiparametric regression models and for functional data analysis.

1. Introducción

La palabra *Estadística* fue introducida en el siglo XVIII por el filósofo y economista alemán Gottfried Achenwall en 1749 y por el político inglés Sir John Sinclair en 1791, para indicar el análisis de datos de estado, es decir, datos completos sobre todos los habitantes de un país, ver van der Zande (2010). Adquirió el significado de la recopilación y clasificación de datos en general a principios del siglo XIX. Cabe mencionar que los datos del censo en esa época representan al día de hoy lo que llamamos *Big Data* y era necesario procesarlos y darles un orden para resumir la información subyacente.

Las bases de la Estadística Matemática surgieron con el apoyo de la Teoría de Probabilidades a principios del siglo XX con Francis Galton, Karl Pearson y Ronald Fisher, entre otros. Más allá de los desarrollos que sirven de fundamento para los procedimientos de inferencia estadística, la implementación de técnicas gráficas que permitió una rápida y sencilla visualización de las observaciones y de sus características tuvo un fuerte impacto. En el campo de la Bioestadística podemos mencionar a Florence Nightingale que fue pionera en el desarrollo de métodos gráficos entre otros para ilustrar el número de muertos por enfermedades o por heridas en la Guerra de Crimea, en relación al número de soldados de cada mes, o sea las tasas de mortalidad. En su honor el período Mayo de 2020 a Julio de 2021 fue declarado el *Año Internacional de la Mujer en Estadística y Ciencia de Datos* por coincidir con el 200 aniversario de su nacimiento, ver https://www.isi-web.org/iywsds. Por otra parte, diversos métodos graficos surgieron en la segunda mitad del siglo XX para detectar observaciones atípicas, o sea, observaciones cuyo comportamiento se aleja de la mayoría de los datos, entre los cuales el más conocido es el boxplot propuesto por Tukey (1970), ver también McGill et al. (1978).

El uso de la palabra *Robustez* fue dado por primera vez en un trabajo de Box (1953) quien escribe sobre la destacable propiedad de robustez a la no-normalidad que poseen los tests para comparación de medias en contraposición con los tests para comparar varianzas. Los desarrollos inciales de procedimientos robustos fueron dados por Tukey (1960) quien destacó la extrema sensibilidad de algunos procedimientos de inferencia estadística convencional a pequeñas desviaciones de los supuestos que los sustentan. Las bases de la inferencia robusta se sentaron en la década del 60/70 con los trabajos, entre otros, de Huber (1964, 1967, 1968) y Hampel (1968, 1971, 1974). En particular, Hampel (1968) caracterizó la robustez de un estimador a través de la continuidad del funcional asociado e introdujo nociones como el punto de ruptura y la función de influencia.

Más de cincuenta años después del trabajo de Huber (1964) sobre M-estimadores de posición, los procedimientos robustos son una elección popular para brindar protección ante la presencia de *outliers*. Como describen los autores antes mencionados, los datos atípicos o *outliers* pueden corresponder

a datos incorrectamente reportados o a errores de medición cuya distribución tiene, por ejemplo, colas más pesadas que la distribución normal. En particular, en el modelo de regresión lineal $y = \alpha + \beta\, x + \varepsilon$ es bien conocido que un pequeño porcentaje de datos atípicos tienen una influencia inusualmente grande en los estimadores de mínimos cuadrados que son los clásicamente utilizados.

Así como el método de mínimos cuadrados se basa en el supuesto de que el error ε tiene distribución normal, y es muy sensible al apartamiento de este supuesto, muchos de los procedimientos estadísticos se fundamentan, directa o indirectamente, en ciertos supuestos sobre lo que se sabe o se supone de los datos. Estos modelos suelen ser simplificaciones o idealizaciones de la realidad y por ello, en muchas ocasiones sólo podemos suponer que el modelo vale aproximadamente, en el sentido, que la distribución de las observaciones se encuentra "cerca" en algún sentido de la postulada por procedimientos como el de máxima verosimilitud o el de mínimos cuadrados, que denominaremos de ahora en más métodos clásicos. Esto puede ser el caso cuando el modelo normal describe el comportamiento de la mayoría de los datos pero algunas observaciones siguen otro patrón y suelen encontrarse lejos del grueso de los datos. Estos comentarios muestran que es necesario desarrollar estrategias para tratar con modelos que se reconocen como imprecisos y es por ello que, desde la década del 80 del siglo XX, en particular en el modelo de regresión lineal, se desarrollaron diversos procedimientos que combinan simultáneamente alta eficiencia en el modelo ideal y resistencia a la presencia de observaciones atípicas. Asimismo, hubo avances en procedimientos estadísticos que suponen modelos más flexibles que los modelos de regresión paramétricos, permitiendo estructuras más generales como es el caso de modelos de regresión noparamétricos, en los que la función de regresión sólo se supone derivable, por ejemplo.

El desarrollo de métodos robustos se amplió a otros ámbitos como el de los modelos lineales generalizados, los modelos de regresión noparamétricos o semiparamétricos, el problema de componentes principales o el de correlación canónica, de modo a obtener resultados confiables, aún en presencia de datos atípicos (*outliers*) y perdiendo, simultáneamente, poca eficiencia si el modelo asumido por la alternativa clásica es válido. Una completa descripción de los últimos avances sobre procedimientos robustos en modelos paramétricos y la teoría subyacente puede verse en Maronna et al. (2019), mientras que algunas propuestas en modelos de regresión noparamétrica pueden verse en Huber (1979) y Härdle (1990).

Como se ha discutido ampliamente en la literatura, ver por ejemplo Galeano y Peña (2019), la explosión de datos de los últimos años hace que el desarrollo y estudio de procedimientos robustos sea muy relevante ya que distintos tipos de errores y *outliers* pueden ocurrir en la era actual del *Big Data*. Este problema es de especial importancia al trabajar con datos funcionales complejos como pueden ser curvas, imágenes o películas donde no

es posible tener una fácil visualización que permita detectar las observaciones espurias debido a los distintos patrones de anomalías que pueden existir. Estos hechos ya habían sido descriptos en Huber (2010) que mencionaba que en conjuntos de datos recopilados por seres humanos se suelen cometer errores groseros, de forma más o menos aleatoria, con una frecuencia global entre el 1% y el 10%. Sin embargo, con la recopilación de datos más o menos automatizada, surgen nuevos tipos de errores groseros que pueden ser sistemáticos y difíciles de identificar por lo que son necesarias nuevas estrategias para lidiar con ellos.

Tanto el reciente libro de Maronna et al. (2019) como los libros de Heritier et al. (2009) y Huber y Ronchetti (2009) se centran en el problema de inferencia robusta cuando tratamos con parámetros finito–dimensionales, como es el caso de los coeficientes en un modelo de regresión lineal o del vector de direcciones principales en el análisis de componentes principales, más aún suponen que las observaciones pertenecen al espacio euclídeo $\mathbb{R}^p$, por lo que dichos procedimientos no se pueden aplicar en forma directa al caso de datos funcionales. Por esta razón, en este trabajo intentaremos describir algunos de los avances que se han obtenido en el ámbito de la inferencia robusta, tanto en datos funcionales como en el problema estimación de funciones de regresión. El resto trabajo está organizado de la siguiente forma. En la Sección 2 se discute el problema de estimación robusta en problemas de regresión noparamétrica así como los desafíos que surgen cuando el número de covariables aumenta. En la Sección 3 se presenta el problema de inferencia cuando los datos provienen de curvas, imágenes o más precisamente, cuando corresponden a objetos de dimensión infinita. Algunos comentarios finales se dan en la Sección 4.

2. El estudio de objetos de dimensión infinita

Como consecuencia del avance de la tecnología ya a partir 1980 era posible representar y estimar funciones, como la función de densidad y de regresión y representarlas de forma relativamente rápida en las pantallas de la computadora. En esa época, aparecieron diversos trabajos relacionados con el estudio de estos objetos que, a diferencia del caso de regresión lineal, eran de dimensión infinita ya que no se suponía un modelo que dependiera de un número finito de parámetros desconocidos.

El modelo de regresión noparamétrica supone tradicionalmente que se tienen observaciones $(y_1, x_1), \dots, (y_n, x_n)$, independientes e idénticamente distribuidas (i.i.d.), tales que las covariables tienen soporte en un intervalo acotado $\mathcal{I}$ y cumplen $y_i = \eta(x_i) + \varepsilon_i$, donde el error ε_i es independiente de x_i y $\mathbb{E}(\varepsilon_i) = 0$. Por lo tanto, $\eta(x)$ representa la esperanza condicional de $y_1|x_1 = x$. En este caso, el objeto de estudio no es un parámetro de dimensión finita sino una función $\eta: \mathcal{I} \to \mathbb{R}$ que se supone suave, por ejemplo, Lipschitz o de clase $\mathcal{C}^2(\mathcal{I})$.

Los estimadores propuestos por Nadaraya–Watson (1964) son promedios locales pesados y por lo tanto, se ven muy afectados por la presencia de datos anómalos, en particular, si estas respuestas atípicas corresponden a las variables independientes cercanas al punto x_0 donde se desea estimar la función η. Ya en el año 1977 Brillinger, en su discusión del trabajo de Stone (1977) mencionaba que estimadores de tipo *M* en este modelo eran necesarios para obtener robustez frente a *outliers*. Härdle (1990) también resalta la importancia de obtener estimaciones resistentes frente a *outliers* desde el punto de vista del análisis de datos ya que un comportamiento errático de dicho estimador puede ocasionar formulaciones paramétricas sesgadas. Estimadores robustos en el contexto noparamétrico pueden definirse como no sensibles a un *outlier* vertical aislado. En este contexto, se definieron distintos procedimientos robustos para estimar la función η cuando los errores no necesariamente tienen primer momento. Podemos mencionar entre otros los trabajos de Härdle y Tsybakov (1988), Boente y Fraiman (1989a y b, 1990) y Boente et al. (2009) que estudian procedimientos basados en núcleos tanto en el caso de observaciones independientes como dependientes y aún cuando hay respuestas faltantes, respectivamente. Una revisión de distintos procedimientos que conducen a estimadores robustos en el problema de regresión noparamétrica puede verse en Härdle (1990), mientras que, para el caso de una sóla variable explicativa, alternativas robustas basadas en splines y sus propiedades fueron consideradas por Huber (1979), Cox (1983), Cunningham et al. (1991) y, más recientemente, por Kalogridis (2020) y Kalogridis y Van Aelst (2021).

Es importante destacar el rol que juega la suavidad de la función a estimar. Como menciona Hampel también en su comentario del trabajo de Stone (1977) "*to talk about robustness is meaningless or, rather, hopeless in the case of a completely arbitrary model; for a model with wild spikes and a nice model with some distant gross errors superimposed are indistinguishable. If we believe in a "smooth" model without spikes, however, then some robustification is possible. In this situation, a clear outlier will not be attributed to some sudden change in the true model, but to a gross error, and hence it may be deleted or otherwise made harmless*."

Para ilustrar el efecto de datos anómalos en la función de regresión consideraremos un conjunto de datos que corresponde a 153 mediciones diarias de calidad de aire en la región de Nueva York entre Mayo y Septiembre de 1973. Dicho conjunto fue analizado en Chambers et al. (1983) y puede encontrarse en el data set airquality del paquete R que es un software abierto de análisis estadístico. El interés es explicar la concentración de ozono ("O_3", medido en partes por billón) a través de una función de 3 variables explicativas potenciales: temperatura ("Temp", en grados Fahrenheit), velocidad del viento ("Wind", en millas por hora) y radiación solar medida en la banda de frecuencias 4000-7700 ("Solar.R", en Langleys). Consideraremos los 111 casos que no tienen datos faltantes y a modo de ilustración

estudiaremos la relación entre Ozono y Temperatura. El diagrama de puntos correspondiente a estas dos variables se presenta en la Fig. 1A.

La Fig. 1B muestra en rojo la estimación de $\eta(x_0)$ obtenida mediante la media local en el punto $x_0 = 75$ con ventana $h = 4.76$, es decir, promediando los valores de ozono correspondientes a temperaturas entre $x_0 - h$ y $x_0 + h$. Por otra parte, la Fig. 1C presenta la estimación obtenida al usar pesos basados en el núcleo normal como un triángulo violeta junto con dicho núcleo en violeta. Dicha estimación corresponde al estimador de Nadaraya–Watson y puede definirse como

$$\hat{\eta}(x_0) = \left\{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right\}^{-1} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right) y_i\,,$$

siendo $(y_i,\ x_i)$ el i −ésimo vector observado con coordenadas el ozono y la temperatura, respectivamente y $K(u) = \exp(-u^2/2)$ el núcleo normal. Finalmente, la Fig. 1D muestra el suavizador lineal en rojo y el M-estimador local en azul. Como vemos los puntos con valores altos de ozono alrededor de 80 grados, afectan el suavizador lineal. El M-estimador puede definirse como la solución en a de la ecuación

$$\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\psi\left(\frac{y_i - a}{\hat{\sigma}}\right) = 0$$

donde $\hat{\sigma}$ es un estimador preliminar robusto de la escala del error, $K: \mathbb{R} \to \mathbb{R}$ es el núcleo (en este caso la densidad normal) y h es la ventana que regula el compromiso entre sesgo y varianza. La función ψ es una función acotada que se comporta como la identidad cerca de 0 y puede elegirse como $\psi = \rho'$ donde ρ es una función par, continua, diferenciable, no-decreciente y tal que si $0 \leq u < v$ son tales que $\rho(v) < \sup_t \rho(t)$ entonces $\rho(u) < \rho(v)$. La Fig. 2 muestra dos posibles elecciones para la función ρ junto con su derivada ψ.

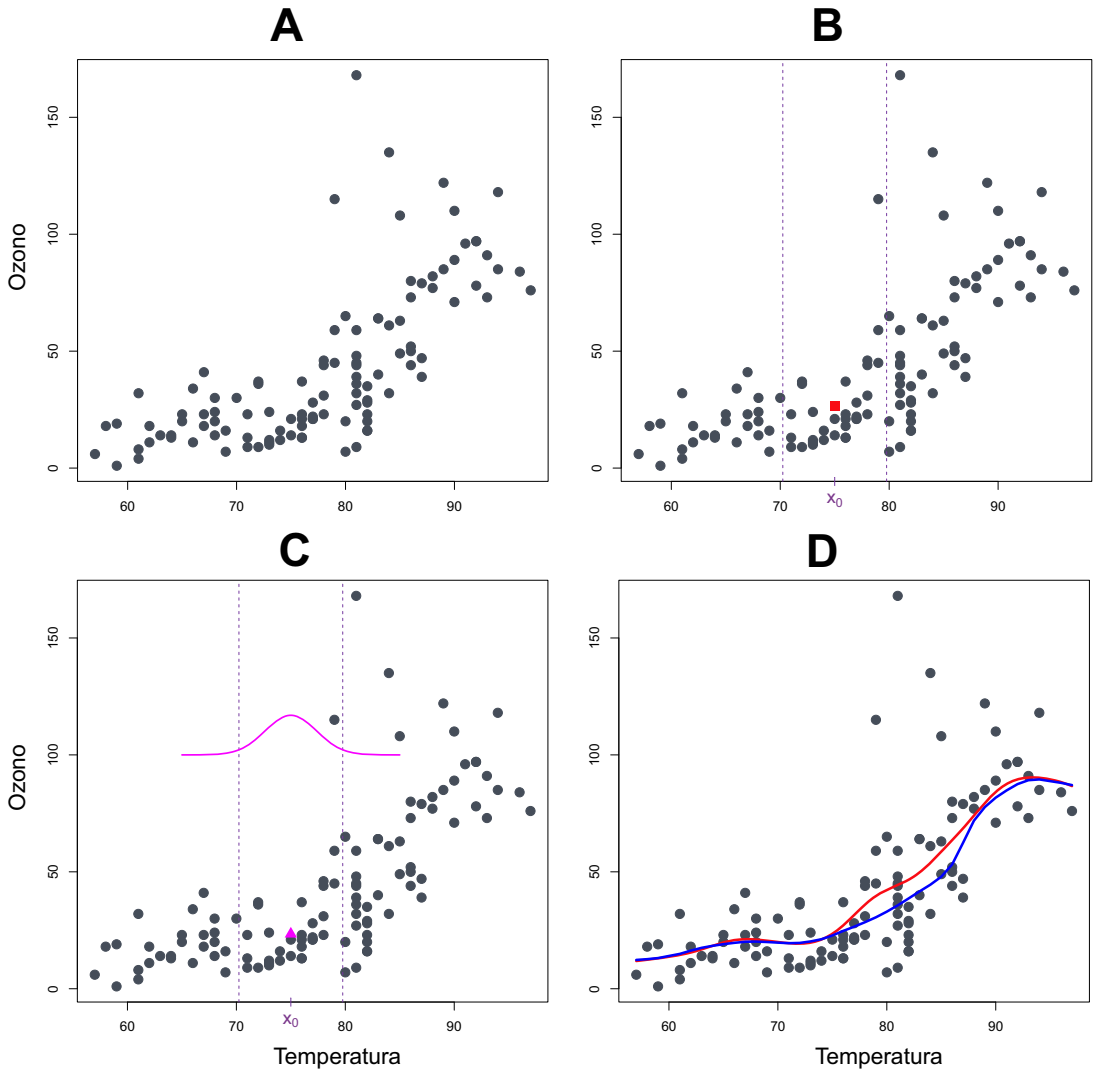

Fig. 1. Datos de ozono versus temperatura (del data set airquality del paquete R). A: Diagrama de puntos. B: El cuadrado rojo indica la media local en $x_0 = 75$ con ventana $h = 4.76$. C: El triángulo violeta indica el estimador de núcleos utilizando el núcleo normal que sobreimpuso en línea violeta; las rectas verticales cortadas corresponden a $x_0 - h$ y $x_0 + h$. D: Las líneas roja y azul corresponden a la estimación de Nadaraya-Watson y al M-estimador local, respectivamente.

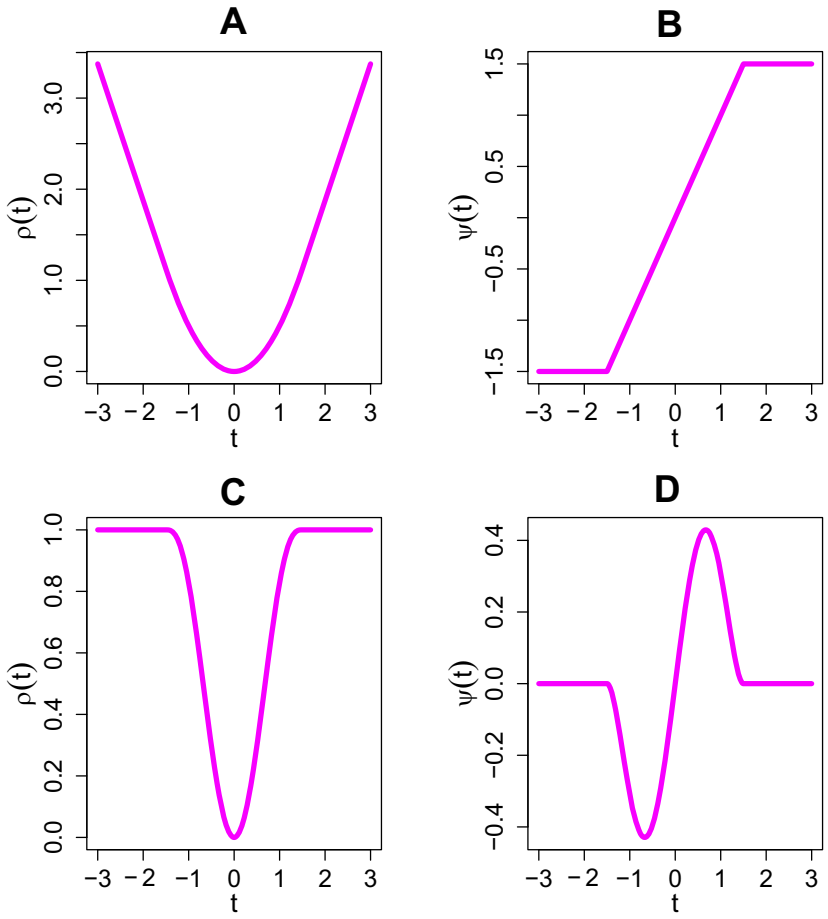

Fig. 2. Función ρ (A y C) con su derivada ψ (B y D). A: Función de Huber $\rho(t) = (t^2/2)\, I_{|t|\le c} + (c\,t - c^2/2) I_{|t|>c}$ con $c = 1.5$ y B $\psi(t) = \min(c, \max(-c, t))$. C: Función de Tukey $\rho(t) = \min(3(t/c)^2 - 3(t/c)^4 + (t/c)^6, 1)$ con $c = 1.5$, y D: $\psi(t) = t\,(1 - (t/c)^2)^2\, I_{[-c,c]}(t)$.

Un problema de estos procedimientos es que no se vuelven viables si la dimensión p de las covariables es alta, por la así llamada *maldición de la dimensión*. Dicha maldición fue discutida en Stone (1982) se explica por las tasas de convergencia que, por ejemplo para funciones de p variables dos veces diferenciables, son del orden $n^{\frac{2}{4+p}}$ siendo n la cantidad de observaciones. Intuitivamente, este fenómeno significa que la cantidad de observaciones necesarias para evitar que los estimadores tengan una varianza inaceptablemente grande crece exponencialmente con la dimensión. Dicho efecto se ve reflejado en las Figs. 3 y 4 que muestran cómo al crecer la dimensión los datos se vuelven cada vez más ralos. La primera muestra los histogramas, correspondientes a 1000 replicaciones, de la distancia entre dos puntos generados con distribución uniforme en el cubo $[0,1]^p$, mientras que la segunda muestra la probabilidad de que una observación en $\mathbb{R}^p$ proveniente de la distribución normal, $N(0_p, \boldsymbol{I}_p)$, pertenezca a la bola unidad. Estos gráficos muestran que al crecer la dimensión los datos se vuelven cada vez más aislados y ralos dificultando el uso de técnicas de suavizado.

Por esta razón, los supuestos de modelado en la práctica deben centrarse en espacios de funciones con dimensión inherente mucho menor que la del espacio de funciones suaves en p variables. Una opción es considerar los así llamados *modelos aditivos*, en los que se restringe la clase de las funciones de regresión suponiendo que la misma es de la forma:

$$\eta(\mathbf{x})=\mu+\sum_{j=1}^{p}\eta_j\left(x_j\right) \quad \text{donde} \quad \mathbb{E}(\eta_j(x_j))=0.$$

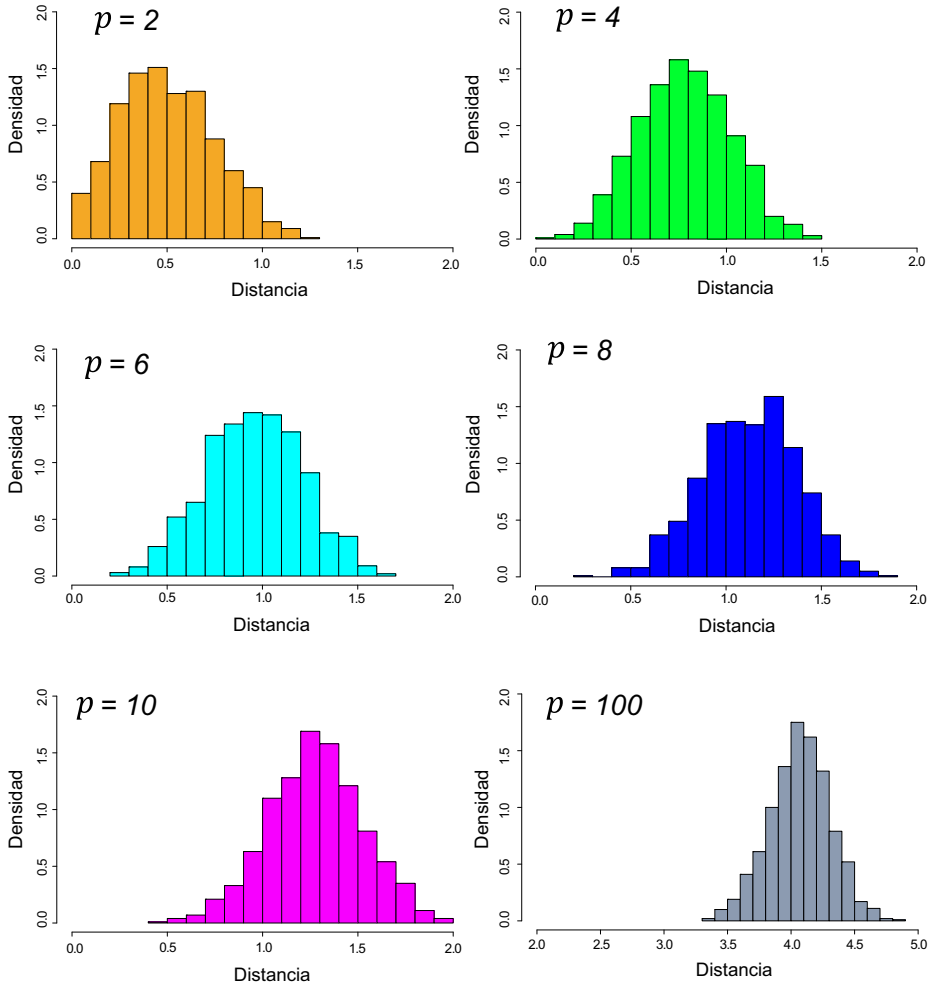

Fig. 3. Histogramas correspondientes a la distancia entre dos puntos generados con distribución uniforme en el cubo $[0,1]^p$, evaluados sobre 1000 replicaciones.

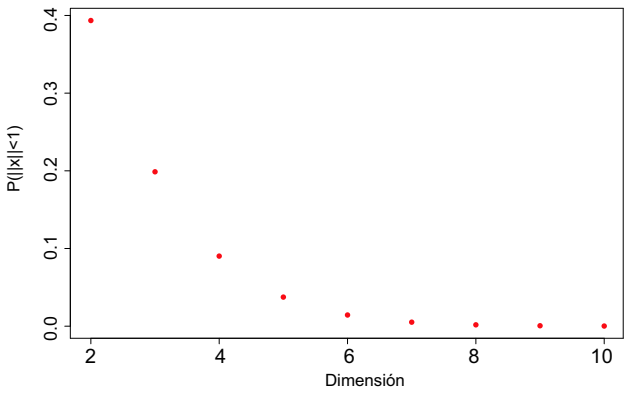

Fig. 4. Gráfico de $P(\| \boldsymbol{x} \| \leq 1)$ como función de la dimensión p cuando $\boldsymbol{x} \sim N(0_p, \boldsymbol{I}_p)$.

Vale la pena destacar que los modelos aditivos generalizan los modelos lineales y son de fácil interpretación pues cada covariable actúa en forma separada sobre el modelo de regresión. La ventaja de estos modelos es que permite hacer suavizados unidimensionales. Más precisamente, Stone (1985) mostró que bajo un modelo aditivo la tasa óptima para estimar η_j es la tasa uno-dimensional. Hablamos entonces de una reducción de la dimensión.

Para estimar las componentes η_j existen dos métodos comúnmente usados, uno basado en un método iterativo llamado backfitting y el otro basado en el procedimiento de integración marginal. Más precisamente, Buja et al. (1989) y Hastie y Tibshirani (1990) propusieron el procedimiento de *backfitting* que consisten en estimar las componentes η_j suavizando los residuos parciales

$$\widehat{\mathrm{R}}_{ij}^{(\ell)} = y_i - \hat{\mu}^{(\ell-1)} - \sum_{s=1}^{j-1} \hat{\eta}_s^{(\ell)}(x_{is}) - \sum_{s=j+1}^{p} \hat{\eta}_s^{(\ell-1)}(x_{is})$$

iterativamente en s y ℓ, utilizando, por ejemplo, polinomios locales como los descriptos en Härdle et al. (2004). Como el estimador de Nadaraya-Watson, estos estimadores son sensibles a observaciones atípicas y por esta razón, Boente et al. (2017) consideraron estimadores robustos basados en M-estimadores polinomiales locales (*backfitting robusto*) obteniendo de esta forma estimadores más resistentes a datos atípicos.

El estimador de integración marginal se basa en la idea intuitiva que, salvo por una constante aditiva, las componentes η_j se pueden recuperar integrando la función η respecto de las demás variables. Linton y Nielsen (1995) mostraron que integrar el estimador de Nadaraya-Watson produce estimadores de las componentes marginales que son asintóticamente normales con tasa de convergencia óptima si $p = 2$. Algunos desarrollos heurísticos, basados en la consistencia del estimador piloto, sugieren que este estimador no convergería con tasa de convergencia óptima en presencia de

más de cuatro covariables. Para resolver este problema, Severance–Lossin y Sperlich (1999) propusieron un estimador basado en polinomios locales adaptivo a la componente a estimar, de modo a obtener tasas uniparamétricas óptimas. Más precisamente, para estimar η_j la propuesta de Severance–Lossin y Sperlich (1999) utiliza polinomios locales solamente para η_j, por lo que en el desarrollo se usan únicamente las covariables x_{ij}. Para resolver la sensibilidad de estos estimadores cuando existen respuestas atípicas en la muestra Boente y Martínez (2017) consideraron una versión robusta del estimador dado en Severance–Lossin y Sperlich (1999). Dicha propuesta robusta se basa en M-estimadores robustos que utilizan un estimador preliminar de escala y polinomios locales en la dirección de interés. Teniendo en cuenta que el procedimiento de integración debe repetirse para cada $1 \leq j \leq p$, este método es numéricamente más costoso que el backfitting robusto si la dimensión de las covariables es grande.

El análisis de los datos de calidad de aire puede verse en Boente et al. (2017), incluímos aquí los resultados obtenidos para ilustrar la sensibilidad del procedimiento de backfitting y la ventaja de usar estimadores robustos. Para este conjunto de datos, Cleveland (1985) encuentra que la relación entre ozono y velocidad de viento es no-lineal, correspondiendo valores bajos de ozono a altas velocidades del viento. Un ajuste robusto basado en un modelo nolineal fue dado en Bianco y Spano (2019) quienes consideraron un modelo de crecimiento exponencial entre el ozono y la velocidad del viento. Como antes consideramos solamente los 111 casos que no contienen observaciones faltantes. La Fig. 5 ilustra que la relación entre ozono y las demás variables no parece ser lineal, por ello consideramos el modelo aditivo $O_3 = \mu + \eta_1(\mathrm{Temp}) + \eta_2(\mathrm{Wind}) + \eta_3(\mathrm{Solar.R}) + \varepsilon$, donde los errores ε se suponen independientes y con distribución simétrica respecto de 0.

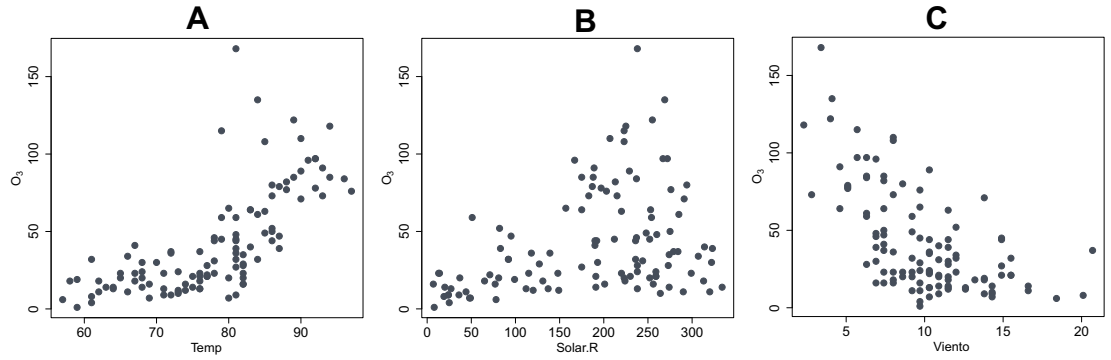

Fig. 5. Datos de ozono versus temperatura (A), velocidad del viento (B) y radiación solar (C). Basado en el data set airquality del paquete R.

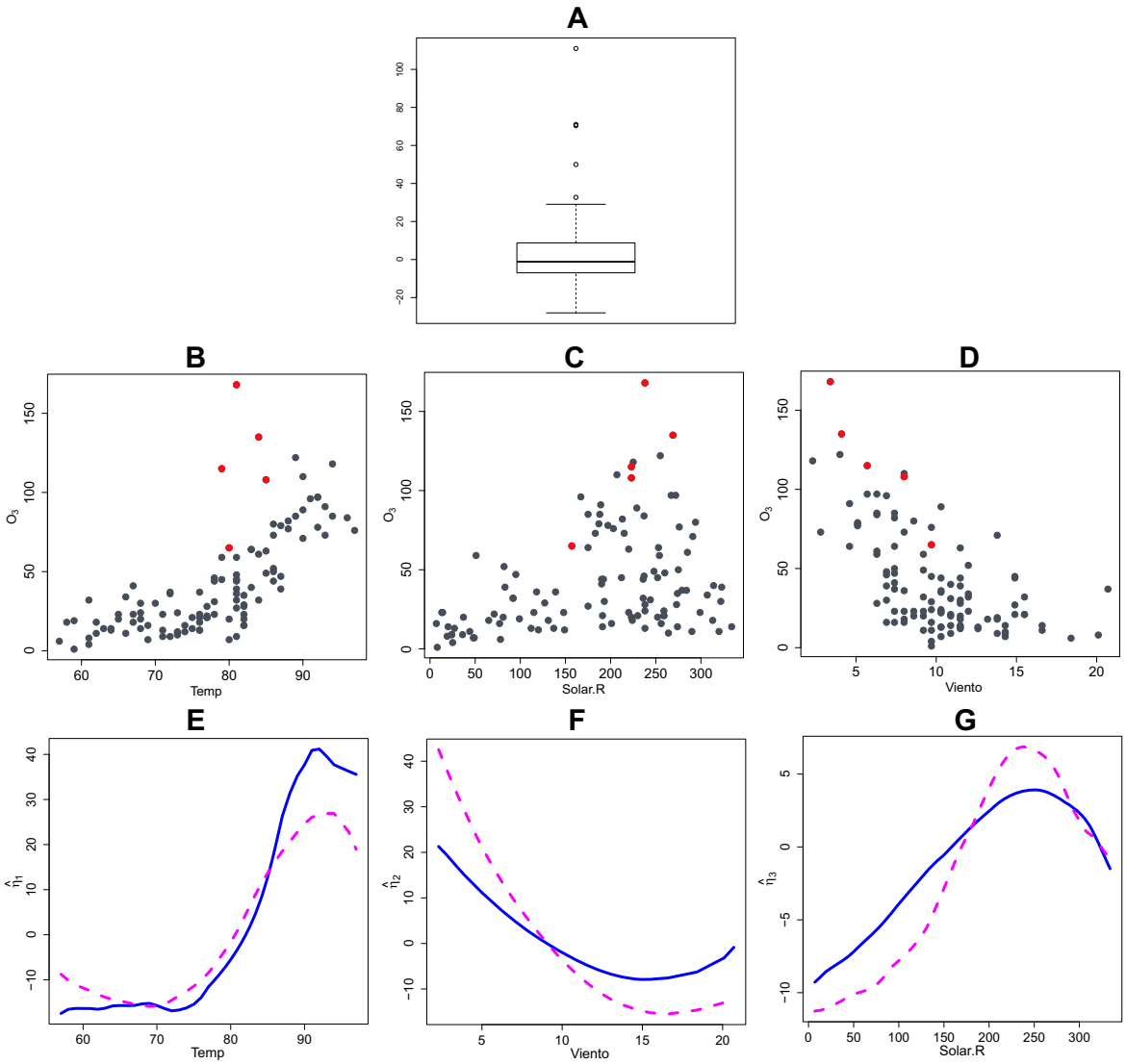

Fig. 6. Datos de calidad de aire. Los datos detectados como atípicos se indican en B-D en rojo. A: Boxplot de los residuos del ajuste robusto. B: Ozono versus temperatura C: Ozono versus velocidad del viento D: Ozono versus radición solar. E-F: Estimaciones clásicas y robustas de las componentes aditivas. El ajuste clásico corresponde a la línea cortada magenta, mientras que el robusto a la línea azul. Basado en el data set airquality del paquete R.

Realizamos un ajuste de los datos utilizando el procedimiento de backfitting clásico propuesto en Buja et al. (1989) y la alternativa robusta definida en Boente et al. (2017) utilizando polinomios locales de grado 1 y en el caso del método robusto consideramos como función de pérdida ρ la función $\rho(t) = \min(3(t/c)^2 - 3(t/c)^4 + (t/c)^6, 1)$ con $c = 4.685$. Los datos atípicos detectados mediante el boxplot de los residuos del ajuste robusto se indican en rojo en las Fig. 6 B a Fig.6D y vemos la influencia que tienen en el estimador clásico que se presenta en linea cortada magenta en las Fig. 6E a Fig. 6G.

Como muestra la Fig. 7, si realizamos un nuevo análisis mediante el procedimiento de backfitting clásico utilizando ahora la muestra limpia, es

decir, sin las observaciones detectadas como atípicas, obtenemos resultados similares a los del estimador robusto (utilizando todos los datos). De esta forma, el ajuste robusto ponderó automáticamente los potenciales datos atípicos dándoles menor peso y produjo estimaciones de las componentes aditivas η_j que son casi idénticas a las del procedimiento de backfitting clásico una vez detectados y eliminados los valores atípicos. Por otra parte, los residuos del ajuste robusto permitieron identificar esos posibles valores atípicos siendo una herramienta útil de diagnóstico.

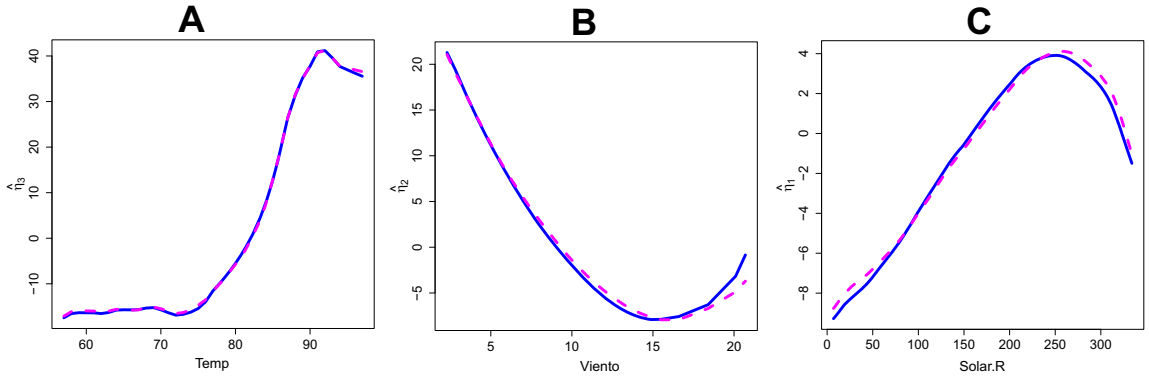

Fig.7. Datos de calidad de aire. Estimaciones clásicas sin los datos atípicos y robustas con todos los datos de las componentes aditivas.El ajuste clásico sin los datos atípicos corresponde a la línea cortada magenta, mientras que el robusto calculado con todos los datos a la línea azul. Basado en el data set airquality del paquete R.

Otra alternativa para lidiar con *maldición de la dimensión* consiste en imponer restricciones paramétricas en algunas variables. Entre este tipo de modelos semiparamétricos se encuentran los modelos parcialmente lineales, modelos de índice simple, modelos parcialmente lineales generalizados o parcialmente lineales de índice simple. Una descripción de distintos procedimientos clásicos para estos modelos pueden verse en Härdle et al. (2004) y Horowitz (2009), mientras que alternativas robustas pueden verse en He y Shi (1996), He et al. (2002), Bianco y Boente (2004) y Bianco et al. (2011) quienes consideraron robustos estimadores en el modelo parcialmente lineal. Por otra parte, Boente et al. (2006) y Boente y Rodriguez (2010) propusieron y estudiaron métodos robustos en modelos parcialmente lineales generalizados. Finalmente, Boente y Rodriguez (2012) y Agostinelli et al. (2020) presentan propuestas robustas en modelos de índice simple. Recientemente, Bravo (2019) propuso estimadores robustos en el contexto de modelos de coeficientes variables cuando hay observaciones faltantes.

3. Los datos son objetos de dimensión infinita

La aparición de las computadoras supuso para la Estadística un cambio con implicaciones en el paradigma metodológico. Ahora es posible trabajar

con conjuntos grandes de datos y almacenar la información correspondiente a datos provenientes de fenómenos climáticos como imágenes satelitales, biológicos como resonancias magnéticas así como espectrogramas de sonidos de utilidad en lingüistica. En estos ejemplos los datos corresponden a observaciones discretizadas de un proceso suave, es decir, los datos ahora son curvas u objetos de dimensión infinita. Por ejemplo, para un fenómeno particular, podríamos tener observaciones $X_1(t), \dots, X_n(t)$ correspondientes a un proceso estocástico indexado en tiempo $\{X(t), t \in \mathcal{I}\}$.

En las propuestas dadas en los últimos 30 años para estimar distintas características o para hacer test de hipótesis, los datos obtenidos se consideraron como trayectorias completamente observadas. En este contexto fue posible definir, entre otros, procedimientos para estimar el parámetro de regresión, las direcciones principales o canónicas, así como, para testear la igualdad de medias o de operadores de covarianza, que son la versión funcional de la matriz de covarianza, cuando tenemos varias poblaciones independientes ver, por ejemplo, Horváth y Kokoszka (2012).

Si bien el interés científico radica en el proceso estocástico subyacente y sus propiedades, en realidad, este proceso está muchas veces latente y no puede observarse directamente, pues los datos sólo pueden medirse en una grilla de puntos que puede ser fija o aleatoria y puede variar o no de individuo a individuo. Una hipótesis algo más general que suponer que se observa toda la trayectoria consiste en suponer que nuestro dato es una observación discretizada del proceso y que todos los datos se registran en la misma grilla de tiempos $\{t_j\}_{j=1,\dots p}$, con $t_1 \leq \cdots \leq t_p$, o sea que observamos $X_{ij} = X_i(t_j)$, $1 \leq i \leq n$, $1 \leq j \leq n$. Si los registros corresponden a un instrumento, como un electroencefalograma o a una imagen por resonancia magnética, la grilla de puntos es usualmente equiespaciada $t_j - t_{j-1} = t_{j+1} - t_j$ para todo j. En tales situaciones, estudiar las componentes de variación mediante las técnicas usuales de componentes principales, correlación canónica o regresión no parece ser lo más indicado. Estas técnicas multivariadas pueden llevar a problemas debido a la falta de regularidad o dar origen a estimadores mal definidos, como en el caso del análisis de correlación canónica donde pueden encontrarse direcciones con correlación canónica muestral 1 si se usan las versiones muestrales multivariadas sin penalizar. Por esta razón, se han desarrollado diversas técnicas de análisis para datos funcionales que sacan provecho del hecho que X o la cantidad a estimar presentan alguna información de regularidad que puede ser explotada de forma funcional, por ejemplo, continuidad o diferenciabilidad.

Cabe mencionar que, en el estudio de las propiedades asintóticas, se supone que el interespaciado $t_{j+1} - t_j$ converge a 0 cuando p crece, de modo que $p = p_n$ es una sucesión que crece a infinito. Como mencionan Wang et al. (2016), a pesar de que valores grandes de p llevan a un problema de alta dimensión, esto también significa que tenemos más datos para obtener información sobre $X(t)$. Por lo tanto, en el estudio de este tipo de objetos, la

alta dimensión es una bendición más que una maldición. Esta bendición se debe justamente a la hipótesis de suavidad de X que permite permite sacar provecho de la información de puntos cercanos mediante alguna técnica de suavizado como las mencionadas en la Sección 2. Por lo tanto, la suavidad impuesta sirve como herramienta de regularización y los así llamados *fat data sets* son en este caso beneficiosos. Un contexto mucho más desafiante corresponde al caso en que las trayectorias son esparsas y que ilustraremos al final de esta sección.

Como mencionamos anteriormente, para este tipo de datos, los errores groseros pueden ser frecuentes y difíciles de visualizar. Por ejemplo, los dispositivos de comunicación y control recopilan datos automáticamente utilizando redes inalámbricas de sensores y en algunos casos, debido a falla de baterías o congestión de comunicación, los nodos sensores pueden no grabar los datos correctamente. Estas fallas producirán valores atípicos en los datos registrados y por ellos es importante detectarlos (o limpiarlos) antes de tratar de construir modelos viables.

Una manera de detectar los datos anómalos es mediante el uso de técnicas robustas en el análisis de datos y el posterior análisis de los residuos, la otra opción es utilizar técnicas gráficas como el boxplot funcional.

El boxplot funcional definido por Sun y Genton (2011) utiliza una de las nociones de profundidad funcional, especificamente, la profundidad de bandas definida en López-Pintado y Romo (2009), para definir un orden entre curvas de acuerdo a sus profundidades, siendo $X_{[1]}(t)$ la curva más profunda (o la curva mediana) e $X_{[n]}(t)$ la más alejada o de menor profundidad. Como esta medida depende de la forma de la curva, se consideran dos tipos de *outliers*: *outliers* de magnitud y de forma. Los primeros consisten en curvas alejadas y los segundos en curvas con patrones diferentes a las demás curvas. Otras medidas de profundidad podrían utilizarse en su construcción. El lector interesado puede consultar el trabajo de Nieto–Reyes y Battey (2016) para una definición formal de profundidad estadística y para una comparación entre las distintas profundidades definidas en este contexto.

En el boxplot clásico definido por Tukey (1970), la caja representa el 50% de los datos y se obtiene a partir del cuartil inferior (Q_I) y superior (Q_S), mientras que los *bigotes* del boxplot son las líneas verticales que se extienden desde la caja indicando la máxima y mínima observación regular (excluyendo los *outliers*), es decir, las observaciones más extremas dentro del intervalo $[Q_I - 1.5d_I, Q_S + 1.5d_I]$ donde $d_I = Q_S - Q_I$ es la distancia intercuartil. En el caso de datos funcionales, el borde del 50% de la región central se define como el *sobre*, y representa lo mismo que en el boxplot univariado, la región que contiene el 50% de las curvas centrales. La curva mediana que es la curva más profunda, es una alternativa robusta de medir centralidad. Al igual que en el boxplot usual, en el boxplot funcional, se utiliza el criterio de dilatar 1.5 la región central, como mecanismo para identificar *outliers*, y así se

construyen las *cercas* del gráfico. Cualquier curva que esté fuera de las cercas, en algún intervalo, puede considerarse como un potencial dato atípico.

En la Fig. 8 se muestra un ejemplo correspondiente al conjunto de datos TECATOR disponible en la librería fda.usc del paquete R y que fue analizado por Ferraty y Vieu (2006).

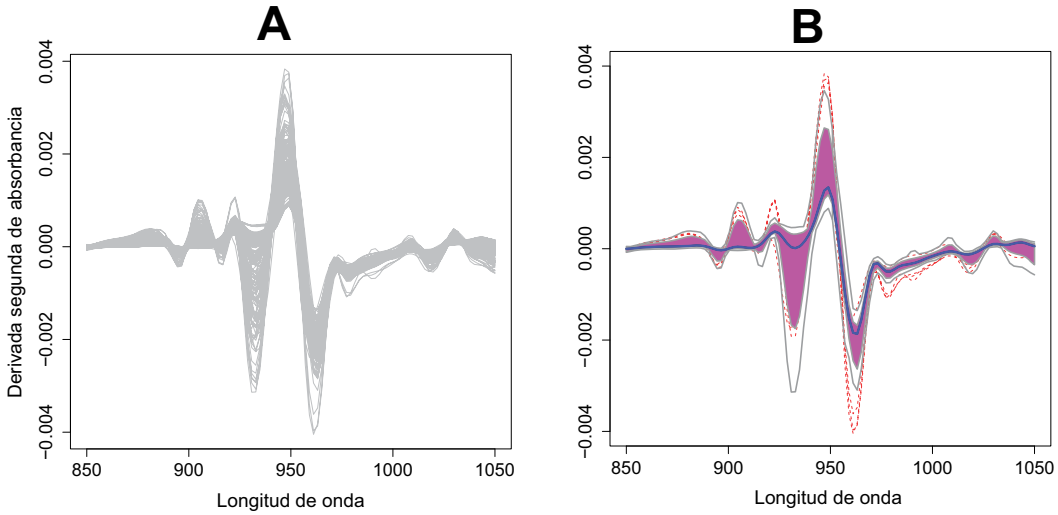

Fig. 8. Conjunto de datos TECATOR disponible en la librería fda.usc del paquete R, derivada segunda de la absorbancia. A: Datos, B: Boxplot funcional.

Este conjunto de datos corresponde a datos de control de calidad de alimentos de 215 muestras de carne finamente picadas con diferentes porcentajes de grasa, proteína y contenido de humedad. Para cada muestra, se midió una curva espectrométrica de absorbancias usando un analizador de alimentos Tecator Infratec en una grilla de 100 longitudes de onda desde 850nm a 1050nm. Los porcentajes de grasa, proteínas y contenido de humedad se determinan por un procedimiento analítico. La Fig. 8 presenta el boxplot funcional de la derivada segunda de la absorbancia. Las líneas cortadas rojas son las curvas detectados como anómalas, la región central o *sobre* se presenta coloreada en magenta y las *cercas* se muestran en gris. La línea azul en la parte central corresponde a la curva mediana.

En el caso de datos funcionales, los datos atípicos pueden presentar distintas estructuras. Los *outliers* no necesitan ser datos "extremos", pueden consistir de curvas que se comportan en forma diferente a las demás o que presentan un comportamiento persistente en nivel, amplitud y/o forma. Hubert et al. (2015) dan una descripción de distintos tipos de datos anómalos que se pueden encontrar en una muestra y los clasifican como los *outliers aislados* o sea, datos que muestran un comportamiento atípico durante un

intervalo de tiempo corto o datos atípicos persistentes, que se definen como datos funcionales que son anómalos en gran parte o en todo el dominio. Entre estos últimos distinguimos *outliers* en nivel, amplitud y/o forma.

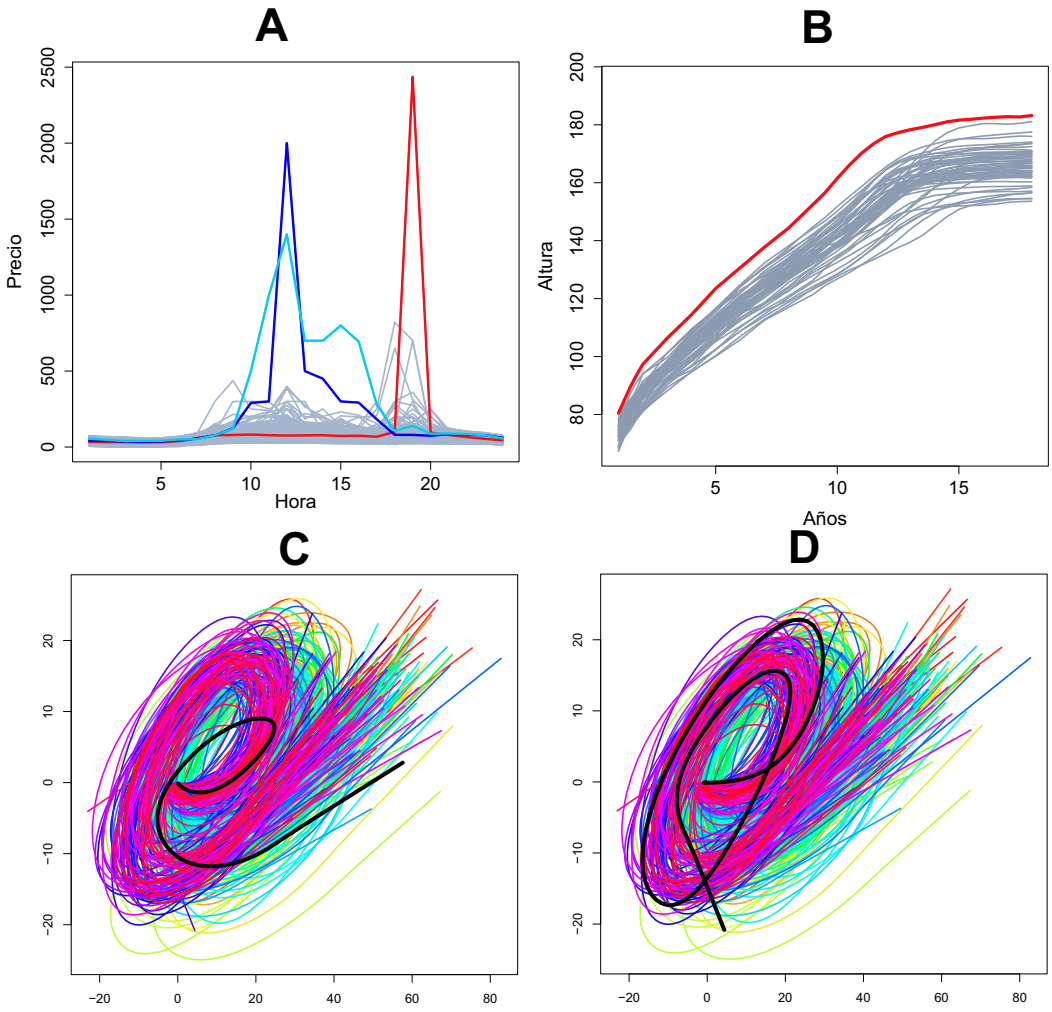

Fig. 9. A: Datos de electricidad en Alemania (de Liebl, 2013) con *outlier* aislado en rojo. B: Datos de altura de niñas (del Berkeley Growth Study, disponibles en la librería fda de R), con *outlier* de nivel en rojo, C y D: Datos de escritura de letra "e" en una tableta WACOM (de Bache y Lichman, 2013). Se destacan en negro un dato atípico de amplitud (C) y uno de forma (D).

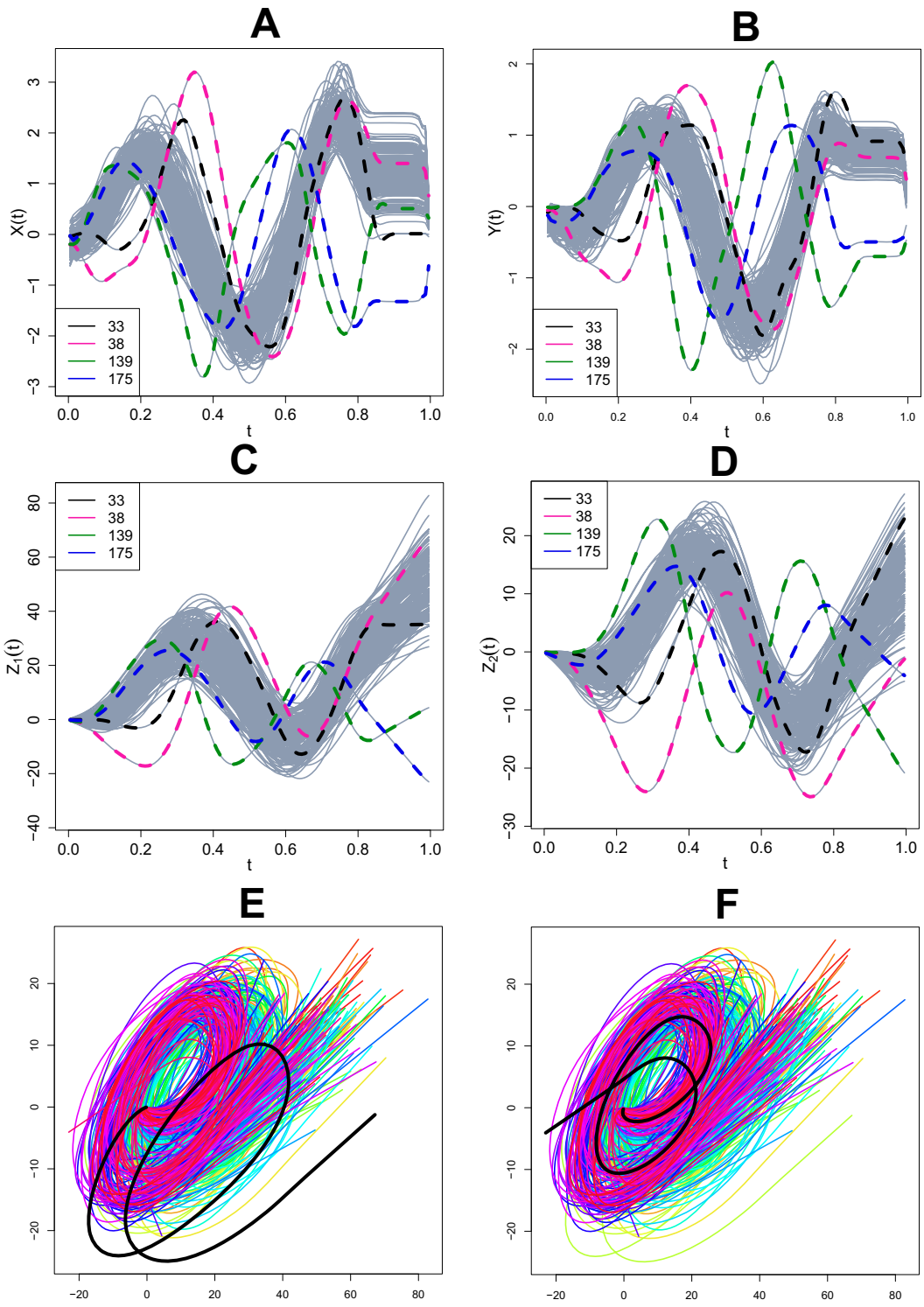

Fig. 10. Datos de letra "e". A y B: Velocidades X(t) e Y(t) en los ejes horizontal y vertical, respectivamente. C y D: Posición de la lapicera los ejes horizontal, $Z_1(t)$, y vertical, $Z_2(t)$, respectivamente. E y F: Letras dibujadas por el lápiz, se resaltan en negro las trayectorias 38 y 175, respectivamente. Basado sobre datos de Bache y Lichman (2013).

Un ejemplo de dato atípico aislado puede verse en rojo la Fig. 9A donde se grafica el precio horario de la electricidad en Alemania entre el 1 de Enero

de 2006 y el 30 de Septiembre de 2008. Estos datos fueron utilizados por Liebl (2013) y pueden encontrarse en el material suplementario de dicho trabajo. Se observan en dicho conjunto dos trayectorias también atípicas realzadas en azul y azul claro que corresponden al 25 y 27 de Julio de 2006. La trayectoria destacada en rojo representa los precios de electricidad el día 7 de Noviembre de 2006. Su comportamiento atípico aislado podría deberse a una acumulación de consumo debido al apagón ocurrido el sábado 4 de Noviembre en muchos de los países de la Unión Europea, incluída Alemania. El efecto de datos anómalos aislados puede dismuirse mediante un M-estimador local como el descripto en la Sección 2.

Un ejemplo de dato anómalo de nivel es fácilmente identificable en la Fig. 9B que muestra la altura de 54 niñas medidas sobre un conjunto de 31 edades (de 1 a 18 años) en el estudio Berkeley Growth Study. Los datos pueden encontrarse en la librería fda de R. Por otra parte, un ejemplo de atipicidad en amplitud es el mostrado en Fig. 9C, donde la forma es parecida a las demás pero en otra escala en su parte central y un *outlier* de forma se ilustra en la Fig. 9D. Esos datos fueron extraídos de de Bache y Lichman (2013) y corresponden a la posición de la punta de una lapicera sobre una tableta WACOM cuando un participante del experimento escribe distintas letras, en este caso, la letra e. Este conjunto de datos fue estudiado por Hubert et al. (2016) para ilustrar el buen desempeño del procedimiento de clasificación basado en profundidades que estos autores proponen.

La Fig. 10 muestra los 186 elementos de la muestra que corresponden a la velocidad en ambos ejes al escribir la letra “e”, así como el movimiento en ambos ejes de la lapicera. Se destacan, con líneas cortadas de color negro, magenta, verde y azul las observaciones 33, 38, 139 y 175, respectivamente, ya que tienen un comportamiento diferente del resto y se eligieron dos de ellas para mostrar el diseño de la letra dibujada.

Como se menciona en Galeano y Peña (2019) el uso de gráficos obtenidos mediante una reduccción de dimensión es de gran ayuda para verificar la homogeneidad de datos y detectar datos atípicos. Una manera de reducir la dimensión es mediante el uso de direcciones aleatorias, ver Cuevas et al. (2007). Otra forma de reducción de dimensión es el análisis de componentes principales funcionales que debe realizarse mediante una método robusto para evitar el enmascaramiento ya descripto, en el caso de observaciones en $\mathbb{R}^p$, por Pison et al. (2000). Diversas propuestas se han dado para estimar las direcciones principales funcionales en forma robusta. Cada una de ellas, se basa en dar un enfoque robusto a las propiedades que caracterizan a las componentes principales funcionales.

Probablemente la primer propuesta de estimadores robustos de las componentes principales funcionales corresponde al trabajo de Locantore et al. (1999) quienes propusieron las llamadas componentes principales esféricas, estudiadas posteriormente en Gervini (2008) y Boente et al. (2019). La idea subyacente de este procedimiento es controlar la influencia de los

posibles datos anómalos dividiendo a las observaciones centradas por su norma. Una propiedad atractiva de este procedimiento es que en el caso en que existan momentos de orden 2, los estimadores de las direcciones principales tienen el mismo límite (en casi todo punto) que los estimadores clásicos, una propiedad conocida como consistencia en el sentido de Fisher.

Hyndman y Ullah (2007) consideraron un enfoque diferente que se basa en el hecho que, en el caso clásico, la primer dirección principal maximiza, sobre la esfera unidad, la varianza de la proyección de X en la dirección α. Estos autores propusieron un procedimiento robusto basado en utilizar un estimador robusto de dispersión, en lugar de la varianza muestral, aplicado a la proyección de las trayectorias observadas y suavizadas. Bali et al. (2011) generalizaron este enfoque combinandolo con una penalización y proyecciones en bases de dimensión creciente de modo a obtener direcciones suaves. Este tipo de procedimientos, que se basan en la noción de encontrar proyecciones interesantes maximizando una cierta función objetivo o índice de proyección, suelen denominarse métodos de *projection–pursuit* y fueron discutidos, en el caso finito-dimensional, por Huber (2010) quien hace una amplia revisión de los mismos.

Finalmente, otra aproximación al problema consiste en definir directamente estimadores robustos de los espacios vectoriales generados por las primeras p direcciones principales, minimizando, por ejemplo, una medida robusta de la distancia entre las observaciones y sus proyecciones ortogonales sobre espacios de dimensión p. Esta aproximación fue utilizada por Lee et al. (2013), Boente y Salibian-Barrera (2015) y Cevallos-Valdiviezo (2016) para dar distintas propuestas robustas.

Como mencionamos anteriormente, procedimientos como el de componentes principales resultan útiles para identificar los datos atípicos mediante el boxplot de los escores o de la norma al cuadrado de los residuos del ajuste $\hat{r}_i(t) = X_i(t) - \hat{X}_i(t)$ con $\hat{X}_i(t) = \hat{\mu}(t) + \sum_{\ell=1}^{p} \hat{\zeta}_{i,\ell}\, \widehat{\phi}_\ell(t)$ la predicción de las trayectorias, donde indicamos por $\widehat{\phi}_\ell$, $1 \le \ell \le p$ a los estimadores robustos de las p primeras direcciones principales y por $\hat{\zeta}_{i,\ell}$, $1 \le i \le n$, $1 \le \ell \le p$ a los escores asociados.

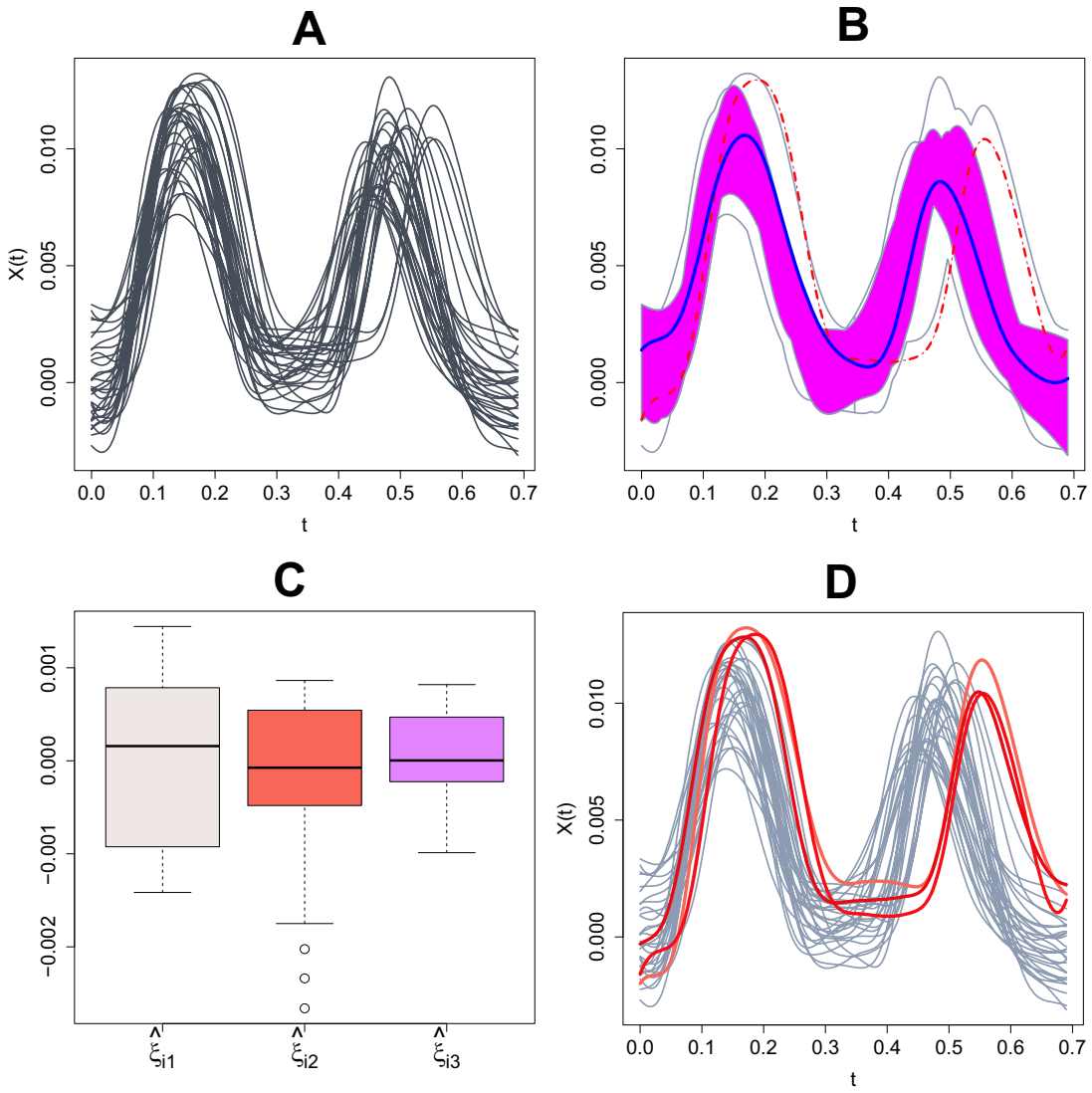

Fig. 11. Datos de movimiento del labio. A: Gráfico de las trayectorias X(t). B: Boxplot funcional. C: En rosa, rojo y violeta se presentan los boxplot de los escores correspondientes a la primera, segunda y tercera dirección principal, respectivamente. D: Movimiento del labio con trayectorias identificadas como atípicas por el boxplot en tonos de rojo. Basado sobre datos de Gervini (2008).

A modo de ejemplo, consideremos los datos de la Fig. 11A que corresponden al movimiento del centro del labio inferior al pronunciar la palabra *bob* en la frase *Say bob again* y que fueron utilizados en Gervini (2008). El movimiento del labio $X(t)$ se registra 32 veces en 501 instantes en el intervalo [0,0.69]. La Fig. 11B muestra el boxplot funcional que permite identificar una trayectoria como anómala. Sin embargo, del gráfico 11D se identifican otras dos trayectorias con el mismo patrón inusual. Para estas tres curvas la recuperación del labio al terminar la letra *o* es más tardía. En este caso, se utilizó la propuesta dada Bali et al. (2011) para estimar las direcciones principales y sus escores, logrando identificar las tres trayectorias con retardo en la recuperación del labio inferior mediante los escores

asociados a la segunda dirección principal cuyo boxplot se presenta en rojo en la Fig. 11C. Dichas trayectorias corresponden a las observaciones 24, 25 y 27.

En vista de estos ejemplos, es claro que es necesario desarrollar procedimientos robustos para resumir y analizar conjuntos de datos funcionales. Entre las técnicas de reducción de dimensión además del análisis de componentes principales funcional descripto más arriba, se encuentran: el análisis de componentes principales comunes funcional cuando tratamos con varias poblaciones y el análisis de correlación canónica funcional que puede utilizarse, por ejemplo, en el caso de la tableta WACOM para encontrar estimadores de las direcciones que maximizan la asociación entre las proyecciones de las velocidades del movimiento del lápiz en el eje horizontal y vertical. Estos problemas han sido abordados por Bali y Boente (2017) y por Alvarez et al. (2019) y Boente y Kudraszow (2020), respectivamente.

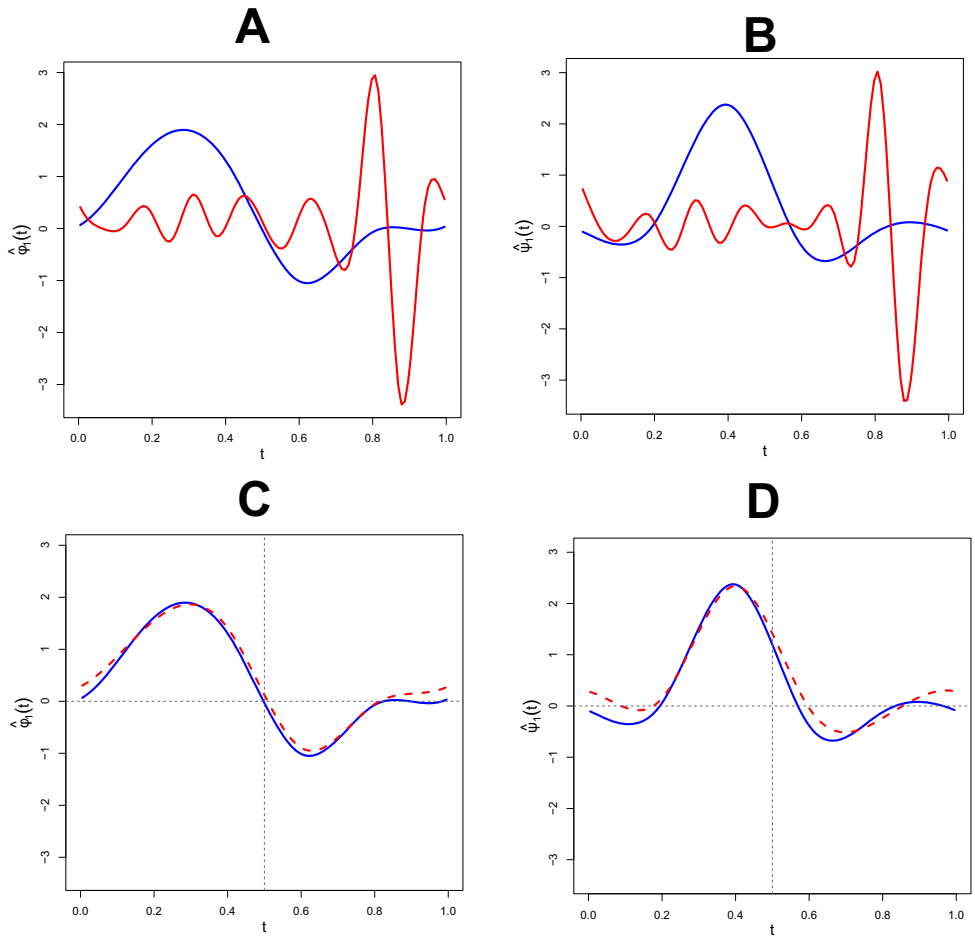

Fig. 12. Datos de la escritura de la letra "e". La línea azul corresponde al estimador robusto y la roja al clásico. A y B: Estimaciones $\hat{\Phi}_1$ y $\hat{\Psi}_1$ de las direcciones canónicas asociadas a X e Y, respectivamente, obtenidas usando todas las observaciones. C y D: Estimaciones de las direcciones canónicas asociadas a X e Y, respectivamente donde en línea cortada se indica la estimación clásica obtenida cuando se eliminan de la muestra las trayectorias detectadas como anómalas. Basado sobre datos de Bache y Lichman (2013).

Para ilustrar el efecto devastador de los datos anómalos consideremos el ejemplo de la letra "e". Las trayectorias correspondientes a la velocidad del lápiz en el eje horizontal (X) y vertical (Y) se muestran en la Fig. 10 con las trayectorias atípicas detectadas en color. En la Fig. 12 se presentan en rojo las estimaciones obtenidas mediante el procedimiento clásico descripto en He et al. (2004) y Ramsay y Silverman (2005). Indicamos por $\hat{\Phi}_1$ y $\hat{\Psi}_1$ a los estimadores de las primeras direcciones canónicas asociadas a X e Y, respectivamente. El panel superior corresponde a las estimaciones obtenidas usando todas las trayectorias. La línea azul corresponde a la estimación obtenida mediante el procedimiento robusto propuesto por Alvarez et al. (2019) utilizando el coeficiente de correlación de Spearman. En el panel inferior se muestra en línea cortada el estimador clásico obtenido cuando se eliminan de la muestra las trayectorias detectadas como anómalas.

Como puede verse en el gráfico, al eliminar las observaciones detectadas como atípicas la estimación clásica da resultados similares a los de la robusta calculada con todos los datos. Por otra parte, las estimaciones robustas permiten visualizar a las direcciones canónicas como un contraste en el rango [0,0.8] con menor peso para valores de tiempo superiores a 0.5 y son casi nulas luego de 0.8, lo que se explica por el hecho que tanto la velocidad en el eje horizontal como en el vertical son casi constantes después de ese instante de tiempo (ver Fig. 10).

El problema de regresión funcional generaliza el problema de regresión lineal muy estudiado en el caso de variables explicativas reales, a la situación en que las covariables son funciones. En este caso, los datos atípicos en las covariables funcionales con alta palanca tienen un efecto aún más devastador que en el caso real. Propuestas robustas para controlar el efecto de datos anómalos han sido dadas por Maronna y Yohai (2013) y por Kalogridis y Van Aelst (2019), mientras que el problema de selección de variables en presencia de varias covariables funcionales fue considerado por Pannu y Billor (2015). Una extensión de este modelo al caso parcialmente lineal en el que además de las covariables funcionales que entran al modelo en forma lineal, existen variables predictoras reales que predicen la respuesta en forma noparamétrica fue considerada por Huang et al. (2015), Qingguo (2015) y Boente et al. (2020).

Finalmente, a continuación presentaremos una situación más desafiante en el análisis de datos funcionales que corresponde al caso en que las trayectorias son esparsas, es decir, para cada individuo los datos se observan en una grilla de tiempo $\{t_{ij}\}_{1\le j\le n_i}$ donde n_i es pequeño y además pueden estar sujetos a errores de medición, o sea, observamos

$$X_{ij} = X_i(t_{ij}) + \varepsilon_{ij}\,,$$

donde ε_{ij} son independientes de los tiempos t_{ij} que pueden ser aleatorios y en el caso en que haya momentos, se supone que $\mathbb{E}\varepsilon_{ij} = 0$. En general, este tipo

de situaciones ocurre en estudios longitudinales donde los sujetos se miden en distintos tiempos y el número de observaciones n_i es usualmente acotado. El análisis de este tipo de datos funcionales requiere más esfuerzo metodológico que el de los observados densamente. Usualmente a pesar de que cada individuo se observa en pocos puntos, el conjunto $\{t_{ij}\}_{1\leq j\leq \mathrm{n}_i, 1\leq i\leq N}$ es un conjunto de puntos que cubre adecuadamente el intervalo $\mathcal{I}$ dominio de X. Este tipo de datos sigue un paradigma distinto ya que sólo permiten obtener tasa noparamétricas al estimar la media y operador de covarianza y ese problema es heredado por las propuestas robustas.

La Fig. 13A muestra un ejemplo de datos esparsos que corresponde a datos de CD4, que es parte del *Multicentre AIDS Cohort Study* (Zeger y Diggle, 1994). Los datos corresponden a 2376 mediciones de recuentos de células CD4, tomadas en 369 hombres. Los tiempos se miden en años desde la seroconversión ($t = 0$). Todo el conjunto de datos está disponible en paquete catdata de R. Para asegurarse de que haya suficientes observaciones para estimar el función de covarianza en cada par de puntos (s, t), nos centramos en el observaciones con $t \geq 0$, y en individuos con más de una medición, obteniéndose $N = 292$ curvas, con el número de observaciones por individuo que oscilan entre 2 y 11 (con una mediana de 5), como se observa en la Fig. 13A. Existen pocas propuestas para el análisis de componentes principales funcional para este tipo de datos. Un enfoque de estimación basado en splines fue dado por James et al. (2000) y propuestas robustas siguiendo este mismo punto de vista fueron consideradas por Gervini (2009) y Maronna (2019). Por otra parte, Yao et al. (2005) utilizan una aproximación que consiste en estimar la función de covarianza $\gamma(s, t) = \mathrm{Cov}(X(s), X(t))$ utilizando los productos cruzados de las observaciones existentes y un suavizado bivariado basado en núcleos. En particular, estos datos fueron analizados por Yao et al. (2005) y la superficie de la Fig. 13B corresponde a la estimación de la función de covarianza obtenida utilizando su propuesta.

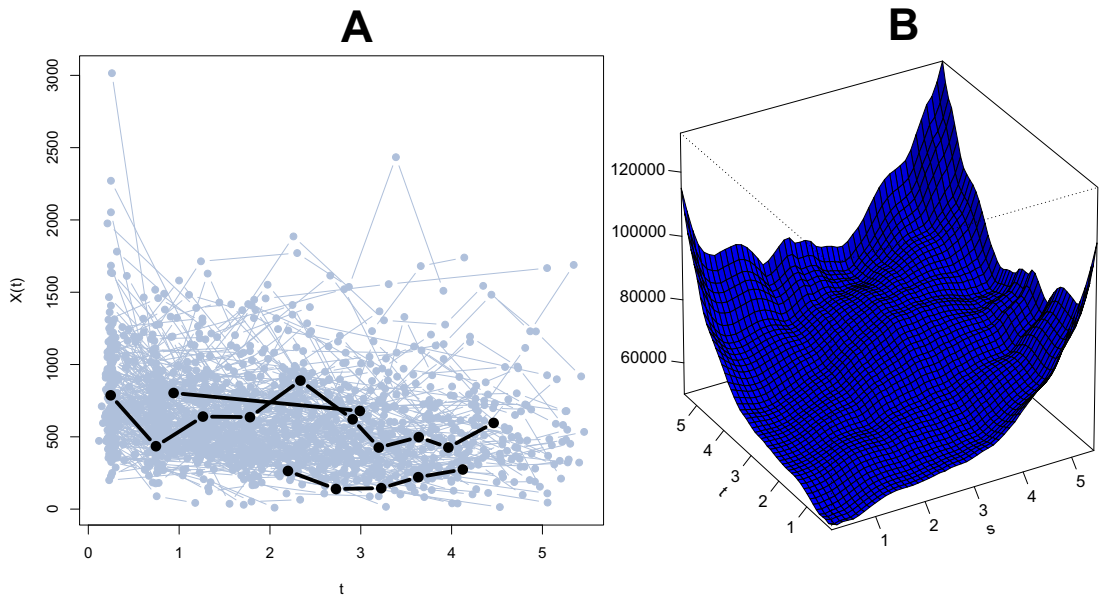

Fig. 13. A: Conteos de CD4 para N = 292 pacientes después de la seroconversión ($t \geq 0$). Se resaltan tres trayectorias elegidas al azar con líneas negras. B: Superficie correspondiente a la estimación de la función de covarianza $\gamma(s,t) = \mathrm{Cov}(X(s), X(t))$. Basado sobre datos del Multicentre AIDS Cohort Study (Zeger y Diggle, 1994).

En el contexto de datos esparsos quedan aún muchos problemas abiertos para los cuales alternativas robustas son necesarias como, por ejemplo, el de correlación canónica funcional o regresión funcional. Referimos al trabajo de Wang et al. (2016) para una discusión sobre este tema en el caso clásico, así como en el importante problema de test de hipótesis para la media y operadores de covarianza. En el caso de datos densos, bandas para el parámetro de posición fueron estudiadas por Lima et al. (2019a) y b).

4. Conclusiones

Esta nueva era donde los desarrollos computacionales generan datos cada vez más complejos, necesita de procedimientos estadísticos nuevos adaptados y diseñados para analizar la complejidad de los datos obtenidos. En este sentido, hay un gran campo en el que la Estadística y en particular, la rama de la Inferencia robusta, tienen que desempeñar un rol destacado. Es fundamental la propuesta de métodos robustos en el contexto del análisis de datos funcionales para obtener estimaciones confiables que permitan además detectar las observaciones atípicas ya que dichas trayectorias anómalas son, en muchos casos, difíciles de identificar a simple vista. El desarrollo de este tipo de procedimientos requieren una eficiente implementación numérica

para que los mismos puedan ser utilizados por la comunidad interesada así como el estudio teórico de sus propiedades que garantice que las propuestas resultan efectivamente consistentes a la cantidad de interés. Estos interesantes problemas seguramente darán lugar en el futuro a innovadores desarrollos con grandes posibilidades de aplicación en diversas ramas de la ciencia.

Agradecimientos

La autora quiere agradecer a los miembros de la Academia Nacional de Ciencias Exactas, Físicas y Naturales por considerarla merecedora de formar parte de dicha institución y por la invitación a publicar este artículo. Los resultados de este trabajo fueron financiados parcialmente por los proyectos PICT 2018-00740 de ANPCYT, 20020170100022BA de la Universidad de Buenos Aires, Buenos Aires, Argentina y por el proyecto MTM2016-76969P del Ministerio de Economía y Competitividad de España (MINECO/ AEI/FEDER, UE).

Referencias

Achenwall G (1749) Abrißder neuesten Staatswissenschaft der vornehmsten Europïschen Reiche und Republiken, Göttingen (Alemania).

Agostinelli C, Bianco A, Boente G (2020). Robust estimation in single index models when the errors have a unimodal density. Annals of the Institute of Mathematical Statistics, 72:855-893.

Alvarez A, Boente G, Kudraszow N (2019). Robust sieve estimators for functional canonical correlation analysis. Journal of Multivariate Analysis, 170:46-62.

Bali L, Boente G (2017) Robust estimators under a functional common principal components model. Computational Statistics and Data Analysis, 113:424-440.

Bali L, Boente G, Tyler D, Wang J-L (2011) Robust functional principal components: a projection-pursuit approach. Annals of Statistics, 39:2852-2882.

Bianco A, Boente G (2004) Robust estimators in semiparametric partly linear regression models. Journal of Statistical Planning and Inference, 122:229-252.

Bianco A, Boente G, González–Manteiga W, Pérez–González A (2011) Asymptotic behavior of robust estimators in partially linear models with missing responses: The effect of estimating the missing probability on the simplified marginal estimators. TEST, 20:524-548.

Bianco A, Spano P (2019) Robust inference for nonlinear regression models. TEST, 28:369-398.

Bickel PJ, Breiman L, Brillinger D, Brunk H, Pierce D, Chernoff H, Cover Th, Cox DR, Eddy W, Hampel F, Olshen R, Parzen E, Rosenblatt M, Sacks J, Wahba G (1977). Discussion: Consistent nonparametric regression. Annals of Statistics, 5:620-640.

Boente G, Fraiman R (1989a) Robust nonparametric regression estimation. Journal of Multivariate Analysis, 29:180-198.

Boente G, Fraiman R (1989b) Robust nonparametric regression estimation for dependent observations. Annals of Statistics, 17:1242-1256.

Boente G, Fraiman R (1990) Asymptotic distribution of robust estimates for nonparametric models from mixing observations. Annals of Statistics, 18:891-906.

Boente G, González–Manteiga W, Pérez–González A (2009) Robust nonparametric estimation with missing data. Journal of Statistical Planning and Inference, 139:571-592.

Boente G, He X, Zhou J (2006) Robust estimates in generalized partially linear models. Annals of Statistics, 34:2856-2878.

Boente G, Kudraszow N (2020) Robust smoothed canonical correlation analysis for functional data. Statistica Sinica, DOI: 10.5705/ss.202020.0084.

Boente G, Martínez A, Salibian–Barrera M (2017) Robust estimators for additive models using backfitting. Journal of Nonparametric Statistics, 29:744-767.
Boente G, Martínez A (2017) Marginal integration M-estimators for additive models. TEST, 26:231-260.
Boente G, Rodríguez D (2010) Robust inference in generalized partially linear models. Computational Statistics and Data Analysis, 54:2942-2966.
Boente G, Rodríguez D (2012) Robust estimates in generalized partially linear single-index models. TEST, 21:386-411.
Boente G, Rodríguez D, Sued M (2019) The spatial sign covariance operator: Asymptotic results and applications. Journal of Multivariate Analysis, 170:115-128.
Boente G, Salibián-Barrera M (2015) S-estimators for functional principal component analysis. Journal of the American Statistical Association, 110:1100-1111.
Boente G, Salibián Barrera M, Vena P (2020) Robust estimation for semi-functional linear regression models. En prensa en Computational Statistics and Data Analysis, 152:107041.
Box GEP (1953) Non-normality and tests on variance. Biometrika, 40:318-335.
Bravo F (2019) Robust estimation and inference for general varying coefficient models with missing observations. TEST, DOI: 10.1007/s11749-019-00692-0.
Buja A, Hastie T, Tibshirani R (1989) Linear smoothers and additive models (with discussion). Annals of Statistics, 17:453-555.
Cevallos-Valdiviezo H (2016) On methods for prediction based on complex data with missing values and robust principal component analysis. Tesis de Doctorado, Universidad de Ghent (Bélgica), pp. 1-157.
Chambers JM, Cleveland WS, Kleiner B, Tukey PA (1983) Graphical Methods for Data Analysis. CRC Press, Boca Raton (USA), pp. 1-395.
Cleveland W (1985) The elements of graphing data. Wadsworth, Monterey (USA), pp. 1-323.
Cox DD (1983). Asymptotics for M-type smoothing splines. Annals of Statistics, 11:530-551.
Cuevas A, Febrero M, Fraiman R (2007) Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics, 22:481-496.
Cunningham JK, Eubank RL, Hsing T (1991) M-type smoothing splines with auxiliary scale estimation. Computational Statistics and Data Analysis, 11:43-51.
Ferraty F, Vieu P (2006) Nonparametric Functional data analysis: Theory and Practice. Springer, New York (USA), pp.1-258.
Galeano P, Peña D (2019) Data science, big data and statistics. TEST, 28:289-329.
Gervini D (2008) Robust functional estimation using the median and spherical principal components. Biometrika, 95:587-600.
Gervini D (2009). Detecting and handling outlying trajectories in irregularly sampled functional datasets. Annals of Applied Statististics, 3:1758-1775.
Hampel FR (1968) Contributions to the theory of robust estimation, PhD Thesis, Dept. Statistics, Univ. California, Berkeley (USA), pp. 1-206.
Hampel F (1971) A general qualitative definition of robustness. Annals of Mathematical Statistics, 42:1887-1896.
Hampel F (1974) The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69:383-393.
Härdle W (1990) Applied nonparametric regression. Cambridge University Press, New York (USA), pp. 1-333.
Härdle W, Müller M, Sperlich S, Werwatz A (2004) Nonparametric and Semiparametric Models. Springer, New York (USA), pp. 1-299.
Härdle W, Tsybakov AB (1988) Robust nonparametric regression with simultaneous scale curve estimation. Annals of Statistics, 16:120-135.
Hastie T, Tibshirani RJ (1990) Generalized Additive Models. Monographs on Statistics and Applied Probability No. 43. Chapman and Hall, London (UK), pp.1-335.
He G, Müller HG, Wang JL (2004) Methods of canonical analysis for functional data. Journal of Statistical Planning and Inference, 122:141-159.
He X, Shi P (1996) Bivariate tensor–product B–spline in a partly linear model. Journal of Multivariate Analysis, 58:162-181.
He X, Zhue ZY, Fung WK (2002) Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika, 89:579-590.
Heritier S, Cantoni E, Copt S, Victoria-Feser M-P (2009) Robust Methods in Biostatistics. Wiley, Chichester (UK), pp. 1-268.
Horowitz J (2009) Semiparametric and Nonparametric Methods in Econometrics. Springer, New York (USA), pp. 1-271.
Horváth L, Kokoszka, P (2012). Inference for functional data with applications. Springer, New York (USA), pp. 1-422

Huang L , Wang H, Cui, H, Wang S (2015). Sieve M-estimator for a semi-functional linear model. Science China, Mathematics, 58: 2421-2434.

Huber P (1964) Robust estimation of a location parameter. Annals of Mathematical Statistics, 35:73-101.

Huber P (1967) The behavior of maximum likelihood estimates under nonstandard conditions. En: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 221–233, University of California Press, Berkeley (USA).

Huber P (1968) Robust confidence limits. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 10:269-278.

Huber PJ (1979) Robust smoothing. En: Launer LR, Wilkinson GN (eds.) Robustness in Statistics, Academic Press, New York (USA), pp. 33-47.

Huber P (1985) Projection pursuit. Annals of Statistics, 13:435-475.

Huber P (2010). Data Analysis: What can be learned from the past 50 years, Wiley, New York (USA), pp. 1-235.

Huber, P, Ronchetti, E. (2009). Robust Statistics. Wiley, New York (USA), pp.1-363.

Hubert M, Rousseeuw PJ, Segaert P (2015) Multivariate functional outlier detection. Statistical Methods, Applications, 24:177-202.

Hyndman RJ, Ullah S (2007) Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics and Data Analysis, 51:4942-4956.

James G, Hastie T, Sugar C (2000) Principal component models for sparse functional data. Biometrika, 87:587-602.

Kalogridis I, Van Aelst S (2019) Robust functional regression based on principal components. Journal of Multivariate Analysis, 173:393-415.

Kalogridis I, Van Aelst S (2021) M-type penalized splines with auxiliary scale estimation. Journal of Statistical Planning and Inference, 212:97-113.

Kalogridis I (2020). Asymptotics for M-type smoothing splines with non-smooth objective functions. https://arxiv.org/pdf/2002.04898.pdf.

Lee S, Shin H, Billor N (2013) M-type smoothing spline estimators for principal functions. Computational Statistics and Data Analysis, 66:89-100.

Liebl D (2013) Modelling and forecasting electricity spot prices: a functional data perspective. The Annals of Applied Statistics, 7:1562-1592.

Lima IR, Cao G, Billor N (2019a) Robust simultaneous inference for the mean function of functional data. TEST, 28:785-803.

Lima IR, Cao G, Billor N (2019b) M-Based simultaneous inference for the mean function of functional data. Annals of the Institute of Statistical Mathematics, 71:577-598.

Linton O, Nielsen J (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika, 82:93-101.

Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K (1999) Robust principal components for functional data. TEST, 8:1-28.

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

Maronna R (2019) Robust functional principal components for irregularly spaced longitudinal data. Statistical Papers. DOI: 10.1007/s00362-019-01147-2

Maronna M, Yohai V. (2013) Robust functional linear regression based on splines. Computational Statistics and Data Analysis, 65:46-55.

Maronna R, Martin R, Yohai V, Salibián-Barrera M (2019) Robust Statistics: Theory and Methods (with R). Wiley, New York (USA), pp.1-430.

McGill R, Tukey JW, Larsen W (1978) Variations of Box Plots. The American Statistician, 32:12-16.

Nadaraya EA (1964) On estimating regression. Theory Probability and Applications, 9:141-142.

Nieto-Reyes A, Battey H (2016) A topologically valid definition of depth for functional data. Statistical Science, 31:61-79.

Pannu J, Billor N (2015) Robust group-Lasso for functional regression model. Communications in Statistics - Simulation and Computation. 46:3356-3374

Pison G, Rousseeuw PJ, Filzmoser P, Croux C (2000) A robust version of principal factor analysis. En: Bethlehem J, van der Heijden P (eds.), Proceedings in Computational Statistics. Physica–Verlag, Heidelberg (Alemania), pp. 385-90.

Qingguo T (2015). Estimation for semi-functional linear regression. Statistics, 49; 1262-1278.

Ramsay JO, Silverman BW (2005). Functional Data Analysis. Springer, New York (USA), pp. 1-426.

Severance-Lossin E, Sperlich S (1999) Estimation of derivatives for additive separable models, Statistics, 33:241-265.

Sinclair J. (ed.) (1791-1799) The statistical account of Scotland. Edinburgh (UK), 21 vols.

Stone C (1977) Consistent nonparametric regression. Annals of Statistics, 5:595-645.
Stone C (1982) Optimal global rates of convergence for nonparametric regression. Annals of Statistics, 10:1040-1053.
Stone C (1985) Additive regression and other nonparametric models. Annals of Statistics, 13:689-705.
Sun Y, Genton MG (2011) Functional boxplots. Journal of Computational and Graphical Statistics, 20:316-334.
Tukey JW (1960) A survey of sampling from contaminated distributions. En: Olkin I, Ghurye, S, Hoefding W, Madow W, Mann, H. (eds.), Contributions to probability and statistics: Essays in honor of Harold Hotelling. Stanford University Press, Stanford (USA), pp. 448-485.
Tukey JW (1970) Exploratory Data Analysis. Addison-Wesley, Massachussets (USA), pp. 1-711.
van der Zande J (2010) Statistik and history in the German enlightenment. Journal of the History of Ideas, 71:411-432.
Wang JL, Chiou J, Müller HG (2016) Functional Data Analysis. Annual Review of Statistics and its Application, 3:257-295.
Watson GS (1964) Smooth regression analysis. Sankhya A, The Indian Journal of Statistics, 26:359-372.
Yao F, Müller HG, Wang JL (2005) Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association, 100:577-590.
Yohai VJ (1987) High breakdown-point and high efficiency robust estimates for regression. Annals of Statistics, 15:642-656.
Zeger SL, Diggle PJ (1994) Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics, 50:689-699.