Universidad Autónoma de Madrid.
Facultad de Ciencias.
Departamento de Matemáticas.

Estimación no paramétrica de conjuntos de nivel. Algunas aplicaciones

Amparo Baíllo Moreno

Tesis doctoral dirigida por
D. Antonio Cuevas González

Madrid, marzo de 2000

A mi abuelo Félix,
ejemplo de trabajo, sacrificio y coherencia

Deseo agradecerle a Antonio todo el esfuerzo y el tiempo que me ha dedicado durante estos cinco años. Su capacidad de trabajo y la atención que me ha prestado han sido para mí un gran estímulo. Me encanta su manera de "vivir" las matemáticas, su visión estadística y lo bien que explica. Le agradezco la minuciosidad y paciencia con la que ha leído cada cosa que escribía y, sobre todo, la tesis.

A mis padres les agradezco las posibilidades que tan generosamente han puesto al alcance de mi mano. Gracias especialmente a mi madre, por ser una compañera de piso y de viaje inmejorable y genial. A mi hermano Álvaro por apreciar tanto las matemáticas y por hacerme creer fervientemente que tienen aplicación. A Gonza por su chispa.

A Pedro porque nuestro proyecto en común es el sentido de mi trabajo y mi ilusión.

A mis compañeros por su buen humor y su tolerancia. Especialmente a Maite, que tantísimo me ha ayudado desde que entramos en el departamento. A Ana, porque nos entendemos fenomenal y el despacho sin ella no será el mismo. A Mayte, porque gracias a ella entiendo mejor algunas cosillas de estadística. A Raúl, por la paciencia con la que me resuelve todas mis dudas de ordenador. A Susi, Yolanda, Chema y Aurora por ser tan entrañables. A María por los festivos compartidos en la uni. A Betito, Pablo y Toño por tantos años de amistad.

A Ana Justel y Juan Cuesta les agradezco su inestimable colaboración en los dos artículos.

Aunque resulta imposible, me encantaría recordar aquí a todos los que han puesto su granito cada día para que esta labor y, en general, mi vida fuese más agradable. A todos ellos, gracias.

Índice

Capítulo 1

Introducción

Al principio de este primer capítulo se presenta de manera muy general el problema de la estimación de conjuntos de nivel, cuya metodología y aplicaciones se estudiarán a lo largo de la memoria desde distintos puntos de vista. A continuación, en la Sección 1.2, se presentan brevemente algunas de las propuestas más representativas de la literatura sobre estimación de conjuntos y se definen dos estimadores, cuyo comportamiento se analizará en los capítulos siguientes. Concluimos la sección con una relación de situaciones prácticas en las que se pueden plantear problemas de este tipo. En la Sección 1.3 se puede encontrar un resumen del contenido de los restantes capítulos. La última parte del capítulo 1 está dedicada a definir notación y enunciar ciertas desigualdades clásicas que se utilizarán más adelante.

1.1 Planteamiento del problema

La teoría de estimación de conjuntos estudia el problema estadístico de estimar un conjunto $S \subset \mathbb{R}^d$ desconocido (generalmente compacto) a partir de una muestra aleatoria de puntos $X_1, \ldots, X_n$ cuya distribución está relacionada con S. El ejemplo más sencillo corresponde al caso en que las X_i tienen la misma distribución de probabilidad y S es el **soporte** de la misma. Una versión más general de

este problema consiste en estimar los **conjuntos de nivel** $S = \{f > c\}$, donde f denota la densidad de X_i (respecto a la medida de Lebesgue) y $c \geq 0$ es una constante prefijada. Estos dos problemas coinciden cuando f es la densidad de una distribución uniforme o, en general, cuando está acotada inferiormente por una constante positiva, ya que, para c suficientemente pequeño, todos los conjuntos de nivel $\{f > c\}$ son precisamente el soporte de f.

El planteamiento anterior considera problemas que están conceptualmente relacionados con los fundamentos de la estereología (*stereology*). El objetivo de ésta es, en un sentido amplio, la reconstrucción de un objeto a partir de la información proporcionada por observaciones de dimensión inferior a la del objeto. Por ejemplo, dado un objeto en dimensión tres, podríamos tratar de estimarlo efectuando en él secciones aleatorias de dimensión dos (planos), uno (líneas rectas) o cero (puntos).

1.2 Una revisión de resultados previos

Se suele citar a Geffroy (1964 a,b) y a Renyi y Sulanke (1963, 1964) como primeras referencias sobre el problema de estimación del soporte. Geffroy (1964 a,b) estudió el caso en que el soporte de la densidad f es el hipografo de una función desconocida g, una hipótesis muy razonable en el análisis de imágenes (ver Korostelev y Tsybakov 1993). Concretamente Geffroy propuso un estimador constante a trozos del soporte, probó la consistencia del mismo, respecto a la métrica funcional L^∞, y obtuvo la distribución asintótica de la distancia del supremo entre estimador y soporte.

1.2.1 El caso convexo

Sin embargo, la mayor parte de la literatura sobre estimación de conjuntos ha estado principalmente orientada a la aproximación de un soporte S convexo. Ba-

jo esta hipótesis, en dimensión $d = 2$, Renyi y Sulanke (1963, 1964) propusieron un estimador muy natural de S: el **cierre convexo** (*convex hull*) de la muestra, $H_n = \text{conv}(X_1, \ldots, X_n)$, que es simplemente el conjunto de todas las combinaciones lineales convexas de las X_i, $i = 1, \ldots, n$. El caso convexo presenta una estructura matemática especialmente manejable debido a la existencia de una isometría que identifica los dominios compactos convexos con sus correspondientes funciones soporte (ver, por ejemplo, Schneider 1993, Gardner 1995). Esto permite aunar resultados de la geometría estocástica y de la teoría de conjuntos aleatorios.

Otras referencias sobre las propiedades de H_n (vértices, volumen, contenido de probabilidad, consistencia,...) y su generalización a dimensiones superiores son Efron (1965), Ripley y Rasson (1977), Moore (1984), Vitale (1987), Groeneboom (1988), Schneider (1988), Molchanov (1993a), Hueter (1994), Dümbgen y Walther (1996).

1.2.2 El caso general: dos estimadores intuitivos

Bajo la hipótesis de que el soporte S sea un conjunto convexo, el estimador natural es H_n, el cierre convexo de la muestra o una dilatación de éste (ver Moore 1984, Ripley y Rasson 1977). En general, cuando S no se supone convexo, no hay un único estimador natural. En esta memoria se estudiará el comportamiento de un estimador del soporte y de un estimador de conjuntos de nivel de densidades, ambos muy fáciles de visualizar y que presentamos a continuación.

Un estimador intuitivo del soporte

Chevalier (1976) y Devroye y Wise (1980) proponen un estimador sencillo e intuitivo, basado en una modificación directa de la muestra,

$$\hat{S}_n = \bigcup_{i=1}^{n} B(X_i, \epsilon_n), \tag{1.1}$$

donde $B(x, a)$ denota la bola de centro x y radio a, y ϵ_n es una sucesión de parámetros de suavizado que debe tender a cero (pero no demasiado rápido) para

que haya consistencia. Para estudiar la consistencia, primero hay que definir distancias adecuadas entre conjuntos. Las más generalizadas son la "distancia en medida", d_μ, que evalúa la μ-medida de las zonas de cada conjunto que no son comunes al otro, y la distancia de Hausdorff, d_H, que es lo que hay que dilatar ambos conjuntos para que se contengan mutuamente (ver Sección 1.4).

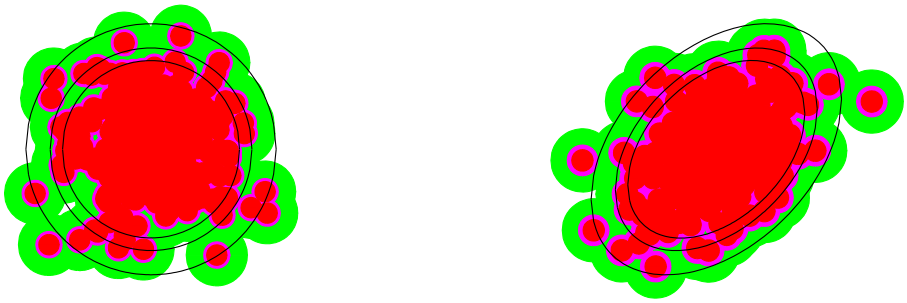

Figura 1: Aspecto del estimador de Devroye y Wise para distintos tamaños del radio ϵ_n (Baíllo, Cuevas y Justel 2000)

Bajo la hipótesis de que S sea ϵ-convexo, Chevalier (1976) consigue tasas de convergencia de $\hat{S}_n$ a S con respecto a una distancia muy parecida a la de Hausdorff. Devroye y Wise (1980) prueban resultados de consistencia universal (es decir, sin restricciones sobre S ni sobre la distribución de los X_i) para el estimador (1.1). Concretamente demuestran que una condición suficiente para que la distancia en medida (respecto a una medida μ) entre $\hat{S}_n$ y S tienda a cero en probabilidad es

$$\epsilon_n \to 0 \quad \text{y} \quad n\epsilon_n^d \to \infty \tag{1.2}$$

cuando $n \to \infty$ (ver también Grenander 1981). Esto se verifica para cualquier medida μ tal que la distribución de X_i, $P_X << \mu$ (en S). También prueban que, si

$$\epsilon_n \to 0 \quad \text{y} \quad \sum_{n=1}^{\infty} \exp(-\alpha n \epsilon_n^d) < \infty \quad \text{para todo } \alpha > 0, \tag{1.3}$$

entonces la distancia en medida converge a cero casi seguro (c.s.).

Observemos que las condiciones (1.2) son precisamente las que se imponen al parámetro de suavizado h_n, en los estimadores no paramétricos de la densidad de

tipo *kernel*, para obtener resultados de consistencia de tipo L^1 (ver Wand y Jones 1995, Simonoff 1996). La condición (1.3) es muy parecida a la que se le pediría a h_n para que el estimador *kernel* fuera consistente casi seguro en L^∞ (ver Nadaraya 1989). En definitiva, la sucesión de radios ϵ_n en el estimador (1.1) juega un papel análogo al de la ventana h_n en los estimadores *kernel* de la densidad.

En el marco del análisis de imágenes y bajo las hipótesis de que las X_i sigan una distribución uniforme en S, que S tenga una frontera Lipschitz a trozos y que ϵ_n sea de orden $O(\log n/n)^{1/d}$, Korostelev y Tsybakov (1993) obtienen tasas de orden $O(\epsilon_n)$ para la convergencia en probabilidad de $d_\mu(\hat{S}_n, S)$ a cero.

El estimador (1.1) puede considerarse como un caso particular de la propuesta *plug-in* para estimar el soporte que aparece en Cuevas y Fraiman (1997). Estos autores obtienen tasas para la convergencia a cero de la distancia de Hausdorff entre el estimador $\hat{S}_n$ y el soporte S (ver también Cuevas 1990). Las hipótesis impuestas sobre S no son muy restrictivas (ver estandaridad en la Sección 1.4), se pide que f esté acotada inferiormente por una constante positiva y que el parámetro de suavizado verifique las siguientes condiciones

$$\epsilon_n \to 0 \quad \text{y} \quad n\epsilon_n^d / \log n \to \infty.$$

La metodología matemática empleada con el estimador (1.1) está más cerca de los planteamientos y resultados de la estimación no paramétrica de densidades que de la teoría relacionada con el cierre convexo. Esto no resulta sorprendente si tenemos en cuenta que estimar un conjunto equivale a estimar una función indicatriz. De hecho, $\hat{S}_n$ es el soporte de un estimador *kernel*, cuyo núcleo es una función de densidad uniforme en la bola unidad y cuya amplitud de banda es ϵ_n. La diferencia más importante respecto a la estimación funcional radica en que aquí necesitaremos distancias entre conjuntos, en lugar de las métricas funcionales que se usan en la estimación de densidades.

Un estimador *plug-in* de los conjuntos de nivel

La estimación no paramétrica de densidades también sugiere un método natural para estimar el soporte S. Bajo condiciones de regularidad sobre f no demasiado

restrictivas, S es esencialmente el conjunto $\{f > 0\}$. Un estimador *plug-in* de $\{f > 0\}$ sería $\{f_n > 0\}$, donde f_n es un estimador no paramétrico de la densidad f. Sin embargo, los estimadores del tipo $\{f_n > c_n\}$ con $c_n \downarrow 0$, son más flexibles. Esta idea, desarrollada en Cuevas y Fraiman (1997), se puede trasladar a la estimación de conjuntos de nivel $\{f > c\}$, donde $c > 0$.

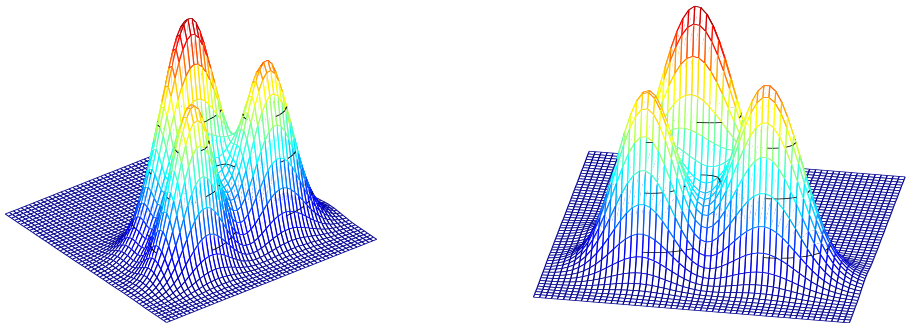

Figura 2: Conjuntos de nivel de una mixtura de tres normales

En general se han utilizado planteamientos de tipo minimax para estimar los conjuntos de nivel $\{f > c\}$. El estimador óptimo desde el punto de vista minimax se define de la manera siguiente: se fija una clase de conjuntos $\mathcal{G}$ (elipsoides, conjuntos estrellados con fronteras suaves,...) y se define una familia $\mathcal{F}$ de densidades f, tal que el conjunto S asociado a f que se desea estimar (el soporte, un conjunto de nivel) pertenezca a $\mathcal{G}$. El objetivo es encontrar un estimador S_n^* de S que sea consistente con la mejor tasa posible, es decir, que satisfaga

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} E_f \left[w\left(\frac{d(S_n^*, S)}{\psi_n} \right) \right] < \infty,$$

donde la sucesión $\psi_n = \psi_n(\mathcal{F})$ es la "tasa óptima de convergencia", pues debe verificar

$$\liminf_{n\to\infty} \inf_{S_n} E_f \left[w\left(\frac{d(S_n, S)}{\psi_n} \right) \right] > 0.$$

$\inf_{S_n}$ denota ínfimo sobre todos los posibles estimadores de S, d denota una distancia entre conjuntos (de la que también dependerá la tasa ψ_n) y w es una función de pérdida.

A pesar de que esta metodología frecuentemente proporciona tasas de convergencia óptimas o casi óptimas, si se decide no asumir la pertenencia de los conjuntos de nivel a una familia específica, es más razonable un estimador del tipo $\{f_n > c\}$.

Para el estimador del soporte $\{\hat{f}_n > c_n\}$, donde $\hat{f}_n$ es un estimador *kernel* y $c_n \downarrow 0$, Cuevas y Fraiman (1997) obtienen tasas de convergencia a S, respecto a las distancias en medida y de Hausdorff. Molchanov (1998) también estudia estimadores *plug-in* de los conjuntos de nivel. En concreto, da condiciones para que se verifique la consistencia fuerte

$$d_H(\{f_n \leq c\} \cap K, \{f \leq c\} \cap K) \to 0 \quad \text{c.s.}, \quad \forall K \subset \mathbb{R}^d \text{ compacto},$$

donde f_n es un estimador genérico de f. Este autor también estudia las hipótesis necesarias para que esta consistencia tenga lugar con la misma tasa con la que el proceso $\{f_n(x) - f(x) : x \in \mathbb{R}^d\}$ converge en distribución.

Tsybakov (1997) menciona un aspecto importante de esta metodología *plug-in*: bajo condiciones de regularidad, los estimadores de los conjuntos de nivel del tipo $\{f_n \leq c\}$ heredan las propiedades de consistencia del estimador f_n de la densidad.

1.2.3 Otras propuestas

Bajo las hipótesis de que el soporte $S \subset \mathbb{R}^2$ sea hipografo de una función suave y que las observaciones sean uniformes en S, Korostelev y Tsybakov (1993) consiguen tasas **minimax** óptimas (respecto a la métrica de Hausdorff y a la distancia en medida) para estimadores de S definidos como hipografos de funciones polinómicas a trozos. Asimismo prueban la optimalidad (respecto a la distancia en medida) de una versión discretizada del estimador de máxima verosimilitud, para soportes que verifican ciertas restricciones de forma (clase de Dudley y convexidad). Bajo la misma hipótesis de que S sea hipografo de una función g suficientemente

regular, Härdle, Park y Tsybakov (1995) prueban que el resultado de optimalidad obtenido por Korostelev y Tsybakov (1993) es generalizable a observaciones X_i que provengan de una distribución no necesariamente uniforme. Sin embargo, estas tasas óptimas sólo se alcanzan si se conocen ciertos parámetros de regularidad que caracterizan a g y a la densidad f de las X_i.

Aquí un aspecto relevante es la nitidez (*sharpness*) de la frontera, es decir, la velocidad con la que f decrece a cero en las proximidades de la frontera de S (ver sección 1.4): cuanto mayor es la nitidez, más rápida es la convergencia.

Korostelev y Tsybakov (1993) y Mammen y Tsybakov (1995) justifican la búsqueda de las tasas minimax óptimas como medio de comparación de diferentes métodos en estimación de conjuntos, más que por su posible aplicación a situaciones reales. A pesar de ello, esta última referencia presenta sus resultados en el contexto del reconocimiento de formas aplicado a imágenes en blanco y negro, siendo su objetivo la estimación de las regiones en negro.

Con el mismo objeto de analizar una imagen en blanco y negro, pero utilizando técnicas computacionales del análisis armónico, Donoho (1999) desarrolla una familia de funciones llamadas *wedgelets*, que vendrían a sustituir a las series de Fourier y a las ondículas (*wavelets*) en la representación y recuperación de cada *pixel* de una imagen definida por una función "no suave".

Hartigan (1987) y Müller y Sawitzki (1991) propusieron un estimador de conjuntos de nivel basado en el concepto de **exceso de masa** (*excess mass*). Por definición, el exceso de masa de un conjunto $A \in \mathcal{B}_{\mathbb{R}^d}$ es una función $M_A : \mathbb{R}^+ \to \mathbb{R}$ definida por

$$M_A(c) = \int_A f(x)dx - c\mu_L(A).$$

Claramente $M_{\{f \geq c\}}(c) \geq M_A(c)$ para cualquier A.

Si definimos el exceso de masa empírico como $M_{A,n}(c) = \frac{1}{n}\sum_{i=1}^n I\{X_i \in A\} - c\mu_L(A)$, un estimador natural de $\{f \geq c\}$ es el conjunto A que maximice $M_{A,n}$ sobre una clase prefijada $\mathcal{G}$ de conjuntos. La restricción $A \in \mathcal{G}$ es necesaria pues, de lo contrario, el estimador resultante sería $A = \{X_1, \ldots, X_n\}$. Hartigan (1987), Müller y Sawitzki (1991), Nolan (1991), Müller (1993) y Polonik (1995, 1999) estudiaron estos estimadores, probaron su consistencia y encontraron tasas de convergencia. Tsybakov (1997) prueba tasas óptimas en el caso de estimadores polinómicos a

trozos basados en una versión local del exceso de masa.

Los conjuntos de nivel también se pueden interpretar como **regiones de tolerancia**. Das Gupta, Ghosh y Zen (1995) presentan un método para construir regiones de confianza de mínimo volumen multivariantes para la moda de la densidad, a la que consideran parámetro de posición. Para el caso particular en que la densidad es "estrellada unimodal" y con una forma totalmente determinada (salvo por el valor del parámetro de posición), ofrecen una fórmula analítica para un conjunto de confianza estrellado.

De todos los procedimientos presentados hasta el momento, muchos son de difícil implementación computacional. Walther (1997) insiste en que la teoría sobre estimación de conjuntos no debería ignorar este aspecto práctico. Con este objeto presenta estimadores de conjuntos de nivel que gozan de excelentes propiedades teóricas y computacionales sobre una clase muy amplia y flexible de conjuntos. Para definir esta clase, se utiliza básicamente el **teorema de rodamiento** (*rolling theorem*) de **Blaschke** (1949), en el que se dan condiciones necesarias y suficientes para que una bola "ruede libremente" (*roll freely*) dentro y fuera de un conjunto compacto, convexo y no vacío de $\mathbb{R}^2$ o $\mathbb{R}^3$. Walther (1997, 1999) extiende este resultado y caracteriza la clase de los conjuntos compactos de $\mathbb{R}^d$ dentro y fuera de los cuales puede rodar libremente una bola de radio menor o igual que una constante r_0. La clase de conjuntos resultante es muy completa e incluye conjuntos no convexos y disconexos. De hecho, se trata esencialmente de los conjuntos r_0-convexos, que ya mencionamos al hablar de Chevalier (1976). En el Teorema 2 de Walther (1997) se dan condiciones suficientes sobre una densidad para que sus conjuntos de nivel estén en dicha clase.

Estas ideas se pueden utilizar para "suavizar" un conjunto de datos multivariante y para diseñar estimadores de conjuntos de nivel $\{f \geq c\}$ y conjuntos de mínimo volumen con un contenido de probabilidad prefijado. Para ello dividamos la muestra $X_1, \ldots, X_n$ en dos subconjuntos

$$\begin{aligned} \mathcal{X}_n^+(c) &:= \{X_i : \hat{f}_n(X_i) \geq c\}, \\ \mathcal{X}_n^-(c) &:= \{X_i : \hat{f}_n(X_i) < c\}, \end{aligned} \tag{1.4}$$

donde $\hat{f}_n$ es un estimador *kernel* de la densidad. Basándose en una versión empírica

del concepto de "granulometría" (ver Matheron 1975) e imponiendo condiciones de regularidad sobre f, Walther (1997) define el siguiente estimador del conjunto de nivel c de f

$$L_n(c) := ((\mathcal{X}_n^-(c) \oplus \epsilon_n B)^c \cap \mathcal{X}_n^+(c)) \oplus \epsilon_n B,$$

donde $\oplus$ denota suma de Minkowski (ver Sección 1.4).

El estimador L_n consiste en tomar bolas de radio ϵ_n en torno a los puntos muestrales X_i en los que $\hat{f}_n(X_i) \geq c$ y que distan por lo menos ϵ_n de los X_j en los que $\hat{f}_n(X_j) < c$. Esencialmente este estimador es una unión de bolas (al igual que (1.1)) en torno a los puntos "más profundos" de la muestra. El radio ϵ_n puede ser constante o una sucesión de parámetros de suavizado.

Mediante un procedimiento análogo, Walther (1997) propone también estimadores para los conjuntos de mínimo volumen y para el caso en que f tiene cierto tipo de discontinuidades. Por último, estudia tasas de convergencia y algoritmos para la implementación de estos métodos. Cuando f es suave, obtiene tasas que oscilan entre $O(\log n/n)^{2/(d+1)}$ y $O(\log n/n)^{1/(d+2)}$, en función de cómo se elija ϵ_n. Cuando f tiene un salto a lo largo de $\partial\{f \geq c\}$, Walther (1997) define estimadores de los conjuntos de nivel y de los de mínimo volumen empleando granulometrías, estimadores *kernel* y las técnicas del exceso de masa. Con ese estimador consigue tasas como las del exceso de masa en Polonik (1995) (salvo un factor logarítmico), pero supera con creces a este último en el aspecto computacional. Precisamente una de las grandes ventajas de los estimadores de Walther (1997), también compartida por el estimador (1.1), es que quedan completamente definidos mediante una lista de centros y de radios. En esta misma línea Walther (1999) analiza el algoritmo de rodamiento, que Sternberg (1986) propuso con objeto de filtrar y suavizar una imagen en escala de grises (*grey-scale image*).

Con objeto de estimar el número de c-conglomerados en el sentido precisado por Hartigan (1975), Cuevas, Febrero y Fraiman (2000) definen un estimador de $\{f > c\}$ parecido al de Walther (1997). Se trata de

$$\tilde{S}_n = \bigcup_{X_i \in \mathcal{X}_n^+} B(X_i, \epsilon_n), \tag{1.5}$$

donde $\mathcal{X}_n^+$ se define igual que en (1.4) y ϵ_n vuelve a ser un parámetro de suavizado. El estimador del número de c-conglomerados es precisamente el número de com-

ponentes conexas de $\tilde{S}_n$. El hecho de que $\tilde{S}_n$ sea una unión finita de bolas facilita la ejecución de un algoritmo que evalúe el número de sus componentes conexas. Si sólo se desea estimar el número de conglomerados se puede tomar ϵ_n constante. Si, por el contrario, lo que interesa es aproximar los conjuntos de nivel de f, para que haya consistencia el parámetro ϵ_n debería verificar la clásica condición (1.2). Bajo condiciones geométricas y de regularidad los autores prueban la consistencia casi segura en medida del estimador (1.5). También sugieren el empleo de *bootstrap* suavizado para estudiar la variabilidad del procedimiento.

En las últimas referencias mencionadas se comprueba una vez más que los métodos de estimación funcional no paramétrica resultan muy útiles en los problemas de estimación de conjuntos.

1.2.4 Aplicaciones

La estimación de conjuntos se puede utilizar como herramienta auxiliar en un procedimiento para detectar posibles comportamientos anómalos de un mecanismo o sistema de producción. Dadas $X_1, \ldots, X_n$ observaciones de una misma población, queremos decidir si una nueva observación Z es anómala, es decir, si no tiene la misma distribución de probabilidad P_X que las X_i. Una sencilla regla de decisión consistiría en calcular S_n, un estimador del soporte S de P_X, y si $Z \notin S_n$ decidir que Z no tiene distribución P_X. Éste es un procedimiento no paramétrico de resolver el problema, ampliamente estudiado en la literatura, de **detección** de un ***change-point*** (ver Carlstein, Müller y Siegmund 1994).

Si P_X tiene una densidad f, el papel del soporte también lo puede desempeñar un conjunto de nivel $\{f > c\}$. De hecho, el uso de conjuntos de nivel puede ser interesante para evitar el efecto de posibles *outliers* y para soslayar la hipótesis de soporte compacto.

Este problema de detección también se podría enmarcar dentro del contexto mucho más amplio del **análisis de imágenes** (ver, por ejemplo, Korostelev y Tsybakov 1993).

Otro método de análisis de imágenes que también emplea la estimación de con-

juntos es la **estimación de la frontera** (*edge estimation*). Esta técnica se puede emplear como un primer paso en la recuperación de una imagen (ver Mammen y Tsybakov 1995, Bertholet, Rasson y Lissoir 1998). El estudio de las fronteras en el contexto de la estimación de imágenes se remonta a Marr (1982), pero fueron Khasminskii y Lebedev (1990) y Korostelev y Tsybakov (1993) quienes le dieron mayor relevancia. En palabras de Donoho (1999), "*in real-world image data, the most interesting aspects of the image are the edges*".

En **econometría** encontramos una aplicación más de la estimación de fronteras. La idea, presentada por Farrell (1957), es medir la eficiencia de una unidad de producción en función de su distancia a una cierta frontera de producción. Esta frontera se ha de estimar a partir de un conjunto de n unidades de producción observadas (ver Kneip, Park y Simar 1996). En el capítulo 4 volveremos sobre este problema lateral de estimar la frontera.

La metodología propuesta para la detección es fácilmente generalizable al problema del **análisis discriminante**, en el que se tienen k poblaciones, cada una de las cuales está caracterizada por un conjunto S^i, $i = 1, \ldots, k$. El objetivo es clasificar una nueva observación Z como procedente de alguna de esas poblaciones. Para ello tomamos una muestra de cada una de las poblaciones, estimamos S^i mediante $\tilde{S}^i$, $i = 1, \ldots, k$, y asociamos Z a la población i tal que $Z \in \tilde{S}^i$.

Otra aplicación interesante de la estimación de conjuntos tiene lugar en el **análisis de conglomerados** (*cluster analysis*). El principal objetivo del análisis de conglomerados es agrupar un conjunto de datos, según su afinidad, en conglomerados (*clusters*). Para una constante $c > 0$ fija, Hartigan (1975) define un c-conglomerado poblacional como una componente conexa del conjunto de nivel $\{f > c\}$. Este concepto está claramente vinculado a la noción de moda. Por un lado, Müller y Sawitzki (1991) han utilizado estimadores de los conjuntos de nivel para definir **tests de multimodalidad** (ver también Hartigan 1987; Müller 1993; Polonik 1995; Cuevas, Febrero y Fraiman 2000). Por otro lado, algunos algoritmos de construcción de conglomerados están basados en la estimación de modas (ver, por ejemplo, Silverman 1986, cap. 6). Típicamente el número de modas (o máximos locales) es mayor o igual que el de c-conglomerados.

Tsybakov (1997) sugiere un nuevo campo en el que aplicar la estimación de

conjuntos de nivel de densidades: una extensión no lineal del **análisis de componentes principales**. Propone estudiar distintos niveles c y las formas de los conjuntos de nivel correspondientes $\{f > c\}$, comparando éstas últimas entre sí y tratando de descubrir una estructura común a todas ellas. Esas siluetas de los conjutos de nivel se podrían interpretar como curvas "principales" no lineales.

Otros problemas estadísticos relacionados con la estimación del soporte se pueden encontrar en Hartigan (1987), Müller y Sawitzki (1991), Mammen y Tsybakov (1995) y Polonik (1995).

1.3 Resumen del contenido de esta memoria

1.3.1 Planteamiento e ideas básicas del capítulo 2

En el capítulo 2 tomamos como punto de partida el artículo de Devroye y Wise (1980). El planteamiento es como sigue: después de observar n vectores aleatorios independientes $X_1, \ldots, X_n$ con la misma distribución de probabilidad P_X desconocida, se toma una nueva observación X_{n+1}. Se ha de decidir si X_{n+1} proviene o no de P_X, es decir, si entre los instantes n y $n+1$ ha habido un cambio en la distribución del proceso, en cuyo caso diremos que $n+1$ es un *change-point*. Devroye y Wise motivan este problema en el contexto del control de calidad en un sistema, cadena de producción o maquinaria, lo cual presupone que P_X sería la distribución que caracteriza un nivel aceptable de calidad en ese sistema. Para indicar que se ha detectado un cambio en la distribución (y, por tanto, en la calidad) del proceso, se utiliza la expresión "el sistema está fuera de control". Se trata por tanto de detectar cambios en ese nivel aceptable de calidad. La idea de Devroye y Wise consiste en decidir que el sistema está fuera de control si $X_{n+1} \notin S$, el soporte de P_X. Como S es desconocido, se puede aproximar mediante el estimador (1.1), definido a partir de la muestra $X_1, \ldots, X_n$, y se decide que la distribución ha cambiado si $X_{n+1} \notin \hat{S}_n$. Con esta regla de decisión, dadas $X_1, \ldots, X_n$, se consideran dos tipos de error: que X_{n+1} tenga distribución con soporte S y, sin embargo,

$X_{n+1} \notin \hat{S}_n$, y el error de "no detección", es decir, que la distribución de X_{n+1} no tenga soporte S y $X_{n+1} \in \hat{S}_{n+1}$. Como ya hemos mencionado en el apartado anterior, Devroye y Wise prueban consistencia (en probabilidad y casi seguro) de la distancia en medida entre $\hat{S}_n$ y S. En términos del problema de detección, esto equivale a probar la convergencia a cero de la probabilidad total de error.

1.3.2 Resultados contenidos en el capítulo 2

Nosotros estamos interesados en lo que denominaremos el error de "falsa alarma": que la distribución de X_{n+1} sea P_X, pero $X_{n+1} \notin \hat{S}_{n+1}$. De los resultados de consistencia de Devroye y Wise se concluye automáticamente la convergencia a cero, bajo condiciones casi universales, de la probabilidad de falsa alarma

$$P_n = P(X_1, \ldots, X_n) = \mathbf{P}\{X_{n+1} \notin \hat{S}_n | X_1, \ldots, X_n\}. \tag{1.6}$$

En los Teoremas 1 y 2 del capítulo 2 obtenemos tasas para la convergencia en probabilidad y casi seguro de P_n. La prueba del Teorema 1 utiliza la metodología de Devroye y Wise, y la del Teorema 2 se basa en la desigualdad de McDiarmid. El uso de uno u otro resultado está en función del orden de convergencia del radio ϵ_n.

Al final del capítulo se discuten algunas maneras de aplicar esta metodología a casos prácticos y se muestra cómo ϵ_n puede elegirse de modo que se controle además la probabilidad de falsa alarma.

1.3.3 Planteamiento e ideas básicas del capítulo 3

El capítulo 3 está dedicado a la estimación de conjuntos de nivel $\{f > c\}$ (donde f es una densidad y c es una constante positiva) mediante el estimador *plug-in* $\{\hat{f}_n > c\}$, siendo $\hat{f}_n$ un estimador *kernel* de la densidad. Sea cual sea el contexto en el que se desarrolle este problema (control de calidad, análisis de conglomerados,...) es interesante conocer la rapidez con la que el "contenido en

probabilidad" (en distintos sentidos) del conjunto de nivel empírico $\{\hat{f}_n > c\}$ tiende a su límite poblacional.

Concretamente, dada una muestra aleatoria $X_1, \ldots, X_n$ de f, podemos decidir no clasificar (es decir, clasificar como no procedente de f) una nueva observación $Z = z$, o decidir que el proceso que estamos supervisando está fuera de control, si $\hat{f}_n(z) \leq c$.

1.3.4 Resultados del capítulo 3

Estudiaremos esa metodología de clasificación desde dos enfoques distintos. En primer lugar, sea

$$P_n(z) := \mathbf{P}\{\hat{f}_n(z) \leq c\} \tag{1.7}$$

la probabilidad de no clasificar el dato z. Estamos interesados en conocer la rapidez de convergencia de la variable aleatoria $P_n(Z)$ a $I_{\{f(Z)\leq c\}}$, cuando la nueva observación Z realmente proviene de la densidad f. En el Teorema 3.1 obtenemos tasas de convergencia de tipo L^1 para $P_n(Z)$.

En segundo lugar estudiamos el caso en que $\{f \leq c\}$ es el conjunto de nivel cuyo contenido de probabilidad verifica

$$P_f\{f \leq c\} = \int_{\{f(z)\leq c\}} f(z)dz = \alpha,$$

siendo $0 < \alpha < 1$ una constante prefijada. Por relacionarlo con los trabajos mencionados anteriormente, $\{f > c\}$ sería el conjunto de mínimo volumen (es decir, de mínima medida de Lebesgue) con un contenido de probabilidad mayor o igual que $1 - \alpha$. En este caso utilizaremos como estimador de $\{f \leq c\}$ el conjunto $\{\hat{f}_n \leq c_n\}$, donde $c_n = c_n(X_1, \ldots, X_n)$ es un valor aleatorio que satisface

$$P_{\hat{f}_n}\{\hat{f}_n \leq c_n\} = \int_{\{z:\hat{f}_n(z)\leq c_n\}} \hat{f}_n(z)dz = \alpha \tag{1.8}$$

En el Teorema 3.2 se obtienen tasas para la convergencia casi segura a α de la sucesión

$$P_f\{\hat{f}_n \leq c_n\} = \int_{\{z:\hat{f}_n(z)\leq c_n\}} f(z)dz. \tag{1.9}$$

Observemos que, en este caso, la probabilidad se evalúa con respecto a la nueva observación Z, de tal manera que (1.9) es una cantidad aleatoria que depende de $X_1, \ldots, X_n$. Aunque la metodología matemática empleada en los dos planteamientos es muy distinta, a grandes rasgos las conclusiones coinciden, en el sentido de que las tasas de convergencia son como mucho de orden $n^{-1/(d+2)}$. En el caso univariante $d = 1$ estas tasas están en la línea de los clásicos resultados *cube-root* ($n^{-1/3}$) que aparecen en algunos contextos no paramétricos (ver, por ejemplo, Kim y Pollard 1990). Como corolario al Teorema 3.2 se obtienen tasas para la convergencia casi segura de c_n a c.

Este segundo enfoque tiene pleno sentido desde la perspectiva del control no paramétrico de la calidad. $X_1, \ldots, X_n$ representaría una muestra piloto (*training sample*) suficientemente grande, formada por v.a.i.i.d. tomadas de una densidad desconocida f en $\mathbb{R}^d$. El objetivo sería decidir si una nueva observación Z proviene también de f. Una posible regla de decisión podría ser aceptar que ha tenido lugar un cambio en la distribución del proceso si $\hat{f}_n(Z) \leq c_n$. La razón por la que sugerimos tomar c_n en lugar de c en la regla de decisión es que, si la probabilidad deseada de falsa alarma (que es la probabilidad de decidir erróneamente que se ha producido un cambio en la distribución) es α, el hecho de elegir c_n como solución de (1.8) hace que la probabilidad "real" de falsa alarma $P_f\{\hat{f}_n \leq c_n\}$ tienda a ese valor ideal α.

1.3.5 El contenido del capítulo 4

La idea central de este capítulo consiste en que, si se tiene alguna información previa acerca de la forma (conexo, estrellado,...) del conjunto S, el estimador de Devroye y Wise, a pesar de su sencillez, puede utilizarse para incorporar esa misma restricción de tipo geométrico. Para ello, se elige el parámetro de suavizado ϵ_n de manera que $\hat{S}_n(\epsilon_n)$ verifique la misma hipótesis de forma que S.

En la Sección 4.2 suponemos que S es un conjunto conexo y elegimos ϵ_n como el ínfimo $\bar{\epsilon}_n$ de los radios ϵ que hacen que $\hat{S}_n(\epsilon) := \bigcup_{i=1}^n B(X_i, \epsilon)$ sea conexo. Entonces $\bar{\epsilon}_n$ es simplemente la mitad de lo que Appel y Russo (1996) llaman la distancia

de conexión (*connectivity distance*) del conjunto $\{X_1, \ldots, X_n\}$. Esta distancia coincide con la longitud M_n del mayor lado del mínimo árbol abarcador definido sobre $\{X_1, \ldots, X_n\}$. Utilizando resultados de Tabakis (1996) y de Penrose (1999) acerca de M_n, probamos consistencia en medida de $\hat{S}_n(\bar{\epsilon}_n)$.

La restricción de forma que se considera en la sección 4.3 es que S sea estrellado. Se toma el ínfimo ϵ_n^* de los radios ϵ_n que hacen que $\hat{S}_n(\epsilon_n)$ tenga forma estrellada. En el Teorema 4.2, bajo una restricción de forma que impide que S tenga infinitos rayos cada vez más estrechos, se prueba que ϵ_n^* converge a cero casi seguro. El Teorema 4.3 es un resultado de consistencia "reforzada": si $\hat{S}_n(\epsilon_n)$ es estrellado y el orden de convergencia a cero del parámetro de suavizado ϵ_n es suficientemente lento, entonces, con probabilidad uno, la frontera del estimador $\hat{S}_n(\epsilon_n)$ converge hacia la frontera de S en distancia de Hausdorff. El calificativo de "reforzada" alude al hecho de que la convergencia de Hausdorff de $\hat{S}_n$ a S no siempre implica la convergencia de las fronteras. Esta última convergencia puede interpretarse, por tanto, como una propiedad complementaria de estimación.

1.4 Notación y resultados auxiliares

Se considera que el espacio $\mathbb{R}^d$ está dotado del producto escalar usual y de la norma euclídea. Para cada conjunto A denotaremos por A^c, $\mathrm{cl}(A)$, $\mathrm{int}(A)$ y ∂A el complementario, la clausura o adherencia, el interior y la frontera de A respectivamente. De ahora en adelante $B(a,r)$ denotará la bola cerrada de centro a y radio r, B denotará $B(0,1)$ y B_d será la medida de Lebesgue de B en $\mathbb{R}^d$.

Dados dos conjuntos $A, C \subset \mathbb{R}^d$, la suma de Minkowski de A y C es $A \oplus C := \{a + c : a \in A, c \in C\}$. Dado $\epsilon > 0$, $A^\epsilon := \bigcup_{x \in A} B(x, \epsilon)$ es el conjunto A "orlado exteriormente" por una banda de anchura ϵ. Obsérvese que $A \oplus \epsilon B = A^\epsilon$.

Se define soporte de una función $f : \mathbb{R}^d \to \mathbb{R}$ como el conjunto cerrado $\mathrm{cl}\{x \in \mathbb{R}^d : f(x) > 0\}$. El ínfimo esencial de f en un conjunto $A \subset \mathbb{R}^d$ es $\operatorname{ess\,inf}_A f := \sup\{a \geq 0 : \mu_L\{x : f(x) < a\} = 0\}$. Por último, dado un conjunto A, denotaremos

por I_A la función indicatriz de A

$$I_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A. \end{cases}$$

Funciones de densidad, variables aleatorias, probabilidades,...

Supondremos que todas las variables aleatorias están definidas en el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbf{P})$ y toman valores en $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$, donde $d \in \mathbb{N}$ y $\mathcal{B}_{\mathbb{R}^d}$ denota la σ-álgebra de Borel en $\mathbb{R}^d$. Dada una variable aleatoria X, denotaremos su distribución de probabilidad por P_X. Si X es absolutamente continua con densidad de probabilidad f, denotaremos su distribución también por P_f. Definimos el soporte de la distribución P_X como el menor subconjunto cerrado de $\mathbb{R}^d$ que tiene probabilidad P_X igual a uno.

Distancias entre conjuntos

Antes hemos mencionado la necesidad de definir un criterio adecuado de proximidad entre conjuntos, de manera que podamos evaluar el comportamiento de los estimadores. Consideramos dos tipos de distancias entre conjuntos: la **distancia en medida** d_μ y la **métrica de Hausdorff** d_H, que juegan un papel análogo al de las métricas L^1 y L^∞, respectivamente, en la estimación de densidades (ver Korostelev y Tsybakov 1993). De hecho, en la literatura de estimación de conjuntos es frecuente que las distancias entre conjuntos aparezcan relacionadas con métricas funcionales.

Dados dos conjuntos $S, T \subset \mathbb{R}^d$, se define su distancia en medida

$$d_\mu(S,T) := \mu(S \Delta T), \tag{1.10}$$

donde Δ denota la diferencia simétrica entre conjuntos, $S \Delta T = (S \backslash T) \cup (T \backslash S)$, y μ es una medida en $\mathcal{B}_{\mathbb{R}^d}$. Por ejemplo, podríamos tomar $\mu = \mu_L$, la medida de Lebesgue o, cuando T sea un estimador de S, otra opción es elegir $\mu = P_X$, la distribución de cada X_i.

Por otro lado,

$$d_H(S,T) := \inf\{\epsilon > 0 : S \subset T^\epsilon, T \subset S^\epsilon\} \tag{1.11}$$

define la distancia de Hausdorff entre S y T. En términos intuitivos, la definición de d_H está basada en una idea de "proximidad física" entre los conjuntos. De hecho, es fácil comprobar que d_H puede expresarse de manera alternativa mediante

$$d_H(S,T) = \max\{\sup_{x\in S}\inf_{y\in T}\|x-y\|, \sup_{x\in T}\inf_{y\in S}\|x-y\|\} \tag{1.12}$$

En Schneider (1993) se puede encontrar un estudio en profundidad de las propiedades de la distancia de Hausdorff. Molchanov (1993b) y Stoyan (1998) utilizan la distancia de Hausdorff en teoría de conjuntos aleatorios. Cuevas y Fraiman (1997) y Walther (1997) trabajan con ambas métricas en el contexto de la estimación de conjuntos. Sin embargo, una cuestión todavía pendiente es un estudio sistemático de la relación entre d_μ y d_H.

Condiciones de forma

En estimación no paramétrica de densidades ocasionalmente se conoce alguna propiedad "de forma" acerca de la densidad f, por ejemplo, el número de modas o de puntos de inflexión. En este caso, una idea natural consiste en incorporar esta información en el estimador no paramétrico de la densidad a través del parámetro de suavizado (ver Cuevas y González-Manteiga 1991). Deseamos plantear el mismo esquema en el contexto de la estimación no paramétrica de conjuntos (ver Sección 1.3 y capítulo 4). Por esta razón definiremos algunas restricciones de forma, distintas de las clásicas como la conexión o la convexidad.

Un concepto más general que el de conjunto **convexo** es el de estrellado. Diremos que el conjunto S es **estrellado** (*star-shaped*) si existe al menos un $x \in S$ (que Toranzos 1979 denomina "punto radiante" de S) tal que, para todo $y \in S$, el segmento que une x con y está contenido en S (ver también Dharmadikari y Joag-Dev 1988, Breen 1992). Una terminología bastante extendida resume esta condición diciendo que "x ve a y vía S". Denominaremos **mirador** (núcleo, *kernel*) de S al

conjunto de tales puntos x y lo denotaremos por $\mathrm{mir}(S)$. En Gardner (1995) se puede encontrar una interesante definición, más general, de conjunto estrellado.

Una condición de regularidad sobre la forma de un conjunto que utilizaremos reiteradamente es la de "estandaridad" (*standardness*). Se dice que un conjunto acotado $S \subset \mathbb{R}^d$ es **estándar** con respecto a una medida μ si, para todo $\lambda > 0$, existe $\delta \in (0,1)$ tal que

$$\mu(B(x,\epsilon) \cap S) \geq \delta \mu_L(B(x,\epsilon)), \quad \forall x \in S,\ 0 < \epsilon \leq \lambda.$$

Si S es el soporte (compacto) de una densidad f, esta hipótesis generalmente se verificará cuando f esté acotada inferiormente por una constante positiva en S y este conjunto tenga una estructura regular que excluya la existencia de picos arbitrariamente agudos.

La condición de estandaridad es muy parecida a la *cone condition* de Korostelev y Tsybakov (1993). De hecho, ambas se utilizan con el objeto de analizar la convergencia en medida del estimador (1.1).

Al presentar los resultados minimax en estimación de conjuntos, se ha explicado de manera informal en qué consiste la hipótesis de **nitidez**. Hay varias maneras posibles de formalizar esta idea. Por ejemplo, en Cuevas y Fraiman (1997) se utiliza la siguiente definición: dada $f : \mathbb{R}^d \to [0,\infty)$ función de soporte S compacto, se dice que $\gamma > 0$ es el orden de nitidez del soporte si $\mu_L(\{f < t\} \cap S)$ es del mismo orden (cuando $t \to 0^+$) que t^γ, es decir, si

$$0 < \liminf_{t \to 0^+} \frac{\mu_L(\{f < t\} \cap S)}{t^\gamma} \leq \limsup_{t \to 0^+} \frac{\mu_L(\{f < t\} \cap S)}{t^\gamma} = C$$

para alguna constante finita C. Se pueden encontrar otras definiciones alternativas que responden a la misma idea en Hall (1982) y Härdle, Park y Tsybakov (1995).

Dado un conjunto finito $C = \{x_1, \ldots, x_n\}$ en $\mathbb{R}^d$, llamaremos **mínimo árbol abarcador** (*minimal spanning tree*) sobre C al grafo conexo que toma como vértices los elementos x_i de C y, como lados, segmentos $[x_i, x_j]$ que unan entre sí los puntos de C de tal manera que la longitud total del grafo sea mínima (ver Penrose 1999).

Estimadores de la densidad

Como ya hemos mencionado previamente, el estimador *kernel* de la densidad es una importante herramienta auxiliar en la estimación de conjuntos. Dada una muestra aleatoria $X_1, \ldots, X_n$ de la densidad de probabilidad f en $\mathbb{R}^d$, se define un estimador *kernel* de f mediante

$$\hat{f}_n(x) := \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \qquad x \in \mathbb{R}^d, \tag{1.13}$$

donde K es una función denominada núcleo (*kernel*), $K_h(\cdot) := h^{-d} K(\cdot/h)$ y $h = h_n > 0$ es una sucesión de parámetros de suavizado (también llamados de "amplitud de ventana"). El parámetro de suavizado h y el núcleo K se eligen en función de las propiedades que deseemos cumpla $\hat{f}_n$ (ver, por ejemplo, Wand y Jones 1995, Simonoff 1996). De ahora en adelante supondremos que K es una función de densidad, para que $\hat{f}_n$ también lo sea.

En lo que sigue se consideran dos distancias para evaluar la calidad del estimador $\hat{f}_n$: la distancia L^1

$$\|f - \hat{f}_n\|_1 = \int |f - \hat{f}_n|$$

y la distancia L^∞

$$\begin{aligned} \|f - \hat{f}_n\|_\infty &= \text{ess sup}|f - \hat{f}_n| \\ &= \inf\{a \geq 0 : \mu_L\{x : |f(x) - \hat{f}_n(x)| > a\} = 0\} \end{aligned}$$

(ver, por ejemplo, Silverman 1986, Devroye 1987). En particular, cuando f y $\hat{f}_n$ son continuas, $d_\infty(f, \hat{f}_n) = \sup_x |f(x) - \hat{f}_n(x)|$.

Devroye (1983) prueba la equivalencia entre la consistencia en L^1 del estimador *kernel*

$$\|f - \hat{f}_n\|_1 \to 0 \quad \text{c.s.} \quad \forall f$$

y la siguiente condición

$$h_n \to 0 \quad \text{y} \quad n h_n^d \to \infty \quad \text{cuando } n \to \infty. \tag{1.14}$$

Además de este resultado universal, se puede probar, bajo hipótesis más restrictivas, que (1.14) es condición necesaria para la consistencia de $\hat{f}_n$ en L^2 (ver Silverman 1986). Por esa razón de ahora en adelante supondremos que el parámetro de suavizado verifica (1.14). En Prakasa Rao (1983) y Nadaraya (1989) se pueden encontrar resultados de consistencia de $\hat{f}_n$ en L^∞.

Algunas desigualdades

Por último, y con el único objeto de que la memoria sea en cierto sentido "autocontenida", enunciaremos algunos resultados clásicos, que serán utilizados en los restantes capítulos.

DESIGUALDAD DE BERRY-ESSEEN. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas (i.i.d.). Supongamos que*

$$EX_1 = 0, \quad EX_1^2 = \sigma^2 > 0, \quad E|X_1|^3 < \infty.$$

Entonces

$$\sup_x \left| \mathbf{P}\left\{ \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} X_j < x \right\} - \Phi(x) \right| \leq A \frac{E|X_1|^3}{\sigma^3 \sqrt{n}},$$

donde A es una constante que satisface $(2\pi)^{-1/2} \leq A < 0.8$ y Φ denota la función de distribución de la normal (0,1).

DESIGUALDAD DE HOEFFDING. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes con media cero y rangos acotados: $a_i \leq X_i \leq b_i$. Para cada $\epsilon > 0$,*

$$\mathbf{P}\left\{ \sum_{i=1}^{n} X_i \geq \epsilon \right\} \leq \exp\left(- \frac{2\epsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

DESIGUALDAD DE MCDIARMID. *(McDiarmid 1989) Sean $X_1, \ldots, X_n$ variables aleatorias independientes que toman valores en un conjunto A. Supongamos que*

$g : A^n \to \mathbb{R}$ *satisface*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_1',\ldots,x_n' \in A}} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i, \qquad (1.15)$$

para $1 \leq i \leq n$. *Entonces*

$$\mathbf{P}\{|g(X_1,\ldots,X_n) - Eg(X_1,\ldots,X_n)| \geq t\} \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

En Devroye (1991) se puede encontrar una breve demostración de esta desigualdad exponencial y algunas aplicaciones de la misma dentro del área de la estimación funcional no paramétrica.

El siguiente resultado proporciona una aproximación de la cola de la normal (0,1), $1 - \Phi(x)$, para valores grandes de x (ver, por ejemplo, Feller 1968).

COMPORTAMIENTO DE LA DISTRIBUCIÓN NORMAL. *Cuando* $x \to \infty$

$$1 - \Phi(x) \sim \frac{\phi(x)}{x},$$

donde ϕ *denota la función de densidad de la normal (0,1). Concretamente, para todo* $x > 0$, *se verifica la siguiente desigualdad*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\phi(x) < 1 - \Phi(x) < \frac{\phi(x)}{x}. \qquad (1.16)$$

A menos que indiquemos lo contrario $\to$ indicará de ahora en adelante convergencia cuando $n \to \infty$.

Capítulo 2

Estimación del soporte y detección no paramétrica

El objetivo de este capítulo es profundizar en la metodología de detección no paramétrica propuesta por Devroye y Wise (1980). Más concretamente, estamos interesados en analizar el comportamiento de la probabilidad de falsa alarma en ese procedimiento de detección. La Sección 2 está dedicada a la obtención de tasas de convergencia de dicha probabilidad a cero. Se demuestran dos teoremas. El primero de ellos se basa en acotaciones que requieren la hipótesis de que el conjunto sea estándar. En el segundo teorema la herramienta básica es la desigualdad de McDiarmid.

En la Sección 3 se sugieren procedimientos para seleccionar el parámetro de suavizado ϵ_n del estimador de Devroye y Wise en situaciones prácticas. Se proponen dos maneras de elegir el radio ϵ_n, automáticamente a partir de los datos, cuando se desea controlar la probabilidad de falsa alarma. Ambos procedimientos están basados en ideas de remuestreo (validación cruzada y *bootstrap* suavizado) y se estudian con mayor detalle junto con otros aspectos prácticos en Baíllo, Cuevas y Justel (2000).

Finalmente, en la última sección, se presenta una relación de problemas que, a corto plazo, pueden orientar la investigación futura.

2.1 El estimador de cubrimiento en el problema de detección

2.1.1 Detección no paramétrica. El método de Devroye y Wise

Supongamos que, de una densidad desconocida f en $\mathbb{R}^d$ de soporte S compacto, tomamos una muestra de observaciones i.i.d. $X_1, \ldots, X_n$. Queremos decidir si una nueva observación X_{n+1} procede o no de f.

Se trata de un problema de contraste de hipótesis no paramétrico, que se puede motivar en el marco del control estadístico de la calidad. En este contexto el objetivo consistiría en decidir si el sistema está fuera de control en la etapa $n+1$, en el sentido de que la distribución de X_{n+1} es diferente de la de las observaciones previas $X_1, \ldots, X_n$.

En este capítulo profundizaremos en la idea propuesta por Devroye y Wise (1980) para abordar este problema. El mecanismo de decisión hace uso de la estimación de conjuntos: dado un estimador del soporte del tipo

$$\hat{S}_n = \hat{S}_n(\epsilon_n) = \bigcup_{i=1}^{n} B(X_i, \epsilon_n), \tag{2.1}$$

decidiremos que en la etapa $n+1$ la distribución del proceso ha cambiado si

$$X_{n+1} \notin \hat{S}_n. \tag{2.2}$$

En lo sucesivo denominaremos a (2.1) estimador de "cubrimiento" o de Devroye y Wise. El primer nombre se debe a que $\{B(X_i, \epsilon_n)\}_{i=1}^n$ es un ejemplo clásico entre los procesos llamados "de cubrimiento" (ver Hall 1988).

En (2.2) el parámetro de suavizado ϵ_n se puede fijar de antemano, de manera que sólo dependa de n (como sucedía en las condiciones suficientes para que $d_\mu(\hat{S}_n, S)$ convergiera a cero), o se puede elegir con objeto de que el estimador $\hat{S}_n$ verifique ciertas restricciones de forma (ver capítulo 4), o bien con objeto de controlar la probabilidad de falsa alarma (ver Baíllo, Cuevas y Justel 2000). En

estas dos últimas situaciones el radio ϵ_n dependería de la muestra $X_1, \ldots, X_n$, no sólo de n.

Este último planteamiento tiene reminiscencias de la clásica metodología de Shewhart (ver, por ejemplo, Montgomery 1985 y Derman y Ross 1997), que está basada en regiones de tolerancia (a partir de las cuales se definen los acostumbrados gráficos de control, *control charts*), construidas con el fin de controlar la probabilidad de falsa alarma. La diferencia fundamental estriba en que el método de Devroye y Wise es **multivariado** y **no paramétrico**.

Un enfoque alternativo (también multivariado y no paramétrico) lo sugiere Liu (1995). La idea consiste en ordenar las observaciones multivariadas en función de su "profundidad simplicial", que es una posible medida de lo lejos que está un punto respecto a la observación más "centrada" o "profunda" de la muestra (para una revisión de las definiciones de profundidad y sus aplicaciones ver también Chakraborty y Chaudhuri 1999, Fraiman y Meloche 1999). De esta manera el problema se reduce a la construcción de un gráfico de control univariado en el que se vayan marcando las profundidades de los puntos que se observan.

El problema del control estadístico de la calidad se puede considerar en un contexto secuencial. Supongamos que observamos una sucesión de variables aleatorias i.i.d. $X_1, X_2, \ldots$ cuya distribución cambia en un instante desconocido m. Consideramos el problema de estimar el *change-point* m. Este planteamiento sugiere el estudio de algunas propiedades secuenciales, como el *average run length* (ARL), que es el número esperado de observaciones que se han de tomar desde que se produce el cambio (es decir, desde m) hasta que se detecta. Se han estudiado algunos métodos de control de calidad desde el punto de vista del ARL y se han obtenido resultados de optimalidad. Es el caso de los procedimientos de detección CUSUM (*CUmulative SUM*) y de Shiryaev-Roberts (ver, por ejemplo, Moustakides 1986, Pollak y Siegmund 1991 y Yakir 1998). El estudio de estas propiedades secuenciales requiere generalmente que la densidad f anterior al cambio (cuando el proceso está bajo control) sea conocida, totalmente o salvo un parámetro. Por esta razón no las consideraremos aquí.

2.1.2 Por qué utilizar el estimador de cubrimiento

Aunque, en principio, se podrían utilizar otros estimadores del soporte en la regla de detección expresada por (2.2), aquí nos centraremos exclusivamente en el estudio del estimador de Devroye y Wise, que Korostelev y Tsybakov (1993) califican como "*simple and rough*". En efecto, tal y como hemos visto en el capítulo 1, se podrían escoger otros estimadores más sofisticados del soporte. Sin embargo, hay al menos tres razones por las que pensamos que es conveniente el uso del estimador de cubrimiento en este problema:

1. La puesta en práctica de las técnicas de estimación de conjuntos multivariados plantea dificultades computacionales (tanto mayores cuanto mayor es la dimensión del espacio), frente a las cuales la relativa sencillez del estimador de cubrimiento aparece como ventaja importante. En cierto sentido, se trata de una situación análoga a la de la estimación de densidades: en el caso multivariado se suelen preferir el histograma o los estimadores de vecinos más próximos (ver, por ejemplo, Scott 1992), aunque el método *kernel* presenta, en ciertos aspectos importantes, mejores propiedades.

2. Como veremos en el capítulo 4, la estructura del estimador de Devroye y Wise permite incorporar sin dificultad algunas restricciones de forma (por ejemplo, que el estimador $\hat{S}_n$ sea conexo). Esto es debido principalmente a que el parámetro ϵ_n tiene una interpretación intuitiva muy directa como factor de dilatación de la muestra.

3. Entender las propiedades básicas del estimador de Devroye y Wise puede servir como paso previo al análisis de ciertos estimadores más sofisticados que, en realidad, guardan muchos puntos en común con el sencillo $\hat{S}_n$ (ver capítulo 3, Cuevas y Fraiman 1997, Walther 1997).

2.2 Tasas de convergencia de la probabilidad de falsa alarma

En el marco del problema de detección planteado en la sección anterior, y empleando la regla de decisión expresada en (2.2), Devroye y Wise (1980) probaron la convergencia a cero de la probabilidad global de cometer un error. Nosotros analizaremos el comportamiento asintótico de la probabilidad de falsa alarma, esto es

$$P_n = P_n(X_1, \ldots, X_n) = \mathbf{P}\{X_{n+1} \notin \hat{S}_n | X_1, \ldots, X_n\}. \tag{2.3}$$

En particular, estamos interesados en obtener resultados del tipo $a_n P_n \to 0$ (en probabilidad o casi seguro) para sucesiones $a_n \uparrow \infty$.

Los dos teoremas que enunciamos a continuación proporcionan tasas de convergencia para la probabilidad de falsa alarma. Ambos resultados son, en cierto sentido, complementarios ya que las restricciones sobre el orden de convergencia del radio ϵ_n son distintas.

TEOREMA 2.1. *Sea $X_1, X_2, \ldots$ una sucesión de vectores aleatorios i.i.d. en $\mathbb{R}^d$ con distribución común P_X, absolutamente continua respecto a la medida de Lebesgue μ_L. Sea S el soporte de P_X y P_n la probabilidad de falsa alarma definida en (2.3). Supongamos que*

i) f, la densidad de P_X respecto a μ_L, está acotada por una constante $M_f > 0$,

ii) S es compacto y estándar con respecto a P_X, con constante de estandaridad δ,

iii) $\epsilon_n \to 0$ y $n\epsilon_n^d \to \infty$.

Entonces

$$a_n P_n \longrightarrow 0, \quad \text{en probabilidad,}$$

donde a_n es una sucesión tal que $a_n = o\left(\exp\{cn\epsilon_n^d\}\right)$, donde $c = \delta B_d 2^{-d}$.

Si, además, $n\epsilon_n^d / \log n \to \infty$, *entonces se da la convergencia completa*

$$\sum_{n=1}^{\infty} \mathbf{P}\{nP_n > \epsilon\} < \infty, \quad \textit{para todo } \epsilon > 0,$$

(y por tanto tambien se verifica que $nP_n \to 0$ *c.s.).*

La demostración es consecuencia directa de la desigualdad de Markov y del siguiente lema:

LEMA 2.1. *Bajo las mismas condiciones del Teorema 2.1,*

$$E(P_n) = O(\exp\{-cn\epsilon_n^d\}), \tag{2.4}$$

siendo $c = \delta B_d 2^{-d}$.

DEMOSTRACIÓN. Para cada n, tomemos un cubrimiento minimal de S formado por bolas $B_j := B(Z_j, \epsilon_n/2)$ (con $j = 1, \ldots, R_n$) centradas en puntos $Z_j \in S$. Tenemos

$$\begin{aligned} E(P_n) &= E\left(P_X\left(\hat{S}_n^c \middle| X_1, \ldots, X_n\right)\right) \\ &= E\left(P_X\left(\hat{S}_n^c \cap \left(\cup_{j=1}^{R_n} B_j\right) \middle| X_1, \ldots, X_n\right)\right) \\ &\leq \sum_{j=1}^{R_n} E\left(P_X\left(\hat{S}_n^c \cap B_j \middle| X_1, \ldots, X_n\right)\right) \end{aligned} \tag{2.5}$$

Definimos

$$A_{n,j} := \left\{\omega : \sum_{i=1}^{n} I_{B_j}(X_i(\omega)) > 0\right\}.$$

Se verifica que

$$A_{n,j} \subset \left\{\omega : \hat{S}_n^c(\omega) \cap B_j = \emptyset\right\}.$$

Por tanto, si X es un vector aleatorio con distribución P_X e independiente de $X_1, \ldots, X_n$, tenemos que $X^{-1}(\hat{S}_n^c \cap B_j) \subseteq A_{n,j}^c \cap X^{-1}(B_j)$ y

$$\begin{aligned} E(P_X(\hat{S}_n^c \cap B_j | X_1, \ldots, X_n)) &= E(\mathbf{P}(X^{-1}(\hat{S}_n^c \cap B_j) | X_1, \ldots, X_n)) \\ &\leq E(\mathbf{P}(A_{n,j}^c \cap X^{-1}(B_j) | X_1, \ldots, X_n)). \end{aligned}$$

Utilizando esta cota en (2.5) obtenemos

$$\begin{aligned} E(P_n) &\leq \sum_{j=1}^{R_n} E(\mathbf{P}(A_{n,j}^c \cap X^{-1}(B_j)|X_1,\ldots,X_n)) \\ &= \sum_{j=1}^{R_n} E(E(I_{A_{n,j}^c} I_{X^{-1}(B_j)}|X_1,\ldots,X_n)) \\ &= \sum_{j=1}^{R_n} E(I_{A_{n,j}^c} E(I_{X^{-1}(B_j)}|X_1,\ldots,X_n)) \\ &= \sum_{j=1}^{R_n} P_X(B_j) E(I_{A_{n,j}^c}) \\ &= \sum_{j=1}^{R_n} P_X(B_j) \left(\mathbf{P}\left\{I_{B_j}(X)=0\right\}\right)^n \\ &= \sum_{j=1}^{R_n} P_X(B_j) \left(1-P_X(B_j)\right)^n \end{aligned}$$

Ahora aplicamos la desigualdad $(1-x)^n \leq e^{-nx}$, válida para $0 \leq x \leq 1$, y obtenemos

$$E(P_n) \leq \sum_{j=1}^{R_n} P_X(B_j) \exp\{-nP_X(B_j)\}. \tag{2.6}$$

Para obtener una cota superior de (2.6) observemos que, como P_X tiene una densidad acotada, $P_X(B_j) \leq M_f B_d \epsilon_n^d$. Por otro lado, de la hipótesis de estandaridad, deducimos que $P_X(B_j) \geq 2^{-d}\delta B_d \epsilon_n^d$. Así pues, (2.6) implica que

$$E(P_n) \leq R_n M_f B_d \epsilon_n^d \exp\{-n2^{-d}\delta B_d \epsilon_n^d\}. \tag{2.7}$$

Finalmente, como $R_n \leq C_2 \epsilon_n^{-d}$, para algún $C_2 > 0$ (dependiente de S), (2.4) se sigue directamente de (2.7).

□

El Teorema 2.1 demuestra que, como era de prever, se pueden conseguir tasas de convergencia tanto más rápidas cuanto mayores sean los valores de ϵ_n. Para conseguir una tasa casi segura $a_n = n$, es suficiente con que ϵ_n converja a cero más

lentamente que $(\log n/n)^{1/d}$. Obviamente, si ϵ_n es demasiado grande, el estimador $\hat{S}_n$ puede ser excesivamente conservador y, por tanto, ineficaz como herramienta de detección. En este sentido, el Teorema 2.2 proporciona información complementaria ya que se aplica a valores de ϵ_n más pequeños. En particular, es este segundo teorema (y no el primero) el que incluye a los ϵ_n cuyo orden exacto, $O(\log n/n)^{1/d}$, coincide con el radio del "cierre conexo" estudiado en el capítulo 4.

TEOREMA 2.2. *Bajo las mismas condiciones y notación del Teorema 2.1, si* $a_n = o(\exp\{cn\epsilon_n^d\})$ *y* $\sum_{n=1}^{\infty} \exp\{-\epsilon^2/na_n^2\epsilon_n^{2d}\} < \infty$ *para todo* $\epsilon > 0$, *entonces*

$$\sum_{n=1}^{\infty} \mathbf{P}\{a_n P_n > \epsilon\} < \infty, \quad \textit{para todo } \epsilon > 0.$$

En particular, si ϵ_n *es de orden exacto* $(\log n/n)^{1/d}$, *entonces se verifica la convergencia completa* $n^\beta P_n \to 0$ *para todo* $\beta \in [0, 1/2)$.

DEMOSTRACIÓN. La herramienta fundamental es la desigualdad de McDiarmid (1989), que aplicaremos a la diferencia entre P_n y $E(P_n)$. Según la notación empleada al enunciar esta desigualdad en el capítulo 1, tomemos $g(X_1, \ldots, X_n) = P_n$. La hipótesis (1.14) de la desigualdad se verifica con

$$c_i = \max\{P_X(B(x_i, \epsilon_n)), P_X(B(x_i', \epsilon_n))\}$$

y $c_i \leq C\epsilon_n^d$ para alguna constante $C > 0$. Obtenemos, pues, que para todo $t > 0$,

$$\mathbf{P}\{|P_n - EP_n| \geq t\} \leq 2\exp\left\{-\frac{2t^2}{nC^2\epsilon_n^{2d}}\right\},$$

y, si $a_n = o(\exp\{cn\epsilon_n^d\})$,

$$P\{a_n|P_n - EP_n| \geq \epsilon\} \leq 2\exp\left\{-\frac{2\epsilon^2}{na_n^2C^2\epsilon_n^{2d}}\right\}, \tag{2.8}$$

para todo $\epsilon > 0$. Por tanto,

$$\sum_{n=1}^{\infty} \mathbf{P}\{a_n|P_n - EP_n| \geq \epsilon\} < \infty \tag{2.9}$$

siempre que

$$\sum_{n=1}^{\infty} \exp\left\{-\frac{2\epsilon^2}{na_n^2 C^2 \epsilon_n^{2d}}\right\} < \infty \tag{2.10}$$

y esto último se cumple por hipótesis. Obsérvese que una condición suficiente para que se dé la convergencia (2.10) es

$$na_n^2 \epsilon_n^{2d} \log n \to 0.$$

Finalmente, por el Teorema 2.1, $a_n E(P_n) \to 0$. Esto y (2.9) implican la convergencia completa (y c.s.) $a_n P_n \to 0$.

□

Observación. La condición de estandaridad en el Teorema 2.1 se puede debilitar suponiendo simplemente que, para todo $\epsilon > 0$, existe algún $\delta(\epsilon) > 0$ tal que $n\epsilon_n^d \delta(\epsilon_n) \to \infty$ y

$$P_X(B(x,\epsilon) \cap S) \geq \delta(\epsilon)\mu_L(B(x,\epsilon)), \quad \forall x \in S. \tag{2.11}$$

Cuando $\epsilon \to 0$, típicamente $\delta(\epsilon)$ decrecerá a cero a la misma velocidad a la que lo hace f. Por ejemplo, para la densidad triangular, $\delta(\epsilon) = \epsilon$. Esto afecta a la tasa de convergencia de $E(P_n)$ ya que, suponiendo (2.11),

$$E(P_n) \leq C \exp(-n\delta(\epsilon_n) B_d \epsilon_n^d).$$

En otras palabras, cuando más lentamente decrezca f en las proximidades de ∂S, más lentas serán las tasas que obtengamos para $E(P_n)$. Conclusiones de este tipo ya habían sido mencionadas en el capítulo 1, al presentar las técnicas minimax de reconstrucción de conjuntos. También aparecen en otros métodos no paramétricos de estimación de conjuntos (ver Cuevas y Fraiman 1997). La medida que estos autores utilizan para cuantificar el decrecimiento de f cerca de ∂S es el "orden de nitidez" de f (ver Sección 1.4).

2.3 El uso práctico de la estimación de conjuntos en detección

El empleo del método de detección de Devroye y Wise (1980) en una situación práctica requiere, dada una muestra piloto cualquiera, elegir de manera razonable el parámetro de suavizado ϵ_n del estimador de cubrimiento $\hat{S}_n$.

En este contexto una de las hipótesis más débiles que se pueden pedir a S es que sea conexo. Por ejemplo, si el vector aleatorio X representa las mediciones efectuadas en un proceso industrial, la no conexión de S querría decir que en realidad tienen lugar dos o más procesos en paralelo. Es muy razonable, pues, suponer que el soporte de la distribución cuando el proceso está bajo control es conexo. Bajo esta hipótesis, podríamos tomar $\epsilon_n = \bar{\epsilon}_n$, el mínimo radio que hace $\hat{S}_n$ conexo. Esto conduce a dar la alarma (es decir, a decidir que ha tenido lugar un cambio en la distribución del proceso) cuando $X_{n+1} \notin \hat{S}_n(\bar{\epsilon}_n)$ o, lo que es equivalente, cuando

$$T_n := \min_{1 \le i \le n} \frac{\|X_{n+1} - X_i\|}{\bar{\epsilon}_n} > 1. \tag{2.12}$$

En el capítulo 4 estudiaremos esta posible elección del radio en un plano más teórico.

Una de las desventajas de la regla (2.12) es que no proporciona ningún control sobre la probabilidad de falsa alarma, salvo la consistencia $P_n \to 0$. Sin embargo, (2.12) sugiere un enfoque más general: dar la alarma cuando la mínima distancia desde la nueva observación X_{n+1} a los datos de la muestra piloto sea "suficientemente grande", con respecto al mínimo radio necesario para que $\hat{S}_n$ sea conexo. Concretamente podemos fijar la probabilidad de falsa alarma α y dar la alarma cuando $T_n > c_\alpha$, donde c_α es el punto crítico de la distribución de T_n correspondiente al nivel α, $\mathbf{P}\{T_n > c_\alpha\} = \alpha$. Esta probabilidad se calcula bajo la hipótesis nula de que no ha habido cambio en la distribución (es decir, que $X_1, \ldots, X_n, X_{n+1}$ son i.i.d.). Como la distribución de T_n es desconocida, es imposible calcular de manera exacta c_α, aunque podemos conseguir una aproximación *bootstrap* de c_α de la manera siguiente:

Aproximación de c_α mediante bootstrap suavizado

1. Generamos un número elevado, B, de muestras *bootstrap* $X_1^*, \ldots, X_n^*, X_{n+1}^*$ remuestreando de los datos originales $X_1, \ldots, X_n$.

2. "Suavizamos" los X_i^* mediante $X_i^0 = X_i^* + Z_i^0$, donde los Z_i^0 son observaciones i.i.d. de una distribución uniforme en $B(0, \bar{\epsilon}_n)$. Obsérvese que estas observaciones artificiales $X_1^0, \ldots, X_n^0, X_{n+1}^0$ en realidad son una muestra obtenida mediante "*bootstrap* suavizado" y tomada de un estimador *kernel* de la densidad (basado en $X_1, \ldots, X_n$), cuyo *kernel* sea uniforme en $B(0,1)$ y cuya amplitud de banda sea $\bar{\epsilon}_n$.

3. Para la j-ésima muestra *bootstrap* $X_1^0, \ldots, X_{n+1}^0$, $(j = 1, \ldots, B)$ se calcula el estadístico $T_n^{(j)}$, dado por (2.12). Luego se emplea la distribución empírica asociada a $T_n^{(1)}, \ldots, T_n^{(B)}$ para estimar la distribución de T_n, y el percentil empírico para estimar el percentil c_α.

La regla de parada $T_n > c_\alpha$ equivale a dar la alarma siempre que $X_{n+1} \notin \hat{S}_n(c_\alpha \bar{\epsilon}_n)$. El punto crítico c_α representa, pues, una dilatación/contracción del mínimo radio para conexión.

Una segunda posibilidad para el cálculo de c_α es seleccionar $\epsilon_n = c_\alpha \bar{\epsilon}_n$ recurriendo a la metodología *leave-one-out*, conceptualmente relacionada con los procedimientos de suavizado mediante validación cruzada (*cross-validation*) (ver, por ejemplo, Wand y Jones 1995, Simonoff 1996).

Suavizado mediante validación cruzada

1. Para cada ϵ de una malla suficientemente fina, se evalúa

$$\hat{P}_n(\epsilon) = \frac{\#\{X_i \notin \hat{S}_{n,i}(\epsilon)\}}{n},$$

donde $\hat{S}_{n,i}(\epsilon)$ es el estimador (1.1) basado en la submuestra $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$.

2. Se selecciona ϵ_n como el valor de ϵ, de entre todos los de la colección, que minimiza la distancia $|\hat{P}_n(\epsilon) - \alpha|$.

Baíllo, Cuevas y Justel (2000) comparan ambos procedimientos de selección del parámetro de suavizado c_α en un estudio en el que se simulan observaciones de la distribución uniforme y de la normal. El comportamiento de ambos métodos de suavizado se evalúa comparando sus respectivas aproximaciones de c_α y sus potencias (el término "potencia" denota aquí la probabilidad de detectar un cambio en la etapa $n+1$) con el correspondiente c_α y la potencia proporcionados por el método de Monte Carlo. El c_α y la potencia calculados con este último método son "ideales", en el sentido de que utiliza la información que proporciona conocer la distribucion del proceso antes y después del cambio, mientras que los procedimientos de validación cruzada y *bootstrap* suavizado son completamente no paramétricos. Teniendo en cuenta este aspecto, los resultados son bastante positivos pues la diferencia entre la potencia aproximada y la ideal oscila, en el caso normal, entre 0.1 y 0.15 y, en el caso uniforme, no supera 0.1 (para un tamaño muestral $n = 100$). Aunque los resultados son claramente favorables al método de validación cruzada, no hay que descartar la posibilidad de que el *bootstrap* suavizado pueda mejorarse sustancialmente, modificando la amplitud de banda que se usa para generar las muestras *bootstrap*.

Una posible explicación heurística acerca de los mejores resultados de la validación cruzada frente al *bootstrap* suavizado, es que éste último funcione peor en una banda externa del soporte. La razón es que, en esta banda externa, las bolas se solapan poco. La distribución de la que se extraen las muestras artificiales tiene como soporte la unión de esas bolas y toma menor valor en la banda externa que en la zona interior del soporte. Por esta razón, el *bootstrap* tiende a producir observaciones X_i^0 abundantes en la zona interior y más dispersas en la banda exterior, y esto da lugar a valores de c_α demasiado pequeños. En el caso uniforme, el contraste entre la calidad en la estimación de la banda externa y la zona interior es todavía más acusado a medida que n aumenta. Por el contrario, en el caso normal la diferencia no tiene tanta importancia, pues la densidad que subyace no está acotada inferiormente por una constante positiva.

2.4 Problemas abiertos

En Devroye y Wise (1980), Korostelev y Tsybakov (1993) y la Sección 2 del presente capítulo la medida considerada para evaluar el error cometido al estimar S mediante $\hat{S}_n$ es d_μ. El comportamiento de la distancia de Hausdorff entre ambos conjuntos es un caso particular de los resultados de Cuevas y Fraiman (1997) (ver también Cuevas 1990). Bajo la hipótesis de que S fuera compacto y estándar y f estuviera acotada inferiormente por una constante positiva, obtuvieron tasas del tipo $o\left((n/\log n)^{1/d}\right)$, para un parámetro de suavizado ϵ_n adecuado. En este contexto, si se piensa en una asignación óptima del parámetro de suavizado, una idea natural es considerar el radio ϵ que minimiza la distancia de Hausdorff entre $\hat{S}_n$ y S.

Los resultados teóricos de la Sección 2 se refieren únicamente al caso de soporte compacto, pero los procedimientos de selección de ϵ_n, presentados en la Sección 3, tienen sentido en un contexto más general. De hecho, en Baíllo, Cuevas y Justel (2000) queda patente su buen funcionamiento en el caso normal. En el caso de soporte no compacto es razonable conjeturar que el ϵ_n, elegido tanto mediante *bootstrap* suavizado como mediante validación cruzada, para una probabilidad de falsa alarma α, digamos $\epsilon_n(\alpha)$, da lugar (al sustituirlo en (1.1)) a un estimador consistente del conjunto de nivel α, $\{f > c\}$ tal que $P_f\{f > c\} = 1 - \alpha$.

En la Sección 3 se han presentado dos métodos para la elección del radio del estimador (1.1), de tal manera que la probabilidad de falsa alarma esté acotada por una constante α prefijada. Como hemos sugerido, conviene realizar un estudio práctico de la metodología basada en el *bootstrap* suavizado. Sin embargo, también habría que analizar desde un punto de vista más teórico el funcionamiento de ambos métodos de elección de c_α.

Capítulo 3

Un estimador *plug-in* de los conjuntos de nivel. Aplicaciones

Se considera la estimación del conjunto de nivel $\{f > c\}$ de una densidad f, mediante un estimador *plug-in* $\{\hat{f}_n > c\}$, en el que $\hat{f}_n$ es un estimador *kernel* de la densidad. Este conjunto de nivel estimado puede utilizarse como herramienta de clasificación en el contexto del **análisis de conglomerados** (ver Hartigan 1975). En la Sección 3.2, utilizando la desigualdad de Berry-Esseen, se obtienen tasas de convergencia para la probabilidad de no clasificar un dato que proviene de la densidad f.

La Sección 3.3 está motivada por otra posible aplicación de los conjuntos de nivel: la **detección no paramétrica**, entendida en el mismo sentido que en el capítulo 2. Con objeto de controlar la probabilidad de falsa alarma en el procedimiento de detección se elige un cierto nivel c, dependiente de la muestra. Una vez definido c, se consiguen tasas para la convergencia de la probabilidad de falsa alarma a una constante prefijada.

En lo esencial el contenido de este capítulo está incluido en Baíllo, Cuesta-Albertos y Cuevas (1999).

3.1 Planteamiento del problema

Dada una medida de probabilidad P_f con función de densidad f, se puede considerar que el conjunto de nivel $\{f > c\}$, para un c suficientemente pequeño, es el soporte "significativo" de P_f. Esta idea se puede utilizar en distintos contextos, muchos de los cuales se encuentran indicados en el apartado del capítulo 1 dedicado a las aplicaciones de la estimación de conjuntos. Por ejemplo, Hartigan (1975) sugiere definir los conglomerados de una población, caracterizada por la densidad f, como las componentes conexas del conjunto de nivel $\{f > c\}$. Cualquier observación que no pertenezca a este conjunto quedará sin clasificar. Como f es desconocida, en el procedimiento de clasificación se podría sustituir $\{f > c\}$ por el estimador *plug-in* $\{\hat{f}_n > c\}$, siendo $\hat{f}_n$ un estimador no paramétrico de f.

Otra posibilidad es diseñar un procedimiento no paramétrico de control de calidad, en la línea de los gráficos de control de Shewhart, pero con un enfoque conjuntista parecido al que propusieron Devroye y Wise (1980): la idea consiste en decidir que el proceso está fuera de control si la nueva observación que tomemos pertenece al conjunto $\{f \leq c\}$. Al igual que ocurre en el contexto del análisis de conglomerados, habría que estimar $\{f \leq c\}$.

En ambas situaciones es interesante saber la rapidez con la que el "contenido en probabilidad" (en distintos sentidos) del conjunto de nivel empírico $\{\hat{f}_n \leq c\}$ tiende a su análogo poblacional.

Para ser más precisos, en la Sección 3.2 se considera el problema de la estimación de conjuntos de nivel en el contexto del **análisis de conglomerados**. Supongamos que $X_1, \ldots, X_n$ es una muestra aleatoria de f. Se decide (ver Cuevas, Febrero y Fraiman 2000) ignorar (no clasificar) una nueva observación $Z = z$ cuando $\hat{f}_n(z) \leq c$, donde $\hat{f}_n(z)$ $(= \hat{f}_n(X_1, \ldots, X_n; z))$ es un estimador *kernel* de f y c es una **constante**. Consideramos la probabilidad de no clasificar el dato z

$$P_n(z) := \mathbf{P}\{\hat{f}_n(z) \leq c\} = \int_{\{(x_1,\ldots,x_n):\hat{f}_n(z)\leq c\}} f(x_1)\cdots f(x_n)dx_1\cdots dx_n. \qquad (3.1)$$

En el Teorema 3.1 se obtienen tasas para la convergencia de la variable aleatoria

$P_n(Z)$ a $I_{\{f(Z)\leq c\}}$, cuando la nueva observación Z también proviene de la densidad f.

En la Sección 3.3 se considera una idea análoga bajo el punto de vista del **control no paramétrico de la calidad**. Para ello, suponemos que $X_1, \ldots, X_n$ es una muestra piloto de observaciones i.i.d. de una densidad desconocida f, que caracterizaría la situación del proceso cuando éste "funciona correctamente". El objetivo es decidir si una nueva observación Z proviene o no de f, lo que equivale a decidir si el proceso sigue bajo control o, por el contrario, ha cambiado su distribución. Para ello hemos propuesto decidir que Z no sigue la distribución dada por f cuando $\hat{f}_n(Z) \leq c$. Si deseamos que la probabilidad de falsa alarma (es decir, la probabilidad de decidir erróneamente que se ha producido un cambio en la distribución) no sea mayor que α, para un $\alpha \in (0,1)$ prefijado, se podría **sustituir la constante c por un valor aleatorio** $c_n = c_n(X_1, \ldots, X_n)$ que satisfaga

$$P_{\hat{f}_n}\{\hat{f}_n \leq c_n\} = \int_{\{z:\hat{f}_n(z)\leq c_n\}} \hat{f}_n(z)dz = \alpha. \tag{3.2}$$

En ese caso la probabilidad "real" de falsa alarma es

$$P_f\{\hat{f}_n \leq c_n\} = \int_{\{z:\hat{f}_n(z)\leq c_n\}} f(z)dz. \tag{3.3}$$

En la Sección 3.3 se obtienen tasas para la convergencia casi segura (hacia α) de esta última probabilidad. Observemos que, en este caso, la probabilidad se evalúa con respecto a la nueva observación Z, luego (3.3) es una cantidad aleatoria que depende de $X_1, \ldots, X_n$. Hacemos notar también que $P_f\{\hat{f}_n \leq c_n\}$ tiene la misma media que la variable aleatoria $P_n(Z)$.

En el marco del control de la calidad, Polansky (1999) propone una manera de llevar esta metodología a la práctica utilizando técnicas *bootstrap*.

Esta misma idea aparece también, desde un punto de vista un poco diferente, en Davies y Gather (1993), donde Z se define como α-*outlier* si $Z \in \{f \leq c\}$, eligiéndose c tal que $P_f\{f \leq c\} = \alpha$. Aunque estos autores trabajan en un contexto paramétrico, la idea también es válida con la metodología no paramétrica. Por ejemplo, un criterio empírico no paramétrico para detectar un α-*outlier* Z vendría dado por la condición $Z \in \{\hat{f}_n \leq c_n\}$ (donde c_n vendría dado por (3.2)).

Por otro lado, desde un punto de vista teórico, este problema se puede plantear de la manera siguiente: $\{\hat{f}_n \geq c_n\}$ es el conjunto de menor volumen (en el sentido de medida de Lebesgue) con un contenido $1-\alpha$ prefijado de probabilidad empírica. ¿Cuánto difiere la probabilidad real $P_f\{\hat{f}_n \geq c_n\}$ de ese valor empírico?

Otros contextos en los que se consideran los conjuntos de nivel son el análisis bayesiano (Das Gupta, Ghosh and Zen 1995, Choudhuri 1999) y, como ya mencionamos en el capítulo 1, la estimación de conjuntos (Cuevas y Fraiman 1997, Tsybakov 1997, Molchanov 1998, Polonik 1999).

3.2 Caso c constante: tasas de convergencia L^1 para la probabilidad de no clasificación

Si Z es una variable aleatoria con densidad f, el límite natural de la sucesión $P_n(Z)$, definida en (3.1), es la variable aleatoria $P_\infty(Z) = I_{\{f(Z)\leq c\}}$.

El siguiente resultado muestra que, tomando adecuadamente el orden del parámetro de suavizado h_n, se puede conseguir que la sucesión $E|P_n - P_\infty|$ converja a 0 a cualquier tasa potencial más lenta que $n^{-1/(d+2)}$. Esta conclusión es válida para densidades f multivariadas, con soporte no acotado, bajo condiciones de cola no muy restrictivas. En Chanda y Ruymgaart (1989) se puede encontrar un resultado relacionado con éste en el campo del análisis discriminante no paramétrico.

TEOREMA 3.1. *Sea f una densidad Lipschitz en $\mathbb{R}^d$. Supongamos que*

(F1) existen $R > r > 0$ tales que $m_f := m_{f,R+r}(0) > 0$ y

$$\int_{B^c(0,R)} \left(\frac{M_{f,h}(x)}{m_{f,h}(x)}\right)^{3/2} f^{1/2}(x)dx < \infty,$$

donde $M_{f,h}(z) := \sup\{f(x) : x \in B(z,h)\}$ *y* $m_{f,h}(z) := \inf\{f(x) : x \in B(z,h)\}$.

(F2) $\mathbf{P}\{|c-f(X)| \leq \epsilon\} = O(\epsilon)$, *cuando* $\epsilon \to 0$, *donde* X *denota un vector aleatorio con densidad* f.

Sea $\hat{f}_n$ *un estimador kernel de* f *tal que*

(K1) el núcleo K *es una función de densidad con soporte compacto, acotada por una constante* $M_K < \infty$.

Supongamos además que

(H1) la amplitud de banda h_n *tiene orden exacto* n^{-s}, *es decir,*

$$0 < \liminf_{n\to\infty} \frac{h_n}{n^{-s}} \leq \limsup_{n\to\infty} \frac{h_n}{n^{-s}} < \infty,$$

para algún $s \in (0, 1/(d+2))$.

Bajo estas condiciones se verifica que

$$E|P_n - P_\infty| = O(n^{-s}), \quad \textit{cuando } n \to \infty. \tag{3.4}$$

DEMOSTRACIÓN. Sea $\sigma_n^2(z) := \mathrm{Var}(K_h(z-X))$ y $E_n(z) := E(K_h(z-X))$, donde X es un vector aleatorio con densidad f. Obsérvese que, como f es Lipschitz, está acotada. Denotemos por M_f una cota de f. Tenemos que

$$EK\left(\frac{z-X}{h}\right) = h\int K(u)f(z-hu)du \leq hM_f. \tag{3.5}$$

Podemos suponer, sin pérdida de generalidad, que $R = 1$ y que $\{K > 0\} \subset B(0,1)$. Entonces

$$\mathbf{P}\{\hat{f}_n(z) \leq c\} = \mathbf{P}\left\{\frac{1}{\sigma_n(z)\sqrt{n}}\sum_{j=1}^{n} Y_j \leq \frac{c - E_n(z)}{\sigma_n(z)/\sqrt{n}}\right\},$$

donde $Y_i = K_h(z - X_i) - E_n(z)$. Al aplicar la desigualdad de Berry-Esseen (ver Sección 4 del capítulo 1) a la última expresión se obtiene

$$\left| \mathbf{P}\{\hat{f}_n(z) \leq c\} - \Phi\left(\frac{c - E_n(z)}{\sigma_n(z)/\sqrt{n}}\right) \right| \leq R_n(z),$$

siendo Φ la función de distribución de la normal (0,1) y

$$R_n(z) = A\frac{E|K_h(z - X) - E_n(z)|^3}{\sigma_n^3(z)\sqrt{n}}.$$

Por tanto,

$$E|P_n - P_\infty| \leq \int \left| \Phi\left(\frac{c - E_n(z)}{\sigma_n(z)/\sqrt{n}}\right) - I_{\{f \leq c\}}(z) \right| f(z)dz + \int R_n(z)f(z)dz. \quad (3.6)$$

En primer lugar consideremos la segunda integral de (3.6). Para simplificar la notación definimos $R_n^*(z) := n^{1/2}A^{-1}R_n(z)$. Entonces, dado $\delta > 0$, existe algún n_0, uniforme en z, tal que, si $n \geq n_0$,

$$\begin{aligned}
R_n^*(z) &= \frac{E\left|K\left(\frac{z-X}{h}\right) - EK\left(\frac{z-X}{h}\right)\right|^3}{E^{3/2}\left|K\left(\frac{z-X}{h}\right) - EK\left(\frac{z-X}{h}\right)\right|^2} \\
&\leq \frac{\delta^3\mathbf{P}\left\{0 < \left|K\left(\frac{z-X}{h}\right) - EK\left(\frac{z-X}{h}\right)\right| \leq \delta\right\}}{\delta^3\mathbf{P}^{3/2}\left\{\left|K\left(\frac{z-X}{h}\right) - EK\left(\frac{z-X}{h}\right)\right| > \delta\right\}} \\
&\qquad + \frac{(2M_K)^3\mathbf{P}\left\{\left|K\left(\frac{z-X}{h}\right) - EK\left(\frac{z-X}{h}\right)\right| > \delta\right\}}{\delta^3\mathbf{P}^{3/2}\left\{\left|K\left(\frac{z-X}{h}\right) - EK\left(\frac{z-X}{h}\right)\right| > \delta\right\}} \\
&\leq \frac{\mathbf{P}\left\{0 < K\left(\frac{z-X}{h}\right) < 2\delta\right\}}{\mathbf{P}^{3/2}\left\{K\left(\frac{z-X}{h}\right) > 2\delta\right\}} + \frac{(2M_K)^3}{\delta^3\mathbf{P}^{1/2}\left\{K\left(\frac{z-X}{h}\right) > 2\delta\right\}},
\end{aligned}$$

donde la última desigualdad se deriva de (3.5). Obsérvese que

$$\mathbf{P}\left\{0 < K\left(\frac{z - X}{h}\right) < 2\delta\right\} \leq P_f[B(z,h)] \leq M_f B_d h^d.$$

Por otro lado, debido a que K es una función de densidad, si tomamos δ suficientemente pequeño, entonces C_δ, la medida de Lebesgue de $\{x : K(x) > 2\delta\}$,

es estrictamente positiva. Además si $x \in B(z,h)$ y $z \in B(0,R)$, se tiene que $x \in B(0,R+r)$ para h suficientemente pequeño. En este caso $f(x) \geq m_f$. Esto se utiliza para acotar la integral definida sobre $B(0,R)$ que aparece en la siguiente desigualdad y se obtiene

$$\begin{aligned}
&\int R_n^*(z)f(z)dz \\
&\leq \int_{B(0,R)} \left(\frac{M_f B_d h^d}{(m_f C_\delta h^d)^{3/2}} + \frac{(2M_K)^3}{\delta^3 (m_f C_\delta h^d)^{1/2}} \right) f(z)dz \\
&\quad + \int_{B^c(0,R)} \left(\frac{M_{f,h}(z) B_d h^d}{(C_\delta h^d m_{f,h}(z))^{3/2}} + \frac{(2M_K)^3}{\delta^3 (C_\delta h^d m_{f,h}(z))^{1/2}} \right) f(z)dz \\
&\leq h^{-d/2} \left(D_1 P_f[B(0,R)] + D_2 \int_{B^c(0,R)} \left(\frac{M_{f,h}(z)}{m_{f,h}(z)} \right)^{3/2} f^{1/2}(z)dz \right),
\end{aligned}$$

donde

$$\begin{aligned}
D_1 &:= \left(\frac{M_f}{m_f C_\delta} + \left(\frac{2M_K}{\delta} \right)^3 \right) (m_f C_\delta)^{-1/2}, \\
D_2 &:= \frac{B_d}{C_\delta^{3/2}} + \frac{(2M_K)^3}{\delta^3 C_\delta^{1/2}}.
\end{aligned}$$

Ahora nos ocuparemos de la primera integral que aparece en (3.6), que denotaremos por $\int Q_n(z)f(z)dz$. Como f es Lipschitz y $\int \|t\|K(t)dt < \infty$, existe una constante $C_0 > 0$ tal que $|f(z) - E_n(z)| \leq C_0 h$, para todo $z \in \mathbb{R}^d$. Separamos dicha integral en dos de la manera siguiente,

$$\int Q_n(z)f(z)dz \leq \int_{\{|c-f(z)| \leq 2C_0 h\}} f(z)dz + \int_{\{|c-f(z)| > 2C_0 h\}} Q_n(z)f(z)dz. \qquad (3.7)$$

Por (F2), la primera integral del término de la derecha es $O(h)$. En lo que se refiere a la segunda, observemos que para los z tales que $|c - f(z)| > 2C_0 h$, $c - f(z)$ y $c - E_n(z)$ tienen el mismo signo, por lo que

$$Q_n(z) = \Phi\left(-\frac{|c - E_n(z)|}{\sigma_n(z)/\sqrt{n}} \right),$$

que se puede acotar utilizando la última desigualdad del capítulo 1. Se obtiene, pues, para n suficientemente grande,

$$\int_{\{|c-f(z)| > 2C_0 h\}} Q_n(z)f(z)dz$$

$$\leq \int_{\{|c-f(z)|>2C_0h\}} \frac{\sigma_n(z)}{\sqrt{2\pi}\sqrt{n}|c-E_n(z)|} \exp\left(-\frac{1}{2}\frac{|c-E_n(z)|^2}{\sigma_n^2(z)/n}\right) dz.$$

Teniendo en cuenta que

$$\sigma_n^2(z) \leq EK_h^2(z-X) \leq \frac{M_f \int K^2(t)dt}{h^d} := \frac{M_2}{h^d};$$

y el hecho de que $|c-E_n(z)| > C_0h$ en el conjunto de integración, tenemos que

$$\int_{\{|c-f(z)|>2C_0h\}} Q_n(z)f(z)dz \leq \frac{\sqrt{M_2}}{\sqrt{2\pi}\sqrt{nh^d}C_0h} \exp\left(-\frac{1}{2}\frac{C_0^2h^2}{M_2/nh^d}\right). \tag{3.8}$$

Por tanto, el miembro de la derecha de (3.8) es de orden $\exp(-Cn^r)$, donde $C > 0$ es una constante y $r = 1 - s(d+2) > 0$. En consecuencia, esta expresión es $o(n^{-q})$ para cualquier $q > 0$.

En resumen, hemos obtenido que la segunda integral que aparece en el término de la derecha de (3.6) es de orden $O\left((nh^d)^{-1/2}\right) = O\left(n^{-(1-sd)/2}\right)$, y la primera es de orden $O(h) = O(n^{-s})$, de lo que se sigue el resultado que deseábamos, pues $n^{-s} > n^{-(1-sd)/2}$.

□

3.2.1 Algunos comentarios

Las siguientes observaciones aclaran el alcance real y el significado intuitivo de las hipótesis que aparecen en el Teorema 3.1, y también ofrecen resultados alternativos en la misma línea.

Observación 1. La conclusión (3.4) sigue siendo válida si reemplazamos (F1) por la hipótesis de que el soporte S de f sea compacto y además que $\inf\{f(x) : x \in S\} > 0$. En este caso la condición Lipschitz impuesta a f debe verificarse exclusivamente sobre S.

Observación 2. La condición de cola (F1) no es muy restrictiva. Por ejemplo, la verifica cualquier densidad f tal que, para $\|x\|$ suficientemente grande, se puede expresar de la forma $f(x) = g(\|x\|)$, donde $g(t)$ es una función decreciente cuyo orden exacto, cuando $t \to \infty$, es, por ejemplo, $g(t) = O(t^{-k})$ (para algún $k > 2d$), o bien $g(t) = O(e^{-at})$ (para algún $a > 0$), o $g(t) = O(e^{-at^2})$ (para algún $a > 0$). Están incluidas, por tanto, las densidades más utilizadas, en particular las de tipo gamma o las gaussianas.

También verifican esta condición las densidades tales que, para $\|x\|$ suficientemente grande, el cociente

$$\frac{M_{f,h}(x)}{m_{f,h}(x)} = \frac{\sup\{f(z) : z \in B(z,h)\}}{\inf\{f(z) : z \in B(z,h)\}}$$

está acotado superiormente por una constante y $\int f^{1/2} < \infty$.

La condición (F1) puede cumplirse incluso en casos en que el anterior cociente tiende a infinito cuando $\|x\| \to \infty$, como ocurre con la distribución normal.

Observación 3. La condición (F2) sólo expresa de manera rigurosa la idea intuitiva de que es difícil estimar la probabilidad de $\{f \leq c\}$ si el conjunto $\{f = c\}$ no está formado únicamente por puntos aislados: si este conjunto incluye un intervalo, $\hat{f}_n$ tendrá ligeras protuberancias que afecten a la convergencia de P_n a P_∞.

Esta condición aparece típicamente en la estimación de conjuntos de nivel. Por ejemplo, en Walther (1997) si el conjunto $\{f = c\}$ tiene "picos", se pide que f no sea plana en $\{f = c\}$ para poder asegurar la consistencia del estimador del conjunto $\{f > c\}$. Tsybakov (1997) también excluye la posibilidad de que la densidad f tenga "mesetas" en el nivel c, imponiendo una condición que denomina α-regularidad. Por último, Polonik (1995) establece la hipótesis

$$\mu_L\{x : |c - f(x)| \leq \epsilon\} = O(\epsilon^{1/\alpha}),$$

cuando $\epsilon \to 0$, para alguna constante $\alpha > 0$. Es una condición más débil que la de Tsybakov (1997) y, para $\alpha = 1$, es muy similar a nuestra (F2). El hecho de que Polonik (1995) trabaje con la medida de Lebesgue μ_L en lugar de con la

probabilidad $\mathbf{P}$, como hacemos en (F2), es debido a que su objetivo es conseguir tasas para la distancia en medida.

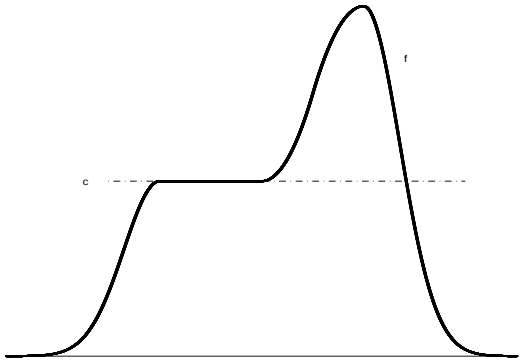

Figura 3: Densidad que no verifica la condición (F2)

Observación 4. Tomando $\alpha \in (0,1)$ y separando la integral que aparece en el miembro de la izquierda de la desigualdad (3.7) en dos integrales definidas sobre los conjuntos $\{|c - f(z)| \leq 2C_0 h^\alpha\}$ y $\{c - f(z) > 2C_0 h^\alpha\}$ respectivamente, (H1) se puede extender a $s \in (0, (d+2)^{-1}]$ y $E|P_n - P_\infty|$ seguiría convergiendo con cualquier tasa potencial más lenta que $n^{-1/(d+2)}$.

Entonces, para $d = 1$, el teorema incluye las clásicas amplitudes de banda que optimizan, en L^2, el comportamiento del estimador *kernel* al estimar la función de distribución ($h = n^{-1/3}$), la función de densidad ($h = n^{-1/5}$) y su derivada r-ésima ($h = n^{-1/(2r+5)}$).

Observación 5. Si $\{c_n\} \subset \mathbb{R}$ verifica $c_n \to c$ y $|c_n - c| = O(h)$ y redefinimos $P_n(z) = \mathbf{P}\{\hat{f}_n(z) \leq c_n\}$, la tasa de convergencia de $E|P_n - P_\infty|$ no se ve afectada.

3.3 Caso c aleatorio: tasas de convergencia para la probabilidad de falsa alarma

3.3.1 Una tasa para la convergencia casi segura

El objetivo de esta sección es obtener tasas para la convergencia a 0 de

$$P_f\{\hat{f}_n \leq c_n\} - \alpha, \tag{3.9}$$

donde $c_n = c_n(X_1, \ldots, X_n)$ es un número positivo tal que $P_{\hat{f}_n}\{\hat{f}_n \leq c_n\} = \alpha$.

En el Teorema 3.2 se prueba que, eligiendo $h_n = (n^{-1}\log n)^{1/3}$, $P_f\{\hat{f}_n \leq c_n\} - \alpha$ alcanza una tasa casi segura de orden $O(h^s)$, para un cierto $0 < s \leq 1$ que depende del comportamiento de f en la cola. La tasa óptima $O(h)$ que corresponde a $s = 1$, se obtiene cuando f tiene soporte compacto. El teorema se establece únicamente en el caso unidimensional, pero parece probable que las mismas ideas se puedan aplicar en el caso multivariado.

A lo largo de la prueba de este resultado, se utilizarán el Teorema 1.3 de Stute (1982) y las observaciones 1.3 y 1.4 de Hall (1990) para acotar la diferencia $|\hat{f}_n(x) - E\hat{f}_n(x)|$ sobre una sucesión creciente de intervalos compactos.

TEOREMA 3.2. *En el caso univariado $d = 1$, supongamos que*

(F3) f es Lipschitz en $\mathbb{R}$.

(F4) $f(x) = O(|x|^{-\gamma})$, cuando $|x| \to \infty$, para algún $\gamma > 1$.

(K2) K es una densidad de soporte compacto y de variación acotada.

Tomemos $h_n = (n^{-1}\log n)^{1/3}$. Entonces

$$\left|P_f\{\hat{f}_n \leq c_n\} - \alpha\right| = O(h^s) \quad c.s., \tag{3.10}$$

donde

$$s = \begin{cases} 1 & \textit{si } f \textit{ tiene soporte compacto} \\ 1-\gamma^{-1} & \textit{en otro caso.} \end{cases}$$

DEMOSTRACIÓN. Consideremos primero el caso en que el soporte de f no está acotado. Sea β un número positivo que fijaremos más adelante. Acotamos el término de la izquierda de (3.10) de la manera siguiente:

$$\begin{aligned} &\left|P_f\{\hat{f}_n \le c_n\} - \alpha\right| \\ &\quad \le \left|P_f\left(\{\hat{f}_n \le c_n\} \cap [-h^{-\beta}, h^{-\beta}]\right) - P_{\hat{f}_n}\left(\{\hat{f}_n \le c_n\} \cap [-h^{-\beta}, h^{-\beta}]\right)\right| \\ &\qquad + P_f\left(\{\hat{f}_n \le c_n\} \cap [-h^{-\beta}, h^{-\beta}]^c\right) + P_{\hat{f}_n}\left(\{\hat{f}_n \le c_n\} \cap [-h^{-\beta}, h^{-\beta}]^c\right) \\ &\quad \le \int_{[-h^{-\beta}, h^{-\beta}]} |f - E\hat{f}_n| + \int_{[-h^{-\beta}, h^{-\beta}]} |\hat{f}_n - E\hat{f}_n| \\ &\qquad + \int_{[-h^{-\beta}, h^{-\beta}]^c} (f + \hat{f}_n) \end{aligned} \tag{3.11}$$

Por (F3) y (K2) existe $C_1 > 0$ tal que $|f(x) - E\hat{f}_n(x)| \le C_1 h$ uniformemente en $\mathbb{R}$. Por tanto,

$$\int_{[-h^{-\beta}, h^{-\beta}]} |f - E\hat{f}_n| = O\left(h^{1-\beta}\right). \tag{3.12}$$

Nuestras hipótesis implican las de Hall (1990, observación 1.3). Del hecho de que $nh \to \infty$, tenemos que, de un n en adelante, $\{|x| \le h^{-\beta}\} \subseteq \{|x| \le n^{\beta}\}$. Luego, por Hall (1990),

$$\sup_{|x| \le h^{-\beta}} |\hat{f}_n(x) - E\hat{f}_n(x)| = O((\log h^{-1}/nh)^{1/2}) \quad \text{c.s.}$$

Entonces hemos obtenido que, para una cierta constante $C_3 > 0$,

$$\int_{[-h^{-\beta}, h^{-\beta}]} |\hat{f}_n - E\hat{f}_n| \le C_3(\log h^{-1}/nh^{2\beta+1})^{1/2} = O\left(h^{1-\beta}\right) \quad \text{c.s.} \tag{3.13}$$

Por otro lado, (F4) implica que

$$\int_{[-h^{-\beta}, h^{-\beta}]^c} f(x)dx \le C_4 h^{\beta(\gamma-1)}. \tag{3.14}$$

Si tomamos $\beta = \gamma^{-1}$ obtenemos el mismo orden de convergencia en (3.12), (3.13) y (3.14). Por esta razón, a partir de ahora, asumiremos que hemos elegido ese β concreto.

Para acotar el último término de (3.11), podemos suponer, sin pérdida de generalidad, que el soporte de K está contenido en $[-1,1]$. A partir de (K2) obtenemos que K está acotado por una constante positiva H_K. Entonces

$$\begin{aligned}\int_{[-h^{-\beta},h^{-\beta}]^c} \hat{f}_n(x)dx &\leq \frac{H_K}{nh}\sum_{i=1}^{n}\int_{[-h^{-\beta},h^{-\beta}]^c} I_{[X_i-h,X_i+h]}(x)dx \\ &\leq \frac{H_K}{nh}\sum_{i=1}^{n} 2hI_{[-h^{-\beta}+h,h^{-\beta}-h]^c}(X_i) \\ &= C_5 P_n\left([-h^{-\beta}+h,h^{-\beta}-h]^c\right), \qquad (3.15)\end{aligned}$$

donde P_n denota la distribución de probabilidad empírica basada en la muestra $X_1,\ldots,X_n$. Por otro lado, utilizando (F4) obtenemos

$$P_f\left([-h^{-\beta}+h,h^{-\beta}-h]^c\right) \leq C_6 h^{\beta(\gamma-1)}, \qquad (3.16)$$

para n suficientemente grande. De (3.16) y la desigualdad de Hoeffding resulta

$$\begin{aligned}&\mathbf{P}\left\{P_n\left([-h^{-\beta}+h,h^{-\beta}-h]^c\right) \geq 2C_6h^{\beta(\gamma-1)}\right\} \\ &\quad\leq \mathbf{P}\left\{\sum_{i=1}^{n}\left[I_{[-h^{-\beta}+h,h^{-\beta}-h]^c}(X_i) - P_f\{[-h^{-\beta}+h,h^{-\beta}-h]^c\}\right] \geq C_6nh^{\beta(\gamma-1)}\right\} \\ &\quad\leq \exp\left[-C_7nh^{2\beta(\gamma-1)}\right].\end{aligned}$$

Como $\beta=\gamma^{-1}$, entonces $\sum_{n=1}^{\infty}\exp[-C_7nh^{2\beta(\gamma-1)}]<\infty$ y, por el lema de Borel-Cantelli,

$$P_n\left([-h^{-\beta}+h,h^{-\beta}-h]^c\right) = O(h^{\beta(\gamma-1)}) \quad \text{c.s.}, \qquad (3.17)$$

que es el mismo orden que habíamos obtenido para los otros términos.

Obsérvese que, si se supone que f tiene soporte compacto, el último término de la expresión (3.11) desaparece al reemplazar el valor $h^{-\beta}$ por una constante suficientemente grande.

□

Observación 6. Es interesante constatar que, a diferencia del Teorema 3.1, el resultado anterior no precisa la hipótesis (F2). Esto se debe a que en el Teorema 3.2 el objetivo no es estimar un conjunto en particular, sino su contenido en probabilidad.

Observación 7. En Baíllo, Cuesta-Albertos y Cuevas (1999) se puede encontrar un corolario al Teorema 3.2 en el que se establece la convergencia de c_n a c (en probabilidad y casi seguro) con cualquier tasa más lenta que h_n^s. En este resultado se impone de nuevo la condición (F2).

3.3.2 Algunos resultados relacionados

Se pueden obtener otros resultados referentes a las tasas de convergencia de $c = c_n$ elegido como en (3.2). Surgen de versiones alternativas del Teorema 3.2 o como consecuencia de las tasas de convergencia L^1 de los estimadores *kernel* multivariantes. A continuación presentamos a grandes rasgos algunos de estos resultados complementarios.

Observación 8. Las densidades f que están acotadas inferiormente por una constante positiva sobre su soporte, por ejemplo un intervalo $[a, b]$, no satisfacen (F3). En lugar de utilizar el teorema que aparece en Hall (1990), se podría simplemente emplear el Teorema 1.2 de Stute (1982). Entonces, para $h_n = (n^{-1} \log n)^{1/3}$, si se verifican

(K3) K se anula fuera de un cierto intervalo finito $[a_1, b_1)$,

(F3') f es continua en un intervalo acotado $[a, b]$ y se anula fuera de este intervalo

y

(F4') f es Lipschitz en su soporte,

podemos modificar ligeramente la prueba del Teorema 3.2 para obtener

$$|P_f\{\hat{f}_n \leq c_n\} - \alpha| = O(h_n) \quad \text{c.s.},$$

es decir, la misma conclusión del Teorema 3.2 pero para f de soporte compacto.

Observación 9. Stute (1984) obtuvo un análogo multivariante del Teorema 1.2 que aparece en Stute (1982). Bajo condiciones similares a las de la observación previa, enunció el siguiente resultado:

$$\sup_{t \in I_0} |\hat{f}_n(t) - E\hat{f}_n(t)| = O((\log h^{-d}/nh^d)^{1/2}) \quad \text{c.s.}, \tag{3.18}$$

donde $I_0 \subset \mathbb{R}^d$ es un cierto conjunto acotado. Por esta razón, de acuerdo con la observación 1.6 de Hall (1990), el Teorema 3.2 sigue siendo válido en el sentido siguiente: la tasa volvería a ser $O(h^s)$, con $h = (n^{-1}\log n)^{1/(d+2)}$, $s = 1$ para f con soporte compacto y $0 < s < 1$ en otro caso.

Observación 10. Holmström y Klemelä (1992) obtuvieron una tasa de convergencia del tipo

$$E\int_{\mathbb{R}^d} |\hat{f}_n - f| = O\left(n^{-2/(d+4)}\right) \tag{3.19}$$

para el error L^1 esperado de un estimador *kernel* multivariado con $h_n = Cn^{-1/(d+4)}$. Si c_n se elige de manera que $P_{\hat{f}_n}\{\hat{f}_n \leq c_n\} = \alpha$ c.s., tenemos que $E|P_f\{\hat{f}_n \leq c_n\} - \alpha| \leq E\int|\hat{f}_n - f|$. Por tanto, por la desigualdad de Markov, la convergencia en probabilidad de $|P_f\{\hat{f}_n \leq c_n\} - \alpha| \to 0$ también se verifica (bajo las condiciones impuestas en Holmström y Klemelä (1992)) con la tasa (3.19).

Capítulo 4

Estimación del soporte bajo restricciones de forma

En este capítulo se considera nuevamente el problema de estimar S, un conjunto en $\mathbb{R}^d$, a partir de una muestra aleatoria de puntos elegidos dentro de él. Cuando se conoce alguna propiedad relativa a la forma de S (por ejemplo, que sea convexo, estrellado, conexo,...) parece razonable utilizar estimadores que posean dicha propiedad.

La **convexidad** es la restricción de forma más ampliamente estudiada en la literatura. Bajo esta hipótesis, el estimador natural de S es el cierre convexo de la muestra, que es un estimador "por defecto": $\text{conv}(X_1, \ldots, X_n) \subset S$ c.s. para todo n. Para propiedades de forma distintas de la convexidad no hay un estimador que surja de manera tan directa. Aquí proponemos utilizar el estimador de Devroye y Wise $\hat{S}_n(\epsilon_n)$, eligiendo el parámetro de suavizado ϵ_n de tal manera que $\hat{S}_n$ verifique la misma propiedad de forma que S. En otras palabras, una elección adecuada de ϵ_n permite, en algunos casos, que $\hat{S}_n$ incorpore la información que poseemos acerca de S.

En la Sección 4.2 la restricción impuesta a S es la de **conexión**. El Teorema 4.1 proporciona condiciones bajo las que el estimador conexo $\hat{S}_n$ es consistente en medida. En la prueba de este resultado se utilizan resultados de Tabakis (1996) y Penrose (1999) sobre árboles abarcadores.

En la Sección 4.3 se considera una hipótesis más restrictiva que la de conexión: que S sea **estrellado**. El Teorema 4.2 prueba la convergencia casi segura a cero del ínfimo de los radios ϵ_n con los que $\hat{S}_n(\epsilon_n)$ es estrellado. El Teorema 4.3 es un resultado sobre convergencia de la frontera del estimador estrellado $\hat{S}_n$ a la frontera de S, para parámetros de suavizado ϵ_n de orden superior o igual a $O(\log n/n)^{1/d_0}$, con $d_0 > d$. Con esta restricción sobre el radio, $\hat{S}_n$ es un estimador "por exceso" para n suficientemente avanzado.

La motivación inicial de este capítulo es más "teórica" que en los anteriores: nuestro principal interés es comprobar la flexibilidad de un estimador tan sencillo como el de cubrimiento. Consideramos, no obstante, que estas ideas son potencialmente aplicables al análisis de imágenes. Por ejemplo, la condición de "forma estrellada" aparece en Donoho (1999) en el contexto del análisis de imágenes con *wedgelets*.

4.1 Estimación de un conjunto conexo

La hipótesis de conexión es una de las restricciones más suaves que se le pueden imponer a la forma de S. Si, por ejemplo, el vector aleatorio X corresponde a observaciones tomadas en un proceso industrial, el hecho de que S no sea conexo quiere decir que, en realidad, estamos observando varios procesos "disjuntos" que funcionan en paralelo.

En esta sección se muestra que el estimador de cubrimiento

$$\hat{S}_n(\epsilon) = \bigcup_{i=1}^{n} B(X_i, \epsilon)$$

es lo suficientemente flexible como para incorporar la hipótesis de conexión y proporcionar una estimación consistente de S. En muchos problemas de control de calidad, si se desea aplicar el método no paramétrico de detección de *change-point* propuesto por Devroye y Wise (1980), este estimador conexo representa una elección muy natural (ver Baíllo, Cuevas y Justel 2000).

La idea central consiste en considerar el estimador $\hat{S}_n$ con el parámetro de suavizado $\bar{\epsilon}_n = \bar{\epsilon}_n(X_1, \ldots, X_n)$, definido por

$$\bar{\epsilon}_n = \inf \left\{ \epsilon > 0 : \hat{S}_n(\epsilon) \text{ es un conjunto conexo} \right\}. \tag{4.1}$$

Un procedimiento iterativo muy sencillo para encontrar el valor $\bar{\epsilon}_n$ es el siguiente (ver también Lebart, Morineau y Warwick 1984; Cuevas, Febrero y Fraiman 2000):

Algoritmo para el cálculo de $\bar{\epsilon}_n$

1. Se comienza por cualquier observación (por ejemplo, X_1) y se calcula la distancia R_1 entre X_1 y el punto muestral más cercano a X_1 (digamos X_2).

2. Se toma la observación (digamos X_3) más cercana al conjunto $\{X_1, X_2\}$ y se calcula $R_2 := \min\{\|X_3 - X_1\|, \|X_3 - X_2\|\}$.

3. Se calcula, por recurrencia, $R_k := \min\{\|X_{k+1} - X_i\|, i = 1, \ldots, k\}$, donde X_{k+1} es la observación más cercana al conjunto $\{X_1, \ldots, X_k\}$ de entre todos los puntos muestrales X_i restantes.

4. Se itera este procedimiento hasta que se han considerado todas las observaciones. Entonces
$$\bar{\epsilon}_n = \max_{k=1,\ldots,n-1} \frac{R_k}{2}.$$

A continuación presentamos unos resultados de consistencia que están en la línea del Teorema 1 de Devroye y Wise (1980). La diferencia fundamental estriba en que ahora $\bar{\epsilon}_n$ es una sucesión estocástica de parámetros de suavizado, en lugar de una sucesión numérica.

TEOREMA 4.1. *Sea $X_1, X_2, \ldots$ una sucesión de vectores aleatorios i.i.d. en $\mathbb{R}^d$ cuya distribución común es absolutamente continua respecto a la medida de Lebesgue μ_L. Sea f la densidad (respecto a μ_L) de cada X_i y S el soporte de f. Supongamos que S es compacto y conexo en $\mathbb{R}^d$ y que $f_0 := \operatorname{ess\,inf}_S f > 0$.*

a) Si $\sup\{|f(x) - f(y)| : x, y \in S, |x - y| < s\} = o((-\log s)^{-1})$ *cuando* $s \to 0$ *entonces*

$$\nu\left(\hat{S}_n(\bar{\epsilon}_n)\Delta S\right) \longrightarrow 0 \quad \textit{en probabilidad,} \tag{4.2}$$

donde $\hat{S}_n(\bar{\epsilon}_n)$ *es el estimador de cubrimiento con la amplitud de banda* $\epsilon_n = \bar{\epsilon}_n$ *definida en (4.1) y* ν *es cualquier medida sobre* $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$ *cuya restricción a* S *sea absolutamente continua respecto a* P_f.

b) Para $d \geq 2$, *supongamos que el conjunto de discontinuidades de la restricción de* f *a* S *tiene medida de Lebesgue nula y no contiene ningún elemento de* ∂S. *Se supone además que* ∂S *es una subvariedad en* $\mathbb{R}^d$, $(d-1)$*-dimensional y* C^2. *Entonces*

$$\nu\left(\hat{S}_n(\bar{\epsilon}_n)\Delta S\right) \longrightarrow 0 \quad \textit{c.s.}$$

DEMOSTRACIÓN. Observemos que

$$\nu\left(\hat{S}_n(\bar{\epsilon}_n)\Delta S\right) = \nu\left(\hat{S}_n(\bar{\epsilon}_n) \setminus S\right) + \nu\left(S \setminus \hat{S}_n(\bar{\epsilon}_n)\right). \tag{4.3}$$

En primer lugar probemos que $\bar{\epsilon}_n \to 0$ c.s. En efecto, tomemos cualquier $\epsilon > 0$. Consideremos un cubrimiento $\mathcal{C}$ de S constituido por cubos cerrados de lado $\epsilon/\sqrt{d}$ e interiores disjuntos. Con probabilidad 1, existe un n a partir del cual cada $C \in \mathcal{C}$ tal que $\mu_L(C \cap S) > 0$ contendrá una observación X_i. Por tanto, $\bar{\epsilon}_n < \text{diag}(C) = \epsilon$ para n suficientemente grande c.s. Por el teorema de la convergencia dominada, esto implica la convergencia casi segura a cero del primer término en el miembro de la derecha de (4.3).

Para estudiar el segundo término de (4.3) observemos que, en el contexto de teoría de grafos, la longitud de $\bar{\epsilon}_n$ es la mitad de la de M_n, el lado más largo del mínimo árbol abarcador (ver, por ejemplo, Lebart, Morineau y Warwick 1984) con vértices $\{X_1, \ldots, X_n\}$. Esto se sigue del algoritmo que hemos presentado antes para el cálculo de $\bar{\epsilon}_n$, ya que este procedimiento iterativo de hecho determina un mínimo árbol abarcador cuyo lado más largo tiene longitud $2\bar{\epsilon}_n$. En efecto, obsérvese que en un árbol abarcador la longitud de cualquier lado de vértice X_1 debe ser, por lo menos, R_1: la longitud de cualquier lado que conecte X_2 con una observación distinta de X_1 debe ser, por lo menos, R_2, etc.

Bajo las hipótesis generales del teorema y la suposición específica de (a), Tabakis (1996) probó que

$$\lim_{n\to\infty} \mathbf{P}\left\{ M_n \geq \left(\frac{k}{A\ B_d}\right)^{1/d} \left(\frac{\log n}{n}\right)^{1/d} \right\} = 1,$$

para todo $k < 1$, donde $A = \sup\{f(x) : x \in S\}$, (obsérvese que f es continua y, por tanto, acotada en el compacto S).

Por esta razón, si denotamos por

$$\tilde{\epsilon}_n = \frac{1}{2}\left(\frac{k}{A\ B_d}\right)^{1/d} \left(\frac{\log n}{n}\right)^{1/d},$$

tenemos que, con probabilidad convergente a 1, $\bar{\epsilon}_n \geq \tilde{\epsilon}_n$. Luego

$$\nu\left(S \setminus \hat{S}_n(\bar{\epsilon}_n)\right) \leq \nu\left(S \setminus \hat{S}_n(\tilde{\epsilon}_n)\right), \tag{4.4}$$

con probabilidad convergente a 1.

Como $\tilde{\epsilon}_n \to 0$ y $n\tilde{\epsilon}_n^d \to \infty$, el miembro de la derecha de (4.4) converge a cero en probabilidad (ver Teorema 1 de Devroye y Wise (1980)). Con esto queda finalmente probada la convergencia a cero en probabilidad de $\nu\left(\hat{S}_n(\bar{\epsilon}_n)\Delta S\right)$.

Bajo una restricción de forma adicional sobre S, Tabakis (1996) obtiene también una cota superior en probabilidad, del tipo $C(\log n/n)^{1/d}$, para el lado más largo del mínimo árbol abarcador. Es decir, que en realidad $(\log n/n)^{1/d}$ es el orden exacto en probabilidad de $\bar{\epsilon}_n$, pero esto no se necesita para probar (4.2).

Por otro lado, bajo las hipótesis generales del teorema y las específicas de (b), Penrose (1999) probó que

$$\lim_{n\to\infty} M_n^d \frac{n}{\log n} = \frac{1}{B_d} \max\left(\frac{1}{f_0}, \frac{2(d-1)}{d\ f_1}\right) \quad \text{c.s.},$$

donde $f_1 := \inf_{\partial S} f$. El resto se prueba de manera completamente análoga a como se hizo en (a), excepto que se debería sustituir $\tilde{\epsilon}_n$ por

$$\tilde{\tilde{\epsilon}}_n := c\left[\frac{1}{B_d} \max\left(\frac{1}{f_0}, \frac{2(d-1)}{d\ f_1}\right)\right]^{1/d} \left(\frac{\log n}{n}\right)^{1/d},$$

siendo c cualquier constante tal que $0 < c < 1$.

□

4.2 Estimación de un conjunto estrellado

En esta sección se considerarán conjuntos $S \subset \mathbb{R}^d$ que verifiquen la condición de ser estrellados. Como ya se expuso en la Sección 1.4, esto quiere decir que existe un "punto radiante" $a \in S$ tal que, para todo $x \in S$, el segmento $[a, x]$ que une a con x está contenido en el conjunto S. Se trata de una condición más fuerte que la de conexión: no se permite la presencia de ningún elemento de S^c en el interior del espacio delimitado por la frontera de S. Pero al mismo tiempo es una hipótesis menos restrictiva que la de convexidad: todo conjunto convexo C es estrellado y su "mirador" (es decir, el conjunto de los puntos radiantes) es el propio C.

Sin embargo, la hipótesis de que S sea estrellado no excluye "situaciones patológicas" asociadas a la presencia de infinitos picos cada vez más agudos. La Figura 4 muestra un ejemplo.

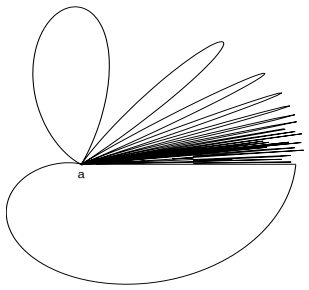

Figura 4: Conjunto estrellado con un único punto radiante a.

A continuación se define una condición de forma, que aparece de manera natural en la demostración de los Teoremas 4.2 y 4.3 y que excluye este tipo de situaciones.

Sea $S \subset \mathbb{R}^d$ un conjunto compacto estrellado, de interior no vacío, y $a \in S$ un punto radiante de S. Sea $\epsilon_0 > 0$. Se define el **índice de apuntamiento** $\rho(S, a, \epsilon_0)$ de S respecto al punto a como el ínfimo de los $\beta \in [0, 1]$ tal que, para todo $x \in S$, $x \neq a$, y para todo $0 < \epsilon \leq \epsilon_0$, la sección circular maximal de la bola $B(x, \epsilon)$ ortogonal al segmento $[a, x]$ (es decir, la intersección de $B(x, \epsilon)$ con el hiperplano afín que pasa por x y es ortogonal a $[a, x]$) está incluida en $S^{\beta\epsilon}$, el conjunto S

"orlado" por una banda de anchura $\beta\epsilon \leq \epsilon$.

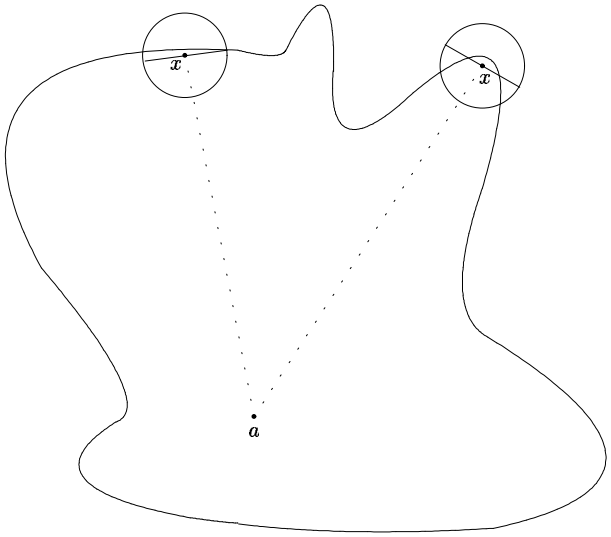

Figura 5: Conjunto estrellado con índice de apuntamiento menor que 1

Por ejemplo, si $S = [-1,1] \times [-1,1]$ y $a = (0,0)$, entonces $\rho(S,a,\epsilon_0) = \dfrac{1}{\sqrt{2}}$ para todo $\epsilon_0 \leq 1$.

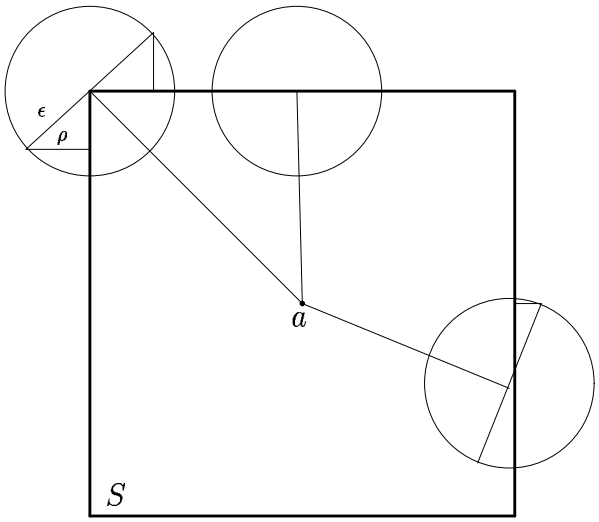

Figura 6

En términos intuitivos, la situación $\rho(S, a, \epsilon_0) = 1$ para todo $\epsilon_0 > 0$ corresponde a un conjunto estrellado rodeado de infinitos picos cada vez más agudos. Para impedir anomalías de este tipo, impondremos condiciones a S que aseguren que su índice de apuntamiento es menor que 1.

En teoría geométrica de conjuntos estrellados se consideran otros casos extremos como, por ejemplo, la presencia de "rayos" de dimensión $d-1$ (ver Figura 7).

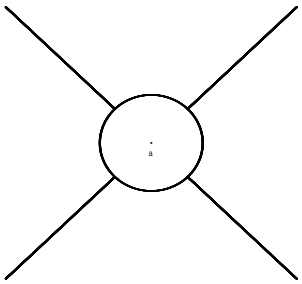

Figura 7

Obsérvese que esta situación queda automáticamente excluida bajo la hipótesis de que S sea el soporte de una distribución absolutamente continua.

En Gardner (1995) se consideran los "cuerpos" estrellados, definidos como conjuntos estrellados $S \subset \mathbb{R}^d$ tales que $S = \mathrm{cl}(\mathrm{int}(S))$. Esta condición excluye situaciones como la de la Figura 7, pero no casos como el de la Figura 4.

A continuación se enuncia un lema que proporciona una condición suficiente muy sencilla para que el índice de apuntamiento sea menor que uno.

LEMA 4.1. *Si $S \subset \mathbb{R}^d$ es un conjunto compacto y estrellado, cuyo mirador contiene una bola $B(a, \delta)$ con $\delta > 0$, entonces existe $\epsilon_0 > 0$ tal que $\rho(S, a, \epsilon_0) < 1$.*

DEMOSTRACIÓN. Podemos suponer, sin pérdida de generalidad, que la bola $B(a, \delta)$ está contenida en el interior de S. Dado que S es estrellado, para cada $x \in S$, el conjunto S contiene al cierre convexo $C(x)$ determinado por $B(a, \delta)$ y x. En $\mathbb{R}^2$ este $C(x)$ sería una especie de "cucurucho" de vértice x. Sea x_0 un

punto de S que maximice la distancia al punto a. Entonces, para todo $x \in S$, el "cucurucho" $C(x)$ no será más "puntiagudo" que $C(x_0)$. Es decir, la sección que un hiperplano ortogonal a $[a,x]$ determine en $C(x)$ a distancia l de x siempre contendrá (tras las oportunas traslaciones y rotaciones) a la sección que originaría en $C(x_0)$ un hiperplano ortogonal a $[a,x_0]$ a distancia l de x_0. Por tanto, para $\epsilon_0 < \delta$, $\rho(S,a,\epsilon_0) = \rho(C(x_0),a,\epsilon_0)$ y claramente $\rho(C(x_0),a,\epsilon_0) < 1$.

□

Bajo la hipótesis de que S fuera conexo, hemos sugerido tomar el parámetro de suavizado $\bar{\epsilon}_n$ dado por (4.1), es decir, el ínfimo de los radios con los que $\hat{S}_n$ era conexo. Análogamente, cuando S se supone estrellado definimos, para cada n fijo,

$$\epsilon_n^* = \inf\{\epsilon > 0 : \hat{S}_n(\epsilon) = \bigcup_{i=1}^{n} B(X_i,\epsilon) \text{ es estrellado}\}. \tag{4.5}$$

Obsérvese que, debido a que la conexión es una condición más débil que ser estrellado, ϵ_n^* está acotado inferiormente c.s. por $\bar{\epsilon}_n$, para cada n. Por otro lado, $\epsilon_n^* \leq \text{diam}(S)$ c.s. para todo n, siendo $\text{diam}(S) = \sup_{x,y\in S} \|x-y\|$.

TEOREMA 4.2. *Sea $S \subset \mathbb{R}^d$ un conjunto compacto y estrellado cuyo mirador contiene una bola $B(a,\delta)$ con $\delta > 0$. Sean $X_1,\ldots,X_n$ observaciones i.i.d. de una distribución con soporte S. Entonces*

$$\epsilon_n^* \to 0 \quad c.s.$$

DEMOSTRACIÓN. Tomemos $\epsilon \in (0,\delta)$. De la demostración del Lema 4.1 se sigue que $\rho(S,a,\epsilon) < 1$. Fijemos un valor cualquiera $\beta \in (\rho(S,a,\epsilon),1)$.

Debido a que S es estrellado y $B(a,\delta) \subset \text{mir}(S)$ con $\delta > 0$, se tiene que para cada $\eta > 0$ existe un recubrimiento $B_1,\ldots,B_M$ de S mediante bolas de radio η tales que

$$\mu_L(B_i \cap S) > 0 \text{ para } i = 1,\ldots,M.$$

Fijados ϵ y β, consideremos un recubrimiento de este tipo donde η ha sido elegido de forma que, para cualesquier $x, x_1, x_2 \in B_i$, se tiene

$$B(x, \beta\epsilon) \subset B(x_1, \epsilon) \cup B(x_2, \epsilon).$$

Por tanto, con probabilidad 1, existe n_0 tal que, para cada $n \geq n_0$, existen al menos dos puntos muestrales en cada B_i y, en consecuencia,

$$\mathbf{P}\{\exists n_0 \text{ tal que } S^{\beta\epsilon} \subset \hat{S}_n(\epsilon), \forall n \geq n_0\} = 1.$$

Obsérvese que $S^{\beta\epsilon}$ es estrellado y que el mirador de $S^{\beta\epsilon}$ también contiene a la bola $B(a, \delta)$.

Ahora veamos que, si $S^{\beta\epsilon} \subset \hat{S}_n(\epsilon)$, entonces $\hat{S}_n(\epsilon)$ es estrellado. En efecto, como $S^{\beta\epsilon}$ es estrellado y su mirador contiene a $B(a, \delta)$, todos los puntos de $S^{\beta\epsilon}$ se pueden ver (vía $S^{\beta\epsilon}$) desde cualquier punto de esta bola. Por tanto, sólo tenemos que considerar los puntos de $\hat{S}_n(\epsilon)$ que no pertenecen a $S^{\beta\epsilon}$. Sea z uno de estos puntos. Entonces existe un X_i ($\in S$ c.s.) tal que $z \in B(X_i, \epsilon)$. Obsérvese que $a \neq X_i$ porque $\epsilon < \delta$. Sea $D(X_i)$ la sección circular maximal de esta bola ortogonal a $[a, X_i]$. Como $\beta \geq \rho(S, a, \epsilon)$, tenemos que $D(X_i) \subset S^{\beta\epsilon}$. Utilizando que $B(a, \delta)$ está contenida en el mirador de $S^{\beta\epsilon}$, obtenemos que el "cono truncado" $H(X_i)$, definido como el cierre convexo de $B(a, \delta)$ y $D(X_i)$, está contenido en $S^{\beta\epsilon}$ (ver Figura 8).

El conjunto $H(X_i) \cup B(X_i, \epsilon)$, que es convexo por construcción, está contenido en $\hat{S}_n(\epsilon)$. Esto prueba que existe un n a partir del cual $\hat{S}_n(\epsilon)$ es estrellado c.s.. En resumen, hemos obtenido que, para todo $\epsilon \in (0, \delta)$, existe n_0 tal que, si $n \geq n_0$, entonces $\epsilon_n^* < \epsilon$ c.s.

□

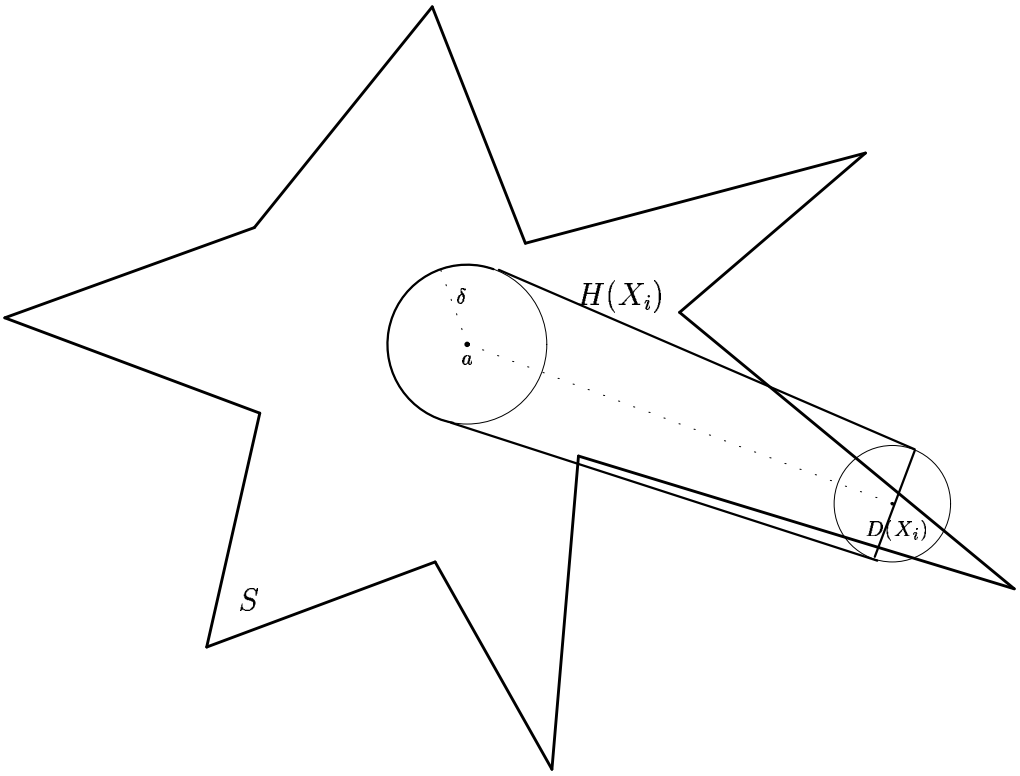

Figura 8

Siempre que $\epsilon_n \to 0$ c.s. (como ocurre con ϵ_n^* bajo las hipótesis del Teorema 4.2) se verifica que

$$d_H(\hat{S}_n(\epsilon_n), S) \to 0 \quad \text{c.s. cuando } n \to \infty.$$

Aunque la distancia de Hausdorff entre dos conjuntos $S, T \subset \mathbb{R}^2$ o $\mathbb{R}^3$ en general se puede visualizar fácilmente, el hecho de que $d_H(S,T)$ sea pequeña en ocasiones no está acompañado de una cercanía "física" entre S y T tan estrecha como en un principio se podría creer. En particular puede ocurrir (incluso en casos muy sencillos) que una sucesión de conjuntos converja a otro en distancia de Hausdorff y, sin embargo, sus fronteras estén muy alejadas de las del límite. La siguiente figura ilustra esta situación.

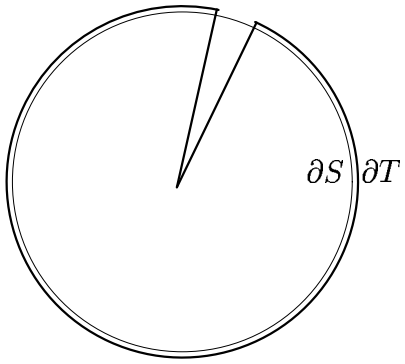

Figura 9

Estas anomalías pueden descartarse imponiendo la convergencia entre las fronteras. Esta idea ha sido considerada en Cuevas y Fraiman (1998) para definir una distancia más flexible que la del supremo entre funciones de densidad.

En el siguiente resultado se prueba la convergencia casi segura, en distancia de Hausdorff, de la frontera del estimador de cubrimiento estrellado $\hat{S}_n$ hacia la frontera de S, cuando el parámetro de suavizado ϵ_n tiende a cero suficientemente despacio.

TEOREMA 4.3. *Sea $S \subset \mathbb{R}^d$ compacto, estrellado y estándar respecto a μ_L. Sean $X_1, \ldots, X_n$ observaciones independientes de una distribución uniforme sobre S. Consideramos el estimador de cubrimiento $\hat{S}_n = \bigcup_{i=1}^n B(X_i, \epsilon_n)$. Supongamos que*

a) $\hat{S}_n$ es estrellado y se verifica que

(i) existe un $d_0 > d$ tal que $\epsilon_n \geq 2\left(\dfrac{\log n}{n}\right)^{1/d_0}$ c.s.

(ii) $\epsilon_n \to 0$ c.s.

b) si denotamos por $\mathcal{N}(\epsilon)$ el mínimo número de bolas con centro en S y radio ϵ que se necesitan para recubrir ∂S, existe un $L > 0$ tal que $\mathcal{N}(\epsilon) \leq \dfrac{L}{\epsilon^{d-1}}$.

Entonces

$$d_H(\partial \hat{S}_n, \partial S) \to 0 \quad c.s. \tag{4.6}$$

DEMOSTRACIÓN. Tomemos, para cada n, un recubrimiento minimal $\{B_{kn}, k = 1, \ldots, \mathcal{N}_n\}$ de ∂S con bolas de radio $\left(\frac{\log n}{n}\right)^{1/d_0}$. Probemos que, con probabilidad 1, existe un n_0 tal que, para todo $n \geq n_0$, se verifica

$$\text{para cada } B_{kn}\,(k = 1, \ldots, \mathcal{N}_n) \text{ existe al menos un } X_i \text{ tal que } X_i \in B_{kn}. \tag{4.7}$$

Esto equivale a probar que

$$\mathbf{P}(\limsup A_n) = 0, \tag{4.8}$$

siendo

$$A_n = \{w : \exists \text{ al menos un } B_{kn} \text{ con } X_i \notin B_{kn},\ i = 1, \ldots, n\}.$$

Por el lema de Borel-Cantelli una condición suficiente para que se verifique (4.8) es que

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty.$$

Observemos en primer lugar que

$$\begin{aligned} \mu_L(B_{kn}) &= C\left(\frac{\log n}{n}\right)^{d/d_0} \\ &= C\left(\frac{\log n}{n}\right)^{\alpha}, \end{aligned}$$

siendo $C > 0$ una constante y $\alpha = \frac{d}{d_0}$ (< 1). Utilizando el hecho de que S es estándar y la hipótesis (b) tenemos que, para una constante $\gamma > 0$,

$$\begin{aligned} \mathbf{P}(A_n) &\leq \mathcal{N}_n \max_{1 \leq k \leq \mathcal{N}_n} \mathbf{P}\{X_i \notin B_{kn}, i = 1, \ldots, n\} \\ &\leq \mathcal{N}_n \left(1 - \gamma\left(\frac{\log n}{n}\right)^{\alpha}\right)^n \\ &\leq 2^{-(d-1)} L \left(\frac{n}{\log n}\right)^{(d-1)/d_0} \left(1 - \gamma\left(\frac{\log n}{n}\right)^{\alpha}\right)^n. \end{aligned}$$

Por tanto,

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) \leq 2^{-(d-1)} L \sum_{n=1}^{\infty} \left(\frac{n}{\log n}\right)^{(d-1)/d_0} \left(1 - \gamma\left(\frac{\log n}{n}\right)^{\alpha}\right)^n.$$

Esta serie es convergente porque puede mayorarse por una del tipo

$$C' \sum n^q \exp(-n^r),$$

con $q, r > 0$, y ésta última converge.

Como consecuencia de (4.7) y de haber supuesto que S y $\hat{S}_n$ son estrellados se verifica que, con probabilidad 1, existe un n_0 a partir del cual $S \subset \hat{S}_n$. En efecto, tomemos n_0 tal que para $n \geq n_0$ se verifica (4.7) y supongamos que existe un $x \in S$ pero $x \notin \hat{S}_n$. Necesariamente $x \in \text{int}(S)$, pues (4.7) implica que todo ∂S está recubierto por bolas del tipo $B(X_i, \epsilon_n)$, es decir, que $\partial S \subset \hat{S}_n$.

Dado que $\hat{S}_n$ es estrellado existe un $y_n \in \text{mir}(\hat{S}_n)$. Prolongamos el segmento $[y_n, x]$ en el sentido $y_n \to x$ hasta que, pasado el punto x, corta a ∂S en un punto z (ver Figura 10).

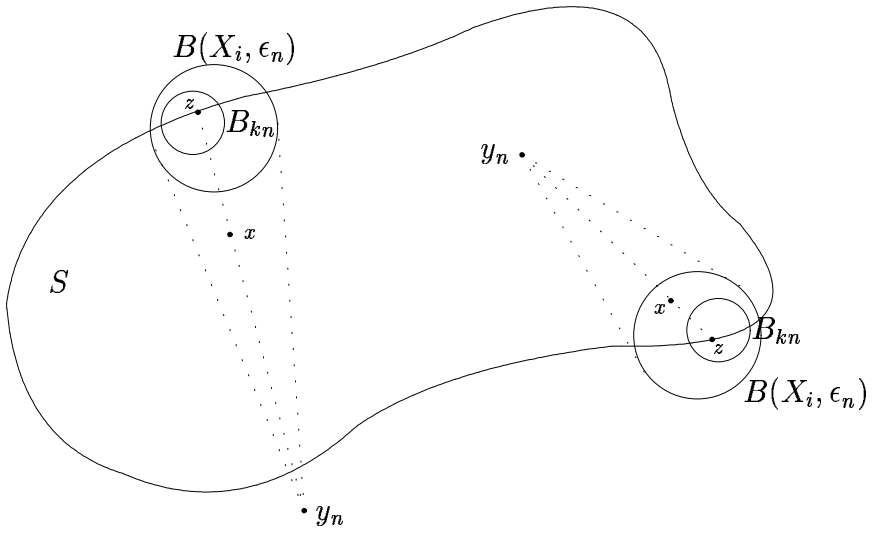

Figura 10

El punto z está contenido en una bola $B_{kn} \subset B(X_i, \epsilon_n) \subset \hat{S}_n$. Por tanto, desde el punto y_n podemos ver, vía $\hat{S}_n$, la bola $B(X_i, \epsilon_n)$ o, lo que es equivalente, el cierre convexo $C(y_n)$ de $\{\{y_n\} \cup B(X_i, \epsilon_n)\}$ está contenido en $\hat{S}_n$ y, por tanto, $x \in \hat{S}_n$, que contradice lo que habíamos supuesto.

Finalmente para terminar de probar que $d_H(\partial\hat{S}_n, \partial S) \to 0$ c.s., supondremos que no se verifica esta convergencia. Entonces hay dos posibilidades:

i) Con probabilidad positiva, existe $\epsilon > 0$ y una subsucesión que denotaremos por $\{x_n\}$, con $x_n \in \partial\hat{S}_n$ y tal que $d(x_n, \partial S) := \sup_{y\in\partial S} \|x_n - y\| > \epsilon$ para todo n.

No puede suceder que existan infinitos $x_n \in \text{int}(S)$ ya que, para $x_n \in \text{int}(S)$, existiría una bola de radio positivo $B(x_n, r_n) \subset S$ y $B(x_n, r_n) \cap \hat{S}_n^c \neq \emptyset$. Esto está en contradicción con el hecho de que, con probabilidad 1, $S \subset \hat{S}_n$ de un n en adelante. Entonces $x_n \in S^c$, que es abierto, a partir de un cierto n y, en consecuencia, $d(x_n, \partial S) \leq \epsilon_n$ c.s. Pero ϵ_n, por hipótesis del teorema, tiende a 0 con probabilidad 1, por lo que hemos llegado a contradicción con lo que supusimos al principio de (i).

ii) Con probabilidad positiva existe $\epsilon > 0$ y una subsucesión que denotaremos $\{x_n\}$ tal que $\{x_n\} \subset \partial S$ y $d(x_n, \partial\hat{S}_n) := \sup_{x\in\partial\hat{S}_n} \|x_n - x\| > \epsilon$. Por (4.7) esto equivale a que $B(x_n, \epsilon) \subset \text{int}(\hat{S}_n)$ para todo n. Como ∂S es un compacto, de $\{x_n\}$ podemos extraer una subsucesión $\{x_{n_j}\}$ convergente a $x \in \partial S$. Por tanto, existe un j_0 tal que si $j \geq j_0$ entonces $\|x_{n_j} - x\| < \epsilon/2$. Es decir, $B(x, \epsilon/2) \subset B(x_{n_{j_0}}, \epsilon)$, luego $B(x, \epsilon/2) \subset \text{int}(\hat{S}_n)$.

Por otro lado, como $x \in \partial S$, existe un $y \in S^c$ tal que $\|x - y\| < \epsilon/4$ y existe $0 < r < \epsilon/4$ tal que $B(y, r) \subset S^c$ y también $B(y, r) \subset B(x, \epsilon/2) \subset \text{int}(\hat{S}_n)$. Tomamos n suficientemente grande para que $\epsilon_n < r/2$ c.s. Entonces, teniendo en cuenta que $B(y, r) \subset S^c$ y que $\hat{S}_n = \bigcup B(X_i, \epsilon_n)$ con $X_i \in S$ c.s., sucede que $y \notin \hat{S}_n$, lo cual está en contradicción con que $B(y, r) \subset \text{int}(\hat{S}_n)$.

Como las dos posibilidades, (i) y (ii), llevan a contradicción, se debe verificar (4.6).

□

Observación 1. Si $S \subset \mathbb{R}^d$ es un conjunto estrellado, una condición suficiente para que S sea estándar respecto a μ_L es que el mirador de S contenga una bola $B(a,\delta)$, con $\delta > 0$.

Para ver esto observemos que, si $x \in S$, entonces el "cucurucho" $C(x)$ determinado por x y $B(a,\delta)$ está contenido en S. Sea $\epsilon > 0$. Si $x \notin B(a,\delta)$ y $\epsilon \leq \delta$, entonces siempre hay un sector de $B(x,\epsilon)$ que está contenido en S. Este sector viene determinado por $B(x,\epsilon) \cap C(x)$. Si x_0 es el punto de ∂S más alejado de a, entonces

$$\begin{aligned}\mu_L(B(x,\epsilon) \cap S) &\geq \mu_L(B(x,\epsilon) \cap C(x)) \geq \mu_L(B(x_0,\epsilon) \cap C(x_0)) \\ &= C_1 \mu_L(B(x_0,\epsilon)) = C_1 \mu_L(B(x,\epsilon))\end{aligned}$$

siendo $C_1 > 0$ una constante.

Si $x \in B(a,\delta)$ y $\epsilon \leq \delta$ entonces, dado cualquier x_1 tal que $\|x_1 - a\| = \delta$,

$$\begin{aligned}\mu_L(B(x,\epsilon) \cap S) &\geq \mu_L(B(x,\epsilon) \cap B(a,\delta)) \geq \mu_L(B(x_1,\epsilon) \cap B(a,\delta)) \\ &= C_2 \mu_L(B(x_1,\epsilon)) = C_2 \mu_L(B(x,\epsilon)),\end{aligned}$$

siendo $C_2 > 0$ una constante.

El recíproco no es cierto, como se puede comprobar en la figura siguiente.

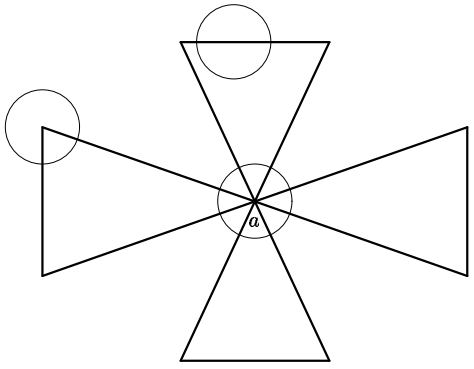

Figura 11: Conjunto estándar cuyo mirador está constituido por un solo punto, a

Observación 2. La hipótesis (a)(i) del Teorema 4.3 no es demasiado restrictiva: por un lado, $\epsilon_n^* \geq \bar{\epsilon}_n$ c.s. ya que la condición de conexión es más débil que la de estrellado y, por otro lado, Penrose (1999) prueba que el orden de convergencia casi segura a cero del radio de conexión $\bar{\epsilon}_n$ es $O\left(\frac{\log n}{n}\right)^{1/d}$. Con (a)(i) sólo estamos pidiendo que el radio ϵ_n del estimador estrellado $\hat{S}_n$ sea un poco más grande que el ínfimo de los radios con los que $\hat{S}_n$ es conexo.

Como se observa en (4.7) la elección de un ϵ_n de estas características conlleva la estimación de S por exceso: con probabilidad uno S estará contenido en $\hat{S}_n$ para n suficientemente grande. En cierto modo es una situación parecida a la que se da en estimación no paramétrica de densidades bajo restricciones de forma. Si se posee información acerca de, por ejemplo, el número de modas o puntos de inflexión de una densidad f, el parámetro de suavizado de un estimador *kernel* que incorpore esa propiedad tiene un orden de convergencia más lento que la amplitud de ventana que minimiza el error L^2 (ver Cao, Cuevas y González-Manteiga 1994). Esta posible pérdida de eficiencia debe interpretarse como el precio que hay que pagar para conseguir estimadores con mejores propiedades de forma y que, al mismo tiempo, verifiquen propiedades adicionales de consistencia (en este caso, convergencia de las fronteras).

Observación 3. La hipótesis (b) del Teorema 4.3 es una condición muy natural. Toranzos (1981) probó que, si $S \subset \mathbb{R}^2$ es un conjunto compacto y estrellado, que contiene a una bola radio positivo, entonces el perímetro de S es finito. Utilizando estas ideas y resultados previos del mismo autor (Toranzos 1967), tal vez se podría establecer una propiedad análoga para los conjuntos $S \subset \mathbb{R}^d$ de las mismas características. En ese caso, parece razonable que la hipótesis (b) pudiera suprimirse.

4.3 Problemas abiertos

En este capítulo se ha presentado la posibilidad de que el estimador de cubrimiento $\hat{S}_n(\epsilon)$ incorpore información geométrica previa acerca de S, simplemente tomando el radio ϵ de manera adecuada. Hemos comprobado que, si S es conexo o estrellado, $\hat{S}_n$ puede heredar cualquiera de estas dos propiedades. Sin embargo, al haber definido $\hat{S}_n$ como una unión de bolas, hay otras propiedades geométricas que nunca va a poder verificar: una de ellas es la convexidad. Un procedimiento alternativo para incorporar la hipótesis de convexidad podría ser tomar el cierre convexo de $\hat{S}_n$, $\mathrm{conv}(\hat{S}_n)$.

Una propiedad de regularidad que tampoco puede cumplir $\hat{S}_n$ es que una bola de radio $r > 0$ (por muy pequeño que sea r) "ruede libremente" por el exterior de $\hat{S}_n$ (ver capítulo 1 y Walther 1997, 1999). Sería interesante utilizar la metodología de Walther (1997), que suaviza $\hat{S}_n$ mediante "granulometrías", para obtener estimadores del tipo:

$$\psi_{-r}(\hat{S}_n) := \hat{S}_n \oplus rB \ominus rB.$$

Parece razonable esperar que un estimador de este tipo (bajo la restricción indicada) verifique un resultado de convergencia de fronteras similar al del Teorema 4.3.

Otra propiedad que no puede incorporarse a $\hat{S}_n$ simplemente variando el radio es la "suavidad" (en el sentido de diferenciabilidad) de la frontera de S. En general tiene sentido preguntarse qué otras propiedades usuales de S, aparte de la conexión y la forma estrellada, pueden trasladarse al estimador de Devroye y Wise mediante cambios sencillos como, por ejemplo, la elección de un radio adecuado.

En la Sección 4.1 aparece un algoritmo para el cálculo de $\bar{\epsilon}_n$, el ínfimo de los radios que proporcionan conexión en el estimador de cubrimiento. Un paso previo a cualquier intento de llevar a la práctica el estimador estrellado $\hat{S}_n(\epsilon_n^*)$ sería determinar un algoritmo eficiente para el cálculo de ϵ_n^*.

El Teorema 4.3 proporciona condiciones suficientes para la convergencia de la

frontera del estimador estrellado $\hat{S}_n(\epsilon_n)$ hacia la frontera de S. Las hipótesis de este teorema excluyen conjuntos S con formas extrañas (por ejemplo, con picos cada vez más agudos). Este resultado motiva el interés por identificar clases amplias de conjuntos S y estimadores S_n en las que sean equivalentes la consistencia en medida $d_\mu(S, S_n) \to 0$, la consistencia en distancia de Hausdorff $d_H(S, S_n) \to 0$ y la consistencia de las fronteras en la métrica de Hausdorff $d_H(\partial S, \partial S_n)$. Un conjunto que perteneciera a una de estas clases no debería tener un aspecto demasiado "complicado".

Una idea relacionada con ésta sería clasificar los conjuntos en función de la tasa con la que converja a cero la distancia de Hausdorff entre la frontera del conjunto y la frontera de un estimador. Parece razonable pensar que, cuanto más lenta sea la tasa, más extraña es la forma del conjunto que queremos estimar.

Bibliografía

[1] J. Aitchison and C. G. G. Aitken, "Multivariate binary discrimination by the kernel method," *Biometrika*, vol. 63, no. 3, pp. 413–420, 1976.

[2] M. J. B. Appel and R. P. Russo, "The connectivity of a graph on uniform points in $[0,1]^d$," 1996. Preprint, Univ. Iowa.

[3] A. Baíllo, J. A. Cuesta-Albertos, and A. Cuevas, "Convergence rates in nonparametric estimation of level sets." Manuscript, 1999.

[4] A. Baíllo, A. Cuevas, and A. Justel, "Set estimation and nonparametric detection," 2000. To appear in *Canad. J. Statist.*

[5] V. Bertholet, J. P. Rasson, and S. Lissoir, "About the automatic detection of training sets for multispectral images classification," in *Advances in Data Science and Classification* (M. V. A. Rizzi and H. H. Bock., eds.), Berlin: Springer, 1998.

[6] W. Blaschke, *Kreis und Kugel.* New York: Chelsea, 1949.

[7] M. Breen, "Illumination for unions of boxes in $\mathbb{R}^d$," *Proc. Amer. Math. Soc.*, vol. 116, no. 1, pp. 197–202, 1992.

[8] R. Cao, A. Cuevas, and W. González-Manteiga, "A comparative study of several smoothing methods in density estimation," *Comput. Statist. Data Anal.*, vol. 17, pp. 153–176, 1994.

[9] E. Carlstein, H.-G. Müller, and D. Siegmund, eds., *Change-point Problems*, vol. 23 of *IMS Lecture Notes.* IMS, 1994.

[10] B. Chakraborty and P. Chaudhuri, "A note on the robustness of multivariate medians," *Stat. Prob. Letters*, vol. 45, pp. 269–276, 1999.

[11] K. C. Chanda and F. H. Ruymgaart, "Asymptotic estimate of probability of misclassification for discriminant rules based on density estimates," *Stat. Prob. Letters*, vol. 8, no. 1, pp. 81–88, 1989.

[12] J. Chevalier, "Estimation du support et du contour du support d'une loi de probabilité," *Ann. Inst. H. Poincaré, sec. B*, vol. 12, pp. 339–364, 1976.

[13] N. Choudhuri, "Bayesian bootstrap credible sets for multidimensional mean functional," *Ann. Statist.*, vol. 26, pp. 2104–2127, 1999.

[14] N. A. C. Cressie, *Statistics for Spatial Data*. New York: Wiley, 1991.

[15] A. Cuevas, "On pattern analysis in the nonconvex case," *Kybernetes*, vol. 19, pp. 26–33, 1990.

[16] A. Cuevas, M. Febrero, and R. Fraiman, "Estimating the number of clusters," 2000. To appear in *Canad. J. Statist.*

[17] A. Cuevas and R. Fraiman, "A plug-in approach to support estimation," *Ann. Statist.*, vol. 25, pp. 2300–2312, 1997.

[18] A. Cuevas and R. Fraiman, "On visual distances in density estimation: the Hausdorff choice," *Statist. Probab. Lett.*, vol. 40, pp. 333–341, 1998.

[19] A. Cuevas and W. González-Manteiga, "Data-driven smoothing based on convexity properties," in *Nonparametric Functional Estimation and Related Topics (Spetses, 1990)* (G. Roussas, ed.), pp. 225–240, Dordrecht: Kluwer Academic Publishers, 1991.

[20] A. DasGupta, J. K. Ghosh, and M. M. Zen, "A new general method for constructing confidence sets in arbitrary dimensions: with applications," *Ann. Statist.*, vol. 23, no. 4, pp. 1408–1432, 1995.

[21] L. Davies and U. Gather, "The identification of multiple outliers," *J. Amer. Statist. Assoc.*, vol. 88, pp. 782–801, 1993.

[22] C. Derman and S. M. Ross, *Statistical Aspects of Quality Control.* San Diego: Academic Press, 1997.

[23] L. Devroye, "The equivalence of weak, strong and complete convergence in L^1 for kernel density estimates," *Ann. Statist.*, vol. 11, pp. 896–904, 1983.

[24] L. Devroye, *A Course in Density Estimation.* Boston: Birkhäuser, 1987.

[25] L. Devroye, "Exponential inequalities in nonparametric estimation," in *Nonparametric Functional Estimation and Related Topics (Spetses, 1990)* (G. Roussas, ed.), pp. 31–44, Dordrecht: Kluwer Academic Publishers, 1991.

[26] L. Devroye and L. Györfi, *Nonparametric density estimation: The L^1 view.* New York: John Wiley, 1985.

[27] L. Devroye and G. L. Wise, "Detection of abnormal behavior via nonparametric estimation of the support," *SIAM J. Appl. Math.*, vol. 38, no. 3, pp. 480–488, 1980.

[28] S. W. Dharmadhikari and K. Joag-dev, *Unimodality, Convexity and Applications.* Boston: Academic Press, 1988.

[29] D. L. Donoho, "Wedgelets: nearly minimax estimation of edges," *Ann. Statist.*, vol. 27, no. 3, pp. 859–897, 1999.

[30] L. Dümbgen and G. Walther, "Rates of convergence for random approximations of convex sets," *Adv. Appl. Prob. (SGSA)*, vol. 28, pp. 384–393, 1996.

[31] B. Efron, "The convex hull of a random set of points," *Biometrika*, vol. 52, pp. 331–343, 1965.

[32] M. J. Farrell, "The measurement of productive efficiency," *J. Roy. Statist. Soc. Ser. A*, vol. 120, no. 3, pp. 253–281, 1957.

[33] W. Feller, *An Introduction to Probability Theory and Its Applications, vol. 1.* Wiley, 1968.

[34] R. Fraiman and J. Meloche, "Multivariate L-estimation," *Test*, vol. 8, no. 2, pp. 255–289, 1999.

[35] R. J. Gardner, *Geometric Tomography*, vol. 58 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1995.

[36] J. Geffroy, "Sur un problème d'estimation géométrique," *Publications de l'Institut de Statistique des Universités de Paris*, vol. 13, pp. 191–210, 1964a.

[37] J. Geffroy, "Sur l'estimation d'une fonction numérique positive à partir d'une échantillon de points pris au hasard dans le domain compris entre son graphe et son intervale de définition," *C. R. Acad. Sci. Paris*, vol. 259, pp. 2762–2764, 1964b.

[38] U. Grenander, *Abstract Inference*. New York: Wiley, 1981.

[39] P. Groeneboom, "Limit theorems for convex hulls," *Probab. Theory Related Fields*, vol. 79, pp. 327–368, 1988.

[40] P. Hall, "On estimating the endpoint of a distribution," *Ann. Statist.*, vol. 10, pp. 556–568, 1982.

[41] P. Hall, *Introduction to the Theory of Coverage Processes*. New York: Wiley, 1988.

[42] P. Hall, "On the law of the logarithm for density estimators," *Statist. Probab. Lett.*, vol. 9, pp. 237–240, 1990.

[43] W. Härdle, B. U. Park, and A. B. Tsybakov, "Estimation of non-sharp support boundaries," *J. Multivariate Anal.*, vol. 55, pp. 205–218, 1995.

[44] J. A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.

[45] J. A. Hartigan, "Estimation of a convex density contour in two dimensions," *J. Amer. Statist. Assoc.*, vol. 82, pp. 267–270, 1987.

[46] L. Holmström and J. Klemelä, "Asymptotic bounds for the expected L^1 error of a multivariate kernel density estimator," *J. Multivariate Anal.*, vol. 42, no. 2, pp. 245–266, 1992.

[47] I. Hueter, "The convex hull of a normal sample," *Adv. in Appl. Probab.*, vol. 26, pp. 855–875, 1994.

[48] R. Z. Khasminskii and V. S. Lebedev, "On the properties of parametric estimators for areas of a discontinuous image," *Problems Control Inform. Theory*, vol. 19, no. 5/6, pp. 375–385, 1990.

[49] J. Kim and D. Pollard, "Cube root aymptotics," *Ann. Statist.*, vol. 18, no. 1, pp. 191–219, 1990.

[50] A. Kneip, B. U. Park, and L. Simar, "A note on the convergence of nonparametric DEA efficiency measures," 1996. Discussion paper. Institut de Statistique. Université Catholique de Louvain.

[51] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*, vol. 82 of *Lecture Notes in Statistics*. New York: Springer-Verlag, 1993.

[52] L. Lebart, A. Morineau, and K. M. Warwick, *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: Wiley, 1984.

[53] R. Y. Liu, "Control charts for multivariate processes," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1380–1387, 1995.

[54] E. Mammen and A. B. Tsybakov, "Asymptotical minimax recovery of sets with smooth boundaries," *Ann. Statist.*, vol. 23, no. 2, pp. 502–524, 1995.

[55] D. Marr, *Vision*. San Francisco: W. H. Freeman and Co., 1982.

[56] G. Matheron, *Random Sets and Integral Geometry*. New York: Wiley, 1975.

[57] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics, 1989 (Norwich, 1989)*, vol. 141 of *London Mathematical Society*

Lecture Notes Series, pp. 148–188, Cambridge: Cambridge University Press, 1989.

[58] I. S. Molchanov, "On distributions of random closed sets and expected convex hulls," *Statist. Probab. Lett.*, vol. 17, pp. 253–257, 1993a.

[59] I. S. Molchanov, *Limit Theorems for Unions of Random Closed Sets.* Berlin: Springer-Verlag, 1993b.

[60] I. S. Molchanov, "A limit theorem for solutions of inequalities," *Scand. J. Statist.*, vol. 25, pp. 235–242, 1998.

[61] D. C. Montgomery, *Introduction to Statistical Quality Control.* New York: Wiley, 1985.

[62] M. Moore, "On the estimation of a convex set," *Ann. Statist.*, vol. 12, pp. 1090–1099, 1984.

[63] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, pp. 1379–1387, 1986.

[64] D. W. Müller, "The excess mass approach in statistics," *Beiträge zur Statistik*, vol. 3, 1993. Univ. Heidelberg.

[65] D. W. Müller and G. Sawitzki, "Excess mass estimates and tests of multimodality," *J. Amer. Statist. Assoc.*, vol. 86, pp. 738–746, 1991.

[66] E. A. Nadaraya, *Nonparametric Estimation of Probability Densities and Regression Curves.* Dordrecht: Kluwer Academic Publishers, 1989.

[67] D. Nolan, "The excess-mass ellipsoid," *J. Multivariate Anal.*, vol. 39, pp. 348–371, 1991.

[68] M. D. Penrose, "A strong law for the longest edge of the minimal spanning tree," *Ann. Probab.*, vol. 27, no. 1, pp. 246–260, 1999.

[69] A. M. Polansky, "Density control charts," 1999. Manuscript.

[70] M. Pollak and D. Siegmund, "Sequential detecting of a change in a normal mean when the initial value is unknown," *Ann. Statist.*, vol. 19, pp. 394–416, 1991.

[71] W. Polonik, "Measuring mass concentration and estimating density contour clusters- an excess mass approach," *Ann. Statist.*, vol. 23, pp. 855–881, 1995.

[72] W. Polonik, "Concentration and goodness-of-fit in higher dimensions: (Asymptotically) distribution-free methods," *Ann. Statist.*, vol. 27, no. 4, pp. 1210–1229, 1999.

[73] B. L. S. Prakasa Rao, *Nonparametric Functional Estimation.* Academic Press, 1983.

[74] A. Rényi and R. Sulanke, "Über die konvexe Hülle von n zufällig gewählten Punkten," *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, vol. 2, pp. 75–84, 1963.

[75] A. Rényi and R. Sulanke, "Über die konvexe Hülle von n zufällig gewählten Punkten (II)," *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, vol. 3, pp. 138–147, 1964.

[76] B. D. Ripley and J. P. Rasson, "Finding the edge of a Poisson forest," *J. Appl. Prob.*, vol. 14, pp. 483–491, 1977.

[77] R. Schneider, "Random approximation of convex sets," *J. Microscopy*, vol. 151, pp. 211–227, 1988.

[78] R. Schneider, *Convex Bodies: the Brunn-Minkowski Theory*, vol. 44 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, 1993.

[79] D. W. Scott, *Multivariate Density Estimation.* New York: Wiley, 1992.

[80] J. Serra, *Image Analysis and Mathematical Morphology.* London: Academic Press, 1982.

[81] A. N. Shiryaev, *Probability.* Berlin: Springer, 2nd ed., 1984.

[82] B. W. Silverman, *Density estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.

[83] J. S. Simonoff, *Smoothing Methods in Statistics.* Berlin: Springer-Verlag, 1996.

[84] S. R. Sternberg, "Grayscale morphology," *Computer Vision, Graphics, and Image Processing*, vol. 35, pp. 333–355, 1986.

[85] D. Stoyan, "Random sets: models and statistics," *Internat. Statist. Rev.*, vol. 66, pp. 1–27, 1998.

[86] W. Stute, "A law of the logarithm for kernel density estimators," *Ann. Probab.*, vol. 10, no. 2, pp. 414–422, 1982.

[87] W. Stute, "The oscillation behavior of empirical processes: the multivariate case," *Ann. Probab.*, vol. 12, no. 2, pp. 361–379, 1984.

[88] E. Tabakis, "On the longest edge of the minimal spanning tree," in *From Data to Knowledge. Studies in Classification, Data Analysis and Knowledge Organization* (W. Gaul and D. Pfeifer, eds.), pp. 222–230, Berlin: Springer-Verlag, 1996.

[89] F. A. Toranzos, "Radial functions of convex and star-shaped bodies," *Amer. Math. Monthly*, vol. 74, pp. 278–280, 1967.

[90] F. A. Toranzos, "Aproximación de conjuntos estrellados compactos por familias especiales," *Rev. Un. Mat. Argentina*, vol. 29, pp. 49–54, 1979.

[91] F. A. Toranzos, "Perímetro de figuras estrelladas," *Math. Notæ*, vol. 29, pp. 95–100, 1981/82.

[92] A. B. Tsybakov, "On nonparametric estimation of density level sets," *Ann. Statist.*, vol. 25, no. 3, pp. 948–969, 1997.

[93] R. Vitale, "Expected convex hulls, order statistics and Banach space probabilities," *Acta Appl. Math.*, vol. 9, pp. 97–102, 1987.

[94] G. Walther, "Granulometric smoothing," *Ann. Statist.*, vol. 25, no. 6, pp. 2273–2299, 1997.

[95] G. Walther, "On a generalization of Blaschke's Rolling Theorem and the smoothing of surfaces," *Math. Meth. Appl. Sci.*, vol. 22, no. 4, pp. 301–316, 1999.

[96] M. P. Wand and M. C. Jones, *Kernel Smoothing.* London: Chapman and Hall, 1995.

[97] B. Yakir, "On the average run length to false alarm in surveillance problems which possess an invariance structure," *Ann. Statist.*, vol. 26, no. 3, pp. 1198–1214, 1998.

Índice de Materias