Trabajo Fin de Máster

Estimación de filamentos

Héctor González Vázquez

Máster en Técnicas Estadísticas

Curso 2024-2025

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación de filamentos
Título en español: Estimación de filamentos
English title: Filament estimation
Modalidad: Modalidad A
Autor: Héctor González Vázquez, Universidade de Santiago de Compostela
Directores: Beatriz Pateiro López, Universidade de Santiago de Compostela; Alberto Rodríguez Casal, Universidade de Santiago de Compostela
Breve resumen del trabajo: En muchas situaciones prácticas los datos se encuentran concentrados en una estructura de dimensión inferior que el espacio ambiente. Esta situación de reducción de la dimensión es la que justifica el empleo de técnicas clásicas del análisis multivariante como el análisis de componentes principales. Si la estructura subyacente no es lineal y se aborda de forma no paramétrica el problema recibe el nombre de manifold estimation. En el caso de que el espacio ambiente sea el espacio euclídeo bidimensional y la estructura a recuperar sea una curva, el problema recibe el nombre de estimación de filamentos. El objetivo de este trabajo fin de máster consistirá en revisar la literatura existente sobre este problema, programar los métodos existentes, y proponer mejoras en los algoritmos. Finalmente, se aplicarán las técnicas estudiadas a un problema del ámbito industrial.
Otras observaciones: Este TFM es una propuesta del estudiante Héctor González Vázquez.

Doña Beatriz Pateiro López, Titular de Universidad de la Universidade de Santiago de Compostela, y don Alberto Rodríguez Casal, Catedrático de Universidad de la Universidade de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Estimación de filamentos

fue realizado bajo su dirección por don Héctor González Vázquez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, doña Beatriz Pateiro López, don Alberto Rodríguez Casal y don Héctor González Vázquez

√ sí ☐ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En Santiago de Compostela, a 13 de enero de 2025.

PATEIRO LOPEZ BEATRIZ - 46898344X
Firmado digitalmente por PATEIRO LOPEZ BEATRIZ - 46898344X
Fecha: 2025.01.13 19:20:53 +01'00'

La directora:
Doña Beatriz Pateiro López

RODRIGUEZ CASAL ALBERTO - 34896829X
Firmado digitalmente por RODRIGUEZ CASAL ALBERTO - 34896829X
Fecha: 2025.01.13 18:44:57 +01'00'

El director:
Don Alberto Rodríguez Casal

Firmado por GONZALEZ VAZQUEZ HECTOR - ***4202** el día 13/01/2025

El autor:
Don Héctor González Vázquez

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el autor declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

A mis directores, Bea y Alberto, por ayudarme y guiarme a lo largo de estos cuatro meses de trabajo y conseguir que al final llegara.

A mis compañeros de la sala π y de café, por acompañarme durante mis inicios en la investigación en tantas mañanas y tardes en la facultad.

A mis amigos en Santiago y en Ourense, porque sin ellos desde que empecé el grado no habría llegado hasta aquí.

A mi familia, especialmente mis padres y mi hermana, por ser una fuente de apoyo permanente e incondicional.

Índice general

Resumen

Resumen en español

La estimación de variedades permite abordar de modo no lineal y no paramétrico el problema de reducción de la dimensión al trabajar con datos en un espacio euclídeo que realmente se distribuyen en (o cerca de) una variedad de dimensión menor, proporcionando una mejor comprensión sobre su estructura subyacente. En el caso particular en el que la variedad es una curva, el problema se denomina estimación de filamentos. El objetivo de este trabajo es proponer un nuevo estimador de filamentos y probar que alcanza la tasa óptima en el sentido minimax de convergencia en distancia de Hausdorff, salvo factor logarítmico, cuando el espacio ambiente es el plano. Primero se realiza una presentación de conceptos, condiciones de forma y estimadores empleados en estimación de conjuntos. A continuación, se revisa un estimador, el llamado estimador EDT (*Euclidean Distance Transform*), en un modelo de estimación de filamentos con ruido aditivo. Además, se presenta un modelo de ruido perpendicular, en un contexto más general de estimación de variedades, en el que se conoce la tasa minimax. Finalmente, se propone el nuevo estimador, denominado estimador EDT con envoltura r–convexa, y se prueba su tasa de convergencia. También se estudia la posible selección del parámetro de forma r a partir de los datos sin afectar a la tasa de convergencia. El estimador propuesto se aplica a un problema de estimación del contorno de secciones de troncos de árboles en inventario forestal.

English abstract

Manifold estimation allows a non–linear and non–parametric dimension reduction when working with data in an euclidean space that are actually supported on (or close to) a lower dimension manifold, providing a better understanding on their underlying structure. In the particular case when the manifold is a curve, the problem is known as filament estimation. The aim of this work is to propose a new filament estimator that achieves the optimal rate in minimax sense of convergence in Hausdorff distance, up to logarithmic factor, when the ambient space is the plane. First, an introduction on concepts, shape conditions and estimators used in set estimation is presented. Next, the so–called EDT (Euclidean Distance Transform) estimator, in a filament estimation model with additive noise, is revised. A perpendicular noise model, in a more general manifold estimation context, in which the minimax rate is known, is also presented. Lastly, the new estimator, called the EDT estimator with r–convex hull, is proposed, and its convergence rate is obtained. We also study a possible choice on the shape parameter r from the data without affecting the convergence rate. The proposed estimator is applied to a tree stem cross section estimation problem in forest inventory.

Capítulo 1

Introducción

La estimación de conjuntos hace referencia al problema estadístico de estimar un conjunto o una característica del mismo a partir de una muestra aleatoria de puntos cuya distribución está relacionada con él. Es un campo de estudio relativamente reciente dentro de la estadística y de gran actividad y aplicación práctica, en el que la geometría juega un papel importante. Para una panorámica general sobre el tema, se pueden consultar los trabajos de Cuevas (2009) y Cuevas y Fraiman (2010), o bien Cholaquidis (2024) para una revisión más reciente.

Posiblemente el problema más abordado dentro de este área es el de estimación del soporte. Si consideramos una distribución de probabilidad $\mathbb{P}_X$ en $\mathbb{R}^d$ y disponemos de una muestra de puntos $\{X_1, \ldots, X_n\}$, habitualmente independientes e idénticamente distribuidos según la distribución anterior, el problema de estimación del soporte consiste en recuperar el soporte S de la distribución, que se define como el menor conjunto cerrado tal que $\mathbb{P}_X(S) = 1$, y que se asume no vacío y compacto. El interés en estimar el soporte de una distribución es claro: es la región donde hay datos.

Otro ejemplo de conjuntos estudiados en la literatura son los llamados conjuntos de nivel. Supongamos que la distribución $\mathbb{P}_X$ es absolutamente continua con densidad f. Para cada t, se define el conjunto de nivel t como $G(t) = \{x \in \mathbb{R}^d \colon f(x) \geq t\}$. Los conjuntos de nivel son zonas que concentran una mayor densidad de probabilidad dentro del soporte S, lo que los hace más adecuados para algunas aplicaciones en las que no son de tanta relevancia las zonas del soporte con poca densidad de datos. Además, son de interés para el análisis clúster, pues los clústeres se pueden definir como las componentes conexas de los conjuntos de nivel.

En ocasiones, no es suficiente con reconstruir el conjunto S, sino que necesitamos ser capaces de estimar también su frontera ∂S. Esto ocurre en aplicaciones prácticas, como en análisis de imágenes, donde la silueta o el borde del objeto es de gran importancia, pues es la que determina la forma del mismo o sus límites. La estimación de la frontera es una tarea que resulta en general más difícil que la estimación del propio conjunto. Aunque dos conjuntos estén próximos (en una cierta distancia entre conjuntos), sus fronteras pueden ser muy distintas.

También hay situaciones donde el objetivo no es el propio conjunto, sino que puede ser alguna característica del mismo. Algunos ejemplos son el volumen del conjunto, es decir, su medida de Lebesgue (área en dos dimensiones, volumen en tres dimensiones,...), o la medida de su frontera (por ejemplo el área de su superficie), cuantificada mediante el llamado contenido de Minkowski.

El problema en el que centraremos nuestra atención en esta memoria es el de estimación de filamentos, un caso particular de estimación de conjuntos. En general, en el contexto de estimación de variedades, también denominado *manifold learning*, habitualmente disponemos de un conjunto de observaciones $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ en un espacio euclídeo de una cierta dimensión d que en realidad se

distribuyen formando una estructura de dimensión menor pues, o bien la distribución que siguen tiene soporte en una variedad de dimensión $k < d$ embebida en el espacio ambiente, lo que se corresponde con un modelo sin ruido, o bien "cerca" de dicha variedad, en el caso de un modelo con ruido. El interés en estimar la variedad viene justificado por la reducción en la dimensión. Los datos pueden encontrarse en un espacio ambiente de dimensión muy alta, de modo que su representación en dimensión menor permite reducir el coste computacional en tiempo y espacio de su análisis, a la vez que ayuda a entender la estructura intrínseca de los mismos manteniendo sus características geométricas más importantes. Clásicamente la reducción de la dimensión se llevaba a cabo de modo lineal mediante el análisis de componentes principales (PCA), proyectando los datos sobre subespacios lineales de dimensión menor. Las técnicas de *manifold learning* permiten abordar esta tarea de modo no lineal y no paramétrico, asumiendo simplemente que la estructura subyacente es una variedad de dimensión menor embebida en el espacio ambiente, véase Meilă y Zhang (2024) para una revisión al respecto.

Cuando la variedad es de dimensión $k = 1$, nos referimos a ella como filamento o curva y el problema se denomina estimación de filamentos. Hay numerosas aplicaciones donde los datos se distribuyen formando estructuras filamentosas. Una de las más importantes surge en cosmología, donde las estrellas y las galaxias en el universo se distribuyen en filamentos que forman la llamada red cósmica, y hay gran cantidad de literatura que investiga sobre las propiedades topológicas de estas estructuras. En sismología, los terremotos se disponen formando una red de filamentos alrededor de las fallas geológicas. En medicina, en imágenes médicas, los vasos sanguíneos y los tejidos forman también estructuras de carácter filamentoso. Por su parte, en aplicaciones de detección remota, son de gran importancia las estructuras en las que se distribuyen los cauces fluviales o las redes de carreteras. Otra de las aplicaciones donde se emplea la detección remota es en inventario forestal, que se puede definir como el proceso de recolección de información forestal para el posterior análisis y estimación de características como el estado del ecosistema, el valor comercial de la madera o el riesgo de incendios. Una de las variables de interés en este ámbito es el diámetro de un árbol a la altura del pecho. Tradicionalmente, este valor se medía con aparatos manuales. Recientemente se han empleado sistemas de detección remota para obtener nubes de puntos de los árboles de una zona de estudio. A partir de estas nubes se pueden obtener secciones horizontales de los troncos de los árboles y aplicar distintos modelos estadísticos, asumiendo una forma circular de estas secciones, para estimar el diámetro. No obstante, la forma de los troncos en muchas ocasiones no es cilíndrica, por lo que puede resultar de interés estimar de modo no paramétrico las curvas que determinan el contorno de estas secciones, véase Figura 1.1.

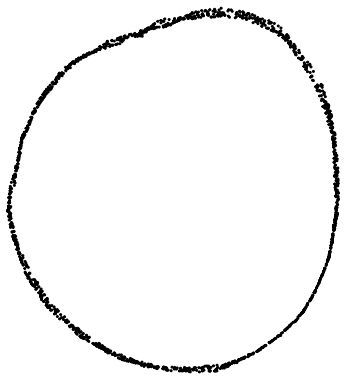

Figura 1.1: Nube de puntos obtenida de una sección horizontal de un tronco.

De entre los muchos métodos estadísticos que se pueden emplear en la estimación de filamentos, Genovese *et al.* (2012a) proponen un estimador, llamado estimador EDT (*Euclidean Distance Transform*), en un modelo de estimación de filamentos con ruido. En este modelo aditivo cada dato se genera sobre el filamento pero después es "distorsionado" por un ruido esférico, es decir, que tiene

como soporte una bola compacta. Este estimador se basa en emplear un estimador del soporte de la distribución de las observaciones X_i conocido como estimador de Devroye–Wise. Por otra parte, los mismos autores, en Genovese *et al.* (2012b), consideran un modelo de estimación de variedades en el que el ruido es perpendicular a la variedad en el punto en el que se genera el dato. En este modelo obtienen la tasa óptima en el sentido minimax de estimación de la variedad en distancia de Hausdorff (un tipo de distancia entre conjuntos), salvo factor logarítmico.

El estimador de Devroye–Wise, que consiste en una versión suavizada de la muestra, tiene la ventaja de ser bastante general, pero en la literatura de estimación del soporte hay otro tipo de estimadores que presentan mejores tasas de convergencia bajo restricciones de tipo geométrico. En particular, hay un estimador que generaliza a la envoltura convexa de la muestra, conocido como envoltura r–convexa, que es flexible y alcanza mejores tasas que el estimador de Devroye–Wise bajo una condición de forma sobre el soporte denominada r–convexidad. El objetivo de este trabajo es proponer una mejora del estimador EDT de Genovese *et al.* (2012a), basada en el uso de la envoltura r–convexa como estimador del soporte, en el modelo de ruido perpendicular de Genovese *et al.* (2012b) cuando la variedad es un filamento cerrado ($k = 1$), y probar que alcanza la tasa óptima minimax, salvo factor logarítmico, cuando los datos están en el plano, es decir, en un espacio ambiente de dimensión $d = 2$. El estimador que proponemos depende de un parámetro $r > 0$ a utilizar en la envoltura r–convexa, que se puede interpretar como un parámetro de forma, y que sería deseable poder seleccionar de forma automática a partir de los datos para que el estimador sea *data–driven*. Probamos que nuestro estimador permite utilizar un selector del parámetro r basado en los datos sin que la tasa de convergencia se vea afectada. En la Sección 4.4 aplicaremos el estimador propuesto en el contexto del problema de estimación de secciones de troncos en inventario forestal, véase Figura 1.2.

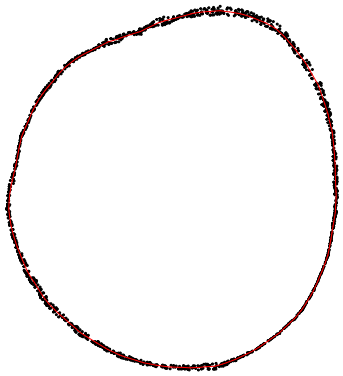

Figura 1.2: Estimación del contorno de una sección horizontal de un tronco a partir del nuevo estimador propuesto.

Esta memoria está organizada de la siguiente forma. El Capítulo 2 está dedicado a una breve revisión de conceptos y resultados teóricos en estimación de conjuntos necesarios para el desarrollo posterior. Tras establecer definiciones y notaciones en la Sección 2.1, la Sección 2.2 define algunas distancias entre conjuntos empleadas en estimación de conjuntos y conceptos de convergencia asociados. En la Sección 2.3 se presentan condiciones de forma sobre conjuntos y estimadores del soporte asociados a dichas restricciones.

En cuanto al Capítulo 3, la Sección 3.1 está dedicada al estimador EDT de Genovese *et al.* (2012a). Se explican el modelo de estimación de filamentos con ruido aditivo y las hipótesis consideradas, para posteriormente presentar el estimador y su tasa de convergencia. Como el estimador EDT es un conjunto de puntos pero no una curva, también se explica la propuesta de Genovese *et al.* (2012a) para extraer una curva dentro de este estimador. En la Sección 3.2 se presenta el modelo de estimación de variedades con ruido perpendicular de Genovese *et al.* (2012b) y la tasa minimax de estimación de la

variedad en distancia de Hausdorff en dicho modelo.

Finalmente, en el Capítulo 4 proponemos el nuevo estimador de filamentos, el estimador EDT con envoltura r–convexa. En la Sección 4.1 explicamos el modelo y las hipótesis que consideramos. En la Sección 4.2 definimos el estimador y probamos su tasa de convergencia, que es la tasa óptima en el sentido minimax cuando la dimensión del espacio ambiente es $d = 2$, salvo factor logarítmico. La Sección 4.3 aborda posibles elecciones del parámetro de forma r y de otro parámetro δ del que depende el estimador EDT con envoltura r–convexa, así como una propuesta de extracción de curva dentro del estimador. Además, se discute el papel de las hipótesis y el modelo en la aplicación práctica del estimador. Para terminar, en la Sección 4.4 aplicamos el nuevo estimador al problema de estimación de secciones de troncos de árboles en inventario forestal.

Capítulo 2

Preliminares en estimación de conjuntos

Este capítulo está dedicado a presentar distintos conceptos y herramientas matemáticas en estimación de conjuntos, así como resultados teóricos sobre ellos. El objetivo no es hacer una revisión exhaustiva sobre el tema, sino simplemente introducir los elementos necesarios para poder abordar en los capítulos posteriores el problema de estimación de filamentos. Buena parte del capítulo sigue conceptos y notación del Capítulo 1 en Pateiro-López (2008).

En primer lugar, en la Sección 2.1 se fija notación y se definen algunas operaciones sobre conjuntos. La Sección 2.2 introduce dos distancias entre conjuntos, la distancia de Hausdorff y la distancia en medida, así como conceptos relacionados con la convergencia de conjuntos. Para terminar, en la Sección 2.3 se presentan distintas restricciones de forma sobre conjuntos y algunos estimadores del soporte asociados a dichas condiciones, junto con resultados teóricos sobre sus tasas de convergencia.

2.1. Notación y definiciones previas

De aquí en adelante nos centraremos en subconjuntos del espacio euclídeo $\mathbb{R}^d$, si bien muchos de los conceptos posteriores se pueden generalizar a otro tipo de espacios, por ejemplo métricos.

Denotamos por $\|\cdot\|$ a la norma en $\mathbb{R}^d$. Dados $a, c \in \mathbb{R}^d$, la distancia entre ambos es $d(a,c) = \|a-c\|$. Definimos la distancia de un punto $a \in \mathbb{R}^d$ a un conjunto no vacío $C \subset \mathbb{R}^d$ como

$$d(a, C) = \text{ínf}\{d(a,c)\colon c \in C\}.$$

Las bolas abierta y cerrada de centro $a \in \mathbb{R}^d$ y radio $r > 0$ se denotan por $\mathring{B}(a,r)$ y $B(a,r)$, respectivamente. En el caso en el que el centro sea 0 y el radio 1, utilizamos la notación $\mathring{B} = \mathring{B}(0,1)$ y $B = B(0,1)$. Dado un conjunto $A \subset \mathbb{R}^d$, A^c, $\text{int}(A)$, $\overline{A}$ y ∂A denotan, respectivamente, el complementario, el interior, la clausura y la frontera de A en $\mathbb{R}^d$.

Comenzamos definiendo las operaciones de suma y resta de Minkowski, que jugarán un papel importante a lo largo de todo el desarrollo.

Definición 2.1 (Operaciones de Minkowski)**.** Sean $A, C \subset \mathbb{R}^d$. Definimos la suma de Minkowski por

$$A \oplus C = \{a + c\colon a \in A, c \in C\}$$

y la resta o sustracción de Minkowski por

$$A \ominus C = \{x \in \mathbb{R}^d\colon \{x\} \oplus C \subset A\}.$$

Dado $\lambda \in \mathbb{R}$, definimos

$$\lambda C = \{\lambda c \colon c \in C\}.$$

Notemos que, según la definición anterior, dado $r > 0$, podemos escribir $r\mathring{B} = \mathring{B}(0, r)$ y $rB = B(0, r)$. Las operaciones de Minkowski están relacionadas con los operadores morfológicos de dilatación y erosión, que tienen su origen en la teoría de la morfología matemática. La morfología matemática es una teoría que trata el análisis de la forma de estructuras espaciales y tiene numerosas aplicaciones en análisis de imágenes, véase Serra (1984) para una referencia clásica sobre este tema. En concreto, los operadores morfológicos actúan sobre un conjunto A haciéndolo interaccionar con otro conjunto C de forma conocida denominado elemento estructurante. Los operadores morfológicos más importantes son los de dilatación y erosión de un conjunto por un elemento estructurante. Aunque se pueden definir para cualquier elemento estructurante, para nuestros objetivos es suficiente definirlos en el caso en el que el elemento estructurante es una bola abierta $r\mathring{B}$ o cerrada rB.

Definición 2.2 (Dilatación y erosión)**.** Dado $r > 0$, se define la dilatación de un conjunto $A \subset \mathbb{R}^d$ por la bola abierta $r\mathring{B}$ como el conjunto

$$A \oplus r\mathring{B} = \bigcup_{a \in A} \mathring{B}(a, r).$$

Se define la erosión del conjunto A por la bola abierta $r\mathring{B}$ como

$$A \ominus r\mathring{B} = \{a \in \mathbb{R}^d \colon \mathring{B}(a, r) \subset A\}.$$

Análogamente, se definen la dilatación y la erosión de A por la bola cerrada rB como $A \oplus rB$ y $A \ominus rB$, respectivamente, esto es, cambiando las bolas abiertas por cerradas en las definiciones anteriores.

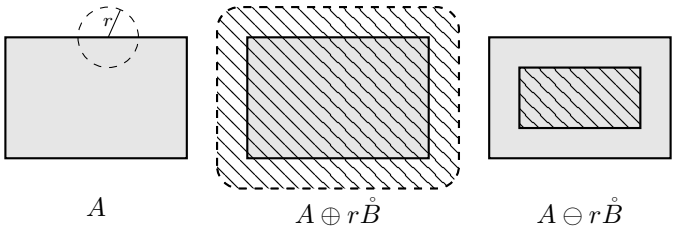

Figura 2.1: Dilatación y erosión de un conjunto A por una bola abierta $r\mathring{B}$. Reproducción de una figura de Pateiro-López (2008).

Informalmente, la dilatación "infla" el conjunto mientras que la erosión lo "contrae", véase Figura 2.1. Cabe mencionar que, en general, la dilatación y la erosión de un conjunto A por un elemento estructurante C no coinciden con $A \oplus C$ y con $A \ominus C$, aunque sí en el caso de que C sea una bola, el que se contempla en la definición anterior y el único que necesitaremos. En ese caso, la dilatación de un conjunto por una bola abierta $\varepsilon\mathring{B}$ o cerrada εB, respectivamente, se conoce asimismo con el nombre de ε–entorno (abierto o cerrado, respectivamente) de A o también con el nombre de conjunto ε–paralelo (abierto o cerrado, respectivamente) de A, y se denota por

$$\mathring{B}(A, \varepsilon) = \{x \in \mathbb{R}^d \colon d(x, A) < \varepsilon\} = A \oplus \varepsilon\mathring{B}$$

o

$$B(A, \varepsilon) = \{x \in \mathbb{R}^d \colon d(x, A) \leq \varepsilon\} = A \oplus \varepsilon B,$$

respectivamente. Como $d(\cdot, A)$ es continua, los conjuntos anteriores son abierto y cerrado, respectivamente.

2.2. Distancias entre conjuntos y convergencia

Para evaluar la calidad de la estimación de un conjunto S mediante un estimador $\widehat{S}_n \equiv \widehat{S}_n(\mathcal{X}_n)$, que depende de la muestra $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, y establecer resultados asintóticos, necesitamos una medida de la discrepancia entre ambos conjuntos. Habitualmente esto se lleva a cabo utilizando una distancia entre conjuntos. A continuación, presentamos dos de las distancias más utilizadas en estimación de conjuntos: la distancia de Hausdorff y la distancia en medida.

Definición 2.3 (Distancia de Hausdorff)**.** Dados $A, C \subset \mathbb{R}^d$ no vacíos y compactos, definimos la distancia de Hausdorff entre ambos como

$$d_H(A, C) = \text{máx}\left\{\sup_{a\in A} d(a, C), \sup_{c\in C} d(c, A)\right\}.$$

Equivalentemente, la distancia de Hausdorff entre A y C es

$$d_H(A, C) = \text{ínf}\{r > 0\colon A \subset C \oplus r\mathring{B} \text{ y } C \subset A \oplus r\mathring{B}\} = \text{ínf}\{r > 0\colon A \subset C \oplus rB \text{ y } C \subset A \oplus rB\}.$$

La distancia de Hausdorff mide de alguna manera la proximidad "visual" o "física" entre dos conjuntos. Informalmente, representa la menor cantidad que tenemos que dilatar cada uno para que contenga al otro. Equivalentemente, es la menor distancia tal que todo punto de uno de los conjuntos está a menos que esa distancia de algún punto del otro conjunto. La Figura 2.2 representa la distancia de Hausdorff entre un punto y un conjunto. Es una distancia habitualmente utilizada en análisis de imágenes o en teoría de fractales. Al restringir d_H a conjuntos no vacíos y acotados aseguramos que está bien definida y, si además se restringe a conjuntos compactos, como en la definición anterior, se puede probar que es una métrica, ver Teorema 2.4.1 en Edgar (1990).

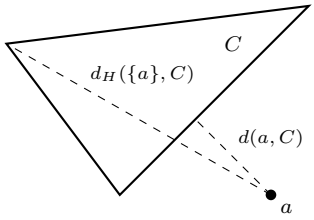

Figura 2.2: Distancia de Hausdorff entre un punto y un conjunto. Reproducción de una figura de Pateiro-López (2008).

Por su parte, la distancia en medida cuantifica como de similar es el contenido de dos conjuntos dados. En particular, cuál es la medida de su diferencia simétrica, definida como

$$A\Delta C = (A \setminus C) \cup (C \setminus A),$$

es decir, los puntos que pertenecen a uno de los conjuntos pero no al otro. Habitualmente se considera el espacio de medida $(\mathbb{R}^d, \mathcal{B}, \mu)$, donde $\mathcal{B}$ es la σ–álgebra de Borel en $\mathbb{R}^d$ y μ es la medida de Lebesgue en $\mathbb{R}^d$, aunque se puede sustituir por otro espacio de medida cualquiera.

Definición 2.4 (Distancia en medida)**.** Dados $A, C \in \mathcal{B}$ acotados, la distancia en medida entre A y C se define como

$$d_\mu(A, C) = \mu(A\Delta C).$$

Cabe mencionar que la distancia en medida no es una métrica, pues, por ejemplo, dos conjuntos que difieran en un conjunto finito de puntos distan cero en medida, a pesar de ser distintos.

En otras ocasiones, resulta más útil considerar, en vez de la medida de Lebesgue, una medida de probabilidad $\mathbb{P}$ en $\mathbb{R}^d$, definiendo $d_{\mathbb{P}}(A, C) = \mathbb{P}(A \Delta C)$. Si la medida de probabilidad tiene asociada una densidad f, la distancia se puede escribir como $d_{\mu_f}(A, C) = \int_{A\Delta C} f(x)dx$. En cualquier caso, representa la probabilidad de que una variable aleatoria con esa distribución de probabilidad tome un valor en uno de los dos conjuntos pero no en el otro.

De aquí en adelante nos interesaremos principalmente por la distancia de Hausdorff a la hora de evaluar la proximidad entre dos conjuntos. La primera condición que podemos exigir a un estimador $\widehat{S}_n$ de un conjunto S es que sea consistente en la distancia escogida. Esto quiere decir que el estimador converja (cuando $n \to \infty$), en dicha distancia, al conjunto que pretende estimar, esto es, que se cumpla

$$d_H(\widehat{S}_n, S) \to 0. \tag{2.1}$$

Como el estimador $\widehat{S}_n$ es aleatorio, pues depende de la muestra, la convergencia en (2.1), puede garantizarse en varios sentidos. Una opción es la convergencia en probabilidad. No obstante, a lo largo del desarrollo teórico, centraremos nuestra atención en resultados de convergencia con probabilidad uno, también denominada de forma casi segura (*almost surely* o *a.s.*). Así, un estimador que verifique (2.1) con probabilidad uno diremos que es consistente en distancia de Hausdorff (o d_H–consistente) casi seguro o con probabilidad uno.

Aunque dos conjuntos estén próximos en distancia de Hausdorff (e incluso aunque lo estén también en distancia en medida), puede que sus formas sean muy diferentes. En muchas de las aplicaciones de la estimación de conjuntos, es importante estimar adecuadamente no solo el propio conjunto, sino también la frontera del mismo, que es la que contiene la información sobre su forma. De este modo, habitualmente se pide al estimador no solo que cumpla (2.1), sino también una convergencia entre fronteras, que normalmente se formaliza nuevamente en términos de la distancia de Hausdorff, pidiendo

$$d_H(\partial\widehat{S}_n, \partial S) \to 0. \tag{2.2}$$

En algunas ocasiones, esta doble condición dada por (2.1) y (2.2) se conoce con el nombre de convergencia completa, véase Cuevas *et al.* (2012).

Una vez garantizada la convergencia de un estimador, es habitual estudiar su tasa de convergencia, que representa la velocidad con la que converge al conjunto a estimar. La formalización de esta idea se hace a través del concepto de O grande. Decimos que una cierta sucesión de números reales no negativos $\{a_n\}$ es del orden de otra sucesión $\{r_n\}$ (cuando $n \to \infty$), y escribimos $a_n = O(r_n)$, si existe una constante $C > 0$ y un índice n_0 tal que $a_n \leq Cr_n$ para todo $n \geq n_0$. Habitualmente estamos pensando que $\{a_n\}$ y $\{r_n\}$ convergen a 0, y en ese caso lo anterior representa que la tasa de convergencia de a_n a 0 es al menos tan rápida como r_n, asintóticamente hablando.

En el caso de estimación de conjuntos, la sucesión en la que estamos interesados es $d_H(\widehat{S}_n, S)$ (o $d_H(\partial\widehat{S}_n, \partial S)$), que es aleatoria, por lo que habitualmente se establecen resultados del tipo

$$d_H(\widehat{S}_n, S) = O(r_n) \quad \text{con probabilidad uno,}$$

y decimos que $\widehat{S}_n$ converge en distancia de Hausdorff (o que es d_H–convergente) a S con tasa r_n casi seguro (o con probabilidad uno). Cuanto más pequeña sea la tasa, más rápida es la convergencia del estimador.

Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, y una sucesión de sucesos $\{A_n\} \subset \mathcal{F}$, decimos que A_n ocurre con probabilidad uno para n suficientemente grande (*eventually almost surely* o e.a.s.) si

$$\mathbb{P}\left(\liminf_{n\to\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \bigcap_{k\geq n} A_k\right) = 1.$$

Informalmente, A_n ocurre e.a.s. si, con probabilidad uno, existe un índice n_0 a partir del cual A_n siempre se cumple. Por ejemplo, decir que $d_H(\widehat{S}_n, S) = O(r_n)$ con probabilidad uno equivale a decir que $d_H(\widehat{S}_n, S) \leq Cr_n$ e.a.s.

De aquí en adelante, en caso de que no se mencione explícitamente, la convergencia de estimadores y las correspondientes tasas se entenderán de modo casi seguro o con probabilidad uno. También se sobreentenderá que estamos considerando la distancia de Hausdorff, salvo que se diga lo contrario.

2.3. Condiciones de forma y estimadores

En estimación de conjuntos, no podemos pretender ser capaces de estimar de forma eficiente cualquier conjunto $S \subset \mathbb{R}^d$ no vacío y compacto a partir de una muestra de puntos usando métodos estadísticos. Dentro de la familia anterior hay conjuntos extremadamente complicados, posiblemente imposibles de representar o imaginar, y por esta razón se suelen imponer algunas restricciones de forma sobre los conjuntos a estimar. Habitualmente, estas condiciones tienen una naturaleza geométrica y son fácilmente interpretables. Establecer este tipo de restricciones nos permite emplear estimadores más sofisticados que aprovechen mejor esta información geométrica y que obtengan mejores tasas de convergencia.

Consideremos el problema de estimación del soporte introducido en el Capítulo 1, esto es, el problema de estimar el soporte $S \subset \mathbb{R}^d$ no vacío y compacto de una distribución de probabilidad absolutamente continua $\mathbb{P}_X$ a partir de una muestra aleatoria simple $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ de la misma. Aunque para el desarrollo teórico sobre estimación de filamentos que haremos en los capítulos siguientes necesitaremos resultados relativos a la estimación del soporte, las condiciones de forma que presentaremos también han sido utilizadas en otras ramas de la estimación de conjuntos, como en estimación de conjuntos de nivel.

2.3.1. Caso general: estimador de Devroye–Wise

Evidentemente, la primera posibilidad es no imponer ninguna condición sobre el conjunto S a estimar. Una idea inicial es considerar como estimador la propia muestra $\mathcal{X}_n$, que se puede probar que es consistente en distancia de Hausdorff, esto es, $d_H(\mathcal{X}_n, S) \to 0$ con probabilidad uno. No obstante, la muestra no es consistente en distancia en medida y tampoco se cumple, en general, que su frontera, que es la propia $\mathcal{X}_n$, converja a la frontera de S.

Por este motivo, un estimador más razonable es el propuesto en Devroye y Wise (1980), que consiste en una versión suavizada de la muestra. Concretamente, el conocido como estimador de Devroye–Wise se define como

$$\widehat{S}_n = \mathcal{X}_n \oplus \varepsilon_n B = \bigcup_{i=1}^{n} B(X_i, \varepsilon_n), \tag{2.3}$$

donde ε_n es una sucesión de parámetros de suavizado o radios que típicamente tiende a 0.

En la Figura 2.3 se representa el estimador de Devroye–Wise en una muestra de $n = 2000$ puntos de una distribución uniforme en la tercera iteración de un copo de nieve de Koch[1] para valores decrecientes del radio ε. Para la obtención del estimador de Devroye–Wise (y también de la envoltura r–convexa que veremos en la Definición 2.11) de una muestra de puntos en el plano en R se puede emplear el paquete `alphahull` (Pateiro-López y Rodríguez-Casal, 2010). Si ε es demasiado grande, obtenemos una estimación sobresuavizada, mientras que si es demasiado pequeño la estimación es infrasuavizada.

[1] https://en.wikipedia.org/wiki/Koch_snowflake

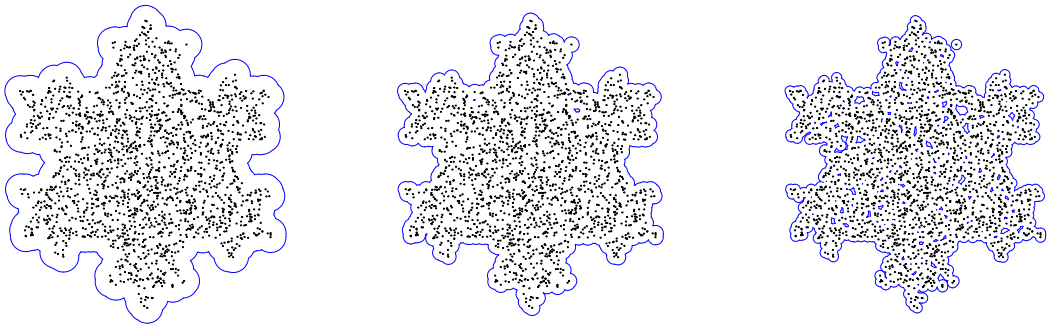

Figura 2.3: De izquierda a derecha, estimador de Devroye–Wise con $\varepsilon = 0.06$, 0.03 y 0.02.

La condición $\varepsilon_n \to 0$ es suficiente para garantizar la d_H–consistencia de $\widehat{S}_n$. Sin embargo, no lo es para asegurar la convergencia entre fronteras pues, si ε_n es demasiado pequeño, el estimador está demasiado fragmentado, tiene demasiados "agujeros" y su frontera no converge a ∂S. En Cuevas y Rodríguez-Casal (2004) se prueba que, además de $\varepsilon_n \to 0$ casi seguro, la condición $S \subset \widehat{S}_n$ e.a.s. permite garantizar la consistencia de $\partial\widehat{S}_n$ casi seguro. Para obtener las tasas de convergencia del estimador de Devroye–Wise y de su frontera, Cuevas y Rodríguez-Casal (2004) piden dos condiciones de forma sobre el soporte S, que definimos a continuación.

Definición 2.5 (Conjunto estándar)**.** Un conjunto de Borel $S \in \mathcal{B}$ se dice (δ, λ)–estándar con respecto a una medida de Borel ν si existen constantes $\delta > 0$ y $\lambda > 0$ tales que

$$\nu(B(x,\varepsilon) \cap S) \geq \delta\mu(B(x,\varepsilon)) \qquad \text{para todos } x \in S \text{ y } 0 < \varepsilon \leq \lambda, \tag{2.4}$$

donde μ es la medida de Lebesgue en $\mathbb{R}^d$.

En el caso particular de que $\nu = \mu$, la condición de estandaridad evita que el conjunto S tenga "picos muy puntiagudos". Si ν es una medida de probabilidad, por ejemplo la de la variable aleatoria según la cual se distribuye la muestra, la condición impide que sea la densidad de esa variable la que presente este tipo de picos o salientes. El supremo δ_S de los $\delta > 0$ para los que se satisface (2.4) para algún $\lambda > 0$ se puede interpretar como una medida de cómo de "puntiagudo" es el conjunto S, y se conoce como constante de estandaridad de S. A menor δ_S, más puntiagudo es el conjunto.

Definición 2.6 (Conjunto parcialmente expandible)**.** Un conjunto de Borel $S \in \mathcal{B}$ acotado se dice (R, r)–parcialmente expandible si existen constantes $R \geq 1$ y $r > 0$ tales que

$$d_H(\partial S, \partial(S \oplus \varepsilon B)) \leq R\varepsilon \qquad \text{para } 0 \leq \varepsilon < r. \tag{2.5}$$

Si se satisface (2.5) para todo $\varepsilon \geq 0$, entonces S se dice R–expandible.

La condición de expandibilidad (parcial) es de algún modo complementaria a la de estandaridad, en el sentido de que impide que el conjunto S presente "entrantes" o "golfos" demasiado profundos. El supremo r_0, posiblemente infinito, de los $r > 0$ para los que se cumple (2.5) para algún $R \geq 1$, por su parte, se interpreta de la siguiente forma: cuanto mayor es r_0, menor profundidad de entrantes presenta S. Un valor menor de la constante R también implica una mayor regularidad del conjunto, de modo que los conjuntos más regulares son aquellos con $r_0 = \infty$ y $R = 1$, que resultan ser los conjuntos convexos. Véase Cuevas y Rodríguez-Casal (2004) para más información sobre la interpretación de estas dos condiciones.

El Teorema 2.7 (Cuevas y Rodríguez-Casal, 2004, Teorema 4) proporciona la tasa de convergencia del estimador de Devroye–Wise y de su frontera bajo las dos restricciones de forma anteriores. En particular, prueba que $d_H(\widehat{S}_n, S)$ y $d_H(\partial\widehat{S}_n, \partial S)$ son $O((\log(n)/n)^{1/d})$ casi seguro.

Teorema 2.7 (Teorema 4 en Cuevas y Rodríguez-Casal, 2004)**.** *Sea $\mathcal{X}_n = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ una muestra aleatoria simple de una distribución $\mathbb{P}_X$, cuyo soporte es S. Supongamos que S es compacto, (R, r)–parcialmente expandible y (δ, λ)–estándar con respecto a $\mathbb{P}_X$. Consideremos el estimador de Devroye–Wise $\widehat{S}_n = \mathcal{X}_n \oplus \varepsilon_n B$, con*

$$\varepsilon_n = C\left(\frac{\log(n)}{n}\right)^{1/d} \qquad \text{para alguna } C > \left(\frac{2}{\delta\omega_d}\right)^{1/d},$$

donde $\omega_d = \mu(B)$ es la medida de Lebesgue de la bola unidad en $\mathbb{R}^d$. Entonces, se verifican e.a.s.

$$d_H(\widehat{S}_n, S) \leq \varepsilon_n \qquad y \qquad d_H(\partial\widehat{S}_n, \partial S) \leq R\varepsilon_n.$$

Notemos que, en el teorema anterior, ε_n es a la vez el radio usado en el estimador de Devroye–Wise y la cota del error que se comete en la estimación del soporte S y de su frontera ∂S. La tasa que proporciona el teorema es una manifestación de la conocida como maldición de la dimensionalidad, habitual en estimación no paramétrica; cuanto mayor es la dimensión d, peor es la tasa de convergencia.

2.3.2. Convexidad y envoltura convexa

Una de las primeras restricciones geométricas que se consideraron en la estimación de conjuntos fue la convexidad, cuya definición recordamos a continuación.

Definición 2.8 (Convexidad)**.** Un conjunto $S \subset \mathbb{R}^d$ se dice convexo si para todos $x, y \in S$ y para todo $\lambda \in [0, 1]$ se cumple $\lambda x + (1 - \lambda)y \in S$.

Definición 2.9 (Envoltura convexa)**.** Dado un conjunto $S \subset \mathbb{R}^d$, se define su envoltura convexa, $\text{conv}(S)$, como la intersección de todos los convexos que lo contienen.

Un conjunto S es convexo si y sólo si $S = \text{conv}(S)$. Si el conjunto S a estimar es convexo, parece razonable emplear un estimador que incorpore esa información y también lo sea, en particular, el estimador más intuitivo es la envoltura convexa de la muestra $\widehat{S}_n = \text{conv}(\mathcal{X}_n)$, pues es el menor convexo que la contiene. En Dümbgen y Walther (1996) se prueba que en este caso la tasa de convergencia de $\widehat{S}_n$ en distancia de Hausdorff es $O((\log(n)/n)^{1/d})$ casi seguro, que coincide con la del estimador de Devroye–Wise. Además, bajo una condición de regularidad en la frontera de ∂S, que resulta ser la misma que se presentará en el Teorema 2.19, la tasa mejora a $O((\log(n)/n)^{2/(d+1)})$. No entramos en mayores detalles, pues la estimación de conjuntos convexos no será de nuestro interés en el desarrollo posterior.

2.3.3. Generalizando la condición de convexidad: envoltura r–convexa

La hipótesis de convexidad sobre el soporte puede ser demasiado restrictiva en muchas aplicaciones prácticas. Cuando S no es convexo, emplear la envoltura convexa de la muestra como estimador es una mala opción, pues si, por ejemplo, el conjunto tiene agujeros en su interior, el estimador tenderá a rellenarlos. Por ello, es preferible emplear condiciones de forma y estimadores más generales y flexibles. A lo largo de este apartado introduciremos distintas restricciones que, de alguna forma, generalizan la noción de convexidad, y que están relacionadas entre ellas. A veces se conocen como condiciones de tipo convexidad (*convexity–type*) o de tipo rodamiento (*rolling–type*).

La primera condición que estudiaremos es la r–convexidad. Antes vamos a definir unos operadores morfológicos relacionados con las operaciones de dilatación y erosión presentadas en la Definición 2.2: el cierre y la apertura de un conjunto con respecto a un elemento estructurante. En particular, cuando el elemento estructurante es una bola, el caso que nos interesa, están relacionados con el concepto de envoltura r–convexa, que veremos en la Definición 2.11.

Definición 2.10 (Cierre y apertura)**.** Dado $r > 0$, se define el cierre de un conjunto $A \subset \mathbb{R}^d$ con respecto a la bola abierta $r\mathring{B}$ como

$$(A \oplus r\mathring{B}) \ominus r\mathring{B}.$$

Se define la apertura de A con respecto a la bola abierta $r\mathring{B}$ como

$$(A \ominus r\mathring{B}) \oplus r\mathring{B}.$$

Análogamente, se definen el cierre y la apertura de A con respecto a la bola cerrada rB como $(A \oplus rB) \ominus rB$ y $(A \ominus rB) \oplus rB$, respectivamente, esto es, cambiando las bolas abiertas por cerradas en las definiciones anteriores.

El cierre intenta recuperar la forma original de un conjunto que ha sido dilatado mediante su erosión, y siempre contiene al conjunto, por ejemplo para la bola abierta se tiene $A \subset (A \oplus r\mathring{B}) \ominus r\mathring{B}$, ver Figura 2.4. Por su parte, la apertura hace lo contrario, dilata un conjunto previamente erosionado, y está contenida en el conjunto, es decir, $(A \ominus r\mathring{B}) \oplus r\mathring{B} \subset A$, ver Figura 2.5. Como la dilatación y la erosión no son operaciones inversas, el cierre y la apertura no coinciden, en general, con el conjunto original. Cuando un conjunto coincide con su cierre con respecto a una bola, decimos que el conjunto es morfológicamente cerrado con respecto a esa bola. Análogamente, si el conjunto coincide con su apertura con respecto a una bola, decimos que es morfológicamente abierto con respecto a esa bola.

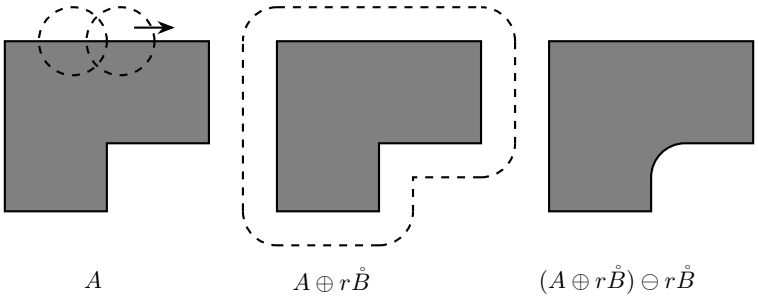

Figura 2.4: Cierre de un conjunto A con respecto a $r\mathring{B}$. Reproducción de una figura de Pateiro-López (2008).

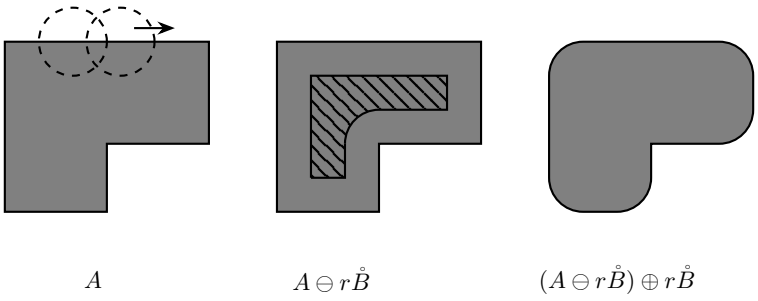

Figura 2.5: Apertura de un conjunto A con respecto a $r\mathring{B}$. Reproducción de una figura de Pateiro-López (2008).

El cierre y la apertura son en cierto modo duales, y verifican diversas propiedades. Por ejemplo, el cierre del complementario coincide con el complementario de la apertura

$$(A^c \oplus r\mathring{B}) \ominus r\mathring{B} = ((A \ominus r\mathring{B}) \oplus r\mathring{B})^c \tag{2.6}$$

y la apertura del complementario coincide con el complementario del cierre

$$(A^c \ominus r\mathring{B}) \oplus r\mathring{B} = ((A \oplus r\mathring{B}) \ominus r\mathring{B})^c. \tag{2.7}$$

Se puede comprobar que la apertura de A con respecto a $r\mathring{B}$ coincide con la unión de las bolas abiertas de radio r contenidas en A, esto es,

$$(A \ominus r\mathring{B}) \oplus r\mathring{B} = \bigcup_{\mathring{B}(y,r)\subset A} \mathring{B}(y,r). \tag{2.8}$$

Por su parte, aplicando (2.7), (2.8) y las leyes de De Morgan, deducimos que el cierre de A con respecto a $r\mathring{B}$ es

$$(A \oplus r\mathring{B}) \ominus r\mathring{B} = \bigcap_{\mathring{B}(y,r)\cap A=\emptyset} \mathring{B}(y,r)^c,$$

es decir, la intersección de complementarios de bolas abiertas que no intersecan al conjunto. Justamente el conjunto anterior es lo que se conoce como envoltura r–convexa de A, y un conjunto r–convexo es aquel que coincide con su envoltura r–convexa.

Definición 2.11 (r–convexidad y envoltura r–convexa)**.** Se dice que $S \subset \mathbb{R}^d$ es r–convexo, para $r > 0$, si $S = C_r(S)$, donde definimos la envoltura r–convexa de S como

$$C_r(S) = \bigcap_{\mathring{B}(y,r)\cap S=\emptyset} \mathring{B}(y,r)^c = (S \oplus r\mathring{B}) \ominus r\mathring{B}.$$

La envoltura r–convexa también se conoce como cierre r–convexo. Notemos que de la definición anterior se deduce que todo conjunto r–convexo es cerrado, pues la envoltura r–convexa es cerrada al ser intersección de cerrados. La r–convexidad generaliza de modo directo la convexidad para conjuntos cerrados. Un conjunto convexo cerrado se puede expresar como la intersección de todos los hiperplanos cerrados que lo contienen, mientras que en el caso de la r–convexidad sustituimos los hiperplanos por complementarios de bolas abiertas de radio r que no intersecan al conjunto, véase Figura 2.6. Dicho de otro modo, un conjunto es r–convexo si cualquier punto que no pertenece al conjunto se puede separar de él por una bola abierta de radio r. Todo conjunto convexo y cerrado es r–convexo para todo $r > 0$, pero no todo conjunto r–convexo para algún $r > 0$ es convexo. No obstante, si $\text{int}(\text{conv}(S)) \neq \emptyset$, sí que se tiene la equivalencia S es convexo y cerrado si y sólo si S es r–convexo para todo $r > 0$, véase Walther (1999).

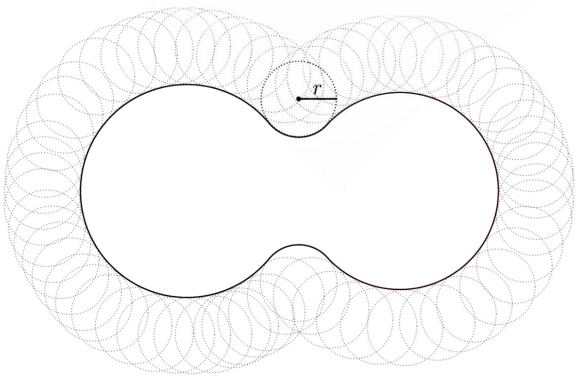

Figura 2.6: Un conjunto r–convexo coincide con su envoltura r–convexa. Figura tomada de Saavedra-Nieves (2015).

De modo análogo a lo que ocurre con la envoltura convexa, la envoltura r–convexa $C_r(S)$ es el menor r–convexo que contiene a S. Así, S es r–convexo si y sólo si $C_r(S) \subset S$, pues el contenido recíproco siempre se cumple. La envoltura r–convexa es monótona respecto a r, en el sentido de que si $r \leq \widetilde{r}$ entonces $C_r(S) \subset C_{\widetilde{r}}(S)$. Por tanto, si S es $\widetilde{r}$–convexo, también es r–convexo para todo $r \leq \widetilde{r}$. A medida que $r \to \infty$, $C_r(S)$ se parece más a la envoltura convexa $\mathrm{conv}(S)$, mientras que cuando $r \to 0$, $C_r(S)$ se acerca a la clausura $\overline{S}$. Cuanto mayor es el valor de r en la condición de r–convexidad, más "regular" es el conjunto S.

Si el soporte S a estimar es r–convexo, el estimador natural es la envoltura r–convexa de la muestra $C_r(\mathcal{X}_n)$, que cumple $\mathcal{X}_n \subset C_r(\mathcal{X}_n) \subset S$. En la práctica, el uso de este estimador depende de la elección del parámetro r, que se puede interpretar como un parámetro de forma, y que habitualmente es desconocido. En la Figura 2.7 se representa la envoltura r–convexa de una muestra de $n = 2000$ puntos de una distribución uniforme en la tercera iteración de un copo de nieve de Koch (c.f. Figura 2.3) para valores decrecientes de r. Si el valor de r es demasiado grande, el estimador se acerca mucho a la envoltura convexa de la muestra, mientras que si es demasiado pequeño produce un estimador con muchos huecos. No obstante, el rol del parámetro r no es el de un parámetro de suavizado, en el sentido de que no debe tender a cero, sino que basta con que sea menor que el correspondiente parámetro poblacional. En la Sección 4.3.1 se profundiza más en este aspecto y se aborda la elección de este parámetro a partir de la muestra; sin embargo, por ahora, supondremos conocido el valor de r.

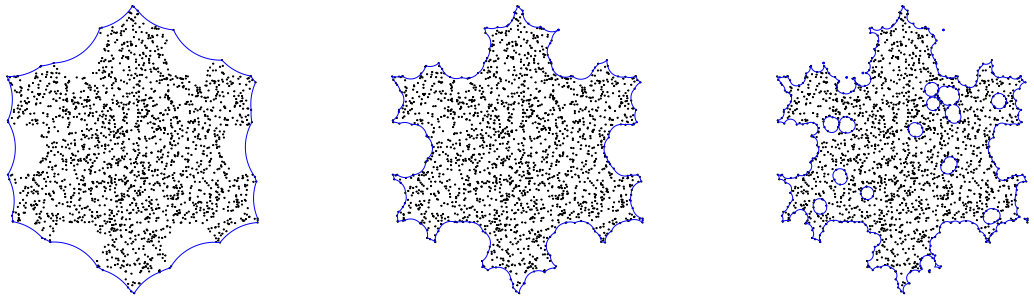

Figura 2.7: De izquierda a derecha, envoltura r–convexa $C_r(\mathcal{X}_n)$ con $r = 0.2$, 0.05 y 0.025.

En Rodríguez-Casal (2007) se prueba que, bajo la hipótesis de que el soporte S sea estándar con respecto a la distribución $\mathbb{P}_X$ de los datos, cualquier estimador $\widehat{S}_n$ que cumpla $\mathcal{X}_n \subset \widehat{S}_n \subset S$ e.a.s. tiene tasa de convergencia $d_H(\widehat{S}_n, S) = O((\log(n)/n)^{1/d})$ casi seguro, en particular la envoltura r–convexa $\widehat{S}_n = C_r(\mathcal{X}_n)$ cuando S es r–convexo. No obstante, al igual que ocurría con la envoltura convexa, la tasa se puede mejorar si se consideran restricciones de suavidad adicionales sobre el soporte. Con este objetivo, a continuación presentamos otras condiciones de forma relacionadas con la r–convexidad.

Definición 2.12 (*Reach*)**.** Dado $S \subset \mathbb{R}^d$, siguiendo la notación de Federer (1959), sea $\mathrm{Unp}(S)$ el conjunto de puntos $x \in \mathbb{R}^d$ que tienen una única proyección en S, es decir, un único punto que minimiza $d(x, S)$. Para $x \in S$, definamos $\mathrm{reach}(S, x) = \sup\{r > 0 \colon \mathring{B}(x, r) \subset \mathrm{Unp}(S)\}$. Se define el *reach* (o alcance) de S como

$$\mathrm{reach}(S) = \text{ínf}\{\mathrm{reach}(S, x) \colon x \in S\}.$$

Informalmente, el *reach* de S es el mayor r tal que todos los puntos a distancia menor que r de S tienen una única proyección en S. Una condición de regularidad sobre un conjunto S es pedir que tenga *reach* positivo (en ocasiones decimos simplemente que S tenga *reach*). Cuanto mayor es el *reach*, más regular es el conjunto. De hecho, los conjuntos convexos tienen *reach* infinito.

Definición 2.13 (r*–rolling*)**.** Dado $r > 0$, se dice que $S \subset \mathbb{R}^d$ satisface la condición de r*–rolling* (por fuera) si para todo $x \in \partial S$ existe $p \in S^c$ tal que $x \in B(p, r)$ y $\mathring{B}(p, r) \cap S = \emptyset$.

La condición de *r–rolling* formaliza la idea de rodamiento por fuera, es decir, que una bola de radio r ruede por el exterior del conjunto S. Hay otra condición de rodamiento también frecuente en la literatura, en este caso por dentro del conjunto, que definimos a continuación.

Definición 2.14 (Rodamiento libre interior)**.** Dado $r > 0$, decimos que la bola rB (la bola de radio r) rueda libremente dentro de un conjunto cerrado $S \subset \mathbb{R}^d$ si para todo $x \in \partial S$ existe $p \in S$ tal que $x \in B(p, r) \subset S$.

Muchas veces nos referiremos a las dos Definiciones 2.13 y 2.14 anteriores como condiciones de rodamiento, aunque cuando sea necesario precisaremos cuál de las dos estamos considerando. Una condición de rodamiento interior puede ser que rB ruede libremente dentro de S, pero también que S^c cumpla la *r–rolling* (véase Figura 2.8), y en general estas dos condiciones no son equivalentes. Análogamente, podemos formalizar el rodamiento exterior mediante la condición de *r–rolling* sobre S o bien pidiendo que rB ruede libremente dentro de $\overline{S^c}$ (tomamos clausura a S^c para garantizar que sea cerrado). El Lema 2.15 establece que, cuando se trata del rodamiento interior, el rodamiento libre es una condición más fuerte que la *r–rolling*.

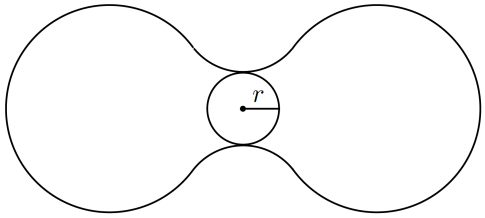

Figura 2.8: Representación del rodamiento interior, que se puede formalizar mediante el rodamiento libre de la bola rB dentro del conjunto S o mediante la condición de *r–rolling* sobre S^c. Figura tomada de Saavedra-Nieves (2015).

Lema 2.15. *Sea $S \subset \mathbb{R}^d$ cerrado y no vacío. Dado $r > 0$, si rB rueda libremente dentro de S entonces S^c satisface la r–*rolling*.*

Demostración. Dado $x \in \partial(S^c) = \partial S$, por la condición de rodamiento libre, existe $p \in S$ tal que $x \in B(p, r) \subset S$. Ahora bien, $S = (S^c)^c$ y $B(p, r) \subset S$ implica $\mathring{B}(p, r) \cap S^c = \emptyset$, de modo que S^c verifica la *r–rolling*. □

Las Proposiciones 1 y 2 en Cuevas *et al.* (2012), que resumimos en el Teorema 2.16, relacionan la *r*–convexidad, el *reach* y la *r–rolling*.

Teorema 2.16 (Proposiciones 1 y 2 en Cuevas *et al.*, 2012)**.** *Sea $S \subset \mathbb{R}^d$ un conjunto compacto.*

1. *Si* $\text{reach}(S) \geq r > 0$*, entonces S es r–convexo.*
2. *Si S es r–convexo para $r > 0$, entonces S satisface la r–*rolling*.*

El segundo punto en el Teorema 2.16 pone de manifiesto que la *r*–convexidad está íntimamente relacionada con el rodamiento exterior. Por otro lado, cabe mencionar que los dos recíprocos del teorema anterior no son ciertos, es decir, existen conjuntos *r*–convexos que no tienen *reach* r y existen conjuntos que cumplen la *r–rolling* que no son *r*–convexos, véase Cuevas *et al.* (2012) para más detalles.

En el desarrollo posterior será importante que se satisfaga una condición de doble rodamiento, es decir, una condición de rodamiento por fuera y por dentro de un conjunto. Siguiendo a Rodríguez-Casal

y Saavedra-Nieves (2016, 2022a,b), definimos la siguiente condición (R^r_λ), que considera un radio de rodamiento exterior, r, y otro de rodamiento interior, λ, ver Figura 2.9.

Definición 2.17 (R^r_λ)**.** Dados $r > 0$ y $\lambda > 0$, decimos que $S \subset \mathbb{R}^d$ satisface la condición (R^r_λ) si S cumple la r*–rolling* y S^c cumple la λ*–rolling*.

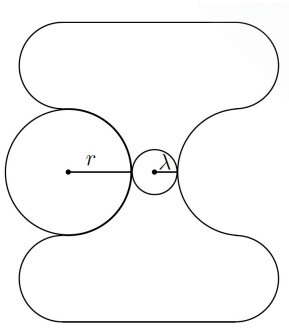

Figura 2.9: Conjunto verificando la condición (R^r_λ). Figura tomada de Saavedra-Nieves (2015).

Finalmente, definimos una clase de conjuntos que cumplen una condición de regularidad relacionada con el cierre y la apertura introducidos en la Definición 2.10: el modelo regular de Serra (Serra, 1984).

Definición 2.18 (Modelo regular de Serra)**.** El modelo regular de Serra se define como la clase de conjuntos $S \subset \mathbb{R}^d$ compactos y morfológicamente abiertos y cerrados con respecto a la bola cerrada rB para algún $r > 0$, es decir, que verifican

$$S = (S \oplus rB) \ominus rB = (S \ominus rB) \oplus rB. \tag{2.9}$$

La primera igualdad en (2.9) nos recuerda a la r–convexidad del conjunto S, cambiando las bolas abiertas por cerradas, mientras que la segunda igualdad está relacionada con la r–convexidad de su complementario mediante la relación (2.6). El Teorema 1 en Walther (1999) caracteriza los conjuntos del modelo regular de Serra en términos de la r–convexidad, del doble rodamiento libre o de la regularidad de su frontera, entre otras condiciones. Más concretamente, a diferencia de la condición (R^r_λ), el doble rodamiento en este caso es con el mismo radio por fuera que por dentro. En Walther (1999) se exige conexión por caminos al conjunto S. Además, se define el rodamiento libre (Definición 2.14) pidiendo que el conjunto $S \ominus rB$ sea conexo por caminos para conservar el significado físico de rodar. No obstante, para nuestros propósitos, es suficiente emplear los resultados sin conexión por caminos, de modo que enunciamos una versión del Teorema 1 en Walther (1999) como aparece en Rodríguez-Casal (2007).

Teorema 2.19 (Teorema 1 en Rodríguez-Casal, 2007)**.** *Sean $S \subset \mathbb{R}^d$ no vacío y compacto. Dado $r > 0$, las siguientes condiciones son equivalentes:*

(I) $S = (S \ominus \lambda B) \oplus \lambda B$ *para todo* $\lambda \in [0, r]$ *y* $S = (S \oplus \lambda B) \ominus \lambda B$ *para todo* $\lambda \in [0, r)$.

(II) S *y* $\overline{S^c}$ *son* r*–convexos e* $\operatorname{int}(S_i) \neq \emptyset$ *para cada componente conexa por caminos* $S_i \subset S$.

(III) La bola λB *rueda libremente dentro de* S *y de* $\overline{S^c}$ *para todo* $\lambda \in [0, r]$.

(IV) ∂S *es una subvariedad de* $\mathbb{R}^d$ $(d-1)$*–dimensional de clase* $\mathcal{C}^1$ *cuyo vector normal unitario exterior,* $n(s)$ *para* $s \in \partial S$*, cumple la condición de Lipschitz*

$$|n(s) - n(t)| \leq \frac{1}{r}|s - t| \qquad \text{para } s, t \in \partial S.$$

Además, S pertenece al modelo regular de Serra si y sólo si existe un $r > 0$ tal que se cumple lo anterior.

El Teorema 2.20 (Rodríguez-Casal, 2007, Teorema 3) prueba que, cuando el soporte S pertenece al modelo regular de Serra, la envoltura r–convexa de la muestra mejora su tasa de convergencia a $O((\log(n)/n)^{2/(d+1)})$. Es necesario poner una hipótesis de acotación inferior sobre la densidad f de la distribución, para evitar que la densidad caiga a 0 en alguna zona del soporte, como por ejemplo cerca de la frontera. Al trabajar con densidades, es habitual considerar acotaciones esenciales, pues la densidad puede modificarse en un conjunto de puntos de medida nula. Recordemos que se definen el ínfimo y el supremo esenciales de una función f en S como

$$\operatorname*{ess\,inf}_{x\in S} f(x) = \sup\big\{a \in \mathbb{R}\colon \mu\big(\{x \in S\colon f(x) < a\}\big) = 0\big\},$$
$$\operatorname*{ess\,sup}_{x\in S} f(x) = \text{ínf}\,\big\{a \in \mathbb{R}\colon \mu\big(\{x \in S\colon f(x) > a\}\big) = 0\big\},$$

respectivamente.

Teorema 2.20 (Teorema 3 en Rodríguez-Casal, 2007)**.** *Sea $\mathcal{X}_n = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ una muestra aleatoria simple de una distribución $\mathbb{P}_X$ absolutamente continua con respecto a la medida de Lebesgue, f la densidad asociada a $\mathbb{P}_X$ y S su soporte. Supongamos que S está en las condiciones del Teorema 2.19 para $r > 0$ y que f está esencialmente acotada inferiormente en S por una constante positiva. Consideremos la envoltura r–convexa de la muestra $\widehat{S}_n = C_r(\mathcal{X}_n)$. Entonces, existe una constante $C_1 > 0$ tal que, si*

$$\varepsilon_n = C_1 \left(\frac{\log(n)}{n}\right)^{2/(d+1)}, \tag{2.10}$$

se verifican e.a.s.

$$d_H(\widehat{S}_n, S) \leq \varepsilon_n, \qquad d_H(\partial\widehat{S}_n, \partial S) \leq \varepsilon_n \qquad y \qquad d_\mu(\widehat{S}_n, S) \leq \varepsilon_n.$$

Capítulo 3

Estimación de filamentos

Como mencionamos en el Capítulo 1, el problema que abordaremos con mayor profundidad en esta memoria es el de estimación de filamentos. En general, en el problema de estimación de variedades, tenemos una variedad M de dimensión k embebida en el espacio euclídeo $\mathbb{R}^d$ y contamos con un conjunto de observaciones $\mathcal{X}_n = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ relacionadas con ella, bien sea porque están directamente sobre la variedad o porque están "cerca" de la misma. En este último caso, habitualmente se considera un modelo de ruido, en el que las observaciones X_i son de la forma

$$X_i = V_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde V_i sigue una cierta distribución en la variedad M y ϵ_i es el ruido. El objetivo es estimar la variedad a partir de las observaciones X_i. Cuando la variedad es una curva de dimensión $k = 1$, estamos en el problema de estimación de filamentos.

En la Sección 3.1 se presenta el modelo de ruido aditivo en estimación de filamentos considerado en Genovese *et al.* (2012a), en el que el ruido tiene como soporte una bola compacta. Además, se define el estimador basado en la EDT (*Euclidean Distance Transform*), un estimador del filamento en tres pasos. Primero, se estima el soporte S de la distribución de las X_i usando el estimador de Devroye–Wise $\widehat{S}_n = \mathcal{X}_n \oplus \varepsilon_n B$ definido en (2.3), y se estima su frontera por $\partial\widehat{S}_n$. Las tasas de convergencia del estimador y de su frontera en distancia de Hausdorff son $O\left((\log(n)/n)^{1/(d+\alpha)}\right)$, donde $\alpha \geq 1/2$ depende de las hipótesis sobre las distribuciones (Teorema 3.7). La constante α indica cómo de rápido decae la densidad de la muestra a cero al aproximarnos a la frontera del soporte, y valdría 0 en el caso más propicio en el que la densidad estuviera acotada inferiormente por una constante positiva, situación que nunca se da en este modelo. En segundo lugar, se construye el estimador EDT $\widehat{\Gamma}_n$, un "conjunto de valores ajustados" definido a partir del estimador del soporte que no es una curva, pero que hereda la tasa de convergencia anterior en distancia de Hausdorff al filamento. Por último, se extrae una curva que está contenida dentro de $\widehat{\Gamma}_n$, que también conserva la tasa de convergencia, y que se propone como estimador del filamento.

En la Sección 3.2 se considera otro modelo, planteado en Genovese *et al.* (2012b), en un contexto de estimación de variedades más general, donde el ruido ϵ_i es perpendicular a la variedad en el punto V_i. En el artículo, bajo las hipótesis consideradas, se obtiene la tasa minimax de estimación de la variedad en distancia de Hausdorff, que resulta del orden de $(1/n)^{2/(k+2)}$, salvo factor logarítmico. Una de las hipótesis del modelo es que la variedad no tenga borde, lo que en el caso de filamentos ($k = 1$) se traduce en que la curva sea cerrada. En tal caso, la tasa minimax para estimar el filamento es $(1/n)^{2/3}$, salvo factor logarítmico.

El estudio de estos dos modelos nos permitirá proponer en el Capítulo 4 un nuevo estimador de filamentos, basado en una mejora del estimador EDT de Genovese *et al.* (2012a), que alcance la tasa

minimax en el modelo de ruido perpendicular de Genovese *et al.* (2012b) cuando la dimensión del espacio ambiente sea $d = 2$.

3.1. El estimador EDT de Genovese *et al.* (2012a)

En esta sección describiremos el modelo aditivo y el estimador basado en la EDT propuestos en Genovese *et al.* (2012a). La mayor parte de la notación sigue la del artículo mencionado, aunque hay otra que se ha cambiado con el objetivo de que sea más coherente con el desarrollo posterior. Las hipótesis y los resultados teóricos también se han estructurado para que sean más fáciles de identificar y de comparar con los de otras secciones de la memoria.

3.1.1. Modelo e hipótesis

Suponemos que tenemos una curva $f : [0,1] \to \mathbb{R}^d$, cuya imagen es el filamento $\Gamma = f([0,1]) \subset \mathbb{R}^d$ que queremos estimar. En ocasiones identificamos la curva con el filamento. Decimos que la curva es abierta si $f(0) \neq f(1)$, mientras que es cerrada si $f(0) = f(1)$. En el modelo considerado disponemos de las observaciones

$$X_i = f(U_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde U_i sigue una distribución H en $[0,1]$, de modo que $V_i = f(U_i)$ está sobre el filamento Γ, y el error ϵ_i sigue una distribución F en $\mathbb{R}^d$, cuya principal característica es que tiene media cero y soporte una bola compacta $B(0,\sigma)$, donde $\sigma > 0$ se conoce como nivel de ruido. Las hipótesis consideradas son las siguientes.

(A) **Sobre la curva**

- (A1) La curva es simple, es decir, f es inyectiva en $(0,1)$.
- (A2) La curva f es continua y tiene gradiente no nulo y finito en cada punto.
- (A3) Se cumple $0 < \sigma < \text{reach}(\Gamma)$.
- (A4) Si f es abierta, entonces $\sigma < \|f(1) - f(0)\|/2$.

(B) **Sobre las distribuciones**

- (B1) Las observaciones son una muestra aleatoria simple

  $$X_i = f(U_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

  donde U_i sigue una distribución H en $[0,1]$ y el error ϵ_i sigue una distribución F en $\mathbb{R}^d$ y es independiente de U_i.
- (B2) La distribución H tiene densidad h con respecto a la medida de Lebesgue en $[0,1]$ que está acotada inferiormente por una constante positiva y superiormente, es decir,

  $$0 < c_1 \leq h(u) \leq c_2 < \infty \qquad \text{para } u \in [0,1].$$

- (B3) La distribución del ruido F cumple lo siguiente.
  - (B3.1) El soporte de F es $B(0,\sigma)$.
  - (B3.2) F tiene densidad acotada ϕ con respecto a la medida de Lebesgue en $\mathbb{R}^d$ que es continua en $\mathring{B}(0,\sigma)$.
  - (B3.3) La densidad ϕ es no creciente, esto es, $\phi(x) \geq \phi(y)$ si $\|x\| \leq \|y\|$.

(B3.4) La densidad ϕ es simétrica, esto es, $\phi(x) = \phi(y)$ si $\|x\| = \|y\|$.

(B3.5) Existen constantes $\beta \geq 0$ y $C_1, C_2 > 0$ tales que

$$C_1(\sigma - \|x\|)^\beta \leq \phi(x) \leq C_2(\sigma - \|x\|)^\beta \qquad \text{cuando } \|x\| \to \sigma.$$

Las hipótesis (A1), que evita que la curva se auto–interseque salvo que los extremos coincidan por ser cerrada, y (A2), que pide suavidad a la curva, permiten definir, en cada punto $f(u)$ de la curva, el vector tangente $T(u) \equiv T(f(u))$ y el conjunto de vectores normales unitarios $\mathcal{N}(u) \equiv \mathcal{N}(f(u))$. Definimos la fibra de tamaño σ en u

$$L(u) \equiv L(f(u)) = \{f(u) + tN(u) \colon N(u) \in \mathcal{N}(u), t \leq \sigma\}, \tag{3.1}$$

que es la porción de espacio normal a la curva en u (un hiperplano de dimensión $d-1$ en $\mathbb{R}^d$) que se encuentra a distancia menor o igual que σ, y el tubo

$$\mathcal{T} = \bigcup_{u \in [0,1]} L(u)$$

como la unión de las fibras. Cuando la curva es cerrada, implícitamente se pide no solo que coincidan las imágenes en los dos extremos, $f(0) = f(1)$, sino también que lo hagan los vectores tangentes $T(0) = T(1)$, de modo que podamos definir los elementos anteriores dependiendo solo del punto del filamento Γ y no del $u \in [0,1]$ asociado; dicho de otra manera, identificamos $u = 0$ con $u = 1$. Cuando la curva es abierta, la hipótesis (A4) evita que los dos extremos de la curva estén muy cerca y que el ruido que "emana" de un extremo se confunda con el del extremo contrario. Definimos los *end caps* inicial y final, respectivamente, como

$$\mathcal{C}_0 = B(f(0), \sigma) \setminus \mathcal{T} \qquad \text{y} \qquad \mathcal{C}_1 = B(f(1), \sigma) \setminus \mathcal{T}, \tag{3.2}$$

que son vacíos cuando f es cerrada y son disjuntos cuando es abierta gracias a la hipótesis anterior. La hipótesis (A3), por un lado, impone una cierta regularidad sobre el filamento, pues exige que tenga *reach* positivo y además mayor que el nivel de ruido σ. Esto evita zonas de mucha curvatura o donde el filamento esté próximo a auto–intersecarse. Además, si denotamos por S al soporte de la distribución marginal de las observaciones X_i, esta hipótesis es crucial para probar que S hereda de alguna forma la regularidad del filamento y cumple las condiciones geométricas necesarias para poder aplicar resultados teóricos sobre estimación del soporte. Cabe mencionar que posiblemente esta hipótesis de que el filamento tenga *reach* ya implique que la curva sea suave y no se auto–interseque, aunque en Genovese *et al.* (2012a) se piden explícitamente las condiciones (A1) y (A2) sobre el filamento.

En cuanto a las hipótesis sobre las distribuciones, (B2) garantiza que la variable $V_i = f(U_i)$ "cubre todo el filamento" y que su densidad no cae a cero en ninguna zona del mismo, es decir, la densidad h es esencialmente una uniforme en $[0,1]$. Por su parte, (B3) define el tipo de ruido del modelo. El hecho de que el soporte de F sea $B(0,\sigma)$ es una hipótesis fundamental en el modelo por dos motivos. Por un lado, como el filamento Γ es compacto por ser f continua, tener un ruido con soporte compacto hace que el soporte de las observaciones S sea también compacto. Por otra parte, el carácter esférico del ruido implica que el soporte S es una dilatación del filamento

$$S = \Gamma \oplus \sigma B = \bigcup_{u \in [0,1]} B(f(u), \sigma),$$

lo que se suele conocer como un entorno tubular de Γ. En la Figura 3.1 se representan sobre el soporte los principales elementos definidos en el modelo presentado. Veremos que esta estructura geométrica del soporte juega un papel esencial en el desarrollo teórico. Las hipótesis (B3.2)–(B3.4) concretan el ruido a uno esféricamente simétrico, con media y moda en 0, cuya densidad va decayendo al acercarnos a la frontera de su soporte. Finalmente, la hipótesis (B3.5) acota la densidad del ruido cerca de la frontera

de $B(0,\sigma)$. A diferencia de la hipótesis (B2), la densidad del ruido sí puede tender a 0 cuando x tiende a la frontera del soporte, y la constante $\beta \geq 0$ controla cómo de rápido decae esta densidad en función de la distancia de x a la frontera, $\sigma - \|x\|$. En concreto, la densidad es del orden de la distancia a la frontera elevada a β. Si $\beta = 0$, tenemos simplemente una acotación de la densidad que impide que tienda a 0, por ejemplo cuando ϕ es uniforme en $B(0,\sigma)$, mientras que, cuanto mayor es $\beta > 0$, más ligeras son las "colas" del ruido, que sí tienden a 0.

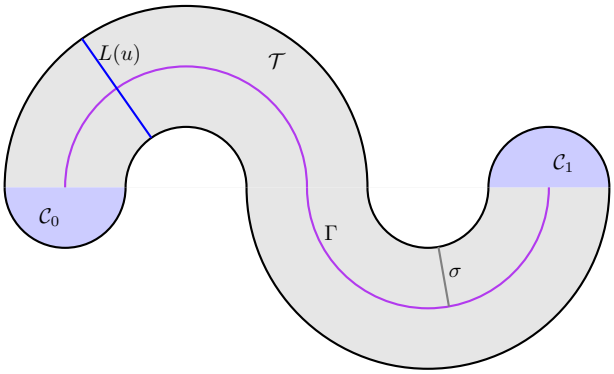

Figura 3.1: Elementos relativos al modelo para una curva abierta.

3.1.2. Estructura geométrica del soporte

Como mencionamos, en este modelo el soporte S de la distribución de los datos tiene una estructura geométrica tubular particular. El siguiente Teorema 3.1 proporciona una descomposición del soporte en términos de las fibras $L(u)$ definidas en (3.1).

Teorema 3.1 (Teorema 1 en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A) y si $S = \Gamma \oplus \sigma B$ (en particular bajo las hipótesis (B)), se cumple lo siguiente.*

(1) $S = \mathcal{T} \cup \mathcal{C}_0 \cup \mathcal{C}_1$. En particular, si f es cerrada, $S = \mathcal{T}$.

(2) Para cada $u \neq v \in [0,1]$, $L(u)$ y $L(v)$ son disjuntos.

(3) Para cada $y \in \mathcal{T}$, existe un único u tal que la fibra $L(u)$ contiene a y, $f(u)$ es el punto más cercano a y del filamento Γ, y además $f(u) + \sigma N(u)$ es el punto más cercano a y de la frontera ∂S. Si $y \in \Gamma$, entonces $N(u)$ puede ser cualquier vector en $\mathcal{N}(u)$; si, en cambio, $y \notin \Gamma$, entonces el punto de la frontera más cercano es único con $N(u) = (y - f(u))/\|y - f(u)\|$.

Los puntos (1) y (2) en el teorema anterior expresan el soporte S como el tubo $\mathcal{T}$, que a su vez es la unión disjunta de las fibras, junto con dos posibles *end caps* $\mathcal{C}_0$ y $\mathcal{C}_1$ si la curva es abierta, véase la Figura 3.1. Por otro lado, (3) relaciona los puntos más cercanos en el filamento Γ y en la frontera del soporte ∂S a un punto del tubo $y \in \mathcal{T}$ mediante los vectores normales.

Otro elemento geométrico fundamental para entender el modelo y el estimador EDT que definiremos es lo que se conoce como el eje medial (*medial axis*) de un conjunto.

Definición 3.2 (Eje medial)**.** Dado $S \subset \mathbb{R}^d$ compacto, decimos que una bola cerrada $B(x,r) \subset S$ es medial si $\mathring{B}(x,r) \cap \partial S = \emptyset$ y $B(x,r) \cap S$ contiene al menos 2 puntos. Definimos el eje medial de S, $M(S)$, como la clausura del conjunto

$$\{x \in S \colon B(x,r) \text{ es medial para algún } r > 0\}.$$

Una bola medial es aquella que, estando contenida en el conjunto S, toca tangencialmente a su frontera en al menos dos puntos, y el eje medial $M(S)$ es la clausura del conjunto de centros de las bolas mediales, ver Figura 3.2. Equivalentemente, se puede definir el eje medial como la clausura del conjunto de puntos de S con al menos dos proyecciones en ∂S. El eje medial se puede interpretar como una especie de "mediana" del conjunto. El Teorema 3.3 afirma que, para nuestro soporte S, el filamento Γ es justamente el eje medial.

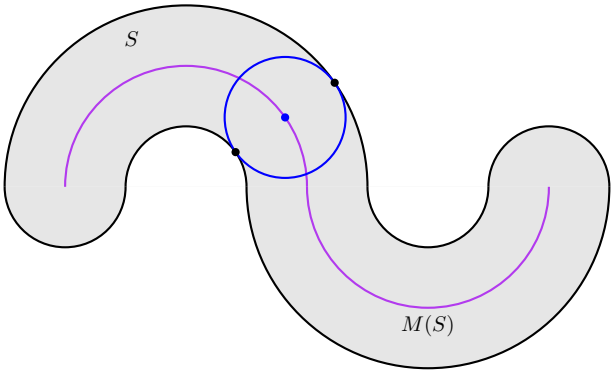

Figura 3.2: Eje medial de un conjunto S. En azul, una bola medial que toca a la frontera en dos puntos.

Teorema 3.3 (Teorema 2 en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A) y si* $S = \Gamma \oplus \sigma B$ *(en particular bajo las hipótesis (B)), se cumple* $\Gamma = M(S)$.

Como se comenta en Genovese *et al.* (2012a), el eje medial de un conjunto no es sencillo de estimar en general, pues no es continuo en distancia de Hausdorff, en el sentido de que dos conjuntos muy próximos en distancia de Hausdorff pueden tener ejes mediales muy distintos. Aún así, la idea para definir el estimador EDT que proponen estos autores se basa en que, en nuestro soporte particular, el filamento maximiza una cierta función llamada *Euclidean distance transform* (EDT). Se define la EDT como la aplicación $\Lambda : \mathbb{R}^d \to [0, \infty)$ dada por

$$\Lambda(y) = d(y, \partial S).$$

El Teorema 3.4 caracteriza el filamento como el conjunto de puntos de S donde la EDT alcanza su máximo, el nivel de ruido σ.

Teorema 3.4 (Teorema 3 en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A) y si* $S = \Gamma \oplus \sigma B$ *(en particular bajo las hipótesis (B)), dado* $y \in S$*, se cumple lo siguiente.*

(1) $y \in \Gamma$ *si y sólo si* $\Lambda(y) = \sigma$.

(2) Si $y \in S \setminus \Gamma$*, entonces* $\Lambda(y) < \sigma$.

(3) $d(y, \Gamma) + \Lambda(y) = \sigma$.

3.1.3. Estimador EDT y tasa de convergencia

El Teorema 3.4 sugiere una forma de definir un estimador del filamento Γ. Ya que Γ maximiza la distancia a la frontera de S, es decir, la EDT, podemos estimar el soporte por $\widehat{S}_n$ y después quedarnos con los puntos de $\widehat{S}_n$ que estén suficientemente lejos de su frontera $\partial\widehat{S}_n$. Así obtendríamos un conjunto de puntos $\widehat{\Gamma}_n$, el llamado estimador EDT, que no es una curva, pero que debería estar próximo, en

distancia de Hausdorff, al filamento. En Genovese *et al.* (2012a) se usa como estimador del soporte el estimador de Devroye–Wise definido en (2.3)

$$\widehat{S}_n = \mathcal{X}_n \oplus \varepsilon_n B = \bigcup_{i=1}^{n} B(X_i, \varepsilon_n).$$

Una vez calculado este estimador $\widehat{S}_n$, se define la EDT empírica $\widehat{\Lambda}$ como la distancia a la frontera del estimador

$$\widehat{\Lambda}(y) = d(y, \partial \widehat{S}_n).$$

Como la EDT alcanza su valor máximo en σ, se estima este valor por el máximo de la EDT empírica, obteniendo

$$\widehat{\sigma} = \max_{y \in \widehat{S}_n} \widehat{\Lambda}(y).$$

Finalmente, se define el estimador EDT como el conjunto de puntos de $\widehat{S}_n$ que están a distancia de $\partial \widehat{S}_n$ mayor o igual que $\widehat{\sigma} - \delta_n$,

$$\widehat{\Gamma}_n = \{y \in \widehat{S}_n \colon d(y, \partial \widehat{S}_n) \geq \widehat{\sigma} - \delta_n\}. \tag{3.3}$$

El parámetro $\delta_n \in (0, \widehat{\sigma})$ controla cuánto "contraemos" el conjunto $\widehat{S}_n$ hacia dentro. Si δ_n fuera 0, nos quedaríamos simplemente con los puntos en los que se maximiza $\widehat{\Lambda}$, que probablemente serían una cantidad finita y no estimarían bien el filamento. Algo similar ocurriría si δ_n fuera demasiado pequeño, el conjunto $\widehat{\Gamma}_n$ estaría demasiado fragmentado y no se parecería a Γ. De hecho, para los resultados teóricos que veremos, es interesante garantizar que el estimador $\widehat{\Gamma}_n$ contenga a Γ para n suficientemente grande, lo que se puede lograr tomando δ_n suficientemente grande, como veremos en el Teorema 3.10. No obstante, si δ_n fuera demasiado grande, por ejemplo próximo a $\widehat{\sigma}$, el estimador EDT sería casi todo el estimador del soporte $\widehat{S}_n$, lo cual tampoco es adecuado porque estaríamos demasiado lejos de Γ en distancia de Hausdorff. Escogiendo un δ_n adecuado, conseguimos que el conjunto $\widehat{\Gamma}_n$ no sea demasiado pequeño ni demasiado grande. En Genovese *et al.* (2012a) se propone tomar $\delta_n = 2\varepsilon_n$, aunque a nivel teórico se puede probar que es suficiente tomar cualquier $\delta_n \geq \varepsilon_n$ del orden de ε_n.

El algoritmo para el estimador EDT, por tanto, es el siguiente. En este caso escribimos el estimador del soporte, el estimador EDT, el parámetro ε, etc., sin subíndice n porque no los estamos pensando como una sucesión en el tamaño muestral n, sino como unos elementos fijos para la muestra concreta que tenemos.

Algoritmo 3.5 (Estimador EDT, Genovese *et al.*, 2012a)**.**

ENTRADA: Las observaciones $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ y el parámetro $\varepsilon > 0$.

SALIDA: El estimador EDT $\widehat{\Gamma}$, un conjunto de puntos.

1. Estimar el soporte con el estimador de Devroye–Wise $\widehat{S} = \mathcal{X}_n \oplus \varepsilon B$.
2. Estimar el nivel de ruido por $\widehat{\sigma} = \max_{y \in \widehat{S}} \widehat{\Lambda}(y)$, donde $\widehat{\Lambda}(y) = d(y, \partial \widehat{S})$ es la EDT empírica.
3. Fijar $\delta = 2\varepsilon$ y obtener $\widehat{\Gamma} = \{y \in \widehat{S} \colon d(y, \partial \widehat{S}) \geq \widehat{\sigma} - \delta\}$.

En la Figura 3.3 se ilustra el Algoritmo 3.5. A efectos de visualización hemos tomado un valor elevado del parámetro ε y el estimador EDT resulta demasiado parecido al estimador del soporte. En este caso sería más adecuado tomar un valor de ε menor o bien, manteniendo ε, disminuir el valor de δ (que ya no sería 2ε) para contraer más el estimador EDT.

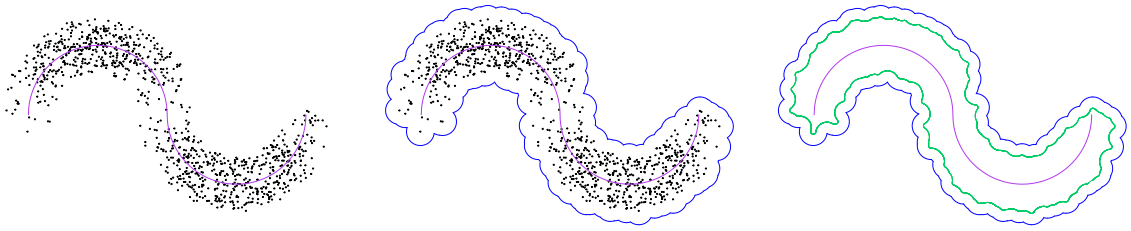

Figura 3.3: A la izquierda, muestra de tamaño $n = 1000$ con ruido aditivo simulada de un filamento (morado) con $\text{reach}(\Gamma) = 0.25$ con nivel de ruido $\sigma = 0.1$. En el centro, en azul, frontera del estimador de Devroye–Wise para $\varepsilon = 0.05$. A la derecha, en verde, frontera del estimador EDT $\widehat{\Gamma}$.

Para obtener la tasa de convergencia del estimador EDT, en Genovese *et al.* (2012a) se siguen esencialmente dos pasos. En primer lugar, se obtiene la tasa de convergencia del estimador de Devroye–Wise $\widehat{S}_n$ y de su frontera $\partial\widehat{S}_n$ adaptando el Teorema 2.7 a las hipótesis del modelo considerado. Después, se prueba que se estiman bien la EDT Λ y el nivel de ruido σ por sus contrapartes empíricas $\widehat{\Lambda}$ y $\widehat{\sigma}$, para finalmente ver que el estimador $\widehat{\Gamma}_n$ también estima bien al filamento y conserva la tasa de convergencia obtenida.

En cuanto a la tasa de convergencia de $\widehat{S}_n$, lo primero que se necesita es probar que el soporte S está en las condiciones del Teorema 2.7, es decir, que es un conjunto estándar y parcialmente expandible (ver Definiciones 2.5 y 2.6).

Teorema 3.6 (Teorema 6 en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A) y si $S = \Gamma \oplus \sigma B$ (en particular bajo las hipótesis (B)), se cumple que el soporte S es (χ, λ)–estándar con respecto a la medida de Lebesgue μ en $\mathbb{R}^d$ para $\chi = 2^{-d}$ y $\lambda = \sigma$ y (R, r)–parcialmente expandible para $R = 1$ y $r = \text{reach}(\Gamma) - \sigma$.*

Sea Q la distribución marginal de X_i y sea q la densidad asociada, cuyo soporte es S. Esta densidad se puede expresar como la convolución

$$q(y) = \int \phi(y - f(u))dH(u) = \int \phi(y - f(u))h(u)du.$$

El siguiente resultado también es necesario para la obtención de la tasa de convergencia. Expresa la velocidad a la que decae a 0 la densidad q, es decir, la densidad de observaciones muestrales, cuando nos aproximamos a la frontera del soporte S.

Teorema 3.7 (Teorema 7 en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A)–(B), existe una constante $c_2 > 0$ tal que, para todo $y \in \text{int}(S)$ suficientemente cerca de ∂S, se cumple*

$$q(y) \geq c_2 d(y, \partial S)^\alpha,$$

siendo $\alpha = \beta + 1/2$, con $\beta \geq 0$ definida en la hipótesis (B3.5).

El Teorema 3.7 nos dice que la densidad q decae a velocidad del orden de la distancia a la frontera elevada a $\alpha = \beta + 1/2 > 0$, donde $\beta \geq 0$ representaba la velocidad a la que decaía la densidad ϕ del ruido en la frontera de la bola $B(0, \sigma)$. Notemos que, aunque la densidad del ruido fuera uniforme en $B(0, \sigma)$, en cuyo caso $\beta = 0$, α siempre es estrictamente positiva (en ese caso $\alpha = 1/2$), de modo que la densidad q de las observaciones no solo no sería uniforme en S, sino que tampoco estaría acotada inferiormente por una constante positiva. Por tanto, bajo estas hipótesis, la densidad $q(y)$ siempre tiende a 0 a medida que y tiende a ∂S.

Ahora ya podemos enunciar el resultado que proporciona la tasa de convergencia del estimador de Devroye–Wise y de su frontera cuando estamos en las condiciones de los Teoremas 3.6 y 3.7. Es una pequeña adaptación del Teorema 2.7.

Teorema 3.8 (Teorema 8 en Genovese *et al.*, 2012a)**.** *Sea $\mathcal{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^d$ una muestra aleatoria simple de una distribución Q con densidad q, cuyo soporte es S. Supongamos que S es compacto, (R, r)–parcialmente expandible y (χ, λ)–estándar con respecto a μ, la medida de Lebesgue en $\mathbb{R}^d$. Supongamos también que existen $C > 0$ y $\alpha \geq 0$ tales que $q(y) \geq Cd(y, \partial S)^\alpha$ para $y \in S$. Consideremos el estimador de Devroye–Wise $\widehat{S}_n = \mathcal{X}_n \oplus \varepsilon_n B$, con*

$$\varepsilon_n = 4 \left(\frac{2}{\chi \omega_d} \right)^{1/d} \left(\frac{\log(n)}{n} \right)^{1/(d+\alpha)}, \tag{3.4}$$

donde $\omega_d = \mu(B)$ es la medida de Lebesgue de la bola unidad en $\mathbb{R}^d$. Entonces, se verifican e.a.s.

$$d_H(\widehat{S}_n, S) \leq \varepsilon_n \qquad y \qquad d_H(\partial \widehat{S}_n, \partial S) \leq \varepsilon_n.$$

Además, $S \subset \widehat{S}_n$ e.a.s.

El Teorema 3.8 adapta el Teorema 2.7 permitiendo que la densidad q no esté acotada inferiormente por una constante positiva (en el caso de que sí, α valdría 0 y recuperaríamos el Teorema 2.7). A cambio, la tasa de convergencia del estimador y de su frontera empeora, en el exponente tenemos $1/(d + \alpha)$ en vez de $1/d$. Volviendo al modelo, en nuestro caso $\alpha = \beta + 1/2 \geq 1/2$ y, cuanto más pesadas sean las "colas" de la distribución del ruido F, menor es β y por tanto α, de modo que mejor es la tasa. Este resultado es intuitivo pues, si tenemos mayor densidad de puntos cerca de la frontera del soporte, estimaremos esta frontera con más precisión.

Una vez obtenida la tasa del estimador del soporte, el siguiente paso es obtener la tasa del estimador EDT $\widehat{\Gamma}_n$. En primer lugar, definamos $\widehat{y}$ como el punto de $\widehat{S}$ en el que la EDT empírica alcanza su máximo $\widehat{\sigma}$, es decir,

$$\widehat{y} = \arg\max_{y \in \widehat{S}} \widehat{\Lambda}(y), \tag{3.5}$$

de modo que $\widehat{\Lambda}(\widehat{y}) = \widehat{\sigma}$. Cabe mencionar que el punto $\widehat{y}$ no tiene por qué ser único. El siguiente resultado acota el error que cometemos, tanto al estimar la EDT Λ por la EDT empírica $\widehat{\Lambda}$ como al estimar el nivel de ruido σ por $\widehat{\sigma}$, así como la distancia de $\widehat{y}$ al filamento Γ. En particular, las tres cosas son del orden del error con el que estimamos la frontera ∂S. Este Teorema 3.9, y también el Teorema 3.10 posterior, están enunciados en general para un estimador $\widehat{S}$ cualquiera que cumpla $S \subset \widehat{S} \subset S \oplus \varepsilon B$ y $d_H(\partial \widehat{S}, \partial S) \leq \varepsilon$, donde ε representa el error que cometemos al estimar la frontera del soporte. En particular, se pueden aplicar al estimador de Devroye–Wise $\widehat{S}_n$ e.a.s. gracias al Teorema 3.8, caso en el que ε_n también coincide con el radio usado para suavizar la muestra.

Teorema 3.9 (Teorema 10 en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A) y si $S = \Gamma \oplus \sigma B$ (en particular bajo las hipótesis (B)), supongamos que $S \subset \widehat{S} \subset S \oplus \varepsilon B$ y que $d_H(\partial \widehat{S}, \partial S) \leq \varepsilon$. Entonces, se cumple lo siguiente.*

(1) $\sup_{y \in \mathbb{R}^d} |\widehat{\Lambda}(y) - \Lambda(y)| \leq \varepsilon$.

(2) $|\widehat{\sigma} - \sigma| \leq \varepsilon$.

(3) $d(\widehat{y}, \Gamma) \leq 2\varepsilon$ *(siendo $\widehat{y} \in \widehat{S}$ tal que $\widehat{\Lambda}(\widehat{y}) = \widehat{\sigma}$).*

El Teorema 3.10 formaliza la idea de que, si estimamos la frontera de S con una cierta tasa de convergencia, el estimador EDT hereda esa misma tasa de convergencia. Dicho de otra manera, cómo

de bien estimamos Γ depende de cómo de bien seamos capaces de estimar la frontera de S. Se puede probar que no es necesario tomar $\delta = 2\varepsilon$, sino que basta tomar cualquier $\delta \geq \varepsilon$ del orden de ε, usando argumentos similares a la prueba que realizaremos en la Proposición 4.17 para el estimador que propondremos en el Capítulo 4.

Teorema 3.10 (Teorema 11–(1) en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A) y si $S = \Gamma \oplus \sigma B$ (en particular bajo las hipótesis (B)), supongamos que $S \subset \widehat{S} \subset S \oplus \varepsilon B$ y que $d_H(\partial\widehat{S}, \partial S) \leq \varepsilon$. Sea $\widehat{\Gamma} = \{y \in \widehat{S} \colon d(y, \partial\widehat{S}) \geq \widehat{\sigma} - \delta\}$ el estimador EDT con $\delta = 2\varepsilon$. Entonces $\Gamma \subset \widehat{\Gamma} \subset \Gamma \oplus (4\varepsilon)B$ y por tanto*

$$d_H(\widehat{\Gamma}, \Gamma) \leq 4\varepsilon.$$

Finalmente, aplicando el Teorema 3.10 al estimador de Devroye–Wise, obtenemos que la tasa de convergencia del estimador EDT es la obtenida en el Teorema 3.8, es decir, $(\log(n)/n)^{1/(d+\alpha)}$.

Teorema 3.11 (Teorema 11–(2) en Genovese *et al.*, 2012a)**.** *Bajo las hipótesis (A)–(B), sea $\widehat{S}_n$ el estimador de Devroye–Wise (2.3) con ε_n dado por (3.4) para $\chi = 2^{-d}$ y $\alpha = \beta + 1/2$, esto es,*

$$\varepsilon_n = 4\left(\frac{2^{d+1}}{\omega_d}\right)^{1/d}\left(\frac{\log(n)}{n}\right)^{1/(d+\alpha)},$$

y sea $\widehat{\Gamma}_n$ el estimador EDT (3.3) con $\delta_n = 2\varepsilon_n$. Entonces, con probabilidad uno,

$$d_H(\widehat{\Gamma}_n, \Gamma) = O\left(\left(\frac{\log(n)}{n}\right)^{1/(d+\alpha)}\right).$$

3.1.4. Extracción de curva

Como ya comentamos, el estimador EDT no es una curva, sino un conjunto de puntos, en palabras de Genovese *et al.* (2012a) un "conjunto de valores ajustados". Como el objeto que estamos estimando es un filamento, parece razonable extraer una curva a partir del estimador EDT, es decir, definir una curva $\widehat{\Gamma}_f$ que esté contenida en $\widehat{\Gamma}$ y que conserve la tasa de convergencia para estimar el filamento.

Genovese *et al.* (2012a) proponen un algoritmo para extraer una curva del estimador EDT con la siguiente idea. Supongamos que la curva es abierta y denotemos por $x_0 \in \widehat{\Gamma} \cap \mathcal{C}_0$ y $x_1 \in \widehat{\Gamma} \cap \mathcal{C}_1$ a dos puntos del estimador EDT que pertenezcan a sendos *end caps* $\mathcal{C}_0$ y $\mathcal{C}_1$ definidos en (3.2). Si tomamos una curva $\widehat{\Gamma}_f$ uniendo x_0 y x_1 que esté contenida en $\widehat{\Gamma}$, gracias al Teorema 3.10, la curva debe cortar a todas las fibras del tubo $\mathcal{T}$ a una distancia menor o igual que 4ε y también estará a una distancia menor o igual que 4ε de los extremos de la curva $f(0)$ y $f(1)$, de modo que se puede probar que $d_H(\widehat{\Gamma}_f, \Gamma) \leq 4\varepsilon$. Como desconocemos los *end caps* $\mathcal{C}_0$ y $\mathcal{C}_1$, usamos unas estimaciones $\widehat{x}_0$ y $\widehat{x}_1$ que calculamos como los puntos que maximizan la longitud del camino más corto entre dos puntos cualesquiera de $\widehat{\Gamma}$, y que se prueba que estiman a los verdaderos extremos $f(0)$ y $f(1)$ con el mismo orden de convergencia. Después definimos la curva extraída como un camino entre $\widehat{x}_0$ y $\widehat{x}_1$.

Cuando la curva es cerrada en lugar de abierta, la idea es "cortar" un trozo de $\widehat{\Gamma}$ alrededor del punto $\widehat{y}$ definido en (3.5) para obtener una curva abierta a la que aplicarle el procedimiento anterior, y luego unir los dos extremos $\widehat{x}_0$ y $\widehat{x}_1$ de la curva abierta extraída para recuperar una curva cerrada.

En la práctica, la obtención de los puntos $\widehat{x}_0$ y $\widehat{x}_1$ que maximizan la longitud del camino más corto se hace generando una *epsilon*–red de puntos en $\widehat{\Gamma}$, obteniendo el árbol de expansión mínima correspondiente y calculando los dos puntos que maximizan la longitud del camino más corto en el árbol. Después se extrae la curva que une $\widehat{x}_0$ y $\widehat{x}_1$ aplicando el algoritmo de Dijkstra. Como la curva extraída es una unión de aristas del árbol, se puede llevar a cabo un último paso de relajación para suavizarla. Para más detalles, véase la Sección 4 en Genovese *et al.* (2012a).

3.2. El modelo de Genovese *et al.* (2012b)

En esta sección presentaremos el modelo de estimación de variedades con ruido perpendicular considerado en Genovese *et al.* (2012b), en el que se calcula la tasa minimax de convergencia en distancia de Hausdorff.

En el modelo, la variedad M a estimar es una variedad de Riemann sin borde de dimensión $k < d$ embebida en $\mathbb{R}^d$, lo que quiere decir que en cada punto de la variedad hay un entorno difeomorfo a un abierto de $\mathbb{R}^k$. Suponemos que M es compacta pero, además, como la tasa minimax que se obtiene se calcula a lo largo de una familia de distribuciones inducidas por una familia de variedades, vamos a suponer que todas las variedades M a estimar están contenidas en un compacto $\mathcal{K} \subset \mathbb{R}^d$ fijado. Aunque en Genovese *et al.* (2012b) no se dice explícitamente, es necesario también pedir que la variedad sea conexa, pues en las demostraciones se trabaja con la distancia geodésica entre dos puntos arbitrarios de M, que solo está bien definida si ambos puntos se pueden conectar mediante una geodésica.

En cada punto $p \in M$ denotamos por T_pM al espacio tangente a p en M, que se puede visualizar como un hiperplano en $\mathbb{R}^d$ de dimensión k, y por $T_pM^\perp$ al espacio normal, que visualizamos como un hiperplano de dimensión $d-k$. De modo similar al modelo de Genovese *et al.* (2012a) visto en la Sección 3.1, vamos a definir la fibra de tamaño σ en p como

$$L(p) = T_pM^\perp \cap B(p, \sigma),$$

donde $\sigma > 0$ es el nivel de ruido, y el tubo

$$\mathcal{T} = \bigcup_{p \in M} L(p).$$

Las observaciones de las que disponemos son de la forma

$$X_i = V_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde V_i sigue una distribución en la variedad M y ϵ_i, el error, tiene distribución condicionada a V_i en la fibra $L(V_i)$. En concreto, se supone que la distribución de V_i es uniforme en M y que la distribución de ϵ_i condicionada a V_i es también uniforme en $L(V_i)$. Por tanto, en este modelo el ruido es perpendicular a la variedad y el soporte S de la distribución marginal de X_i es el tubo, es decir,

$$S = \mathcal{T}. \tag{3.6}$$

Según se remarca en Genovese *et al.* (2012b), la hipótesis de uniformidad en la variedad no es crítica, y se puede sustituir por una distribución con densidad acotada inferiormente por una constante positiva, de modo similar a la hipótesis (B2) en el modelo aditivo, pero la asunción de uniformidad en el ruido perpendicular sí parece ser crítica para la obtención de la tasa minimax.

Finalmente, hay que imponer una condición geométrica sobre la variedad, que igual que en Genovese *et al.* (2012a) es pedir que el *reach* de M sea superior al nivel de ruido σ. El *reach* de M es pequeño si la variedad presenta una alta curvatura o tiene zonas en las que está próxima a auto–intersecarse. La condición $\sigma < \text{reach}(M)$ garantiza que el tubo $\mathcal{T}$ es la unión disjunta de las fibras de tamaño σ que lo forman, un resultado análogo al Teorema 3.1–(2). De hecho, otra definición equivalente de *reach* es el mayor valor tal que las fibras perpendiculares no se intersecan si no se prolongan más que ese tamaño, véase la Figura 3.4. El inverso de este valor también se conoce como número de condición de la variedad, véase Niyogi *et al.* (2008). Por otro lado, como la variedad no tiene borde, el dilatado $M \oplus \sigma B$ coincide con $\mathcal{T}$ y por tanto con el soporte S, esto es,

$$M \oplus \sigma B = \mathcal{T} = S,$$

un resultado similar al Teorema 3.1–(1) pero sin los *end caps* $\mathcal{C}_0$ y $\mathcal{C}_1$, como sucedía cuando la curva era cerrada. Al igual que con la condición $M \subset \mathcal{K}$, a la hora de estudiar la tasa minimax se necesita

una cota inferior del *reach*, κ, común a todas las variedades de la familia que definiremos, por lo que en realidad la hipótesis que se pide es $\kappa \leq \text{reach}(M)$ y $\sigma < \kappa$. En la práctica, simplemente estamos pensando en una variedad compacta y con $\text{reach}(M) > \sigma$.

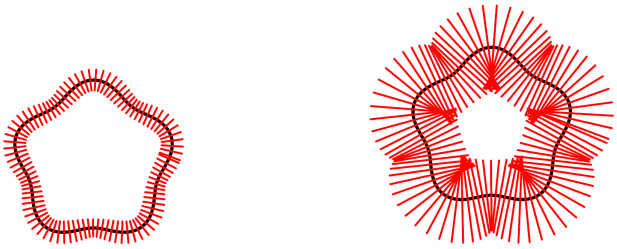

Figura 3.4: En la figura izquierda, las fibras perpendiculares de tamaño $\sigma < \text{reach}(M)$ no se intersecan. En la figura derecha, algunas fibras perpendiculares de tamaño $\sigma > \text{reach}(M)$ se intersecan. Figura tomada de Genovese *et al.* (2012b).

A continuación recogemos todas las hipótesis consideradas.

(C) **Sobre la variedad**

(C1) $M \subset \mathcal{K} \subset \mathbb{R}^d$ es una variedad de Riemann sin borde, de dimensión $k < d$, compacta y conexa.

(C2) Se cumple $0 < \kappa \leq \text{reach}(M)$.

(C3) Se cumple $0 < \sigma < \kappa$.

(D) **Sobre las distribuciones**

(D1) Las observaciones son una muestra aleatoria simple

$$X_i = V_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde V_i sigue una distribución en la variedad M y el error ϵ_i sigue una distribución condicionada a V_i en la fibra $L(V_i)$.

(D2) La distribución de V_i es uniforme en la variedad M.

(D3) La distribución de ϵ_i condicionada a V_i es uniforme en la fibra $L(V_i)$.

Denotemos por $\mathcal{M} \equiv \mathcal{M}(\mathcal{K}, \kappa)$ a la familia de todas las variedades M que cumplen las hipótesis (C1)–(C2), donde el compacto $\mathcal{K} \subset \mathbb{R}^d$ y la constante $\kappa > 0$ están fijadas. Fijemos ahora un nivel de ruido $\sigma < \kappa$. Dada una variedad $M \in \mathcal{M}$, las hipótesis (D) inducen una distribución $Q \equiv Q_M$ de las observaciones X_i con soporte $S \equiv S_M = M \oplus \sigma B$. Denotamos por

$$\mathcal{Q} \equiv \mathcal{Q}(\mathcal{K}, \kappa, \sigma) = \{Q_M \colon M \in \mathcal{M}(\mathcal{K}, \kappa)\}$$

a la familia de todas las distribuciones inducidas por las variedades de la familia $\mathcal{M}$.

El Lema 3.12 prueba que la densidad $q \equiv q_M$ de la distribución Q_M con respecto a la medida de Lebesgue está esencialmente acotada inferiormente por una constante positiva y superiormente. Más aún, la acotación es uniforme en toda la familia $\mathcal{M}$ si la consideramos "normalizada" con respecto a la densidad uniforme en $S_M = M \oplus \sigma B$, que es $u_M = 1/\mu(S_M)$.

Lema 3.12 (Lema 4 en Genovese *et al.*, 2012b)**.** *Bajo las hipótesis (C)–(D), existen constantes* $0 < C_* \leq C^* < \infty$*, que dependen solo de* κ *y* d*, tales que*

$$C_* \leq \inf_{M \in \mathcal{M}} \operatorname*{ess\,inf}_{y \in S_M} \frac{q_M(y)}{u_M(y)} \leq \sup_{M \in \mathcal{M}} \operatorname*{ess\,sup}_{y \in S_M} \frac{q_M(y)}{u_M(y)} \leq C^*.$$

Por tanto, fijada una variedad M, la distribución marginal de X_i es "casi" una uniforme en su soporte S_M, ya que

$$0 < \frac{C_*}{\mu(M \oplus \sigma B)} \leq \operatorname*{ess\,inf}_{y \in S_M} q_M(y) \leq \operatorname*{ess\,sup}_{y \in S_M} q_M(y) \leq \frac{C^*}{\mu(M \oplus \sigma B)} < \infty.$$

3.2.1. Tasa minimax

A la hora de determinar la tasa óptima que puede tener un estimador $\widehat{M}_n$ de la variedad M, el criterio de riesgo que consideraremos es el minimax. Supongamos que tenemos un estimador $\widehat{M}_n$, esto es, una función medible de $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ que toma valores en el conjunto de las variedades. Sea una variedad $M \in \mathcal{M}$ y la correspondiente distribución inducida $Q \in \mathcal{Q}$. Si la función de pérdida que usamos para cuantificar la proximidad entre $\widehat{M}_n$ y M es la distancia de Hausdorff $d_H(\widehat{M}_n, M)$, la esperanza de esta pérdida con respecto a la distribución Q es la función de riesgo $\mathbb{E}_Q\big[d_H(\widehat{M}_n, M)\big]$. Para determinar la bondad del estimador $\widehat{M}_n$, podemos ponernos en el peor caso y calcular cuál es el mayor valor que puede tomar la función de riesgo a lo largo de todas las posibles variedades a estimar

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q\big[d_H(\widehat{M}_n, M)\big].$$

El riesgo minimax se define como el ínfimo de los valores anteriores a lo largo de todos los posibles estimadores

$$\inf_{\widehat{M}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q\big[d_H(\widehat{M}_n, M)\big].$$

Informalmente, el riesgo minimax captura cómo de bien lo hace el estimador que mejor lo hace en el peor caso.

Cuando hablamos de tasa minimax, nos referimos al orden asintótico del riesgo minimax cuando $n \to \infty$. Genovese *et al.* (2012b) consiguen acotar la tasa minimax para el modelo presentado en esta sección. El Teorema 3.13 proporciona una cota inferior de la tasa minimax.

Teorema 3.13 (Teorema 1 en Genovese *et al.*, 2012b)**.** *Bajo las hipótesis (C)–(D), existe una constante* $C_1 > 0$ *tal que, para* n *suficientemente grande,*

$$\inf_{\widehat{M}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q\big[d_H(\widehat{M}_n, M)\big] \geq C_1 \left(\frac{1}{n}\right)^{2/(k+2)}.$$

Por su parte, el Teorema 3.14 proporciona una cota superior de la tasa minimax. Su prueba se basa en construir un estimador que alcanza dicha tasa, aunque esta construcción es teórica y no es utilizable en la práctica.

Teorema 3.14 (Teorema 2 en Genovese *et al.*, 2012b)**.** *Bajo las hipótesis (C)–(D), existen un estimador* $\widehat{M}_n$ *y una constante* $C_2 > 0$ *tales que, para* n *suficientemente grande,*

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q\big[d_H(\widehat{M}_n, M)\big] \leq C_2 \left(\frac{\log(n)}{n}\right)^{2/(k+2)}.$$

Los Teoremas 3.13 y 3.14 establecen que la tasa minimax está entre $(1/n)^{2/(k+2)}$ y $(\log(n)/n)^{2/(k+2)}$. Como, asintóticamente, n domina frente a $\log(n)$, ambas cotas están próximas y se suele decir que la tasa minimax es $(1/n)^{2/(k+2)}$, salvo factor logarítmico. Notemos que la tasa solo depende de la dimensión de la variedad a estimar, k, y no de la del espacio ambiente, d, a pesar de que el soporte de la distribución de los datos es un conjunto de dimensión d. En caso de que la variedad a estimar sea un filamento, de dimensión $k = 1$, la tasa minimax es $(1/n)^{2/3}$, salvo factor logarítmico.

Capítulo 4

Una nueva propuesta de estimador minimax

El objetivo de este capítulo es proponer un nuevo estimador para el modelo de ruido perpendicular presentado en la Sección 3.2 en el contexto de estimación de filamentos, es decir, cuando la variedad es de dimensión $k = 1$, que alcance la tasa minimax, salvo factor logarítmico, cuando el espacio ambiente sea el plano ($d = 2$).

El estimador consiste en una mejora del estimador EDT (Algoritmo 3.5) visto en la Sección 3.1 usando como estimador del soporte la envoltura r–convexa de la muestra $\widehat{S}_n = C_r(\mathcal{X}_n)$ (Definición 2.11) en lugar del estimador de Devroye–Wise. Probamos que en este modelo se cumplen las hipótesis que garantizan que la tasa de convergencia de $\widehat{S}_n$ y $\partial\widehat{S}_n$ es $O\left((\log(n)/n)^{2/(d+1)}\right)$ gracias al Teorema 2.20, que mejora la tasa $O((\log(n)/n)^{1/d})$ del estimador de Devroye–Wise dada por el Teorema 2.7. Vemos que el estimador EDT $\widehat{\Gamma}_n$ definido a partir de la envoltura r–convexa también hereda la misma tasa de convergencia al filamento. Cuando la dimensión del espacio ambiente es $d = 2$, la tasa $O\left((\log(n)/n)^{2/3}\right)$ es, salvo factor logarítmico, la tasa minimax (Teoremas 3.13 y 3.14).

La Sección 4.1 está dedicada a describir con detalle el modelo y las hipótesis que consideramos, así como las razones que motivan su elección. En la Sección 4.2 presentamos el nuevo estimador y estudiamos su tasa de convergencia. La Sección 4.3 aborda posibles elecciones del parámetro de forma r a emplear en la envoltura r–convexa y de un parámetro δ del que depende el estimador EDT con envoltura r–convexa, así como una propuesta de extracción de curva dentro de dicho estimador. También se discute la necesidad o no de las hipótesis pedidas en el modelo a la hora de la aplicación práctica del estimador. Finalmente, en la Sección 4.4 se ilustra el estimador propuesto mediante una aplicación a datos reales de secciones de troncos de árboles en inventario forestal.

En este capítulo, la dimensión de la variedad siempre se asumirá que es $k = 1$ cuando hagamos referencia al modelo de ruido perpendicular de la Sección 3.2.

4.1. Modelo e hipótesis

Aunque los modelos de ruido aditivo presentado en la Sección 3.1 y de ruido perpendicular presentado en la Sección 3.2 son distintos, es claro que ambos comparten elementos fundamentales, como la estructura geométrica tubular del soporte o la condición de *reach*. Restringiendo el modelo de ruido perpendicular a variedades de dimensión $k = 1$, hay básicamente tres diferencias entre ambos modelos. Por un lado, el tipo de ruido del modelo, que en el primer modelo es esférico y en el segundo perpen-

dicular. Por otra parte, el tratamiento matemático del filamento. En el modelo aditivo, el filamento Γ estaba parametrizado globalmente mediante una curva $f : [0,1] \to \mathbb{R}^d$ tal que $f([0,1]) = \Gamma$, mientras que en el de ruido perpendicular el filamento es una variedad, de modo que en principio solo tenemos parametrizaciones locales del mismo por abiertos. Por último, en el modelo aditivo permitíamos tanto curvas abiertas como cerradas, mientras que en el de ruido perpendicular, al ser la variedad sin borde, la curva tiene que ser cerrada.

Si bien el estimador EDT de Genovese *et al.* (2012a) se propone en el modelo aditivo, propondremos nuestro estimador en el modelo de ruido perpendicular por dos motivos. El primero es que en ese modelo se conoce la tasa minimax, de modo que podemos probar que nuestro estimador la alcanza en el caso de espacio ambiente de dimensión $d = 2$. El segundo tiene que ver con la obtención de la tasa de convergencia de la envoltura r–convexa como estimador del soporte. Con ruido esférico, el Teorema 3.7 recogía que la densidad q tiende a 0 a medida que nos aproximamos a la frontera del soporte S a una velocidad del orden de la distancia a ∂S elevada a $\alpha \geq 1/2$. Sin embargo, para poder aplicar el Teorema 2.20 y obtener la tasa de convergencia de la envoltura r–convexa, necesitamos que la densidad no tienda a 0 en la frontera del soporte, sino que esté esencialmente acotada inferiormente por una constante positiva, condición que sí se cumple en el modelo de ruido perpendicular gracias al Lema 3.12. Si no considerásemos el ruido perpendicular, deberíamos generalizar el Teorema 2.20 a densidades que tiendan a cero para poder obtener las tasas de convergencia, tarea que resultaría mucho más complicada a nivel técnico que la prueba del Teorema 3.8 que se realiza en Genovese *et al.* (2012a), que es una adaptación inmediata del Teorema 2.7, y que está fuera del objetivo de este trabajo. Además, tampoco podemos admitir curvas abiertas pues, si el ruido es perpendicular, el soporte de las observaciones es el tubo $\mathcal{T}$ (Ecuación (3.6)), y en una curva abierta este tubo no coincide con el dilatado $\Gamma \oplus \sigma B$, ya que el dilatado contiene también los dos *end caps* $\mathcal{C}_0$ y $\mathcal{C}_1$, véase Teorema 3.1–(1) y Figura 3.1. Necesitamos que el soporte de las observaciones S coincida con $\Gamma \oplus \sigma B$ para poder probar que S está en las condiciones del Teorema 2.19 y así poder aplicar el Teorema 2.20. Si la curva fuera abierta, el tubo $\mathcal{T}$ no cumpliría la condición de rodamiento libre por dentro del conjunto, que probaremos en el Lema 4.9. Véase la Sección 4.3.3 para una discusión del papel del modelo y las hipótesis en la aplicación práctica.

En resumen, el modelo que consideraremos es el de ruido perpendicular descrito en la Sección 3.2 en el caso particular en el que la variedad es un filamento ($k = 1$), que se corresponde con las hipótesis (C*)–(D*) recogidas a continuación. Las hipótesis (C*) simplifican las hipótesis (C) eliminando el compacto $\mathcal{K}$ y la constante κ, que solo eran necesarias para la obtención de la tasa minimax, mientras que las hipótesis (D*) no son más que las hipótesis (D) reescritas para el caso particular de filamentos.

(C*) **Sobre el filamento**

- (C1*) $\Gamma \subset \mathbb{R}^d$ es una variedad de Riemann sin borde, de dimensión $k = 1$, compacta y conexa.
- (C2*) Se cumple $0 < \sigma < \text{reach}(\Gamma)$.

(D*) **Sobre las distribuciones**

- (D1*) Las observaciones son una muestra aleatoria simple

$$X_i = V_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

  donde V_i sigue una distribución en el filamento Γ y el error ϵ_i sigue una distribución condicionada a V_i en la fibra $L(V_i)$.
- (D2*) La distribución de V_i es uniforme en el filamento Γ.
- (D3*) La distribución de ϵ_i condicionada a V_i es uniforme en la fibra $L(V_i)$.

Al igual que en el modelo de ruido perpendicular de la Sección 3.2, las hipótesis (C*)–(D*) implican que el soporte S de la distribución marginal de X_i es el tubo $\mathcal{T}$, y que este tubo a su vez coincide

con el dilatado $\Gamma \oplus \sigma B$, de modo que $S = \Gamma \oplus \sigma B$. Aunque el estimador EDT se definía en el modelo aditivo y no en el de ruido perpendicular, hay resultados teóricos y procedimientos de demostración que nos gustaría emplear para nuestro estimador en este segundo modelo, aunque estén probados para el primero. Si nos fijamos, por ejemplo, en los Teoremas 3.1, 3.3 y 3.4 en el modelo aditivo, solo usan las hipótesis (A) sobre el filamento y la condición $S = \Gamma \oplus \sigma B$, pero no necesitan que el ruido sea el esférico. Para poder emplear dichos resultados aquí, vamos a ver que todo filamento en las hipótesis (C*) se puede expresar como una curva $f : [0,1] \to \mathbb{R}^d$ en las hipótesis (A*) recogidas a continuación, que no son más que las hipótesis (A) cambiando (A4) por (A4*), pues nuestra curva ahora siempre es cerrada.

(A*) **Sobre la curva**

- (A1*) La curva es simple, es decir, f es inyectiva en $(0,1)$.
- (A2*) La curva f es continua y tiene gradiente no nulo y finito en cada punto.
- (A3*) Se cumple $0 < \sigma < \text{reach}(\Gamma)$.
- (A4*) La curva es cerrada, esto es, $f(0) = f(1)$ y $T(0) = T(1)$.

Para ello, haremos uso del siguiente teorema de clasificación de variedades de dimensión 1, cuya prueba puede encontrarse en Milnor (1965).

Teorema 4.1 (Apéndice en Milnor, 1965)**.** *Toda variedad diferenciable de dimensión $k = 1$ conexa es difeomorfa o bien al círculo $\mathbb{S}^1 \subset \mathbb{R}^2$ o bien a uno de los tres intervalos de números reales $[0,1]$, $(0,1]$ o $(0,1)$.*

Como la hipótesis (C1*) exige que la variedad sea además compacta y sin borde, del Teorema 4.1 deducimos que todo filamento en la hipótesis (C1*) es difeomorfo al círculo $\mathbb{S}^1$. Tomando una parametrización del círculo $g : [0,1] \to \mathbb{R}^2$ diferenciable, por ejemplo $g(u) = (\cos(2\pi u), \sin(2\pi u))$, y componiendo con el difeomorfismo anterior, obtenemos la curva $f : [0,1] \to \mathbb{R}^d$ deseada. Ahora es claro que todo filamento en las hipótesis (C*) verifica las hipótesis (A*), pues (A1*) se cumple por la inyectividad de g en $(0,1)$ y (A2*) gracias a su suavidad.

Como mencionamos, los Teoremas 3.1, 3.3 y 3.4 en el modelo aditivo solo dependían de las hipótesis sobre la curva y de la condición $S = \Gamma \oplus \sigma B$, por lo que también se verifican en este modelo, particularizados al caso en el que la curva es cerrada. A continuación los enunciamos.

Teorema 4.2. *Bajo las hipótesis (A*) y $S = \Gamma \oplus \sigma B$, en particular bajo las hipótesis (C*)–(D*), se cumple lo siguiente.*

(1) $S = \mathcal{T}$.

(2) Para cada $u \neq v \in [0,1]$, $L(u)$ *y* $L(v)$ *son disjuntos.*

(3) Para cada $y \in \mathcal{T}$, *existe un único* u *tal que la fibra* $L(u)$ *contiene a* y, $f(u)$ *es el punto más cercano a* y *del filamento* Γ, *y además* $f(u) + \sigma N(u)$ *es el punto más cercano a* y *de la frontera* ∂S. *Si* $y \in \Gamma$, *entonces* $N(u)$ *puede ser cualquier vector en* $\mathcal{N}(u)$; *si, en cambio,* $y \notin \Gamma$, *entonces el punto de la frontera más cercano es único con* $N(u) = (y - f(u))/\|y - f(u)\|$.

Teorema 4.3. *Bajo las hipótesis (A*) y $S = \Gamma \oplus \sigma B$, en particular bajo las hipótesis (C*)–(D*), se cumple* $\Gamma = M(S)$.

Teorema 4.4. *Bajo las hipótesis (A*) y $S = \Gamma \oplus \sigma B$, en particular bajo las hipótesis (C*)–(D*), dado $y \in S$, se cumple lo siguiente.*

(1) $y \in \Gamma$ *si y sólo si* $\Lambda(y) = \sigma$.

(2) Si $y \in S \setminus \Gamma$, *entonces* $\Lambda(y) < \sigma$.

(3) $d(y, \Gamma) + \Lambda(y) = \sigma$.

4.2. El estimador

El Teorema 4.4 afirma que, para este modelo de ruido perpendicular y curva cerrada, el filamento también se caracteriza como el conjunto de puntos que maximiza la EDT. El estimador que proponemos consiste en una mejora del estimador EDT de Genovese *et al.* (2012a) (Algoritmo 3.5), que también se basa en la propiedad anterior, de la siguiente manera.

En primer lugar, estimamos el soporte S de la distribución de X_i. En vez de emplear el estimador de Devroye–Wise, utilizamos la envoltura r–convexa de la muestra (Definición 2.11)

$$\widehat{S}_n = C_r(\mathcal{X}_n) = \bigcap_{\mathring{B}(y,r)\cap\mathcal{X}_n=\emptyset} \mathring{B}(y,r)^c = (\mathcal{X}_n \oplus r\mathring{B}) \ominus r\mathring{B}.$$

A partir de este estimador del soporte también definimos la EDT empírica $\widehat{\Lambda}(y) = d(y, \partial\widehat{S}_n)$ y estimamos el nivel de ruido σ por $\widehat{\sigma} = \max_{y\in\widehat{S}_n} \widehat{\Lambda}(y)$. Finalmente, definimos el estimador EDT con envoltura r–convexa como el conjunto de puntos de $\widehat{S}_n$ suficientemente lejos de su frontera

$$\widehat{\Gamma}_n = \{y \in \widehat{S}_n : d(y, \partial\widehat{S}_n) \geq \widehat{\sigma} - \delta_n\}. \tag{4.1}$$

En ocasiones, cuando no haya peligro de confusión con el estimador de Genovese *et al.* (2012a), podremos denotar a este estimador simplemente por "estimador EDT". El parámetro $\delta_n \in (0, \widehat{\sigma})$ también regula cómo de grande es $\widehat{\Gamma}_n$; cuanto menor es δ_n, más contraemos $\widehat{S}_n$ para obtener $\widehat{\Gamma}_n$. Si δ_n es demasiado pequeño el estimador $\widehat{\Gamma}_n$ puede fragmentarse y no converger a Γ, mientras que si δ_n es demasiado grande la convergencia a Γ puede ser lenta, véase la Sección 4.3.2 para más detalles.

Algoritmo 4.5 (Estimador EDT con envoltura r–convexa)**.**

ENTRADA: Las observaciones $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, el parámetro de forma $r > 0$ y el parámetro $\delta > 0$.

SALIDA: El estimador EDT con envoltura r–convexa $\widehat{\Gamma}$, un conjunto de puntos.

1. Estimar el soporte con la envoltura r–convexa de la muestra $\widehat{S} = C_r(\mathcal{X}_n)$.
2. Estimar el nivel de ruido por $\widehat{\sigma} = \max_{y\in\widehat{S}} \widehat{\Lambda}(y)$, donde $\widehat{\Lambda}(y) = d(y, \partial\widehat{S})$ es la EDT empírica.
3. Obtener $\widehat{\Gamma} = \{y \in \widehat{S} : d(y, \partial\widehat{S}) \geq \widehat{\sigma} - \delta\}$.

Como vemos, para calcular el estimador $\widehat{\Gamma}$ necesitamos seleccionar el valor de dos parámetros; $r > 0$ para la envoltura r–convexa y $\delta > 0$.

La Figura 4.1 ilustra el funcionamiento del Algoritmo 4.5 para estimar la circunferencia unidad. El valor escogido para r es la diferencia entre el *reach* de la curva poblacional, que es el radio de la circunferencia, y el nivel de ruido, que es $\sigma = 0.5$. La justificación teórica para esta elección es que el soporte S verifica el rodamiento exterior (relacionado con la r–convexidad) con este valor de r gracias al Lema 4.9 que probaremos más adelante. Por otro lado, el valor empleado de δ es $0.2\widehat{\sigma}$. En muchas ocasiones, en lugar de seleccionar un valor de $\delta \in (0, \widehat{\sigma})$ en términos absolutos, resulta más práctico fijar una proporción de $\widehat{\sigma}$.

En la siguiente sección estudiaremos los valores de r y δ_n a nivel teórico, mientras que en la Sección 4.3 veremos qué podemos emplear en la práctica cuando los desconocemos.

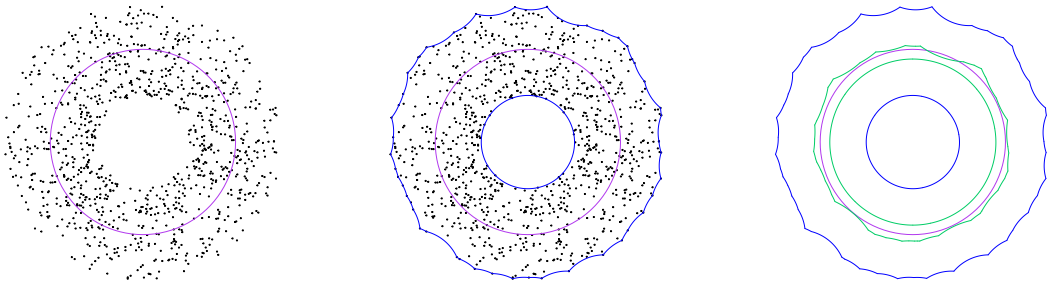

Figura 4.1: A la izquierda, muestra de tamaño $n = 1000$ con ruido perpendicular simulada de la circunferencia unidad (morado) ($\text{reach}(\Gamma) = 1$) con nivel de ruido $\sigma = 0.5$. En el centro, en azul, frontera de la envoltura r–convexa $C_r(\mathcal{X}_n)$ para $r = \text{reach}(\Gamma) - \sigma = 0.5$. A la derecha, en verde, frontera del estimador EDT con envoltura r–convexa $\widehat{\Gamma}$ para $\delta = 0.2\widehat{\sigma}$.

4.2.1. Tasa de convergencia

Para obtener la tasa de convergencia del estimador EDT basado en la envoltura r–convexa $\widehat{\Gamma}_n$ seguiremos el mismo esquema de pasos que para el estimador EDT de Genovese *et al.* (2012a). En primer lugar, obtendremos la tasa de convergencia del estimador del soporte, la envoltura r–convexa $\widehat{S}_n = C_r(\mathcal{X}_n)$. Para ello, probaremos que estamos en las hipótesis del Teorema 2.20.

La primera hipótesis a verificar es que la densidad q de la distribución Q de X_i está esencialmente acotada inferiormente por una constante positiva. Como ya comentamos, cuando el ruido es perpendicular, este hecho se cumple como consecuencia del Lema 3.12, como recoge la siguiente proposición.

Proposición 4.6. *Bajo las hipótesis (C*)–(D*), existen constantes* $0 < K_1 \leq K_2 < \infty$ *tales que*

$$K_1 \leq \operatorname*{ess\,inf}_{y \in S} q(y) \leq \operatorname*{ess\,sup}_{y \in S} q(y) \leq K_2.$$

La otra hipótesis necesaria para aplicar el Teorema 2.20 es que el soporte S esté en las condiciones del Teorema 2.19. Según la condición (III) en dicho teorema, necesitamos probar una condición de rodamiento libre por dentro y otra por fuera del conjunto S. Para el rodamiento por fuera del soporte, usaremos el siguiente resultado debido a Federer (1959).

Lema 4.7 (Corolario 4.9 en Federer, 1959)**.** *Sean* $A \subset \mathbb{R}^d$ *un conjunto cerrado no vacío y* $\sigma > 0$*. Entonces,*

$$\text{reach}(A \oplus \sigma B) \geq \text{reach}(A) - \sigma.$$

Además, también emplearemos el siguiente lema, cuya prueba puede consultarse en Pateiro-López (2008).

Lema 4.8 (Lema A.0.2 en Pateiro-López, 2008)**.** *Sean* $S \subset \mathbb{R}^d$ *un conjunto cerrado no vacío y* $r > 0$*. Supongamos que la bola* rB *rueda libremente dentro de* S*. Entonces,*

$$\partial S = \partial(\overline{S^c}).$$

La importancia del Lema 4.8 radica en que relaciona la frontera del conjunto S al que se refiere el rodamiento libre por dentro con la del conjunto $\overline{S^c}$ al que se refiere el rodamiento libre por fuera. Nos servirá para relacionar las dos condiciones de rodamiento presentadas en el Capítulo 2, a saber, la

r–rolling (Definición 2.13) y el rodamiento libre (Definición 2.14), que en general no son equivalentes. Ahora ya podemos enunciar y probar el siguiente resultado, que recoge que el soporte S cumple tanto por fuera como por dentro las dos condiciones de rodamiento.

Lema 4.9. *Bajo las hipótesis (A*) y $S = \Gamma \oplus \sigma B$, en particular bajo las hipótesis (C*)–(D*), sea $\alpha = \text{reach}(\Gamma) - \sigma > 0$. Entonces, se cumple lo siguiente.*

(1) La bola σB rueda libremente dentro de S.

*(2) S^c cumple la σ–*rolling.

*(3) S cumple la α–*rolling.

(4) La bola αB rueda libremente dentro de $\overline{S^c}$.

Demostración.

(1) Basta notar que, dado $x \in \partial S \subset S = \bigcup_{p\in\Gamma} B(p,\sigma)$, existe $p \in \Gamma \subset S$ tal que $x \in B(p,\sigma) \subset S$.

(2) Es consecuencia de (1) y del Lema 2.15.

(3) Por el Lema 4.7,
$$\text{reach}(S) = \text{reach}(\Gamma \oplus \sigma B) \geq \text{reach}(\Gamma) - \sigma = \alpha > 0.$$
Por el Teorema 2.16, S es α–convexo y por tanto cumple la α–*rolling*.

(4) Por el punto (3) anterior, como S cumple la α–*rolling*, para todo $x \in \partial S$ existe $p \in S^c$ tal que $x \in B(p,\alpha)$ y $\mathring{B}(p,\alpha) \cap S = \emptyset$. Ahora bien, $p \in S^c$ implica $p \in \overline{S^c}$, $\mathring{B}(p,\alpha) \cap S = \emptyset$ implica $B(p,\alpha) \subset \overline{S^c}$ y, como σB rueda libremente dentro de S gracias al punto (1), por el Lema 4.8, $\partial(\overline{S^c}) = \partial S$, con lo que αB rueda libremente dentro de $\overline{S^c}$.

□

Como consecuencia, obtenemos que el soporte S está en las condiciones del Teorema 2.19.

Proposición 4.10. *Bajo las hipótesis (A*) y $S = \Gamma \oplus \sigma B$, en particular bajo las hipótesis (C*)–(D*), si $r > 0$ cumple*
$$0 < r \leq \text{mín}\{\text{reach}(\Gamma) - \sigma, \sigma\}, \tag{4.2}$$
entonces S está en las condiciones del Teorema 2.19 para ese valor de r.

Demostración. Es consecuencia de la condición (III) del Teorema 2.19 gracias a los puntos (1) y (4) del Lema 4.9 y al hecho de que, si una bola de un cierto radio rueda libremente dentro de un conjunto, cualquier bola de radio menor también lo hace. □

Cabe notar que la Proposición 4.10 no asegura que el lado derecho en (4.2) sea el mayor valor de r para el que S está en las condiciones del Teorema 2.19. Simplemente nos da un valor para el cual se cumple. Además, como veremos en la Sección 4.3.1, en realidad el valor de r a utilizar en la envoltura r–convexa no tiene por qué ser tal que se verifique una condición de rodamiento por fuera y por dentro, sino que viene determinado por la r–convexidad, que está relacionada únicamente con el rodamiento exterior.

Una vez que tenemos la tasa de convergencia del estimador del soporte, el siguiente paso, al igual que en la Sección 3.1.3, es probar que el estimador EDT con envoltura r–convexa $\widehat{\Gamma}_n$ conserva dicha tasa. Para ello, queremos probar resultados análogos a los Teoremas 3.9 y 3.10. En esos resultados

pedíamos que el estimador del soporte cumpliera $S \subset \widehat{S} \subset S \oplus \varepsilon B$ y $d_H(\partial\widehat{S}, \partial S) \leq \varepsilon$. A diferencia del estimador de Devroye–Wise, que contiene al soporte S e.a.s. (véase Teorema 3.8), a la envoltura r–convexa de la muestra $\widehat{S}_n = C_r(\mathcal{X}_n)$ le ocurre lo contrario, es decir, está contenida en el soporte siempre que S sea r–convexo. En lugar de $S \subset \widehat{S} \subset S \oplus \varepsilon B$, la condición que vamos pedir sobre el estimador del soporte para probar la tasa de convergencia del estimador EDT es $\Gamma \subset \widehat{S} \subset S$, es decir, además de que el estimador $\widehat{S}$ esté contenido en el soporte, pedimos también que contenga al filamento. Esto nos permitirá probar más adelante que el estimador $\widehat{\Gamma}$ también contiene al filamento (Proposición 4.17). La Proposición 4.13 prueba que, para la envoltura r–convexa, en efecto se cumple $\Gamma \subset \widehat{S}_n$ e.a.s. Para su demostración utilizaremos dos resultados que se emplean en Rodríguez-Casal (2007) como resultados intermedios para probar el Teorema 2.20.

Proposición 4.11 (Proposición 2 en Rodríguez-Casal, 2007)**.** *Bajo las hipótesis del Teorema 2.20, existe una constante $c > 0$ suficientemente grande tal que*

$$\partial S \subset B(\mathcal{Z}_n, 2\xi_n) \quad \textit{e.a.s.},$$

siendo $\xi_n = (c\log(n)/n)^{1/(d+1)}$ y $\mathcal{Z}_n = \{X_i \in \mathcal{X}_n \colon d(X_i, \partial S) \leq \xi_n^2\}$.

El segundo resultado intermedio, el Lema 4.12, proporciona una especie de "cota inferior" para la envoltura r–convexa $C_r(\mathcal{X}_n)$, que usaremos para probar que $\Gamma \subset C_r(\mathcal{X}_n)$ e.a.s.

Lema 4.12 (Lema 3 en Rodríguez-Casal, 2007)**.** *Bajo las hipótesis del Teorema 2.20, sea $\xi_n > 0$. Supongamos que $\partial S \subset B(\mathcal{Z}_n, 2\xi_n)$, siendo $\mathcal{Z}_n = \{X_i \in \mathcal{X}_n \colon d(X_i, \partial S) \leq \xi_n^2\}$, y que $d_H(\mathcal{X}_n, S) < r$. Entonces,*

$$S \ominus (L\xi_n^2)B \subset C_r(\mathcal{X}_n) \qquad \textit{para } L \geq \text{máx}\left\{2, \frac{8}{r}\right\}.$$

Proposición 4.13. *Bajo las hipótesis (C*)–(D*), sean $r > 0$ tal que S está en las condiciones del Teorema 2.19 (en particular cualquier r dado por (4.2)) y $\widehat{S}_n = C_r(\mathcal{X}_n)$. Entonces, se cumple e.a.s.*

$$\Gamma \subset \widehat{S}_n \subset S.$$

Demostración. Por la condición (II) del Teorema 2.19, S es r–convexo, de modo que $\widehat{S}_n \subset S$.

Por otra parte, gracias a la Proposición 4.6, estamos en las hipótesis del Teorema 2.20. Por la Proposición 4.11, $\partial S \subset B(\mathcal{Z}_n, 2\xi_n)$ e.a.s., siendo ξ_n y $\mathcal{Z}_n$ los definidos en dicha proposición. Además, como con probabilidad uno $d_H(\mathcal{X}_n, S) \to 0$, tenemos que $d_H(\mathcal{X}_n, S) < r$ e.a.s. Aplicando el Lema 4.12 obtenemos que, e.a.s.,

$$S \ominus (L\xi_n^2)B \subset \widehat{S}_n \qquad \text{para } L \geq \text{máx}\left\{2, \frac{8}{r}\right\}.$$

Como $L\xi_n^2 \to 0 < \sigma$, para n suficientemente grande tenemos $S \ominus \sigma B \subset S \ominus (L\xi_n^2)B$. Finalmente, como $\Gamma \subset \big(\Gamma \oplus \sigma B\big) \ominus \sigma B = S \ominus \sigma B$, se deduce $\Gamma \subset \widehat{S}_n$ e.a.s. □

Las Proposiciones 4.16 y 4.17 adaptan los Teoremas 3.9 y 3.10, respectivamente, al caso de un estimador que cumpla $\Gamma \subset \widehat{S} \subset S$ y $d_H(\partial\widehat{S}, \partial S) \leq \varepsilon$. Antes vamos a probar un par de lemas que serán de utilidad al trabajar con distancias.

Lema 4.14. *Si $A, C \subset \mathbb{R}^d$ son compactos no vacíos e $y \in \mathbb{R}^d$, entonces $|d(y, A) - d(y, C)| \leq d_H(A, C)$.*

Demostración. Sean z_* un punto más cercano a y en C, $\widehat{z}$ un punto más cercano a y en A y $\widetilde{z}$ un punto más cercano a z_* en A. Entonces

$$d(y, A) = \|y - \widehat{z}\| \leq \|y - \widetilde{z}\| \leq \|y - z_*\| + \|z_* - \widetilde{z}\| = d(y, C) + d(z_*, A) \leq d(y, C) + d_H(C, A).$$

Intercambiando los papeles de A y C se obtiene $d(y, C) - d(y, A) \leq d_H(A, C)$, de donde se concluye $|d(y, A) - d(y, C)| \leq d_H(A, C)$. □

Lema 4.15. *Si $A \subset C$ son compactos en $\mathbb{R}^d$ y $a \in A$, entonces $d(a, \partial A) \leq d(a, \partial C)$.*

Demostración. Si $a \in \partial A$ el resultado es inmediato, así que supongamos $a \in \text{int}(A)$. Como ∂C es compacto, sea $x \in \partial C$ tal que $d(a, x) = d(a, \partial C)$. Si $x \in \partial A$ de nuevo el resultado es inmediato, así que supongamos $x \notin \partial A$. En ese caso, $x \in A^c$, pues en caso contrario x pertenecería a $A \setminus \partial A = \text{int}(A) \subset \text{int}(C)$, en contra de que $x \in \partial C$. Sea L el segmento que une a con x y probemos que existe $y \in L$ tal que $y \in \partial A$, con lo que deduciremos $d(a, \partial A) \leq d(a, y) \leq d(a, x) = d(a, \partial C)$. Supongamos por el contrario que $L \cap \partial A = \emptyset$, entonces $L \subset \text{int}(A) \cup A^c$. Definiendo $U = \text{int}(A) \cap L$ y $V = A^c \cap L$, ambos conjuntos son abiertos en la topología de L, no vacíos ($a \in U$ y $x \in V$), disjuntos y su unión es L, lo que contradice la conexidad del segmento. □

Recordemos que definíamos $\widehat{y} = \arg\max_{y \in \widehat{S}} \widehat{\Lambda}(y)$, posiblemente no único, de modo que $\widehat{\Lambda}(\widehat{y}) = \widehat{\sigma}$.

Proposición 4.16. *Bajo las hipótesis (A*) y $S = \Gamma \oplus \sigma B$, en particular bajo las hipótesis (C*)–(D*), supongamos que $\Gamma \subset \widehat{S} \subset S$ son compactos y que $d_H(\partial\widehat{S}, \partial S) \leq \varepsilon$. Entonces, se cumple lo siguiente.*

(1) $\sup_{y \in \mathbb{R}^d} |\widehat{\Lambda}(y) - \Lambda(y)| \leq \varepsilon$.

(2) $\widehat{\sigma} \leq \sigma \leq \widehat{\sigma} + \varepsilon$.

(3) $d(\widehat{y}, \Gamma) \leq \varepsilon$ *(siendo $\widehat{y} \in \widehat{S}$ tal que $\widehat{\Lambda}(\widehat{y}) = \widehat{\sigma}$).*

Demostración.

(1) Es consecuencia del Lema 4.14.

(2) Para la primera desigualdad, gracias al Lema 4.15 y a que $\widehat{y} \in S$,

$$\widehat{\sigma} = d(\widehat{y}, \partial\widehat{S}) \leq d(\widehat{y}, \partial S) \leq \max_{y \in S} d(y, \partial S) = \sigma.$$

Para la segunda, sea y_* un punto de Γ más cercano a $\widehat{y}$. Usando el Teorema 4.4–(1), el apartado (1) anterior y que $y_* \in \widehat{S}$,

$$\sigma = \Lambda(y_*) \leq \widehat{\Lambda}(y_*) + \varepsilon \leq \widehat{\Lambda}(\widehat{y}) + \varepsilon = \widehat{\sigma} + \varepsilon.$$

(3) Como $\widehat{y} \in S$, usando el Teorema 4.4–(3), el Lema 4.15 y el apartado (2) anterior,

$$d(\widehat{y}, \Gamma) = \sigma - d(\widehat{y}, \partial S) \leq \sigma - d(\widehat{y}, \partial\widehat{S}) = \sigma - \widehat{\sigma} \leq \varepsilon.$$

□

Es interesante notar respecto del apartado (2) de la Proposición 4.16 que, a diferencia del Teorema 3.9–(2), como ahora nuestro estimador está por dentro del soporte S, se cumple $\widehat{\sigma} \leq \sigma$.

En el siguiente resultado, para poder probar que $\widehat{\Gamma}$ contiene a Γ y acotar la distancia de Hausdorff entre ambos conjuntos, necesitamos suponer que el parámetro δ es mayor o igual que ε. En el Teorema 3.10, Genovese *et al.* (2012a) suponían $\delta = 2\varepsilon$, pero nosotros lo enunciamos con más generalidad.

Proposición 4.17. *Bajo las hipótesis (A*) y $S = \Gamma \oplus \sigma B$, en particular bajo las hipótesis (C*)–(D*), supongamos que $\Gamma \subset \widehat{S} \subset S$ son compactos y que $d_H(\partial\widehat{S}, \partial S) \leq \varepsilon$. Sea $\widehat{\Gamma} = \{y \in \widehat{S} \colon d(y, \partial\widehat{S}) \geq \widehat{\sigma} - \delta\}$ el estimador EDT con $\delta \geq \varepsilon$. Entonces $\Gamma \subset \widehat{\Gamma} \subset \Gamma \oplus (2\delta)B$ y por tanto*

$$d_H(\widehat{\Gamma}, \Gamma) \leq 2\delta.$$

Demostración. Veamos primero que si $y \in \widehat{\Gamma}$ entonces $d(y,\Gamma) \leq 2\delta$, con lo que probaremos que $\widehat{\Gamma} \subset \Gamma \oplus (2\delta)B$. Sea $y \in \widehat{\Gamma}$, usando que $\widehat{\Gamma} \subset \widehat{S} \subset S$, el Teorema 4.4–(3), el Lema 4.15 y la Proposición 4.16–(2),

$$d(y,\Gamma) = \sigma - d(y,\partial S) \leq \sigma - d(y,\partial\widehat{S}) \leq \sigma - (\widehat{\sigma} - \delta) \leq \sigma - (\sigma - \varepsilon - \delta) = \varepsilon + \delta \leq 2\delta.$$

Por otro lado, veamos que $\Gamma \subset \widehat{\Gamma}$. Sea $y \in \Gamma$, entonces también $y \in \widehat{S}$ y además, usando la Proposición 4.16–(1), el Teorema 4.4–(1) y la Proposición 4.16–(2),

$$d(y,\partial\widehat{S}) \geq d(y,\partial S) - \varepsilon = \sigma - \varepsilon \geq \widehat{\sigma} - \varepsilon \geq \widehat{\sigma} - \delta,$$

de modo que $y \in \widehat{\Gamma}$. □

Finalmente, aplicando la Proposición 4.17 al estimador EDT con envoltura r–convexa que proponemos, obtenemos el Teorema 4.18, que proporciona su tasa de convergencia. Como ya comentamos en la Proposición 4.17, el parámetro δ_n, que visto como una sucesión en n debe tender a cero, debe ser mayor o igual que ε_n, que representa la tasa de convergencia de la envoltura r–convexa como estimador del soporte. Si δ_n es "demasiado pequeño", $\widehat{\Gamma}_n$ podría fragmentarse y no contener a Γ, y la Proposición 4.17 no sería válida. De este modo, el parámetro δ_n limita la velocidad de convergencia del estimador pero, si no es tampoco "demasiado grande", es decir, si es del mismo orden que ε_n, el estimador EDT no empeora (conserva) la tasa de convergencia de la envoltura r–convexa.

Teorema 4.18. *Bajo las hipótesis (C*)–(D*), sean $r > 0$ tal que S está en las condiciones del Teorema 2.19 (en particular cualquier r dado por (4.2)), $\widehat{S}_n = C_r(\mathcal{X}_n)$ y $\widehat{\Gamma}_n$ el estimador EDT con envoltura r–convexa definido en (4.1). Si $\delta_n \to 0$ es tal que $\delta_n \geq \varepsilon_n$ e.a.s., donde ε_n es el dado por (2.10), entonces con probabilidad uno*

$$d_H(\widehat{\Gamma}_n,\Gamma) = O(\delta_n).$$

En particular, si además $\delta_n = O(\varepsilon_n)$ con probabilidad uno, entonces con probabilidad uno

$$d_H(\widehat{\Gamma}_n,\Gamma) = O\left(\left(\frac{\log(n)}{n}\right)^{2/(d+1)}\right).$$

Demostración. Por la Proposición 4.6, aplicando el Teorema 2.20, obtenemos que $d_H(\partial\widehat{S}_n,\partial S) \leq \varepsilon_n$ e.a.s., siendo ε_n dado por (2.10). Por la Proposición 4.13, se cumple $\Gamma \subset \widehat{S}_n \subset S$ e.a.s. Aplicando la Proposición 4.17 se deduce el resultado. □

Del Teorema 4.18 deducimos que, cuando la dimensión del espacio ambiente es $d = 2$, el estimador EDT con envoltura r–convexa alcanza la tasa $O\left((\log(n)/n)^{2/3}\right)$, que es, salvo factor logarítmico, la tasa óptima en el sentido minimax para estimar el filamento en este modelo (Teoremas 3.13 y 3.14 para $k = 1$).

4.3. Consideraciones prácticas

Como ya resaltamos, en la práctica, el estimador EDT con envoltura r–convexa depende de dos parámetros r y δ, que sería deseable poder calcular a partir de la muestra, de modo que el estimador sea totalmente basado en los datos o *data–driven*. En esta sección estudiaremos posibles elecciones de ambos parámetros. Además, el estimador EDT es un conjunto de puntos (no una curva), por lo que también proponemos una forma de extraer una curva a partir de él. Finalmente, discutimos el papel de las hipótesis consideradas a la hora de aplicar el estimador en la práctica.

4.3.1. Selector del parámetro de forma r

El parámetro de forma $r > 0$ a utilizar en la envoltura r–convexa $\widehat{S} = C_r(\mathcal{X}_n)$ puede tener una gran influencia en la estimación del soporte. Valores demasiado pequeños de r devuelven un estimador demasiado fragmentado, con demasiados huecos y componentes conexas, mientras que valores demasiado grandes de r producen lo contrario, un estimador demasiado parecido a la envoltura convexa de la muestra y que podría no estimar el verdadero soporte, véase la Figura 2.7. A nivel teórico, en la Sección 4.2.1 vimos que es posible utilizar cualquier valor de r tal que el soporte S está en las condiciones del Teorema 2.19 (ver Teorema 4.18), que en esencia es una condición de doble rodamiento, por dentro y por fuera del soporte, y probamos que tales valores existen en la Proposición 4.10. El rodamiento interior y exterior en realidad no juegan un papel simétrico. El rodamiento interior solo es necesario para garantizar una regularidad sobre el soporte S que permita obtener la tasa de convergencia, mientras que el parámetro r de rodamiento exterior, el que está relacionado con la r–convexidad (recuérdese el Teorema 2.16), es el que realmente determina los valores que podemos usar en la envoltura r–convexa.

¿Cuál es entonces el valor óptimo del parámetro r que debemos aspirar a estimar a partir de la muestra con un selector, pongamos $R_n \equiv R_n(\mathcal{X}_n)$, que después usaremos para calcular la envoltura R_n–convexa $C_{R_n}(\mathcal{X}_n)$? Siguiendo a Rodríguez-Casal y Saavedra-Nieves (2016, 2022a,b), definimos el valor óptimo como el mayor valor de r tal que el soporte es r–convexo, que denotamos por r_0.

Definición 4.19. Dado $S \subset \mathbb{R}^d$ compacto, no convexo y r–convexo para algún $r > 0$, definimos

$$r_0 = \sup\{\gamma > 0 \colon C_\gamma(S) = S\}. \tag{4.3}$$

En la Definición 4.19 y en adelante, para facilitar la exposición, suponemos que S es no convexo, pues en caso de que lo fuera r_0 sería infinito y sería más apropiado usar la envoltura convexa en lugar de la r–convexa. Esto no supone un problema en nuestro problema de estimación de filamentos, pues el soporte, que es un entorno tubular de una curva cerrada, nunca va a ser convexo en la práctica. La Proposición 4.20, debida a Rodríguez-Casal y Saavedra-Nieves (2016), establece que, bajo la condición de doble rodamiento (R_λ^r) introducida en la Definición 2.17, que permite un radio de rodamiento distinto por fuera que por dentro del conjunto, el supremo en (4.3) es en realidad un máximo y $C_{r_0}(S) = S$.

Proposición 4.20 (Proposición 2.4 en Rodríguez-Casal y Saavedra-Nieves, 2016)**.** *Sea $S \subset \mathbb{R}^d$ no vacío, compacto y no convexo y sea r_0 definido en (4.3). Si S verifica la condición (R_λ^r), entonces $C_{r_0}(S) = S$.*

Si consideramos un valor de r mayor que r_0, $C_r(S)$ no coincide con S y por tanto $C_r(\mathcal{X}_n)$ no estima el soporte S. Por otra parte, bajo la condición de que r_0 sea un máximo, si r es menor que r_0, $C_r(\mathcal{X}_n)$ no es un estimador admisible, pues $C_r(\mathcal{X}_n) \subset C_{r_0}(\mathcal{X}_n) \subset S$, de modo que $C_{r_0}(\mathcal{X}_n)$ siempre está a menor distancia del soporte S que pretende estimar. Esto justifica la elección de r_0 como el parámetro óptimo a estimar.

Para el caso que nos ocupa, el estimador EDT con envoltura r–convexa, vamos a suponer que disponemos de un estimador R_n de r_0, en principio dependiente de la muestra y por tanto aleatorio, tal que

$$0 < r_* \leq R_n \leq r_0 \quad \text{e.a.s.},$$

siendo $r_* > 0$ una constante. Lo que estamos diciendo es que, con probabilidad uno, para n suficientemente grande, R_n está por debajo del parámetro r_0 que pretende estimar, pues si estuviera por encima la estimación de S por $C_{R_n}(\mathcal{X}_n)$ podría ser inconsistente, pero de tal manera que R_n no tiende a cero (R_n no se corresponde con la idea habitual de parámetro de suavizado). Bajo esta condición, en el Teorema 4.21 probamos que la tasa de convergencia del estimador EDT con envoltura R_n–convexa se mantiene con respecto a la probada en el Teorema 4.18.

Teorema 4.21. *Bajo las hipótesis (C*)–(D*), supongamos que S no es convexo. Sean r_0 definido en (4.3), $R_n > 0$ tal que $0 < r_* \leq R_n \leq r_0$ e.a.s., $\widehat{S}_n = C_{R_n}(\mathcal{X}_n)$ y $\widehat{\Gamma}_n$ el estimador EDT con envoltura R_n–convexa definido en (4.1). Si $\delta_n \to 0$ es tal que $\delta_n \geq \varepsilon_n$ e.a.s., donde ε_n es el dado por (2.10) para $r = \text{mín}\{\text{mín}\{\text{reach}(\Gamma) - \sigma, \sigma\}, r_*\} > 0$, entonces con probabilidad uno*

$$d_H(\widehat{\Gamma}_n, \Gamma) = O(\delta_n).$$

En particular, si además $\delta_n = O(\varepsilon_n)$ con probabilidad uno, entonces con probabilidad uno

$$d_H(\widehat{\Gamma}_n, \Gamma) = O\left(\left(\frac{\log(n)}{n}\right)^{2/(d+1)}\right).$$

Demostración. Por los apartados (2) y (3) del Lema 4.9, S verifica la condición (R_σ^α), siendo $\alpha = \text{reach}(\Gamma) - \sigma > 0$. Por la Proposición 4.20, $C_{r_0}(S) = S$. Sea $r = \text{mín}\{\widetilde{r}, r_*\} > 0$, siendo $\widetilde{r} = \text{mín}\{\text{reach}(\Gamma) - \sigma, \sigma\} > 0$. Se tiene $0 < r \leq R_n \leq r_0$ e.a.s., y por tanto

$$C_r(\mathcal{X}_n) \subset C_{R_n}(\mathcal{X}_n) \subset S \quad \text{e.a.s.}$$

Además, por la Proposición 4.10, S está en las condiciones del Teorema 2.19 para $\widetilde{r}$ y por tanto también para r. Como se verifica la Proposición 4.6, por el Teorema 2.20, $d_H(\partial C_r(\mathcal{X}_n), \partial S) \leq \varepsilon_n$ e.a.s., siendo ε_n el dado por (2.10) para r. Argumentando de modo análogo a la prueba del Teorema 2.20 en Rodríguez-Casal (2007), es posible deducir que $d_H(\partial C_{R_n}(\mathcal{X}_n), \partial S) \leq d_H(\partial C_r(\mathcal{X}_n), \partial S) \leq \varepsilon_n$ e.a.s. Por otra parte, gracias a la Proposición 4.13, $\Gamma \subset C_r(\mathcal{X}_n) \subset S$ e.a.s., de modo que $\Gamma \subset C_{R_n}(\mathcal{X}_n) \subset S$ e.a.s. Aplicando la Proposición 4.17 a $\widehat{S}_n = C_{R_n}(\mathcal{X}_n)$ se deduce el resultado. □

Observación 4.22. Si consideramos un valor de r con $0 < r \leq r_0$, como caso particular del Teorema 4.21 tomando $R_n = r$ constante, deducimos que el Teorema 4.18 es válido para cualquier $r \leq r_0$, aunque S no esté en las condiciones del Teorema 2.19 para r. Es decir, como ya hemos comentado, el parámetro r a utilizar en la envoltura r–convexa viene determinado por la r–convexidad del soporte S, esto es, el rodamiento exterior con radio r. Aunque la bola rB no ruede por dentro del conjunto S, es suficiente con que lo haga una bola de radio menor, pues el rodamiento interior simplemente tiene como objetivo garantizar una cierta regularidad del soporte.

En relación a la existencia de un estimador R_n que verifique $0 < r_* \leq R_n \leq r_0$ e.a.s., referimos al lector al estimador propuesto en Rodríguez-Casal y Saavedra-Nieves (2022b). Este estimador, denotado por $\widehat{r}_0$, que se basa en la idea de *maximal spacings*, se prueba que converge en probabilidad a r_0 (Teorema 3 en Rodríguez-Casal y Saavedra-Nieves, 2022b) bajo unas ciertas hipótesis. En concreto, la hipótesis adicional que tendríamos que pedir con respecto al modelo que consideramos aquí sería que la densidad q de las observaciones X_i fuese Lipschitz continua en el soporte S. Multiplicando este estimador por un valor $\nu \in (0, 1)$, es decir, definiendo $R_n = \nu\widehat{r}_0$, se cumple

$$\mathbb{P}(0 < r_* \leq R_n \leq r_0) \to 1$$

que, aunque no es una condición de modo casi seguro, permitiría obtener las correspondientes tasas que hemos estudiado en probabilidad. Está fuera de los propósitos de este trabajo estudiar en mayor profundidad el selector del parámetro de forma r.

4.3.2. Sobre el parámetro δ y la extracción de curva

El otro parámetro del que depende el estimador EDT basado en la envoltura r–convexa es $\delta > 0$, que controla cuánto contraemos el estimador del soporte $\widehat{S}$ hacia dentro para obtener el estimador EDT $\widehat{\Gamma}$. Si δ es demasiado pequeño, nos metemos mucho hacia dentro y el estimador se fragmenta, puede

no contener al verdadero filamento Γ y por tanto podríamos no tener consistencia de $\widehat{\Gamma}$, no aplicaría la Proposición 4.17. El Teorema 4.18 (o el Teorema 4.21) nos dice cómo de grande al menos ha de ser δ_n para garantizar la consistencia de $\widehat{\Gamma}$; δ_n, que debe tender a 0, debe ser mayor o igual que ε_n, que es una cota de error teórica que representa la precisión con la que estimamos la frontera de S. Ahora bien, en ese caso δ_n determina la velocidad de convergencia de $\widehat{\Gamma}_n$ a Γ, por lo que el valor óptimo de δ_n es el menor posible, del mismo orden que ε_n, es decir, $O\left((\log(n)/n)^{2/(d+1)}\right)$. En la práctica desconocemos ε_n, y no está relacionado con ningún otro parámetro, a diferencia del estimador EDT de Genovese *et al.* (2012a) con Devroye–Wise, donde ε_n era también el parámetro de suavizado de la muestra.

Por otra parte, en cuanto a la extracción de una curva $\widehat{\Gamma}_f$ dentro del estimador EDT con envoltura r–convexa $\widehat{\Gamma}$, proponemos proceder de forma distinta a Genovese *et al.* (2012a). En la Sección 3.1.4 la extracción de curva se basaba en la construcción de una *epsilon*–red y el correspondiente árbol de expansión mínima para después buscar el camino más corto en dicho árbol entre los dos extremos estimados. Este procedimiento depende de la precisión de la *epsilon*–red y además puede dar lugar a caminos poco suaves si no se aplica el paso de relajación posterior.

Ya que el filamento es el eje medial (Definición 3.2) del soporte S gracias al Teorema 4.3, parece razonable extraer una curva en base al eje medial del soporte estimado $\widehat{S}$. No obstante, como ya hemos resaltado, el eje medial es inestable en el sentido de que no es continuo en distancia de Hausdorff. Al calcular el eje medial de la envoltura r–convexa, típicamente obtenemos una curva interior cerrada que se encuentra próxima al filamento, junto con una gran cantidad de segmentos que unen esta curva interior con la frontera de la envoltura, véase el panel superior derecho en la Figura 4.2. Para evitar tomar estos segmentos "espurios", proponemos definir la curva extraída como la unión de aquellos segmentos que forman el eje medial de $\widehat{S}$ que están completamente contenidos en el estimador EDT $\widehat{\Gamma}$. La Figura 4.2 ilustra este procedimiento de extracción de curva.

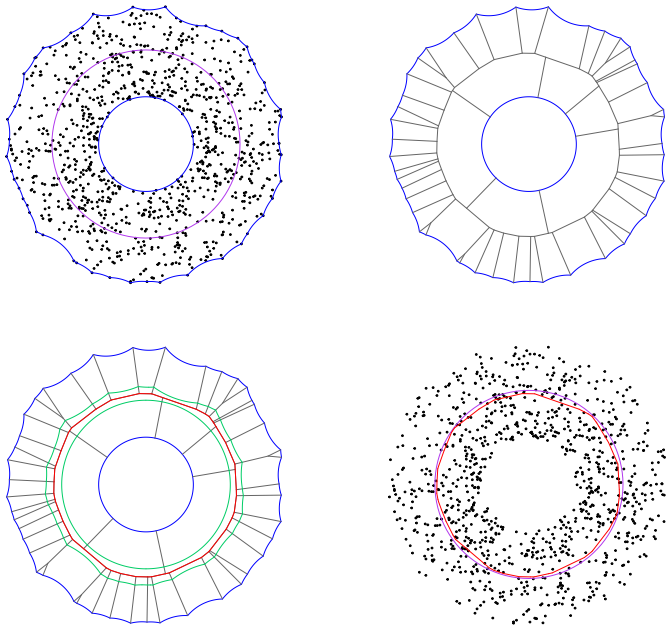

Figura 4.2: Continuación del ejemplo en la Figura 4.1. Arriba a la izquierda, en morado, filamento poblacional y, en azul, frontera de la envoltura r–convexa $C_r(\mathcal{X}_n)$ para $r = \text{reach}(\Gamma) - \sigma = 0.5$. Arriba a la derecha, en gris, segmentos que forman el eje medial de $C_r(\mathcal{X}_n)$. Abajo a la izquierda, en verde, frontera del estimador EDT con envoltura r–convexa $\widehat{\Gamma}$ para $\delta = 0.2\widehat{\sigma}$ y, en rojo, curva extraída (segmentos que forman el eje medial de $C_r(\mathcal{X}_n)$ completamente contenidos en $\widehat{\Gamma}$).

Volviendo a la elección del parámetro δ, por lo expuesto anteriormente, debemos tener cuidado de que δ no sea demasiado grande, pues en ese caso el estimador EDT sería muy ancho y la curva extraída podría contener varios segmentos espurios, véase mitad izquierda en la Figura 4.3. No obstante, δ tampoco debe ser demasiado pequeño para evitar que el estimador se fragmente y la curva extraída no sea cerrada, como ocurre en la mitad derecha en la Figura 4.3. Una posible elección sería el menor valor de δ tal que la curva extraída contenga una curva cerrada. Esto debería asegurar que el estimador EDT no se rompe en varios trozos, de modo que la curva extraída se cierra sobre si misma, imitando el carácter cerrado del filamento poblacional, y por tanto lo estima con el mismo orden de convergencia que el estimador EDT. Esta propuesta parece funcionar bien en los ejemplos simulados, pero el estudio de la convergencia teórica de la curva extraída supera los objetivos de esta memoria.

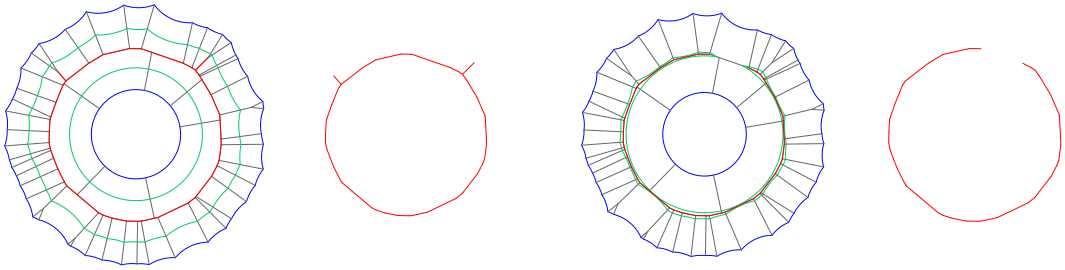

Figura 4.3: Continuación del ejemplo en la Figura 4.2. En azul, frontera de la envoltura r–convexa $C_r(\mathcal{X}_n)$ para $r = \text{reach}(\Gamma) - \sigma = 0.5$. En gris, segmentos que forman el eje medial de $C_r(\mathcal{X}_n)$. En verde, frontera del estimador EDT con envoltura r–convexa $\widehat{\Gamma}$ para $\delta = 0.5\widehat{\sigma}$ (mitad izquierda) o $\delta = 0.1\widehat{\sigma}$ (mitad derecha). En rojo, curva extraída en cada caso.

4.3.3. Sobre el modelo y las hipótesis

En la Sección 4.1 detallamos las hipótesis a considerar a la hora de proponer el nuevo estimador. Es importante mencionar que, en la práctica, hay en esencia dos hipótesis importantes a la hora de poder aplicar el nuevo estimador con un buen comportamiento: que el soporte de los datos sea una dilatación del filamento, $S = \Gamma \oplus \sigma B$, y la hipótesis de que el *reach* del filamento sea mayor que el nivel de ruido, $\sigma < \text{reach}(\Gamma)$. Que se cumpla $S = \Gamma \oplus \sigma B$ es fundamental porque el estimador EDT se basa en estimar la frontera del soporte y después "contraerse" hacia dentro una cierta distancia. Si los puntos se distanciaran más del filamento en unas zonas que en otras, esto es, si el nivel de ruido no fuese el mismo en todo el filamento, el estimador no funcionaría bien. Por otra parte, la hipótesis de *reach* es importante porque garantiza, junto con la condición anterior, que las fibras perpendiculares no se intersecan, que el eje medial del soporte es el filamento y que éste maximiza la EDT, si no se cumplieran realmente no tendría sentido emplear el estimador EDT para estimar el filamento.

En cuanto al resto de hipótesis, la suavidad y la no auto–intersección de la curva, como ya comentamos en la Sección 3.1.1, es posible que ya se cumplan gracias a la hipótesis de *reach*. Por su parte, el hecho de que la curva sea cerrada y las hipótesis sobre el modelo generador de los datos y las correspondientes distribuciones, como el ruido perpendicular, pueden incumplirse y aun así el estimador funcionar adecuadamente, siempre que se verifiquen las dos hipótesis mencionadas. Realmente solo son necesarias para la obtención de las tasas minimax, pero cualquier mecanismo generador de los datos que provoque que el soporte sea una dilatación del filamento de la forma $S = \Gamma \oplus \sigma B$ permite el uso del estimador EDT con envoltura r–convexa. De hecho, Genovese *et al.* (2012b) comentan que un revisor del artículo señala como posibilidad considerar que las observaciones tienen distribución uniforme en $M \oplus \sigma B$. Bastaría cualquier densidad esencialmente acotada inferiormente en $\Gamma \oplus \sigma B$ para mantener

la tasa de convergencia de la envoltura r–convexa, y aunque la densidad cayese a cero en la frontera del soporte el estimador seguiría funcionando, solo que con peores tasas, que aun así creemos que mejorarían a las del estimador EDT con Devroye–Wise. De este modo, el estimador se puede aplicar, por ejemplo, en el modelo de ruido aditivo visto en la Sección 3.1 (sea la curva cerrada o abierta), o en cualquier otro modelo en el que los datos se distribuyan directamente sobre el soporte $\Gamma \oplus \sigma B$.

4.4. Aplicación a datos reales: estimación de secciones de troncos de árboles en inventario forestal

En esta sección ilustraremos el uso del estimador EDT con envoltura r–convexa propuesto en este capítulo, aplicándolo a un problema del ámbito forestal, más concretamente en la estimación del contorno de secciones de troncos de árboles.

Se puede definir el inventario forestal como el proceso sistemático de recolección de datos e información forestal para su posterior análisis y tratamiento con el objetivo de estimar el valor de la madera en un cierto área forestal, evaluar el estado de la biodiversidad o determinar el riesgo de incendio, entre otros propósitos. Una de las variables más importantes a tener en cuenta en este contexto es el diámetro a la altura del pecho (dbh, *diameter at breast height*), que mide el diámetro que tiene el tronco de un árbol a una altura de aproximadamente 1.3 metros (o 4.5 pies) sobre el suelo, lo que corresponde a la altura aproximada a la que se encuentra el pecho de un adulto. Esta variable es de interés porque, por ejemplo, a partir de ella se calcula el área basal (ba, *basal area*), que se define como el área de la sección horizontal del tronco cortada a la altura del pecho (aproximadamente 1.3 m) asumiendo que la sección transversal del árbol es circular, es decir, aplicando la fórmula del área de un círculo $\text{ba} = \pi(\text{dbh}/2)^2$. En ocasiones, en lugar del área basal absoluta de uno o varios árboles, interesa el área basal relativa de los árboles de una arboleda o zona de estudio con respecto al área o extensión de la misma, que se expresa por ejemplo en metros cuadrados por hectárea de terreno (m^2/ha). A partir de esta medida se puede estimar el volumen de masa arbórea, por ejemplo para conocer el valor comercial de la madera que contiene o para evaluar el estado de una cierta zona forestal.

Tradicionalmente, el inventario forestal era una tarea realizada "manualmente". Por ejemplo, el dbh se puede obtener midiendo la circunferencia del tronco del árbol a la altura del pecho con una cinta métrica y dividiendo entre π, asumiendo forma de circunferencia, o bien midiendo directamente su grosor mediante un calibrador. Realizar estas medidas árbol por árbol es arduo, lleva mucho tiempo y es más susceptible a errores. En las últimas décadas se han desarrollado sistemas de detección remota que permiten ayudar a automatizar el proceso, como los PLS (*Personal Laser Scanner*), un tipo de escáner que transporta la persona que realiza la recogida de datos y que genera una nube tridimensional de puntos a medida que la persona se va desplazando con el escáner a lo largo del terreno, véase Figura 4.4(d). Aunque este sistema ahorra una gran cantidad de tiempo y produce unos resultados precisos, no se usa ampliamente por el coste que puede tener este equipamiento. En Gollob *et al.* (2021b) se explora otra posibilidad: el uso de la tecnología LiDAR (*Light Detection and Ranging*) que incorporan los sensores de los dispositivos iPad de Apple. Usando aplicaciones para el sistema operativo iOS se pueden conseguir nubes de puntos 3D similares a las que se obtienen con los PLS (Figura 4.4(a)–(c)) pero empleando un dispositivo de uso más común como es el iPad, aunque el tiempo de recolección de datos y el error en las mediciones sea superior.

Gollob *et al.* (2021b) comparan el uso de un PLS con el de tres aplicaciones de iPad, a saber, 3D Scanner App, Polycam y SiteScape, en la detección de la posición de los árboles y la estimación del dbh en una muestra de parcelas localizadas en el área de entrenamiento e investigación de la *University of Natural Resources and Life Science* de Viena, en Austria. Se llevan a cabo una serie de pasos de procesado de las nubes de puntos descritos en Gollob *et al.* (2020), entre ellos métodos de *clustering* basados en el algoritmo DBSCAN para identificar los árboles. El tronco de cada árbol (clúster) se

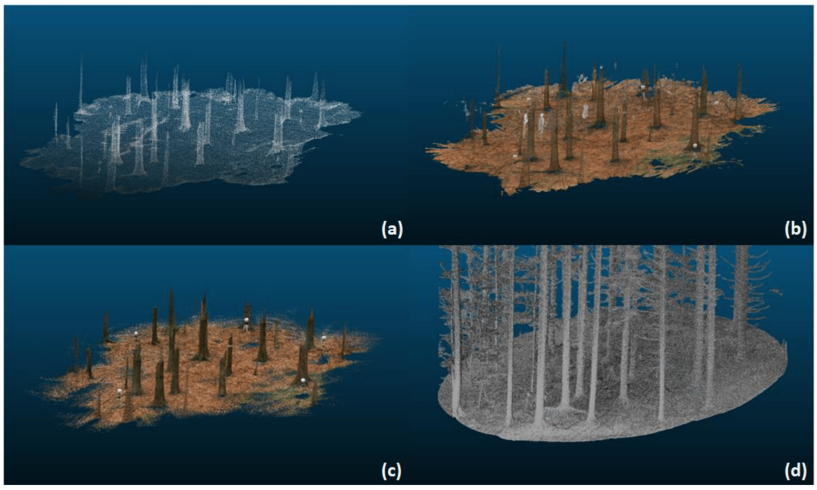

Figura 4.4: Ejemplos de nubes de puntos tridimensionales obtenidas con (a) 3D Scanner App, (b) Polycam, (c) SiteScape y (d) PLS. Figura tomada de Gollob *et al.* (2021b).

estratifica en varias capas horizontales, que comienzan aproximadamente a altura 1 m sobre el suelo y terminan a altura 2.6 m, y de grosor vertical 15 cm. En cada capa se proyectan los puntos sobre un plano horizontal para obtener una sección bidimensional del tronco del árbol, véase Figura 4.5. En estas secciones bidimensionales se aplican 5 modelos distintos, paramétricos o no paramétricos, para estimar el diámetro del tronco a la altura de esa sección, ver Figura 4.6. Dos modelos se basan en calcular el diámetro de un ajuste circular de los puntos, mientras que un tercero ajusta una elipse y calcula el diámetro como la media cuadrática de los dos radios de la elipse. Finalmente, se ajustan también dos modelos GAM, uno con *splines* de suavización y otro con *tensor product*, de la siguiente manera. Se pasan los puntos de la sección bidimensional a coordenadas polares usando como centro alguno de los ajustes anteriores, se ajusta un modelo GAM de la distancia al centro en función del ángulo, se vuelven a pasar los puntos y el ajuste a coordenadas cartesianas, a partir de este ajuste se estima el área que encierra y por último se estima el diámetro como el que corresponde a un círculo con ese área. Cabe mencionar también que se usa como medida del ruido en la nube de puntos la

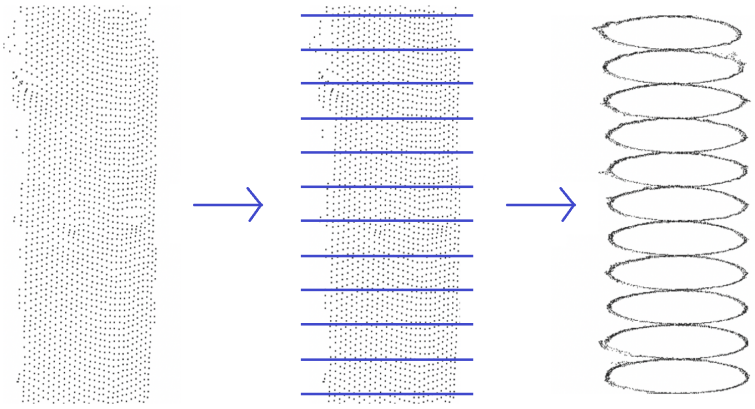

Figura 4.5: Esquema de obtención de las secciones bidimensionales horizontales. Modificación de una figura de Eto *et al.* (2020).

desviación típica de los residuos de los modelos. Finalmente, a partir de los centros y de los cinco valores obtenidos para el diámetro en cada una de las secciones horizontales del tronco, se determinan, mediante una serie de criterios, la posición y el dbh estimados del árbol. Véase Gollob *et al.* (2020, 2021b) para más detalles.

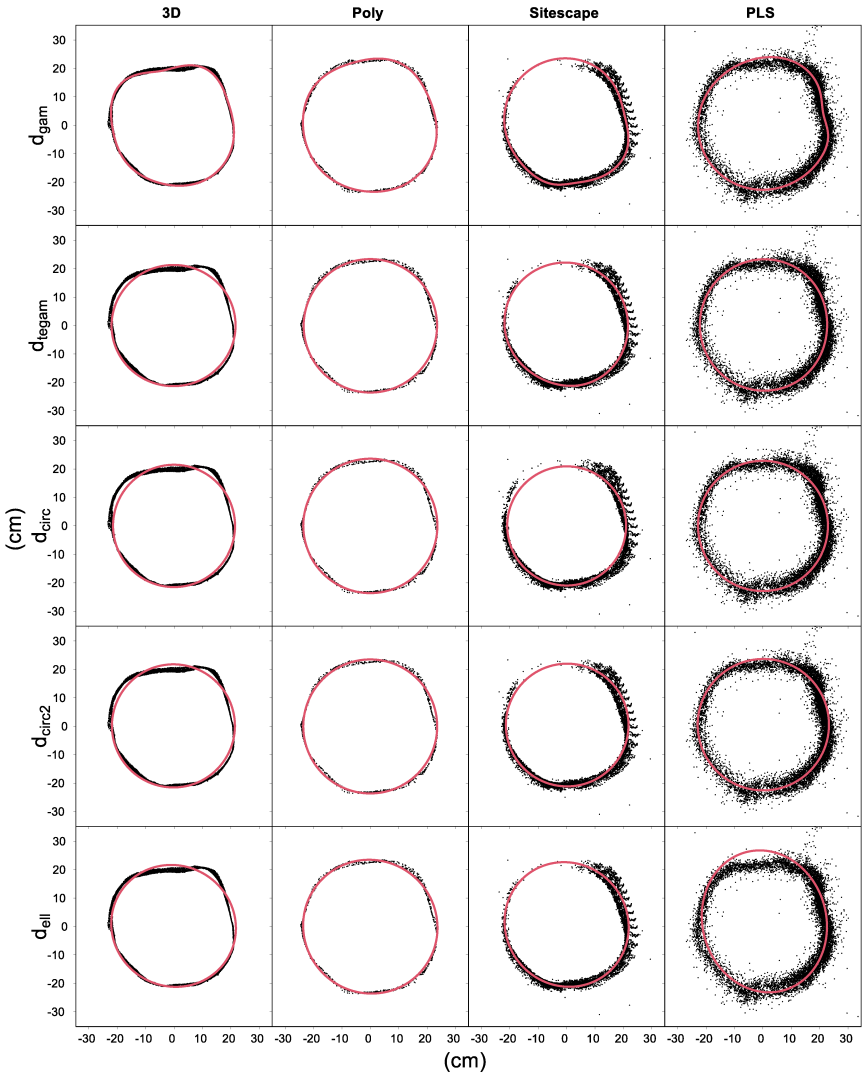

Figura 4.6: Ajustes con cada uno de los 5 modelos (por filas, de arriba abajo, GAM con *splines* de suavización y con *tensor product splines*, ajustes circulares y ajuste elíptico) para estimar el diámetro en una sección horizontal a la altura del pecho (1.3 m) obtenida con las tres aplicaciones de iPad o el PLS (columnas). Figura tomada de Gollob *et al.* (2021b).

En el contexto de estimación de filamentos, podemos estar interesados en estimar no paramétricamente la propia curva cerrada que determina el contorno de una sección horizontal del tronco de un árbol, independientemente de una posterior estimación de diámetros, áreas o volúmenes. Además, el enfoque *plug–in* de estimar el área de la sección o el volumen del tronco a partir del diámetro podría no ser adecuado, a mayores de que se asume una forma circular o cilíndrica que podría no ajustarse bien a los troncos, ver Figura 4.7. Por ejemplo, Hunčaga *et al.* (2020), en un contexto similar, resaltan "*The shape of cross sections was not circular and, therefore, influenced the estimation errors of the*

circle–fitting algorithm". Por este motivo, puede ser más útil considerar el estimador EDT con envoltura r–convexa propuesto en este capítulo para estimar las secciones. Además, en caso de que fuera de interés estimar el nivel de ruido de la nube de puntos, podríamos usar $\widehat{\sigma}$ como estimador, en lugar de la desviación típica de los residuos de los modelos GAM.

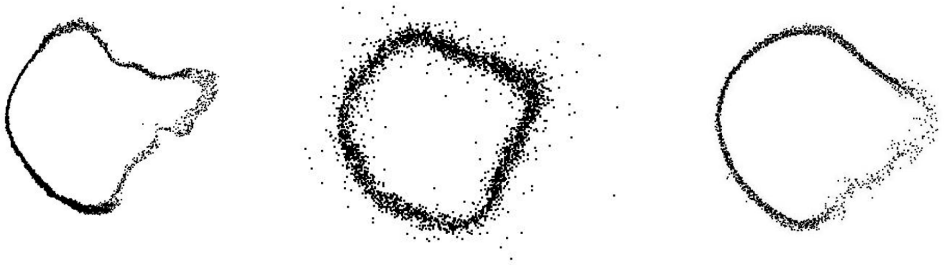

Figura 4.7: Secciones de tronco con forma irregular. Figura tomada de Hunčaga *et al.* (2020).

Vamos a aplicar el estimador EDT con envoltura r–convexa a unas secciones horizontales de troncos que hemos extraído de los mismos datos empleados en Gollob *et al.* (2021b), que están libremente disponibles en Gollob *et al.* (2021a). Hemos usado las nubes de puntos correspondientes a la aplicación 3D Scanner App, pues son las que producen secciones que más se ajustan a las hipótesis de la Sección 4.1, sobre todo al hecho de que el soporte de los datos sea una dilatación del filamento $S = \Gamma \oplus \sigma B$. Otro tipo de aplicaciones producen secciones con grosores del soporte que varían mucho en las distintas zonas (comparar columnas en Figura 4.6) y aplicar nuestro estimador en esos casos produciría malos resultados, pues el estimador EDT depende de modo crítico de que el nivel de ruido σ sea el mismo en todos los puntos del filamento, como discutimos en la Sección 4.3.3. Para el grosor vertical de las capas horizontales, que después se proyectan para conformar las secciones bidimensionales horizontales, hemos optado por 5 cm en lugar de los 15 cm sugeridos en Gollob *et al.* (2021b) pues, cuando el tronco tiene una cierta curvatura, tomar capas más gruesas puede producir distribuciones de puntos irregulares y con distinto nivel de ruido en distintas zonas del soporte S, que nuevamente provocarían un mal comportamiento del estimador.

En las Figuras 4.8, 4.9 y 4.10 se puede observar el resultado de aplicar el estimador EDT con envoltura r–convexa y la correspondiente curva extraída a tres secciones horizontales. Se ha usado un valor de $r = 0.1$ para la envoltura r–convexa $\widehat{S} = C_r(\mathcal{X}_n)$ en los tres casos. Además, para calcular el estimador EDT, se ha tomado $\delta = 0.97\widehat{\sigma}$. El valor de δ es muy alto, casi igual a $\widehat{\sigma}$, porque en las secciones el nivel de ruido varía bastante en los distintos puntos de la curva. Si tomáramos un valor de δ más pequeño, como $\widehat{\sigma}$ queda determinado por las zonas más gruesas del estimador del soporte, el estimador EDT se obtendría contrayendo (erosionando) más hacia dentro $C_r(\mathcal{X}_n)$ y en las zonas más estrechas no habría ningún punto en $\widehat{\Gamma}$, lo que produciría un estimador fragmentado en muchos trozos.

En la Figura 4.8 observamos un ejemplo de sección en la que la estimación obtenida es buena. El nivel de ruido varía, aunque no excesivamente, pero además la curva no presenta zonas con demasiada curvatura, es decir, su *reach* no es demasiado pequeño, de modo que el estimador da un buen resultado.

La Figura 4.9 corresponde a una sección con una forma más irregular. En particular, hay dos zonas, una en la parte superior del dibujo y otra en la inferior, donde la sección presenta dos "entrantes" bastante pronunciados. Estas zonas, de muy alta curvatura, provocan que el *reach* de la curva poblacional sea muy pequeño, o incluso nulo si la curva pierde su diferenciabilidad, y posiblemente se incumple la hipótesis de *reach*. Vemos que la envoltura r–convexa no es capaz de capturar estos entrantes con un valor tan alto de r en comparación con el *reach*, lo que provoca que tanto $C_r(\mathcal{X}_n)$ como $\widehat{\Gamma}$ tengan demasiado grosor en esas dos zonas y que la curva extraída no se ajuste bien en esas dos regiones. Si tomáramos un valor menor de r conseguiríamos una mejor estimación del soporte en esas zonas, pero a cambio en otras $C_r(\mathcal{X}_n)$ sería demasiado delgado y el estimador se fragmentaría. Volviendo al valor de

r empleado en la figura, también podemos observar, en el entrante situado en la zona inferior, que la curva extraída contiene a un segmento espurio de los que forman el eje medial de $C_r(\mathcal{X}_n)$. Esto se debe a que en esa zona el estimador EDT es tan grueso que este segmento está completamente contenido dentro de él y la extracción de curva no consigue eliminarlo.

Por último, en la Figura 4.10 tenemos un ejemplo de sección a lo largo de la cual el nivel de ruido varía mucho. La parte más gruesa, que determina $\widehat{\sigma}$, se encuentra en la zona superior izquierda. Como en otras zonas, especialmente las situadas en la parte inferior izquierda, el nivel de ruido es mucho menor, el estimador EDT se queda sin puntos, se fragmenta, y la curva extraída no contiene una curva cerrada, se divide en varios trozos. En este caso sería más adecuado intentar utilizar un valor mayor de δ aunque, como comentamos antes, el valor escogido ya es muy elevado.

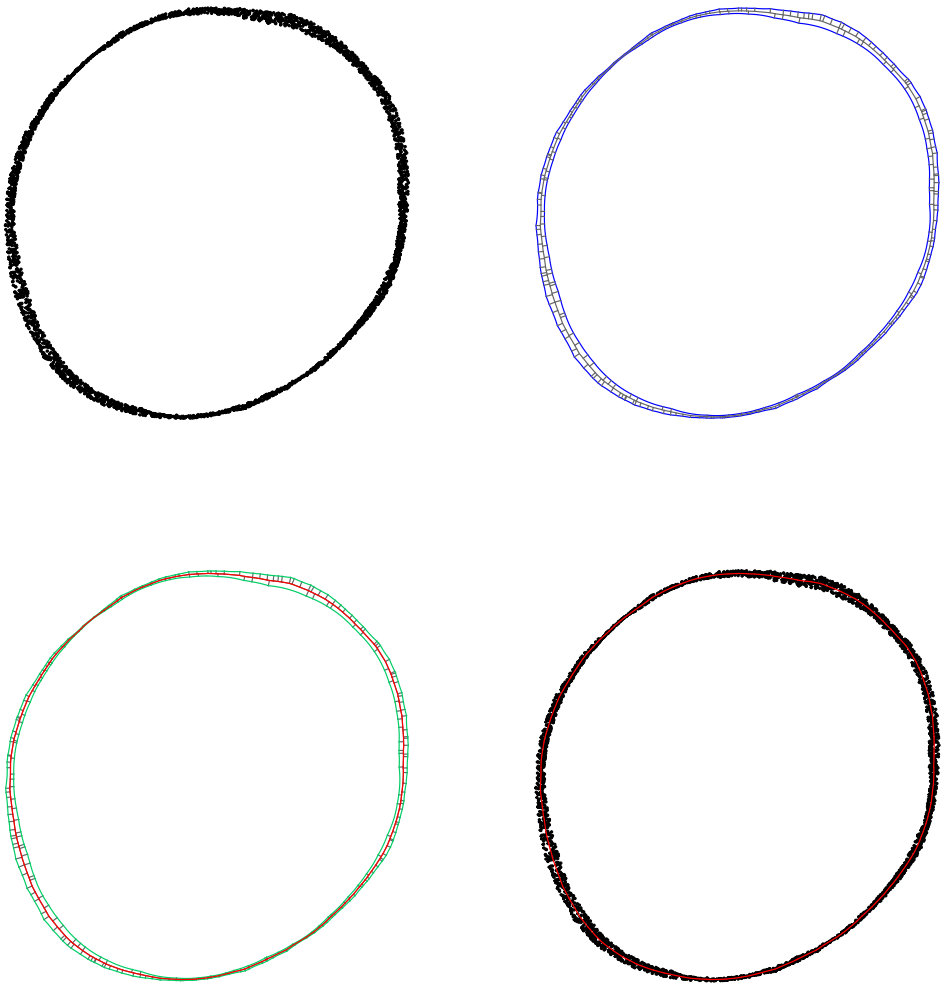

Figura 4.8: Arriba a la izquierda, muestra de puntos correspondiente a una sección horizontal de tronco. Arriba a la derecha, en azul, frontera de la envoltura r–convexa $C_r(\mathcal{X}_n)$ para $r = 0.1$ y, en gris, segmentos que forman el eje medial de $C_r(\mathcal{X}_n)$. Abajo a la izquierda, en verde, frontera del estimador EDT con envoltura r–convexa $\widehat{\Gamma}$ para $\delta = 0.97\widehat{\sigma}$ y, en rojo, curva extraída.

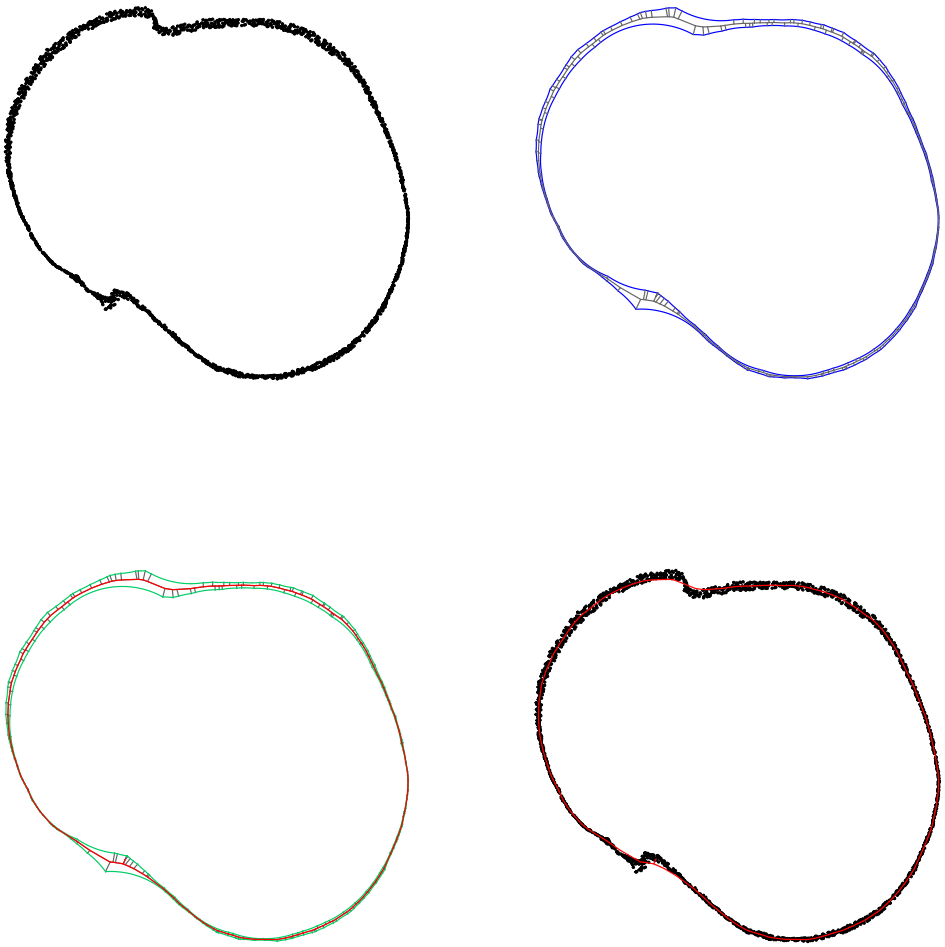

Figura 4.9: Arriba a la izquierda, muestra de puntos correspondiente a una sección horizontal de tronco. Arriba a la derecha, en azul, frontera de la envoltura r–convexa $C_r(\mathcal{X}_n)$ para $r = 0.1$ y, en gris, segmentos que forman el eje medial de $C_r(\mathcal{X}_n)$. Abajo a la izquierda, en verde, frontera del estimador EDT con envoltura r–convexa $\widehat{\Gamma}$ para $\delta = 0.97\widehat{\sigma}$ y, en rojo, curva extraída.

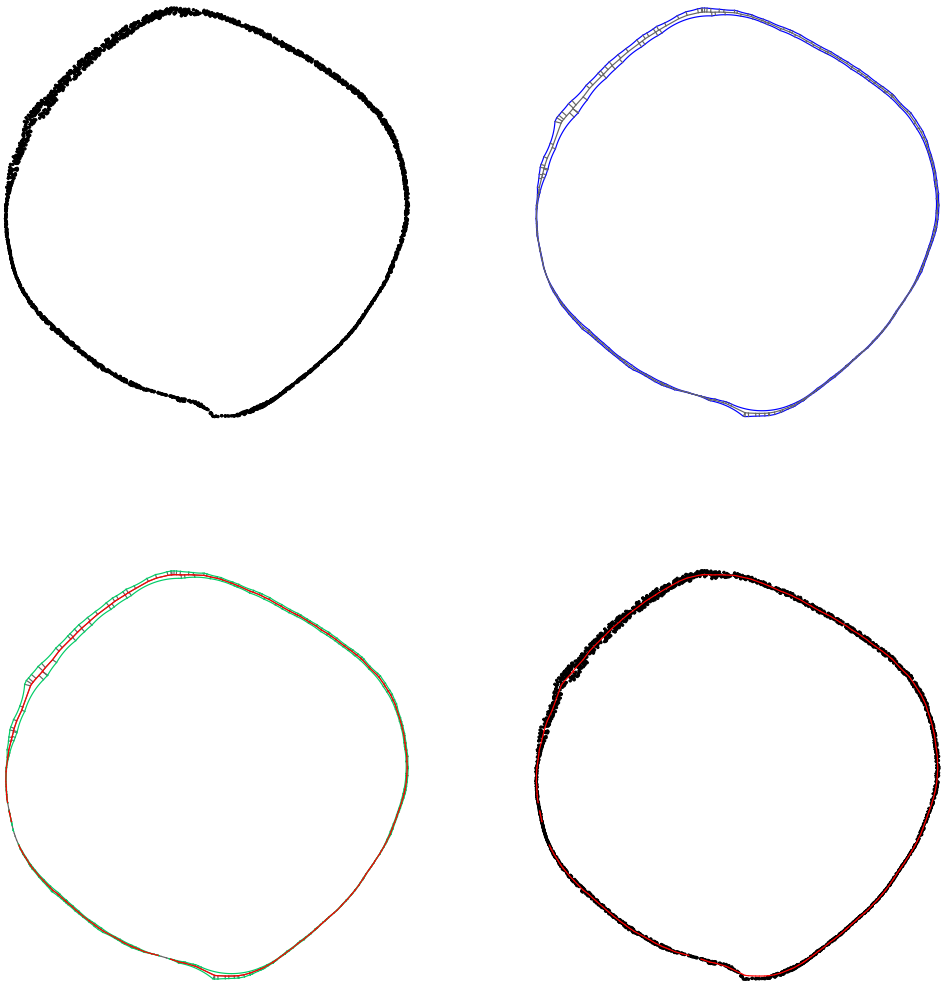

Figura 4.10: Arriba a la izquierda, muestra de puntos correspondiente a una sección horizontal de tronco. Arriba a la derecha, en azul, frontera de la envoltura r–convexa $C_r(\mathcal{X}_n)$ para $r = 0.1$ y, en gris, segmentos que forman el eje medial de $C_r(\mathcal{X}_n)$. Abajo a la izquierda, en verde, frontera del estimador EDT con envoltura r–convexa $\widehat{\Gamma}$ para $\delta = 0.97\widehat{\sigma}$ y, en rojo, curva extraída.

Bibliografía

Cholaquidis, A. (2024). New advances in set estimation. *Boletín de Estadística e Investigación Operativa*, 40(1):5–21.

Cuevas, A. (2009). Set estimation: Another bridge between statistics and geometry. *Boletín de Estadística e Investigación Operativa*, 25(2):71–85.

Cuevas, A. y Fraiman, R. (2010). Set Estimation. En Kendall, W. S. y Molchanov, I., editores, *New Perspectives in Stochastic Geometry*, capítulo 11, pp. 374–397. Oxford University Press.

Cuevas, A., Fraiman, R., y Pateiro-López, B. (2012). On statistical properties of sets fulfilling rolling–type conditions. *Advances in Applied Probability*, 44(2):311–329.

Cuevas, A. y Rodríguez-Casal, A. (2004). On boundary estimation. *Advances in Applied Probability*, 36(2):340–354.

Devroye, L. y Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488.

Dümbgen, L. y Walther, G. (1996). Rates of convergence for random approximations of convex sets. *Advances in Applied Probability*, 28(2):384–393.

Edgar, G. A. (1990). *Measure, Topology, and Fractal Geometry*. Springer–Verlag, New York.

Eto, S., Masuda, H., Hiraoka, Y., Matsushita, M., y Takahashi, M. (2020). Precise calculation of cross sections and volume for tree stem using point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:205–210.

Federer, H. (1959). Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., y Wasserman, L. (2012a). The geometry of nonparametric filament estimation. *Journal of the American Statistical Association*, 107(498):788–799.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., y Wasserman, L. (2012b). Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291.

Gollob, C., Ritter, T., Kraßnitzer, R., Tockner, A., y Nothdurft, A. (2021a). iLAUT - iPad laser scanner data from Austrian forest Inventory plots (1.0). Conjunto de datos, Zenodo. https://doi.org/10.5281/zenodo.5070671.

Gollob, C., Ritter, T., Kraßnitzer, R., Tockner, A., y Nothdurft, A. (2021b). Measurement of forest inventory parameters with Apple iPad pro and integrated LiDAR technology. *Remote Sensing*, 13(16):3129.

Gollob, C., Ritter, T., y Nothdurft, A. (2020). Forest inventory with long range and high-speed personal laser scanning (PLS) and simultaneous localization and mapping (SLAM) technology. *Remote Sensing*, 12(9):1509.

Hunčaga, M., Chudá, J., Tomaštík, J., Slámová, M., Koreň, M., y Chudỳ, F. (2020). The comparison of stem curve accuracy determined from point clouds acquired by different terrestrial remote sensing methods. *Remote Sensing*, 12(17):2739.

Meilă, M. y Zhang, H. (2024). Manifold learning: what, how, and why. *Annual Review of Statistics and Its Application*, 11:393–417.

Milnor, J. W. (1965). *Topology from the Differentiable Viewpoint*. University Press of Virginia, Charlottesville.

Niyogi, P., Smale, S., y Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39:419–441.

Pateiro-López, B. (2008). *Set estimation under convexity type restrictions*. Tesis doctoral, Universidade de Santiago de Compostela.

Pateiro-López, B. y Rodríguez-Casal, A. (2010). Generalizing the convex hull of a sample: the R package alphahull. *Journal of Statistical Software*, 34:1–28.

Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions. *Annales de l'Institut Henri Poincaré – Probabilités et statistiques*, 43(6):763–774.

Rodríguez-Casal, A. y Saavedra-Nieves, P. (2016). A fully data–driven method for estimating the shape of a point cloud. *ESAIM: Probability and Statistics*, 20:332–348.

Rodríguez-Casal, A. y Saavedra-Nieves, P. (2022a). A data–adaptive method for estimating density level sets under shape conditions. *The Annals of Statistics*, 50(3):1653–1668.

Rodríguez-Casal, A. y Saavedra-Nieves, P. (2022b). Spatial distribution of invasive species: an extent of occurrence approach. *TEST*, 31(2):416–441.

Saavedra-Nieves, P. (2015). *Nonparametric data–driven methods for set estimation*. Tesis doctoral, Universidade de Santiago de Compostela.

Serra, J. (1984). *Image Analysis and Mathematical Morphology*. Academic Press, London.

Walther, G. (1999). On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. *Mathematical Methods in the Applied Sciences*, 22(4):301–316.