

UN NOVO ESTIMADOR ÓPTIMO MINIMAX PARA VARIEDADES

Héctor González-Vázquez¹, Beatriz Pateiro-López¹ e Alberto Rodríguez-Casal¹

¹Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela; Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMaga).

RESUMO

A estimación de variedades ten como obxectivo, a partir dun conxunto de datos de alta dimensión, recuperar a estrutura de variedade de dimensión inferior que subxace, arredor da cal realmente se distribúen eses datos. O propósito deste traballo é propoñer un novo estimador de variedades mediante dúas contribucións. A primeira consiste en xeneralizar o estimador EDT (*Euclidean Distance Transform*), inicialmente proposto por Genovese *et al.* (2012a) para estimar filamentos, a variedades de dimensión arbitraria e coa posibilidade de empregar calquera estimador piloto do soporte. A segunda achega reside en empregar como estimador do soporte a envoltura r -convexa da mostra, que fai uso dunha condición que xeneraliza a convexidade, coñecida como r -convexidade, e que proporciona maior flexibilidade ao estimador. En ambos casos próbase a taxa de converxencia do estimador en distancia de Hausdorff para estimar a variedade e amósase que, ao empregar a envoltura r -convexa, o estimador acada a taxa óptima no sentido minimax no modelo de ruído estudado por Genovese *et al.* (2012c) cando a dimensión da variedade a estimar é unha unidade menos que a do espazo ambiente. Finalmente, inclúese unha aplicación a datos reais do ámbito forestal, onde se estima a curva que determina o contorno de seccións horizontais de troncos de árbores.

Palabras e frases chave: Estimación de variedades; Taxa minimax; Envoltura r -convexa; Estimación de filamentos.

1. INTRODUCCIÓN

A chamada *manifold hypothesis*, formalizada por Fefferman *et al.* (2016), é unha hipótese que establece que un conxunto de datos de posible alta dimensión $\mathcal{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$, pese a se atoparen nun espazo de dimensión D , en realidade tenden a concentrarse arredor dunha estrutura de subvariedade $\mathcal{M}$ de dimensión menor $d < D$ dentro dese espazo ambiente. A *estimación de variedades* (*manifold estimation*) ten como obxectivo, a partir dos datos, reconstruír ou estimar a variedade $\mathcal{M}$ dentro do espazo ambiente, revelando a estrutura que subxace. A literatura estatística de estimación de variedades, que cobrou grande importancia nos últimos anos, centrouse, por unha banda, en obter ou limitar as taxas minimax de converxencia en distancia de Hausdorff para estimar a variedade $\mathcal{M}$ baixo distintos modelos (isto é, distintas hipóteses sobre a variedade ou sobre o tipo de ruído que se asume no proceso xerador dos datos) (Aamari *et al.*, 2023; Aamari e Levrard, 2019; Genovese *et al.*, 2012b,c; Kim e Zhou, 2015) e, por outra banda, en construír estimadores procurando nalgúns casos acadar ditas taxas óptimas (Aamari e Levrard, 2018; Divol, 2021; Fefferman *et al.*, 2025; Puchkin e Spokoiny, 2022; Yao e Xia, 2025).

O problema de *estimación de filamentos* pódese ver como un caso particular da estimación de variedades cando a variedade $\mathcal{M}$ a estimar é un filamento, é dicir, unha curva de dimensión $d = 1$ embebida en $\mathbb{R}^D$, habitualmente con $D = 2$ (aínda que non necesariamente). A estimación de filamentos atopa aplicación práctica nunha variedade de campos científicos onde aparecen de xeito natural estruturas filamentosas. En astrofísica e cosmoxía, a física solar estuda as protuberancias, estruturas gasosas de plasma en forma de filamento que se forman na superficie do Sol (Bernasconi *et al.*, 2005), mentres que a teoría da rede cósmica (*cosmic web*) establece a distribución da materia no universo nunha sorte de tecido formado por filamentos galácticos interconectados (Libeskind *et al.*, 2017). En sismoloxía, os epicentros dos terremotos concéntranse seguindo as liñas das fallas xeolóxicas (Li e Ghosal, 2020), mentres que na análise de imaxes médicas resulta de interese identificar estruturas como os vasos sanguíneos ou os neurofilamentos (Liu *et al.*, 2022). Aínda que en xeral as curvas a estimar poden ser abertas ou pechadas, polas hipóteses que estableceremos na Sección 2, ao longo de todo o traballo consideraremos unicamente curvas pechadas.

O obxectivo deste traballo é propoñer un novo estimador de variedades. O noso estimador baséase no coñecido como estimador EDT (*Euclidean Distance Transform*), proposto inicialmente por Genovese

et al. (2012a) no contexto máis particular de estimación de filamentos, e o xeneraliza. O estimador EDT, que explicaremos con máis detalle na Sección 3, emprega un estimador piloto $\hat{S}$ do soporte S da distribución dos datos $\mathcal{X}_n$. En particular, Genovese *et al.* (2012a) propuxeron empregar o coñecido como estimador de Devroye-Wise (Devroye e Wise, 1980). A nosa contribución é dobre. En primeiro lugar, xeneralizamos o estimador en dous sentidos: o novo estimador EDT permite estimar variedades de calquera dimensión e non só curvas, idea que xa aparece en Genovese *et al.* (2012c), e ademais empregando calquera estimador do soporte e non só o estimador de Devroye-Wise. Baixo estas condicións máis xerais probamos a taxa de converxencia en distancia de Hausdorff do estimador EDT, que depende da taxa de converxencia do estimador do soporte que se empregue. En segundo lugar, propoñemos usar como estimador do soporte a envoltura r -convexa da mostra, orixinalmente empregada no contexto de estimación de conxuntos (Rodríguez-Casal, 2007), que se basea nunha restrición de forma, a r -convexidade, que xeneraliza á convexidade e proporciona maior flexibilidade ao estimador. Cando empregamos a envoltura r -convexa e $d = D - 1$, é dicir, cando a variedade ten unha dimensión menos que o espazo ambiente, amosamos que o estimador EDT con envoltura r -convexa acada a taxa óptima no sentido minimax para estimar a variedade en distancia de Hausdorff probada por Genovese *et al.* (2012c).

En particular, no caso $d = 1$ e $D = 2$, temos un estimador de filamentos no plano que é óptimo no sentido minimax, ademais de computacionalmente factible. Este estimador aplicáremolo na Sección 5 a uns datos reais que xorden no campo da ecoloxía, en particular en inventario forestal. Neste contexto, algunhas tecnoloxías de detección remota, como LiDAR (*Light Detection And Ranging*), permiten a obtención de nubes de puntos tridimensionais que recollen información estrutural do ambiente, como a elevación do terreo e obxectos como a vexetación. Estas nubes de puntos utilízanse para extraer características clave das árbores, como a xeometría das seccións transversais horizontais dos troncos. Para elo, trátanse ditas nubes de puntos e obtéñense a partir delas outras nubes de puntos, agora bidimensionais, que se corresponden co contorno das seccións transversais do tronco. Na Figura 1 represéntanse dúas destas seccións horizontais, obtidas a partir dos datos de Gollob *et al.* (2021a), tomados en parcelas localizadas na área de adestramento e investigación da *University of Natural Resources and Life Science* de Viena, en Austria (Gollob *et al.*, 2021b). Pódese observar como os puntos se distribúen en cada caso arredor da curva (pechada) que determina o contorno da sección transversal do tronco da árbore. Os métodos habituais de medición forestal asumen que as seccións dos troncos teñen forma circular, pero esta hipótese incúmprese habitualmente na natureza (Hunčaga *et al.*, 2020; Matérn, 1956; Pulkkinen, 2012). Polo tanto, propoñemos empregar un método non paramétrico como é o novo estimador EDT con envoltura r -convexa para estimar a xeometría das seccións transversais. Este exemplo pon de manifesto a utilidade que pode ter na práctica estimar un filamento no plano ou, de modo máis xeral, unha variedade dunha dimensión menos que a do espazo ambiente, situación onde o noso estimador é óptimo minimax: a variedade pode verse como o borde dun obxecto ou conxunto que si ten dimensión completa.

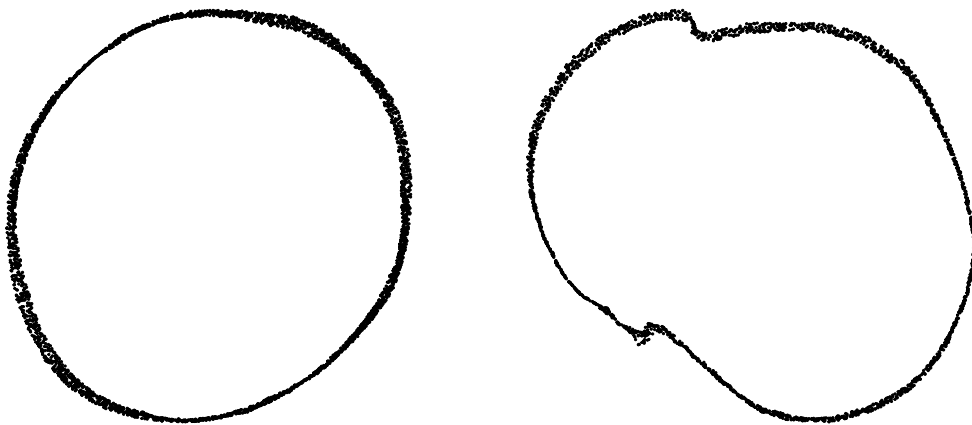


Figura 1: Dúas nubes de puntos correspondentes ao contorno de seccións horizontais de troncos de árbores.

Este traballo está organizado do seguinte xeito. Na Sección 2 defínense o modelo e as hipóteses que consideramos ao longo do traballo e na Sección 3, baseándonos nas propiedades xeométricas do modelo, definimos o novo estimador, primeiro xeneralizando o estimador EDT de Genovese *et al.* (2012a) e despois empregando a envoltura r -convexa da mostra como estimador do soporte. A Sección 4 está adicada aos resultados sobre a converxencia asíntótica e á optimalidade minimax do estimador. Na Sección 5 aplicamos o estimador ao conxunto de datos reais presentados, mentres que na Sección 6 recolleemos as principais conclusións do traballo.

2. UN MODELO EN ESTIMACIÓN DE VARIEDADES

En estimación de variedades o obxectivo é, a partir da mostra $\mathcal{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$, estimar a variedade $\mathcal{M} \subset \mathbb{R}^D$ arredor da cal se distribúen os datos. De modo moi xeral, pódense considerar dous grandes tipos de modelos, a saber, sen ruído ou con ruído. Os modelos sen ruído asumen que os datos dos que dispoñemos seguen unha distribución cuxo soporte é exactamente a variedade $\mathcal{M}$ que queremos estimar. En canto aos modelos con ruído, aínda que se poden clasificar en distintos tipos, caracterízanse polo feito de que as observacións están “preto” da variedade $\mathcal{M}$, pero non necesariamente sobre a mesma.

O modelo que consideraremos é o modelo con ruído estudado por Genovese *et al.* (2012c), no que calcularon a taxa minimax. Primeiro recolleemos todas as hipóteses do modelo e a continuación explicamos cada unha xunto coa notación introducida.

(M) Sobre a variedade

(M1) $\mathcal{M} \subset \mathbb{R}^D$ é unha variedade de Riemann sen borde, de dimensión d , compacta e conexa, con $d < D$.

(M2) Cúmrese $0 < \sigma < \text{reach}(\mathcal{M})$.

(D) Sobre o proceso xerador dos datos

As observacións X_i , cuxo soporte denotamos por S , son unha mostra aleatoria simple da forma

$$X_i = V_i + Z_i, \quad i = 1, \dots, n,$$

onde V_i segue unha distribución uniforme na variedade $\mathcal{M}$ e a distribución de Z_i , condicionada a V_i , é uniforme en $T_{V_i}\mathcal{M}^\perp \cap B(0, \sigma)$.

A hipótese (M1) establece que a estrutura que queremos recuperar a partir dos datos é unha variedade sen borde de dimensión $d < D$ dentro do espazo euclidiano $\mathbb{R}^D$. Isto quere dicir que, en cada punto da variedade, existe un entorno da mesma que é difeomorfo a unha bóla aberta en $\mathbb{R}^d$. É importante notar que a hipótese de que a variedade non teña borde tradúcese, no caso de estimación de filamentos, en que a curva non pode ter extremos e polo tanto debe ser pechada, de xeito que non consideraremos curvas abertas ao longo deste traballo. Pola súa banda, a compacidade e conexidade da variedade son condicións técnicas habituais na literatura e necesarias para a obtención das taxas minimax e para a súa estimación.

A hipótese (D) establece o proceso xerador dos datos. En primeiro lugar, xérase V_i segundo unha distribución uniforme na variedade $\mathcal{M}$. Despois, engádese un ruído que é perpendicular á variedade nese punto V_i da seguinte forma. Dado o punto $V_i \in \mathcal{M}$, denotamos por $T_{V_i}\mathcal{M} \subset \mathbb{R}^D$ ao espazo tanxente a $\mathcal{M}$ en V_i , que se interpreta como o hiperplano de dimensión d (que pasa pola orixe de coordenadas) formado polos vectores tanxentes á variedade nese punto, mentres que $T_{V_i}\mathcal{M}^\perp \subset \mathbb{R}^D$ representa o espazo perpendicular, que se interpreta como o hiperplano de dimensión $D - d$ (tamén pasando pola orixe) formado polos vectores ortogonais a $\mathcal{M}$ en V_i . Tendo isto en conta, o ruído Z_i que se engade segue unha distribución uniforme en $T_{V_i}\mathcal{M}^\perp \cap B(0, \sigma)$ (onde $B(x, r) \subset \mathbb{R}^D$ denota a bóla pechada de centro $x \in \mathbb{R}^D$ e raio $r > 0$), é dicir, estamos sumando a V_i un vector perpendicular á variedade nese punto cuxa lonxitude é como moito σ . Este parámetro $\sigma > 0$ coñécese como *nivel de ruído*, e regula canto se poden afastar as observacións da variedade $\mathcal{M}$. Nótese que é un nivel de ruído fixo, a diferenza doutros modelos onde é necesario que este nivel de ruído converxa a cero a medida que o número de datos n tende a infinito (Aamari e Levrard, 2018; Divol, 2021).

Por último, a hipótese (M2) é unha condición de regularidade sobre a variedade $\mathcal{M}$ a estimar, en particular, que o *reach* de $\mathcal{M}$ sexa maior que o nivel de ruído σ . O *reach* (ou *alcance*) dun conxunto $A \subset \mathbb{R}^D$, denotado por $\text{reach}(A)$, defínese como o maior $r > 0$ tal que todos os puntos en $A \oplus r$ teñen unha única proxección métrica en A (Federer, 1959), onde

$$A \oplus r = \bigcup_{x \in A} B(x, r) = \{x \in \mathbb{R}^D : d(x, A) \leq r\}$$

coñécese como a *r-dilatación* de A , e a proxección métrica dun punto $x \in \mathbb{R}^D$ sobre A defínese como calquera $y \in A$ que realiza a distancia de x a A , isto é, $d(x, y) = d(x, A)$, sendo $d(x, y) = \|x - y\|$ a distancia euclidiana e $d(x, A) = \inf_{y \in A} d(x, y)$ a distancia punto-conxunto. O *reach* é unha cantidade que mide simultaneamente dúas cousas: a curvatura da variedade e o ancho da estrutura tipo colo de botella máis estreita que presente a mesma. Canto maior sexa $\text{reach}(\mathcal{M})$, menos curvada é a variedade e menos estreitos son os seus colos de botella, polo que máis regular é. De feito, na literatura de estimación de variedades, é unha hipótese moi habitual asumir que a variedade ten *reach* positivo, véxase Aamari *et al.* (2019) e as referencias que contén.

A xeometría do soporte

Dada a variedade $\mathcal{M}$, definimos a *fibra perpendicular de tamaño* $\sigma > 0$ en $p \in \mathcal{M}$ como

$$L_\sigma(p) = \{p\} \oplus \left(T_p \mathcal{M}^\perp \cap B(0, \sigma) \right) = \left(\{p\} \oplus T_p \mathcal{M}^\perp \right) \cap B(p, \sigma),$$

onde $A \oplus C = \{a + c : a \in A, c \in C\}$ denota a suma de Minkowski de dous conxuntos $A, C \subset \mathbb{R}^D$. Podemos visualizar a fibra $L_\sigma(p)$ como a porción de espazo perpendicular a $\mathcal{M}$ en p , neste caso desprazada para que estea centrada no propio punto p e non na orixe, que está a distancia como moito σ de p . Definimos o *tubo de raio* $\sigma > 0$ *arredor de* $\mathcal{M}$ como a unión de todas as fibras ao longo da variedade

$$\mathcal{T}_\sigma = \bigcup_{p \in \mathcal{M}} L_\sigma(p),$$

que se pode visualizar como un entorno tubular que envolve á variedade $\mathcal{M}$. Baixo as hipóteses (M), pódese probar que o tubo de raio σ coincide coa σ -dilatación de $\mathcal{M}$, isto é,

$$\mathcal{M} \oplus \sigma = \mathcal{T}_\sigma,$$

véxase por exemplo Genovese *et al.* (2012c). Cabe mencionar que, se non se verifica a hipótese de *reach* (M2), o anterior non é certo.

De aquí en adiante denotaremos por S ao soporte de X_i . Baixo as hipóteses (M) e (D), o soporte das observacións é o tubo $\mathcal{T}_\sigma$, que polo anterior á súa vez coincide con $\mathcal{M} \oplus \sigma$, de modo que

$$S = \mathcal{M} \oplus \sigma = \mathcal{T}_\sigma.$$

Porén, o recíproco do anterior non é certo, no sentido de que, se baixo as hipóteses (M) o soporte das observacións X_i é $S = \mathcal{M} \oplus \sigma = \mathcal{T}_\sigma$, o proceso xerador dos datos non ten por que ser (D). Ao longo deste traballo, os resultados que obtivemos relativos á xeneralización do estimador EDT só precisan das hipóteses (M) e $S = \mathcal{M} \oplus \sigma = \mathcal{T}_\sigma$, o que inclúe ao modelo coas hipóteses (M) e (D) de Genovese *et al.* (2012c) pero tamén a calquera outro modelo no que os datos teñan como soporte ese entorno tubular da variedade, que na literatura en ocasións reciben o nome de modelo de ruído tubular ou modelo de ruído paralelo (véxase por exemplo Aaron *et al.*, 2017). Pola súa banda, os resultados relativos ao uso da envoltura r -convexa como estimador do soporte para acadar as taxas minimax si que están enunciados coas hipóteses (M) e (D), pois é o modelo no que Genovese *et al.* (2012c) calcularon ditas taxas.

3. A NOVA PROPOSTA DE ESTIMADOR

Nesta sección presentamos a nosa nova proposta de estimador de variedades.

Xeneralizando o estimador EDT

A primeira contribución deste traballo consiste en xeneralizar a variedades de calquera dimensión d o estimador EDT, inicialmente proposto por Genovese *et al.* (2012a) no contexto de estimación de filamentos ($d = 1$). O estimador baséase no feito de que, baixo as hipóteses (M) e $S = \mathcal{M} \oplus \sigma = \mathcal{T}_\sigma$, para todo $y \in S$ verifícase

$$d(y, \mathcal{M}) + d(y, \partial S) = \sigma \quad (1)$$

(onde ∂A denota a fronteira topolóxica en $\mathbb{R}^D$ dun conxunto $A \subset \mathbb{R}^D$), véxase Genovese *et al.* (2012c) ou Proposición B.5 en Bolón (2024). En particular, (1) implica que

$$\sigma = \max_{y \in S} d(y, \partial S) \quad (2)$$

e tamén que

$$\mathcal{M} = \{y \in S : d(y, \partial S) = \sigma\}. \quad (3)$$

Se definimos a aplicación EDT (*Euclidean Distance Transform*) como a distancia á fronteira do soporte

$$\Lambda(y) = d(y, \partial S) \quad \text{para } y \in S,$$

a ecuación (2) significa que o máximo da EDT no soporte S é o nivel de ruído σ , mentres que a ecuación (3) supón que a variedade $\mathcal{M}$ é precisamente o conxunto de puntos do soporte que maximizan a EDT. Aproveitando este feito xorde o estimador EDT, que definimos no Algoritmo 1.

Algoritmo 1. *Estimador EDT*

ENTRADA: Un estimador do soporte $\hat{S} \subset \mathbb{R}^D$ e o parámetro $\delta > 0$.

SAÍDA: O estimador EDT $\hat{\mathcal{M}} \subset \mathbb{R}^D$.

1. Estimar o soporte S mediante o estimador $\hat{S}$ e a súa fronteira ∂S mediante $\partial \hat{S}$.
2. Estimar a EDT $\Lambda(y) = d(y, \partial S)$ mediante a EDT empírica $\hat{\Lambda}(y) = d(y, \partial \hat{S})$.
3. Estimar o nivel de ruído $\sigma = \max_{y \in S} d(y, \partial S)$ mediante $\hat{\sigma} = \max_{y \in \hat{S}} d(y, \partial \hat{S})$.
4. Estimar a variedade $\mathcal{M} = \{y \in S : d(y, \partial S) = \sigma\}$ mediante o estimador EDT

$$\hat{\mathcal{M}} = \{y \in \hat{S} : d(y, \partial \hat{S}) \geq \hat{\sigma} - \delta\}.$$

A idea detrás do estimador EDT é a seguinte. Primeiro necesitamos estimar o soporte mediante un estimador piloto $\hat{S}$. Genovese *et al.* (2012a) propuxeron empregar un estimador concreto coñecido como o estimador de Devroye-Wise, pero o noso enfoque é máis xeral ao permitir calquera estimador do soporte, o que complica algo máis a análise das súas propiedades estatísticas e a obtención da súa taxa de converxencia. Dito estimador non só o empregamos para estimar S , senón que tamén empregamos a súa fronteira $\partial \hat{S}$ para estimar ∂S . Grazas a isto, podemos a continuación estimar a EDT empregando un enfoque *plug-in*, isto é, substituíndo ∂S por $\partial \hat{S}$ na súa definición, obtendo unha versión empírica da mesma. Unha vez que temos estimada a EDT, podemos estimar σ mediante o máximo que alcance esta EDT empírica no soporte estimado, que denotamos por $\hat{\sigma}$. Finalmente, dado que a variedade $\mathcal{M}$ está formada polos puntos que maximizan a EDT, definimos o estimador EDT $\hat{\mathcal{M}}$ como o conxunto de puntos de $\hat{S}$ cuxa EDT empírica sobrepasa un certo limiar $\hat{\sigma} - \delta$.

O que estamos facendo intuitivamente é partir do estimador do soporte $\hat{S}$ e “contraelo” para obter $\hat{\mathcal{M}}$. O parámetro $\delta \in (0, \hat{\sigma})$ regula como de severa é esta contracción. Canto menor sexa δ , máis contraemos e máis “estreito” é o estimador EDT. De feito, se $\delta = \hat{\sigma}$, temos que o estimador $\hat{\mathcal{M}}$ coincide con $\hat{S}$ e polo tanto é demasiado grande con respecto á variedade $\mathcal{M}$ que pretende estimar, mentres que, se $\delta = 0$, $\hat{\mathcal{M}}$ estaría formado unicamente polos puntos nos que se maximice a EDT empírica, que serían un conxunto demasiado pequeno e tampouco estimarían ben $\mathcal{M}$. Como probamos nos resultados teóricos da Sección 4, δ debe tender a 0 a unha taxa axeitada, non debe ser nin demasiado grande nin demasiado pequeno.

O uso da envoltura r -convexa como estimador do soporte

Con respecto a que estimador do soporte empregar no Paso 1 do estimador EDT, Genovese *et al.* (2012a) propuxeron usar o coñecido como estimador de Devroye-Wise (Devroye e Wise, 1980). O *estimador de Devroye-Wise* defínese como

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

onde $\varepsilon > 0$ é un parámetro de suavizado que tipicamente tende a 0. Este estimador consiste polo tanto nunha versión suavizada ou dilatada da mostra, unindo bólas de raio ε arredor de cada unha das observacións mostrais.

O estimador de Devroye-Wise ten a vantaxe de ser un estimador do soporte que se pode empregar baixo condicións moi xerais pero, a cambio, ten unha taxa de converxencia lenta. Porén, se coñecemos que o soporte verifica algunha condición de forma adicional, resultaría máis eficiente empregar un estimador que incorpore esa condición. Por exemplo, na literatura de estimación do soporte unha das primeiras restricións de forma que se consideraron foi a convexidade. Baixo a suposición de que o soporte sexa convexo, o estimador natural a empregar é a envoltura convexa da mostra. A convexidade é unha condición demasiado restritiva en moitas situacións, como é o caso do noso modelo, onde o soporte S non a verifica, polo que propoñemos empregar unha restrición de forma máis flexible que a xeneraliza, coñecida como a r -convexidade.

Dicimos que un conxunto $A \subset \mathbb{R}^D$ é r -convexo para $r > 0$ se $A = C_r(A)$, é dicir, se coincide coa súa envoltura r -convexa, definida como

$$C_r(A) = \bigcap_{\dot{B}(y, r) \cap A = \emptyset} \dot{B}(y, r)^c,$$

onde $\dot{A}$ denota o interior dun conxunto $A \subset \mathbb{R}^D$ e A^c o seu complementario, de xeito que $\dot{B}(y, r)^c$ denota o complementario da bóla aberta de centro $y \in \mathbb{R}^D$ e raio $r > 0$. A r -convexidade xeneraliza á convexidade da seguinte forma. É coñecido que todo conxunto convexo pechado pode expresarse como a intersección

de todos os semiespazos que o conteñen. No caso da r -convexidade, substituímos estes semiespazos por complementarios de bólas abertas de raio r . Dito de outro modo, un conxunto é r -convexo se todo punto que non pertence ao conxunto se pode separar del mediante unha bóla aberta de raio r , en contraposición cos hiperplanos separadores dos conxuntos convexas. Intuitivamente, podemos pensar nunha bóla de raio r que roda polo exterior da fronteira do conxunto A e vai determinando a forma da súa envoltura r -convexa. O seguinte resultado garante que o soporte $S = \mathcal{M} \oplus \sigma = \mathcal{T}_\sigma$ do noso modelo é r -convexo.

Proposición 1. *Baixo as hipóteses (M) e $S = \mathcal{M} \oplus \sigma = \mathcal{T}_\sigma$, S é α -convexo para $\alpha = \text{reach}(\mathcal{M}) - \sigma > 0$.*

Sabendo que o soporte a estimar é r -convexo, o estimador natural é a *envoltura r -convexa da mostra*

$$\hat{S} = C_r(\mathcal{X}_n) = \bigcap_{\dot{B}(y,r) \cap \mathcal{X}_n = \emptyset} \dot{B}(y,r)^c.$$

Na Figura 2 represéntase a envoltura r -convexa dunha nube de puntos para valores decrecentes de r (véxase Pateiro-López e Rodríguez-Casal, 2010 para a súa implementación en R). O valor do parámetro r proporciona flexibilidade ao estimador e debe ser escollido de forma axeitada. Pódese considerar un parámetro de forma, pero non é un parámetro de suavizado no sentido habitual, xa que non debe converxer a 0. Valores demasiado grandes de r devolven un estimador moi parecido á envoltura convexa, mentres que valores demasiado pequenos levan a un estimador moi fragmentado e con ocos baleiros. Na Sección 4 faise unha breve discusión sobre o papel deste parámetro e a posibilidade de seleccionalo de forma automática a partir dos datos.

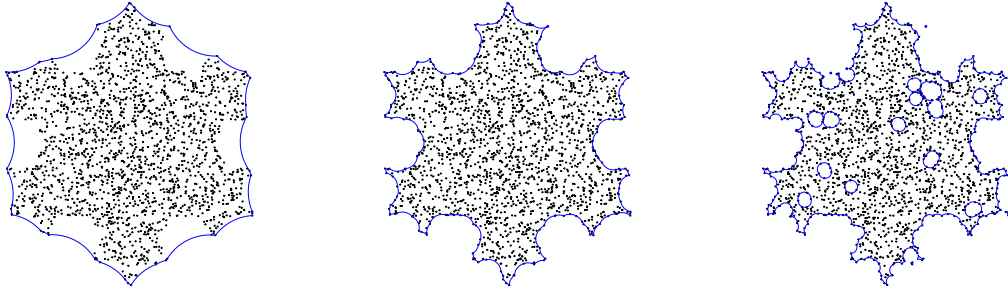


Figura 2: De esquerda a dereita, fronteira da envoltura r -convexa $C_r(\mathcal{X}_n)$ para $r = 0.2, 0.05$ e 0.025 .

A segunda contribución deste traballo consiste na proposta de empregar, no primeiro paso do estimador EDT, a envoltura r -convexa da mostra $\hat{S} = C_r(\mathcal{X}_n)$ como estimador do soporte, dando lugar ao estimador EDT con envoltura r -convexa, que recolleemos no Algoritmo 2. A Figura 3 ilustra o funcionamento deste algoritmo no caso de estimación de filamentos no plano ($d = 1, D = 2$).

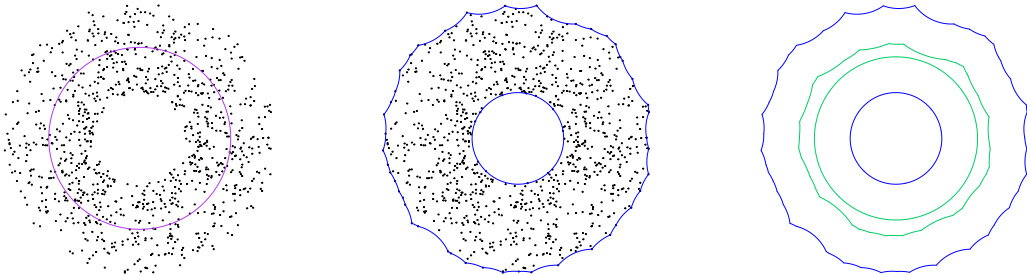


Figura 3: Á esquerda, mostra de tamaño $n = 1000$ simulada do modelo con hipóteses (M) e (D) sendo a variedade $\mathcal{M}$ o filamento circunferencia unidade en morado ($\text{reach}(\mathcal{M}) = 1$) e con nivel de ruído $\sigma = 0.5$. No centro, en azul, fronteira da envoltura r -convexa $C_r(\mathcal{X}_n)$ para $r = \text{reach}(\mathcal{M}) - \sigma = 0.5$. Á dereita, en verde, fronteira do estimador EDT con envoltura r -convexa $\hat{\mathcal{M}}$ para $\delta = 0.2\hat{\sigma}$.

Algoritmo 2. *Estimador EDT con envoltura r -convexa*

ENTRADA: As observacións $\mathcal{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ e os parámetros $r > 0$ e $\delta > 0$.

SAÍDA: O estimador EDT con envoltura r -convexa $\widehat{\mathcal{M}} \subset \mathbb{R}^D$.

- Aplicar o estimador EDT (Algoritmo 1) empregando como estimador do soporte a envoltura r -convexa da mostra $\widehat{S} = C_r(\mathcal{X}_n)$.

4. RESULTADOS ASINTÓTICOS E OPTIMALIDADE

Nesta sección recolleemos os resultados principais do traballo relativos ao comportamento asintótico dos estimadores propostos. Neste tipo de resultados empregaremos unha notación na que os distintos conxuntos ou parámetros levarán un subíndice n cando queiramos salientar que dependen do tamaño mostral n , como por exemplo $\widehat{S}_n$ ou $\widehat{\mathcal{M}}_n$.

Ao analizar o comportamento dun estimador $\widehat{\mathcal{M}}_n$, precisamos unha función de perda que mida como de preto está ese estimador da variedade $\mathcal{M}$ a estimar, para o que se emprega unha distancia entre conxuntos. No contexto da estimación de variedades, a distancia de Hausdorff é a distancia amplamente usada na literatura. Dados $A, C \subset \mathbb{R}^D$ non baleiros e compactos, defínese a *distancia de Hausdorff* entre ambos como

$$d_H(A, C) = \max \left\{ \sup_{a \in A} d(a, C), \sup_{c \in C} d(c, A) \right\} = \inf \{r > 0: A \subset C \oplus r \text{ e } C \subset A \oplus r\}.$$

Informalmente, para que dous conxuntos estean próximos en distancia de Hausdorff é preciso que todo punto de cada un dos dous conxuntos teña algún punto do outro conxunto preto.

Cando falamos da *taxa de converxencia en distancia de Hausdorff* dun estimador $\widehat{\mathcal{M}}_n$, ou simplemente taxa de converxencia, referímonos á orde asintótica á que $d_H(\widehat{\mathcal{M}}_n, \mathcal{M})$ converge a 0 cando $n \rightarrow \infty$. Para formalizar isto, dadas dúas sucesións de números reais deterministas $\{a_n\}$ e $\{r_n\}$, dicimos que $a_n = O(r_n)$ se existe unha constante $C > 0$ tal que $a_n \leq C r_n$ para todo n a partir dun certo índice n_0 . Cando a sucesión a_n que queiramos estudar sexa aleatoria porque depende da mostra, preocuparémonos por resultados de converxencia case segura do tipo “ $a_n = O(r_n)$ con probabilidade un”, onde se entende que a constante $C > 0$ que aparece na definición da O grande non depende da mostra.

Estimador EDT

En relación ao estimador EDT presentado no Algoritmo 1, o noso resultado principal é o seguinte teorema, que dá a súa taxa de converxencia en distancia de Hausdorff. Como comentamos na Sección 2, este resultado é válido para calquera modelo de ruído no que o proceso xerador dos datos teña como soporte un entorno tubular da variedade, o que en particular inclúe ao proceso descrito pola hipótese (D). O resultado é de carácter determinista, no sentido de que se asume que o estimador do soporte $\widehat{S}_n$ é unha sucesión de conxuntos que non é aleatoria, e polo tanto tampouco son aleatorios o resto de elementos asociados que interveñen no resultado. Porén, este teorema pódese aplicar tamén cando si dispoñemos dunha mostra aleatoria e dun estimador $\widehat{S}_n \equiv \widehat{S}_n(\mathcal{X}_n)$ obtido a partir dela pois, se as hipóteses do teorema se verifican de modo case seguro, como consecuencia as teses tamén.

Teorema 1. *Baixo as hipóteses (M) e $S = \mathcal{M} \oplus \sigma = \mathcal{T}_\sigma$, sexa $\widehat{S}_n \subset \mathbb{R}^D$ un compacto tal que se cumpre $\max\{d_H(\widehat{S}_n, S), d_H(\partial \widehat{S}_n, \partial S)\} \leq \varepsilon_n$ para n suficientemente grande, con $\varepsilon_n \rightarrow 0$. Sexan $\widehat{\sigma}_n = \max_{y \in \widehat{S}_n} d(y, \partial \widehat{S}_n)$ e $\widehat{\mathcal{M}}_n = \{y \in \widehat{S}_n: d(y, \partial \widehat{S}_n) \geq \widehat{\sigma}_n - \delta_n\}$ o estimador EDT. Se $\delta_n \rightarrow 0$ é tal que $\delta_n \geq 2\varepsilon_n$ para n suficientemente grande, entón*

$$d_H(\widehat{\mathcal{M}}_n, \mathcal{M}) = O(\delta_n). \quad (4)$$

En particular, se ademais $\delta_n = O(\varepsilon_n)$, entón

$$d_H(\widehat{\mathcal{M}}_n, \mathcal{M}) = O(\varepsilon_n). \quad (5)$$

O Teorema 1 establece que a taxa de converxencia do estimador EDT ao empregar un estimador do soporte $\widehat{S}_n$ arbitrario depende de ε_n , que se pode interpretar, asintoticamente, como o máximo entre a taxa de converxencia de $\widehat{S}_n$ como estimador de S e a taxa de converxencia da súa fronteira $\partial \widehat{S}_n$ como estimador de ∂S . En xeral, que dous conxuntos estean próximos en distancia de Hausdorff non implica que as súas fronteiras tamén o estean, polo que, á hora de elixir o estimador do soporte, unha elección axeitada non debe ter en conta unicamente a velocidade de converxencia de $\widehat{S}_n$, senón tamén da súa fronteira. En canto ao parámetro δ_n , o teorema establece que debe converxer a 0 e que non debe facelo a unha taxa máis

rápida que ε_n . Baixo estas condicións, a taxa de converxencia do estimador EDT coincide coa taxa á que converxa δ_n a 0, ecuación (4). Polo tanto, a elección óptima de δ_n é do mesmo orde que ε_n , para que o estimador EDT tamén converxa a dita taxa, ecuación (5).

Estimador EDT con envoltura r -convexa e optimalidade minimax

Por outra banda, á hora de determinar a optimalidade dun estimador, como mencionamos na introdución, tamén seguiremos o criterio maioritariamente empregado na literatura, que é o criterio minimax. O *risco minimax* para estimar a variedade $\mathcal{M}$ nun certo modelo defínese como

$$\inf_{\widehat{\mathcal{M}}_n} \sup_{\mathcal{M}} \mathbb{E}_{\mathcal{M}}[d_H(\widehat{\mathcal{M}}_n, \mathcal{M})],$$

onde o ínfimo percorre todos os posibles estimadores, é dicir, todas as posibles funcións medibles da mostra $\widehat{\mathcal{M}}_n \equiv \widehat{\mathcal{M}}_n(\mathcal{X}_n)$, o supremo percorre todas as variedades do modelo a estimar, e $\mathbb{E}_{\mathcal{M}}$ denota a esperanza baixo a distribución que seguen os datos X_i que induce a variedade $\mathcal{M}$ nese modelo. Informalmente, o risco minimax captura como de ben o pode facer un estimador no peor caso. A nivel teórico, o interese céntrase en estudar a *taxa minimax*, que é a orde asíntótica do risco minimax cando $n \rightarrow \infty$.

Baixo as hipóteses (M) e (D), Genovese *et al.* (2012c) limitaron (acoutaron) a taxa minimax en distancia de Hausdorff para estimar a variedade $\mathcal{M}$, obtendo que dita taxa está entre

$$O\left(\left(\frac{1}{n}\right)^{2/(d+2)}\right) \text{ (límite inferior)} \quad \text{e} \quad O\left(\left(\frac{\log(n)}{n}\right)^{2/(d+2)}\right) \text{ (límite superior)}. \quad (6)$$

Como ambos límites só difiren nun factor logarítmico, adoita dicirse que a taxa está axustada salvo factores logarítmicos.

A nosa proposta para que o estimador EDT acade a taxa minimax é o uso da envoltura r -convexa da mostra $\widehat{S}_n = C_r(\mathcal{X}_n)$ como estimador do soporte, tal e como presentamos no Algoritmo 2 na Sección 3. A taxa de converxencia da envoltura r -convexa da mostra e da súa fronteira foron obtidas por Rodríguez-Casal (2007) baixo unha condición de forma que ten que ver coa r -convexidade tanto de S como da clausura do seu complementario $\overline{S^c}$ (onde $\overline{A}$ denota a clausura de $A \subset \mathbb{R}^D$), á que nos referimos como *condición de dobre rodamento*, e que o soporte do noso modelo verifica. Baixo esta condición e algunha hipótese técnica adicional, Rodríguez-Casal (2007) proba que, con probabilidade un para n suficientemente grande, verificase $d_H(C_r(\mathcal{X}_n), S) \leq \varepsilon_n$ e $d_H(\partial C_r(\mathcal{X}_n), \partial S) \leq \varepsilon_n$, sendo

$$\varepsilon_n = C_1 \left(\frac{\log(n)}{n}\right)^{2/(D+1)} \quad (7)$$

para certa constante $C_1 > 0$.

O uso da envoltura r -convexa engade agora un novo parámetro a escoller no estimador EDT: que valor de r debemos empregar? A resposta a nivel teórico é que serve calquera valor tal que o soporte sexa r -convexo. A envoltura r -convexa é monótona respecto de r no sentido de que, se $r < \tilde{r}$, entón $C_r(S) \subset C_{\tilde{r}}(S)$. Esta monotonía provoca que, se S é $\tilde{r}$ -convexo para un certo valor $\tilde{r} > 0$, tamén sexa r -convexo para todo $r \in (0, \tilde{r})$. Polo tanto, seguindo a Rodríguez-Casal e Saavedra-Nieves (2016, 2022a,b), existe un valor de r “canónico” que representa como de regular é o soporte con respecto á r -convexidade, que é o maior valor tal que S é r -convexo. Así, dado $S \subset \mathbb{R}^D$ compacto, non convexo e r -convexo para algún $r > 0$, definimos

$$r_0 = \sup\{\gamma > 0: C_\gamma(S) = S\}. \quad (8)$$

Na definición anterior e en adiante, para facilitar a exposición, supoñemos que S non é convexo, pois no caso convexo r_0 sería infinito e resultaría máis axeitado empregar a envoltura convexa da mostra. Rodríguez-Casal e Saavedra-Nieves (2016) proban que, baixo a condición de dobre rodamento anteriormente mencionada, o supremo r_0 definido en (8) é un máximo e polo tanto $C_{r_0}(S) = S$. Baixo estas condicións, o valor de r óptimo a empregar na envoltura r -convexa é r_0 .

Na práctica, porén, o valor de r_0 é descoñecido, de xeito que o ideal sería seleccionar o valor de r de xeito automático usando un selector $R_n \equiv R_n(\mathcal{X}_n)$ a partir dos propios datos, para despois empregalo na envoltura R_n -convexa $C_{R_n}(\mathcal{X}_n)$ e así ter un estimador *data-driven*. Rodríguez-Casal e Saavedra-Nieves (2022b) propuxeron un estimador $\widehat{r}_0 \equiv \widehat{r}_0(\mathcal{X}_n)$, baseado no concepto de *maximal spacings*, que converxe en probabilidade a r_0 . Se multiplicamos este estimador por unha constante $\nu \in (0, 1)$, definindo un selector $R_n = \nu \widehat{r}_0$, cúmprese que existe unha constante $r_* > 0$ tal que

$$\mathbb{P}(0 < r_* \leq R_n \leq r_0) \rightarrow 1. \quad (9)$$

Neste traballo supoñemos que dispoñemos dun selector R_n que cumpra que, con probabilidade un para n suficientemente grande,

$$0 < r_* \leq R_n \leq r_0$$

para unha certa constante $r_* > 0$. A súa interpretación é que o selector R_n non debe sobrepasar o valor r_0 , pois nese caso o soporte non sería R_n -convexo e a correspondente envoltura R_n -convexa non estimaría a S , pero por outra parte R_n tampouco debe tender a 0, xa que o seu papel non é o dun parámetro de suavizado. Esta condición é algo máis forte que (9) ao ser de xeito case seguro, o que tamén nos permite obter a taxa de converxencia do estimador case seguro, pero cabe comentar que relaxar as hipóteses e empregar a condición en probabilidade (9) permitiría obter a correspondente taxa en probabilidade.

O Teorema 2, que se pode ver como unha adaptación do Teorema 1 ao caso da envoltura r -convexa, proporciona a taxa de converxencia do estimador EDT con envoltura r -convexa ao empregar un selector R_n nas condicións anteriores. Neste teorema o selector R_n e o estimador do soporte $\hat{S}_n$ si que son elementos aleatorios que dependen da mostra, así como o resto de elementos que dependen deles. O outro parámetro δ_n a usar no estimador EDT tamén pode depender da mostra e ser aleatorio, de xeito que, aínda que neste traballo non estudamos a súa elección a partir da mostra, senón só a orde óptima á que debe converxer a cero, este resultado poderíase aplicar no caso de dispoñer dun selector do mesmo.

Teorema 2. *Baixo as hipóteses (M) e (D), supoñamos que S non é convexo. Sexan r_0 definido na ecuación (8), $R_n > 0$ tal que, para unha constante $r_* > 0$, con probabilidade un para n suficientemente grande tense $0 < r_* \leq R_n \leq r_0$, e $\hat{S}_n = C_{R_n}(\mathcal{X}_n)$. Sexan $\hat{\sigma}_n = \max_{y \in \hat{S}_n} d(y, \partial \hat{S}_n)$ e $\hat{\mathcal{M}}_n = \{y \in \hat{S}_n : d(y, \partial \hat{S}_n) \geq \hat{\sigma}_n - \delta_n\}$ o estimador EDT con envoltura R_n -convexa. Se $\delta_n \rightarrow 0$ é tal que $\delta_n \geq 2\varepsilon_n$ con probabilidade un para n suficientemente grande, onde ε_n é o dado pola ecuación (7) para $C_1 > 0$ suficientemente grande, entón con probabilidade un*

$$d_H(\hat{\mathcal{M}}_n, \mathcal{M}) = O(\delta_n).$$

En particular, se ademais $\delta_n = O((\log(n)/n)^{2/(D+1)})$ con probabilidade un, entón con probabilidade un

$$d_H(\hat{\mathcal{M}}_n, \mathcal{M}) = O\left(\left(\frac{\log(n)}{n}\right)^{2/(D+1)}\right). \quad (10)$$

Se comparamos a taxa do estimador EDT con envoltura r -convexa en (10) coa taxa minimax en (6), vemos que ambas coinciden, salvo factor logarítmico, cando $2/(d+2) = 2/(D+1)$, é dicir, cando $d = D - 1$. Polo tanto, baixo o modelo que definen as hipóteses (M) e (D), o estimador EDT con envoltura r -convexa é óptimo no sentido minimax de converxencia en distancia de Hausdorff para estimar variedades de dimensión unha unidade menos que o espazo ambiente.

5. APLICACIÓN A DATOS REAIS

Nesta sección aplicamos o novo estimador EDT con envoltura r -convexa aos datos de seccións horizontais de troncos presentados na introdución, véxase Figura 1. Estes datos enmarcábanse no contexto de estimación de filamentos no plano, é dicir, cando $d = 1$ e $D = 2$, situación na que, baixo o modelo presentado, o noso estimador é óptimo no sentido minimax de converxencia en distancia de Hausdorff á vista dos resultados da Sección 4.

O estimador EDT con envoltura r -convexa que se obtén ao aplicar o Algoritmo 2 aos datos deste exemplo é un subconxunto do plano $\hat{\mathcal{M}} \subset \mathbb{R}^2$, pero non unha curva. Como a variedade que queremos estimar é unha curva, é habitual que se propoña como estimador da mesma outra curva. Isto pódese facer definindo o que se chama *curva extraída*, que é unha curva contida dentro do estimador EDT, e que se pode probar baixo certas condicións que non abordamos neste resumo que conserva a taxa de converxencia óptima.

Aplicando o estimador EDT con envoltura r -convexa ás nubes de puntos da Figura 1, e despois extraendo unha curva dentro do estimador en cada caso, obtemos os resultados que se amosan na Figura 4, na que xa se representan directamente as curvas extraídas en vermello sobre cada unha das nubes. Escollemos o valor $r = 0.1$ para o parámetro da envoltura r -convexa de xeito manual, aínda que tamén se podería empregar o estimador $\hat{r}_0$ de Rodríguez-Casal e Saavedra-Nieves (2022b) como se explicou na Sección 4. En canto ao parámetro δ , tivemos que escoller un valor moi alto, en concreto $\delta = 0.97\hat{\sigma}$, o que significa que contraemos moi pouco $\hat{S}$ para obter $\hat{\mathcal{M}}$. A razón desta elección é que o nivel de ruído non é constante ao longo da curva en ningunha das dúas nubes de puntos, senón que hai zonas que son máis grosas e outras máis delgadas. Tomar un valor pequeno de δ , contraendo moito, levaría a que nas zonas máis delgadas da nube non habería puntos do estimador EDT, de xeito que este estimador non rodearía por completo ao filamento, senón que se “fragmentaría” en varios anacos. De feito, un posible criterio para seleccionar o parámetro δ a partir da mostra sería escoller o menor valor deste parámetro tal que o estimador EDT está formado por un único “anaco”. Explorar este selector, que está fóra dos propósitos deste traballo, podería ser interesante no futuro.

Na sección da esquerda vemos que o estimador consegue capturar axeitadamente a forma. Na sección da dereita o estimador faino ben na maior parte do contorno, pero hai dúas zonas nas que o tronco parece

que presenta uns entrantes bastante pronunciados que o estimador non consegue capturar. Nestas dúas zonas o que está a acontecer é que se viola a hipótese de *reach* (M2), pois ao ser zonas de moita curvatura o *reach* é moi pequeno, ou incluso nulo se se perde a diferenciabilidade da curva por un pico. En todo caso, o estimador en ambos casos capta ben a xeometría subxacente ás seccións horizontais, que claramente non son circulares.

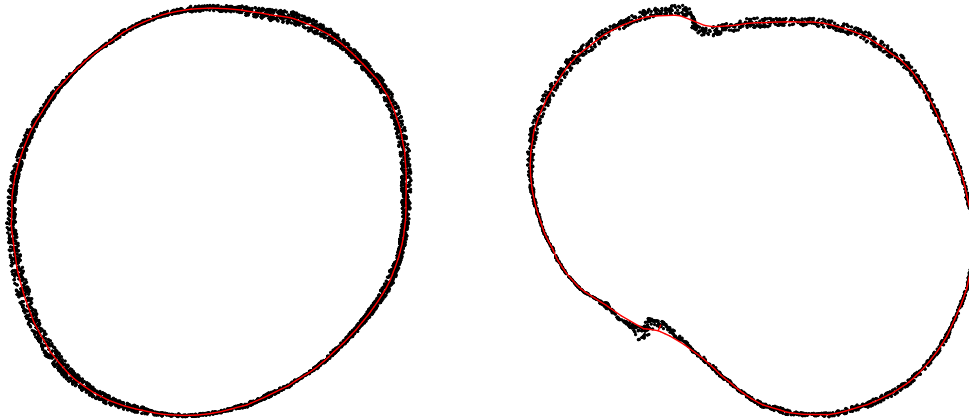


Figura 4: En vermello, curva extraída dentro do estimador EDT con envoltura r -convexa con $r = 0.1$ e $\delta = 0.97\hat{\sigma}$ para cada unha das dúas seccións horizontais de troncos.

6. CONCLUSIÓN

Neste traballo propoñemos un novo estimador para variedades. O noso estimador supón dúas contribucións. Por unha banda, xeneralizamos o estimador EDT, orixinalmente proposto por Genovese *et al.* (2012a), para estimar variedades de calquera dimensión e empregando calquera estimador do soporte, e probamos a súa taxa de converxencia en distancia de Hausdorff, que depende da taxa de converxencia do estimador do soporte que se empregue. Por outra banda, propoñemos empregar como estimador do soporte a envoltura r -convexa da mostra, dando lugar ao estimador EDT con envoltura r -convexa, e probamos a súa taxa de converxencia, que resulta ser a taxa óptima no sentido minimax no modelo estudado por Genovese *et al.* (2012c) para estimar variedades de dimensión unha unidade menos que a do espazo ambiente. O estimador depende de dous parámetros, un parámetro r que non converxe a 0 a empregar na envoltura r -convexa, que discutimos como se podería seleccionar de xeito automático a partir da mostra, e un parámetro δ , do que sabemos a orde óptima á que debe converxer a 0, pero que non abordamos como poder seleccionar. Finalmente, aplicamos o novo estimador a un problema do ámbito forestal para estimar o contorno das seccións horizontais de troncos de árbores.

AGRADECEMENTOS

O primeiro autor agradece o apoio económico da axuda de apoio á etapa predoutoral da Xunta de Galicia ED481A-2025-173, cofinanciada parcialmente pola Unión Europea no marco do programa FSE+ Galicia 2021-2027. Este traballo é parte do proxecto de I+D+i PID2020-116587GB-I00 financiado por MICIU/AEI/10.13039/501100011033 e do proxecto ED431C 2025/03 financiado pola Consellería de Educación, Ciencia, Universidades e Formación Profesional.

REFERENCIAS

- Aamari, E., Aaron, C., e Levrard, C. (2023). Minimax boundary estimation and estimation with boundary. *Bernoulli*, 29(4):3334–3368.
- Aamari, E., Kim, J., Chazal, F., Michel, B., Rinaldo, A., e Wasserman, L. (2019). Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399.
- Aamari, E. e Levrard, C. (2018). Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry*, 59(4):923–971.
- Aamari, E. e Levrard, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204.

- Aaron, C., Cholaquidis, A., e Cuevas, A. (2017). Detection of low dimensionality and data denoising via set estimation techniques. *Electronic Journal of Statistics*, 11(2):4596–4628.
- Bernasconi, P. N., Rust, D. M., e Hakim, D. (2005). Advanced automated solar filament detection and characterization code: description, performance, and results. *Solar Physics*, 228(1):97–117.
- Bolón, D. (2024). *Object oriented inference methods*. Tese doutoral, Universidade de Santiago de Compostela.
- Devroye, L. e Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488.
- Divol, V. (2021). Minimax adaptive estimation in manifold inference. *Electronic Journal of Statistics*, 15(2):5888–5932.
- Federer, H. (1959). Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491.
- Fefferman, C., Ivanov, S., Lassas, M., e Narayanan, H. (2025). Fitting a manifold of large reach to noisy data. *Journal of Topology and Analysis*, 17(02):315–396.
- Fefferman, C., Mitter, S., e Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049.
- Genovese, C. R., Perone-Pacífico, M., Verdinelli, I., e Wasserman, L. (2012a). The geometry of nonparametric filament estimation. *Journal of the American Statistical Association*, 107(498):788–799.
- Genovese, C. R., Perone-Pacífico, M., Verdinelli, I., e Wasserman, L. (2012b). Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941–963.
- Genovese, C. R., Perone-Pacífico, M., Verdinelli, I., e Wasserman, L. (2012c). Minimax manifold estimation. *Journal of Machine Learning Research*, 13(43):1263–1291.
- Gollob, C., Ritter, T., Kraßnitzer, R., Tockner, A., e Nothdurft, A. (2021a). iLAUT - iPad laser scanner data from Austrian forest Inventory plots (1.0). Conjunto de dados, Zenodo. <https://doi.org/10.5281/zenodo.5070671>.
- Gollob, C., Ritter, T., Kraßnitzer, R., Tockner, A., e Nothdurft, A. (2021b). Measurement of forest inventory parameters with Apple iPad pro and integrated LiDAR technology. *Remote Sensing*, 13(16):3129.
- Hunčaga, M., Chudá, J., Tomašík, J., Slámová, M., Koreň, M., e Chudý, F. (2020). The comparison of stem curve accuracy determined from point clouds acquired by different terrestrial remote sensing methods. *Remote Sensing*, 12(17):2739.
- Kim, A. K. H. e Zhou, H. H. (2015). Tight minimax rates for manifold estimation under Hausdorff loss. *Electronic Journal of Statistics*, 9(1):1562–1582.
- Li, W. e Ghosal, S. (2020). Posterior contraction and credible sets for filaments of regression functions. *Electronic Journal of Statistics*, 14(1):1707–1743.
- Libeskind, N. I., van de Weygaert, R., Cautun, M., Falck, B., Tempel, E., Abel, T., Alpaslan, M., Aragón-Calvo, M. A., Forero-Romero, J. E., Gonzalez, R., Gottlöber, S., Hahn, O., Hellwing, W. A., Hoffman, Y., Jones, B. J. T., Kitaura, F., Knebe, A., Manti, S., Neyrinck, M., Nuza, S. E., Padilla, N., Platen, E., Ramachandra, N., Robotham, A., Saar, E., Shandarin, S., Steinmetz, M., Stoica, R. S., Sousbie, T., e Yepes, G. (2017). Tracing the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 473(1):1195–1217.
- Liu, Y., Wang, G., Ascoli, G. A., Zhou, J., e Liu, L. (2022). Neuron tracing from light microscopy images: automation, deep learning and bench testing. *Bioinformatics*, 38(24):5329–5339.
- Matérn, B. (1956). On the geometry of the cross-section of a stem. *Meddelanden från Statens skogsforskningsinstitut*, 46(11).
- Pateiro-López, B. e Rodríguez-Casal, A. (2010). Generalizing the convex hull of a sample: the R package alphahull. *Journal of Statistical Software*, 34(5):1–28.
- Puchkin, N. e Spokoiny, V. (2022). Structure-adaptive manifold estimation. *Journal of Machine Learning Research*, 23(40):1–62.
- Pulkkinen, M. (2012). On non-circularity of tree stem cross-sections: effect of diameter selection on cross-section area estimation, Bitterlich sampling and stem volume estimation in Scots pine. *Silva Fennica*, 46(5B):747–986.
- 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. e 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. e 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. e Saavedra-Nieves, P. (2022b). Spatial distribution of invasive species: an extent of occurrence approach. *TEST*, 31(2):416–441.
- Yao, Z. e Xia, Y. (2025). Manifold fitting under unbounded noise. *Journal of Machine Learning Research*, 26(45):1–55.