UNIVERSIDAD DE SONORA

DIVISIÓN DE CIENCIAS EXACTAS Y NATURALES

Programa de Licenciatura en Matemáticas

Titulo de la tesis

T E S I S

Que para obtener el título de:

Licenciado en Matemáticas

Presenta:

Nombre del tesista

Director de tesis: Prof. Jesús Francisco Espinoza Fierro

Hermosillo, Sonora, México ... de 20xx

Sinodales

...
...,
...

Dr. Fulano de tal
Departamento de Matemáticas,
Universidad de Sonora

Sutano...
Departamento de Matemáticas,
Universidad de Sonora

Mangano...
Departamento de Matemáticas,
Universidad de Sonora

...

Agradecimientos

...

Índice general

Introducción ix

1. Espacios Métricos, Coberturas y Complejos Simpliciales 1

2. Utilizando Cubiertas y Nervios para el Análisis de Datos Exploratorio y Visualización: El Algoritmo Mapper. 7

3. Reconstrucción Geométrica e Inferencia Homológica 11
- 3.1. Inferencia Homológica . . . 14
- 3.2. Aspectos estadísticos de la inferencia homológica . . . 18
- 3.3. Más allá de la Distancia de Hausdorff: Distancia a una Medida . . . 19

4. Homología Persistente 21
- 4.1. Filtraciones . . . 21
- 4.2. Algunos Ejemplos . . . 22
- 4.3. Módulos y Diagramas de Persistencia . . . 26
- 4.4. Paisajes de persistencia . . . 27
- 4.5. Representaciones Lineales de la Persistencia Homológica . . . 28
- 4.6. Métricas en el Espacio de Diagramas de Persistencia . . . 30
- 4.7. Propiedades de Estabilidad en los Diagramas de Persistencia . . . 31

5. Aspectos Estadísticos de la Homología Persistente 35
- 5.1. Resultados de Consistencia para la Homología Persistente . . . 35
- 5.2. Estadísticos de la Homología Persistente Calculados en una Nube de Puntos 37
- 5.3. Estadísticos para una Familia de Diagramas de Persistencia y Otras Representaciones . . . 39
- 5.4. Otros Acercamientos Estadísticos al Análisis Topológico de Datos . . . 40
- 5.5. Homología Persistente y el Aprendizaje Autoḿatico . . . 41

6. Análisis Topológico de datos para Ciencia de Datos con la Libreria GUDHI 45
- 6.1. Bootstrap y Comparasión de Configuraciones de Unión de Proteínas . . . 45
- 6.2. Clasificación de Datos de Sensores . . . 46

7. Discusión 49

Apéndices **51**

A. Cosas que no deberían ir en el texto principal **53**

Bibliografía **55**

Introducción

El análisis topológico de datos (ATD) es un campo reciente que emerge de varios trabajos en topología (algebraica) aplicada y la geometría computacional durante la primera década del siglo **XXI**. Aunque es posible encontrarse con acercamientos geométricos al análisis de datos desde mucho antes, el ATD comenzó a desarrollarse como un campo con los trabajos de Edelsbrunner et al. (2002) [64] y Zomorodian y Carlsson (2005) [131] en homología persistente, el campo fue popularizado en un destacado artículo en 2009 [22]. El ATD es motivado principalmente por la idea que la topología y la geometría brindan un acercamiento poderoso para inferir de manera robusta características cualitativas y cuantitativas sobre la estructura de un conjunto de datos [e.g., Chazal (2017) [30]].

El objetivo del ATD es generar métodos matemáticos, estadísticos y algorítmicos bien fundamentados para inferir, analizar y explotar las complejas estructuras topológicas y geométricas subyacentes a datos que usualmente son representados como nubes de puntos en espacios Euclideanos o espacios métricos más generales. En el transcurso de los últimos años se ha realizado un esfuerzo considerable para proporcionar estructuras de datos robustas y eficientes, además de algoritmos para ATD que actualmente son implementados y facilitados a través de paqueterías estándar como la paquetería GUDHI[1] (C++ y Python) Maria et al. (2014) [91] y su interfaz para el software R, Fasy et al. (2014a) [65], Dionysus[2], PHAT[3], DIPHA[4] o Giotto[5]. Aunque evoluciona con rapidez, el ATD proporciona un conjunto de herramientas maduras y eficientes que pueden ser usadas de manera complementaria o conjunta a otras herramientas de la ciencia de datos.

Estructura General del Análisis Topológico de Datos

El ATD se ha desarrollado recientemente en múltiples direcciones y campos de aplicación. Actualmente existe una variedad de métodos inspirados por acercamientos topológicos y geométricos. Dar un resumen que cubra con entereza de los acercamientos existentes se encuentra fuera del alcance de esta introducción. Sin embargo, muchos métodos estándar siguen la siguiente secuencia:

1. Suponemos que la entrada de datos es un conjunto finito de puntos con una noción de distancia o similitud entre ellos. Esta puede ser inducida por una métrica en el espacio de entrada (e.g. la métrica Euclidiana si se trata de datos inmersos en $\mathbb{R}^d$) o ser una métrica intrínseca definida por una matriz de distancia por pares. La definición de la

[1] https://gudhi.inria.fr/
[2] https://www.mrzv.org/software/dionysus/
[3] https://bitbucket.org/phat-code/phat
[4] https://github.com/DIPHA/dipha
[5] https://giotto-ai.github.io/gtda-docs/0.4.0/library.html

métrica en los datos normalmente es parte de la entrada o es guiada por la aplicación. No obstante, es importante notar que la elección de dicha métrica puede ser crítica para revelar características topológicas y geométricas interesantes de los datos.

2. Se construye una figura "continua" sobre los datos con el propósito de resaltar las estructuras topológicas y geométricas subyacentes. Usualmente se trata de un complejo simplicial o una familia anidada de complejos simpliciales, llamada filtración, la cual refleja la estructura de los datos en diferentes escalas. Los complejos simpliciales pueden ser vistos como generalizaciones de gráficas vecinales que clásicamente son construidas sobre los datos en muchos tipos de análisis o algoritmos de aprendizaje. El desafío aquí es definir tales estructuras de tal manera que sean capaces de reflejar información relevante acerca de la estructura de los datos y que puedan ser construidas de manera efectiva y manipuladas en la práctica.

3. Información topológica y geométrica es extraída de las estructuras construidas sobre los datos. Esto puede resultar en una reconstrucción completa, típicamente una triangulación, de la forma subyacente de los datos de los cuales se pueden extraer fácilmente propiedades topológicas y geométricas en forma de resúmenes o aproximaciones las cuales requieren métodos específicos, como la homología persistente, para la extracción de información relevante. Más allá de la identificación de información topológica/geométrica interesante y su visualización e interpretación, el desafío en este paso es mostrar su relevancia, en particular su estabilidad con respecto a las perturbaciones o la presencia de ruido en los datos de entrada. Es por ello que entender el comportamiento estadístico de las propiedades inferidas es también una cuestión importante.

4. La información topológica y geométrica proporciona una nueva familia de característicos y descriptores de los datos. Estos pueden ser usados para entender mejor los datos (en particular a través de visualización) o pueden ser combinados con otros tipos de característicos para un análisis posterior o tareas de aprendizaje automático. Esta información también puede ser utilizada para diseñar modelos bien ajustados para el análisis de datos o el aprendizaje automático. Mostrar el valor añadido y complementario (con respecto a otras características) de la información proporcionada por las herramientas del ATD es un punto importante en este paso.

El Análisis Topológico de Datos y la Estadística

Hasta hace poco, los aspectos teóricos del TDA y la inferencia topológica recaían principalmente en acercamientos determinísticos. Estos acercamientos no tomaban en cuenta la naturaleza aleatoria de los datos y la variabilidad intrínseca de las cantidades topológicas que infieren. Así, la mayoría de los métodos correspondientes son de carácter explicativo, sin ser capaces de distinguir eficientemente entre información y lo que normalmente es llamado "ruido topológico".

Un acercamiento estadístico al ATD implica considerar los datos como generados de una distribución desconocida y a su vez que las propiedades topológicas inferidas utilizando métodos del ATD son vistos como estimadores de cantidades topológicas que describen un objeto subyacente. Bajo este acercamiento, el objeto desconocido usualmente corresponde al soporte de la distribución de los datos (o parte del mismo). Los objetivos principales de

un acercamiento estadístico al análisis topológico de datos pueden ser abreviados como la siguiente lista de problemas:

Tópico 1: Demostrar consistencia y estudiar la tasa de convergencia de los métodos del ATD.

Tópico 2: Proporcionar regiones de confianza para características topológicas y discutir la significacia de las cantidades topológicas estimadas.

Tópico 3: Seleccionar escalas relevantes en las cuales el fenómeno topológico debe ser considerado, en función de los datos observados.

Tópico 4: Lidiar con valores atípicos y brindar métodos robustos para el ATD.

Aplicaciones del Análisis Topológico de Datos en la Ciencia de Datos.

Desde el punto de vista de las aplicaciones, recientemente hay muchos resultados prometedores que han demostrado la eficacia de acercamientos topológicos y geométricos en una multitud de campos, tales como la ciencia de materiales (Kramar et al., 2013 [80]; Nakamura et al., 2015 [96]; Pike et al., 2020 [103]), análisis de formas 3D (Skraba et al., 2010 [119]; Turner et al., 2014b [123]), análisis de imágenes (Qaiser et al., 2019 [106]; Rieck et al., 2020 [110]), análisis de series de tiempo multivariadas (Khasawneh y Munch, 2016 [77]; Seversky et al., 2016 [115]; Umeda, 2017 [124]), medicina (Dindin et al., 2020 [61]), biología (Yao et al., 2009 [129]), genómica (Carrière y Rabadán, 2020 [28]), química (Lee et al., [86]; Smith et al., 2021 [120]), redes sensoriales (De Silva y Ghrist, 2007 [57]) y transportación (Li et al., 2019 [89]), entre otros. Dar una lista exhaustiva de las aplicaciones del ATD esta fuera del alcance de esta introducción. Por otra parte, la mayoría de los resultados del ATD son fruto de su combinación con otras técnicas de análisis y aprendizaje. De esta manera vemos que clarificar la posición y complementariedad del ATD con respecto a otros acercamientos y herramientas en la ciencia de datos es una cuestión importante y un campo de investigación activo.

Así, los objetivos generales de este documento son los siguientes. Primero, se intenta proporcionar a los analistas de datos con una breve pero exhaustiva introducción a los fundamentos matemáticos y estadísticos del ATD. Con este propósito, nos enfocamos en una selección de herramientas y tópicos, los complejos simplicales y su uso para el análisis topológico de datos exploratorio, la inferencia geométrica y la homología persistente, los cuales juegan un rol central en el ATD. Segundo, se apunta a demostrar como, gracias al reciente progreso del software, herramientas del ATD pueden ser fácilmente aplicadas en la ciencia de datos. En particular, mostraremos como la versión de Python de la paquetería GUDHI permite una sencilla implementación y uso de las herramientas presentadas. Nuestro objetivo es proporcionar al analista de datos referencias relevantes de manera que se obtenga un comprensión clara de los elementos básicos del ATD y sea capaz de utilizar sus métodos y software en un conjunto propio de problemas y datos.

Otros estudios del ATD, complementarios a este trabajo, pueden ser encontrados en la literatura. Wasserman (2018) [127] presenta una perspectiva estadística al ATD, y se concentra, en particular, en las conexiones entre el ATD y el agrupamiento por densidad. Sizemore et al. (2019) [118] propuso un estudio acerca de las aplicaciones del ATD a las neurociencias. Finalmente, Hensel et al. (2021) [73] presenta un resumen de las aplicaciones del ATD al aprendizaje automático.

Capítulo 1

Espacios Métricos, Coberturas y Complejos Simpliciales

Debido a que las características topológicas y geométricas suelen ser asociadas con espacios continuos, datos representados como una conjunto finito de observaciones no revelan información topológica directamente. Una manera natural de revelar algún tipo de estructura topológica en los datos es "conectar" puntos de datos que se encuentren cerca con el propósito de exhibir una forma continua global subyacente en los datos. Usualmente cuantificamos la noción de cercania entre puntos utilizando una distancia (o medida de disimilaridad), y muchas veces resulta conveniente considerar conjuntos de datos como espacios métricos discretos o muestras de espacios métricos. Esta sección introduce conceptos generales para la inferencia geométrica y topológica; una presentación más completa del tema se encuentra en el estudio por Boissonnat et al. (2018) [9].

Espacios Métricos

Recordemos que un espacio métrico (M, ρ) es un conjunto M con una función $\rho : M \times M \to \mathbb{R}_+$, llamada distancia, tal que para cualquier $x, y, z \in M$, se tiene lo siguiente:

I) $\rho(x, y) \geq 0$ y $\rho(x, y) = 0$ si y sólo si $x = y$,

II) $\rho(x, y) = \rho(y, x)$, y

III) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

Dado un espacio métrico (M, ρ), el conjunto de subconjuntos compactos de (M, ρ) denotado por $\mathcal{K}(M)$, puede ser dotado con la distancia de Hausdorff; dados dos subconjuntos compactos $A, B \subseteq M$, la distancia de Hausdorff $d_H(A, B)$ entre A y B es definida como el número no negativo más pequeño δ, tal que para cualquier $a \in A$, existe $b \in B$ de manera que $\rho(a, b) \leq \delta$ (Figura 1.1). En otras palabras, si dado cualquier subconjunto compacto $C \subseteq M$, denotamos por $d(\cdot, C) : M \to \mathbb{R}_+$ a la función distancia de C definida por $d(x, C) \coloneqq \text{ínf}_{c \in C} \rho(x, c)$ para cualquier $x \in M$, entonces se puede probar que la distancia de Hausdorff entre A y B esta definida por una de las siguientes igualdades:

$$d_H(A,B) = \max\left\{\sup_{b\in B} d(b,A), \sup_{a\in A} d(a,B)\right\}$$
$$= \sup_{x\in M} |d(x,A) - d(x,B)| = \|d(\cdot,A) - d(\cdot,B)\|_\infty$$

Es un resultado clásico que la distancia de Hausdorff es en efecto una distancia en el conjunto de subconjuntos compactos de un espacio métrico. Desde la perspectiva del ATD, esta distancia brinda una manera conveniente de cuantificar la proximidad entre diferentes conjuntos de datos que provienen del mismo espacio métrico. Sin embargo, a veces es necesario comparar conjuntos de datos que no son muestreados del mismo espacio. Por fortuna la noción de la distancia de Hausdorff puede ser generalizada para comparar cualquier par de espacios métricos compactos, esta es la idea de la distancia de Gromov-Hausdorff.

Dados dos espacios métricos compactos, (M_1, ρ_1) y (M_2, ρ_2), decimos que son isométricos si existe una biyección $\phi : M_1 \to M_2$ que preserva distancias, esto es, $\rho_2(\phi(x), \phi(y)) = \rho_1(x, y)$ para cualquier $x, y \in M_1$. La distancia de Gromov-Hausdorff mide cuan lejos están dos espacios métricos de ser isométricos.

Definición 1.1. La distancia de Gromov-Hausdorff $d_{GH}(M_1, M_2)$ entre dos espacios métricos compactos es el ínfimo de los números reales $r \geq 0$ tal que existe un espacio métrico (M, ρ) y dos subespacios compactos C_1 y $C_2 \subset M$ que son isométricos a M_1 y M_2 y que cumplen $d_H(C_1, C_2) \leq r$.

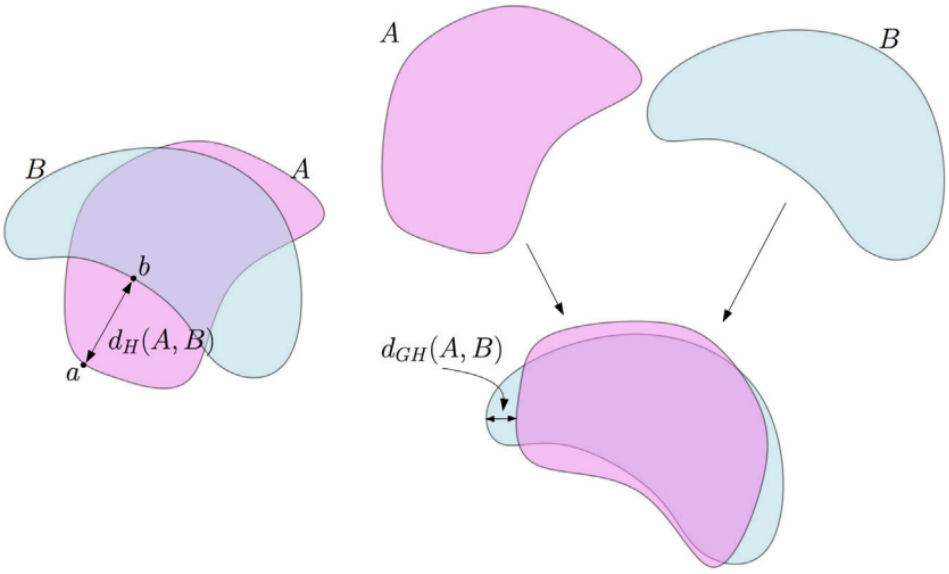


Figura 1.1: Izquierda: la distancia de Hausdorff entre dos subconjuntos A y B en el plano. en este ejemplo, $d_H(A,B)$ es la distancia entre el punto a en A que es el más lejano a B y su vecino más cercano b en B. Derecha: la distancia de Gromov-Hausdorff entre A y B. A puede ser rotado para reducir su distancia de Hausdorff a B. Así, $d_{GH}(A,B) \leq d_H(A,B)$.

Usaremos la distancia de Gromov-Hausdorff más adelante para el estudio de las propiedades de estabilidad de los diagramas de persistencia.

Conectar pares de puntos de datos cercanos mediante aristas lleva a la noción estándar de una gráfica simple de la cual la conectividad de los datos puede ser analizada usando, por ejemplo, algoritmos de agrupamiento. Para ir más allá de la conectividad, una idea central en el ATD es construir nociones equivalentes a las gráficas simples pero de dimensión más alta, utilizando no sólo pares sino $(k+1)$-tuplas de puntos de datos cercanos. El resultado son objetos llamados complejos simpliciales, los cuales nos ayudan a identificar nuevas características topológicas tales como ciclos, huecos, y sus correspondientes de dimensiones superiores.

Complejos Simpliciales Geométricos y Abstractos

Los complejos simplicales pueden considerarse como gráficas generalizadas a dimensiones superiores. Son objetos matemáticos que son de naturaleza topológica y combinatoria a la vez, una propiedad que los hace particularmente útiles para el ATD.

Dado un conjunto $\mathbb{X} = \{x_0, \ldots, x_k\} \subset \mathbb{R}^d$ con $k+1$ puntos afínmente independientes, el simplejo k-dimensional $\sigma = [x_0, \ldots, x_k]$ generado por $\mathbb{X}$ es la envolvente convexa de $\mathbb{X}$. Los puntos de $\mathbb{X}$ son llamados vértices de σ, y los simplejos generados por los subconjuntos de $\mathbb{X}$ son llamados caras de σ. Un complejo simplicial geométrico K en $\mathbb{R}^d$ es una colección de simplejos que cumplen lo siguiente:

I) Cualquier cara de un simplejo de K es un simplejo de K y,

II) La intersección de cualesquiera dos simplejos de K es el conjunto vacío o una cara común de ambos simplejos.

La unión de los simplejos de K es un subconjunto de $\mathbb{R}^d$ llamado el espacio subyacente de K que hereda la topología de $\mathbb{R}^d$. Así, K puede ser visto como un espacio topológico a través de su espacio subyacente. Es de notar que una vez que se conocen los vértices, K se encuentra completamente caracterizado por la descripción combinatoria de una colección de simplejos que satisfacen ciertas reglas de incidencia.

Dado un conjunto V, un complejo simplicial abstracto con un conjunto de vértices V es un conjunto $\tilde{K}$, de subconjuntos finitos de V tales que los elementos de V pertenecen a $\tilde{K}$ y que para cualquier $\sigma \in \tilde{K}$, cualquier subconjunto de σ pertenece a $\tilde{K}$. Los elementos de $\tilde{K}$ son llamados las caras o los simplejos de $\tilde{K}$. La dimensión de un simplejo abstracto es su cardinalidad menos 1 y la dimensión de $\tilde{K}$ es la mayor de las dimensiones de sus simplejos. Es de notar que los complejos simpliciales de dimensión 1 son gráficas.

La descripción combinatoria de cualquier complejo simplicial geométrico K da lugar a un complejo simplicial abstracto $\tilde{K}$. El inverso también es cierto; siempre es posible asociar con un complejo simplical abstracto $\tilde{K}$ un cierto espacio topológico $|\tilde{K}|$ tal que si K es un complejo simplicial geométrico cuya descripción combinatoria es la misma que la de $\tilde{K}$, entonces el espacio subyacente de K es homeomorfo a $|\tilde{K}|$. Dicha K es llamada una realización geométrica de $\tilde{K}$. Como consecuencia de esto, los complejos simpliciales abstractos pueden ser vistos como espacios topológicos y los complejos simpliciales geométricos pueden ser vistos como realizaciones geométricas de la estructura combinatoria subyacente. Así, se puede considerar a los complejos simpliciales como objetos combinatorios que se ajustan bien a cálculos computacionales efectivos y a su vez como espacios topológicos de los cuales se pueden inferir propiedades topológicas.

Construcción de Complejos Simpliciales a partir de Datos

Dado un conjunto de datos, o más generalmente, un espacio métrico o topológico, existen varias maneras de construir complejos simpliciales. Esta es una presentación de algunos ejemplos clásicos que son usados con frecuencia en la práctica.

Comenzando con una extensión inmediata de la noción de una gráfica, Supóngase que tenemos un conjunto de puntos $\mathbb{X}$ en un espacio métrico (M, ρ) y un número real $\alpha \geq 0$. El complejo de Vietoris-Rips $Rips_\alpha(\mathbb{X})$ es el conjunto de simplejos $[x_0, \ldots, x_k]$ tal que $\rho_{\mathbb{X}}(x_i, x_j) \leq \alpha$ para todo (i, j), ver Figura 1.2. De aquí vemos que el complejo de Vietoris-Rips es efectivamente un complejo simplicial abstracto. Aunque, en general, incluso cuando $\mathbb{X}$ es un subconjunto finito de $\mathbb{R}^d$, $Rips_a(\mathbb{X})$ no admite una realización geométrica en $\mathbb{R}^d$; en particular, puede ser de una dimensión mayor a d, por ejemplo, si se tienen $d+2$ puntos en R^d que cumplen $\rho_{\mathbb{X}}(x_i, x_j) \leq \alpha$ para todo (i, j), entonces $Rips_a(\mathbb{X})$ es de dimensión $d+1$, podemos ver un caso similar en el tetraedro formado en el complejo derecho de la Figura 1.2.

Estrechamente relacionado al complejo de Vietoris-Rips está el complejo de Čech $Cech_a(\mathbb{X})$ el cual se define como el conjunto de simplejos $[x_0, \ldots, x_k]$ tales que las $k+1$ bolas cerradas $B(x_i, \alpha)$ tienen intersección no vacía, ver Figura 1.2. Estos dos complejos estan relacionados por

$$Rips_\alpha(\mathbb{X}) \subseteq Cech_\alpha(\mathbb{X}) \subseteq Rips_{2\alpha}(\mathbb{X})$$

y que si $\mathbb{X} \subset \mathbb{R}^d$, entonces $Cech_\alpha(\mathbb{X})$ y $Rips_{2\alpha}(\mathbb{X})$ tienen el mismo esqueleto 1-dimensional, esto es, comparten el mismo conjunto de vértices y aristas.

El Teorema del Nervio

El complejo de Čech es un caso particular de una familia de complejos asociados con cubiertas. Dada una cubierta $\mathcal{U} = (U_i)_{i \in I}$ de $\mathbb{M}$, conjunto de puntos en $\mathbb{R}^d$, es decir, una familia de conjuntos U_i tales que $\mathbb{M} = \cup_{i \in I} U_i$, el nervio de $\mathcal{U}$ es el complejo simplicial abstracto $C(\mathcal{U})$ cuyos vértices son los U_i's y que cumple

$$\sigma = [U_{i_0}, \ldots, U_{i_k}] \in C(\mathcal{U}) \text{ si y sólo si } \cap_{j=0}^{k} U_{i_j} \neq \varnothing.$$

Dada una cubierta de un conjunto de datos, donde cada conjunto de la cubierta es, por ejemplo, una agrupación de los puntos de los datos que tienen ciertas propiedades en común, su nervio proporciona una descripción combinatoria, compacta y global, de las relaciones entre estos conjuntos a través de sus patrones de intersección. Ver Figura (1.3).

Un teorema fundamental en topología algebraica se encarga de relacionar, bajo ciertas condiciones, la topología del nervio de una cubierta con la topología de la unión de los conjuntos de dicha cubierta. Es necesario introducir algunas nociones adicionales para ser formales a la hora de enunciar este resultado conocido como el teorema del nervio.

Dos espacios topológicos, X y Y, usualmente son considerados iguales desde un punto de vista topológico si son homeomorfos, esto es, si existen dos funciones, biyectivas y continuas, $f : X \to Y$ y $g : Y \to X$ tales que $f \circ g$ y $g \circ f$ son las funciones identidad de Y y X, respectivamente. En muchas ocasiones, pedir que X y Y sean homeomorfos resulta ser una condición demasiado fuerte para asegurar que X y Y compartan propiedades topológicas de interés para el ATD. Dos funciones continuas $f_0, f_1 : X \to Y$ se dicen ser homotópicas si

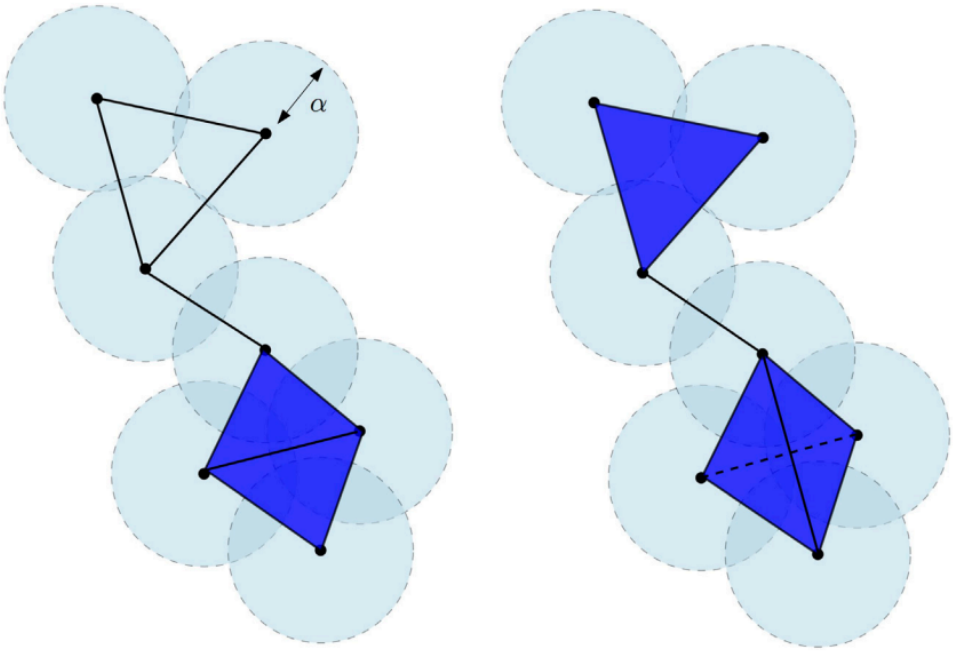


Figura 1.2: El complejo de Čech, $Cech_{\alpha}(\mathbb{X})$ (izquierda) y el de Vietoris-Rips $Rips_{2\alpha}(\mathbb{X})$ (derecha) en una nube finita de puntos en $\mathbb{R}^2$. La parte inferior de $Cech_{\alpha}(\mathbb{X})$ es la unión de dos triángulos adyacentes, mientras que la parte inferior de $Rips_{2\alpha}(\mathbb{X})$ es el tetraedro generado por los cuatro vértices y todas sus caras. La dimensión del complejo de Čech es 2. La dimensión del complejo de Vietoris-Rips es 3. Es de notar que el complejo de Vietoris-Rips, en este caso, no puede ser inmerso en $\mathbb{R}^2$.

existe una función continua $H : X \times [0,1] \to Y$ tal que para cualquier $x \in X$, $H(x,0) = f_0(x)$ y $H(x,1) = g(x)$. Los espacios X y Y se dicen ser homotópicamente equivalentes si existen dos funciones, $f : X \to Y$ y $g : Y \to X$, tales que $f \circ g$ y $g \circ f$ son homotópicas a la funcón identidad de Y y X, respectivamente. Las funciones f y g son llamadas homotópicamente equivalentes. La noción de equivalencia homotópica es más débil que la de homeomorfismo; si X y Y son homeomorfos, entonces son homotópicamente equivalentes, pero el recíproco no es cierto. Sin embargo, espacios que son homotópicamente equivalentes aún comparten muchos invariantes topológicos, como la conexidad por caminos, los grupos de homotopía y, en particular, tienen la misma homología.

Un espacio se dice ser contraíble si es homotópicamente equivalente a un punto. Las bolas, y en general los conjuntos convexos en $\mathbb{R}^d$, son ejemplos básicos de espacios contraíbles. Las cubiertas abiertas, para las cuales se tiene que todos sus elementos e intersecciones son contraíbles, tienen la siguiente propiedad.

Teorema 1.2 (Teorema del Nervio)**.** Sea $\mathcal{U} = (U_i)_{i \in I}$ una cubierta abierta de un espacio topológico X tal que la intesección de cualquier subcolección de los U_i's es contraíble o vacía. Entonces, X y el nervio $C(\mathcal{U})$ son homotópicamente equivalentes.

Es fácil verificar que subconjuntos convexos de espacios euclidianos son contraíbles. Como consecuencia, si $\mathcal{U} = (U_i)_{i \in I}$ es una colección de subconjuntos convexos de $\mathbb{R}^d$, entonces

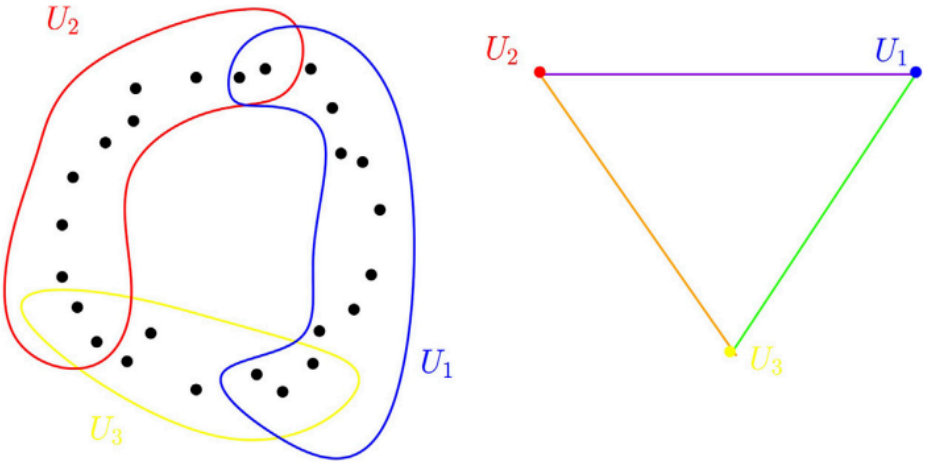


Figura 1.3: Nube de puntos muestreada en el plano y una cubierta de conjuntos abiertos para esta nube (izquierda). El nervio de esta cubierta es un triágulo (derecha). Los vértices corresponden a uno de los conjuntos de la cubierta mientras que las aristas corresponden a una de las intersecciones no vacías entre dos conjuntos de la cubierta.

$C(\mathcal{U})$ y $\cup_{i\in I} U_i$ son homotópicamente equivalentes. En particular, si $\mathbb{X}$ es un conjunto de puntos en $\mathbb{R}^d$, entonces el complejo de Čech $Cech_\alpha(\mathbb{X})$ es homotópicamente equivalente a la unión de bolas $\cup_{x\in\mathbb{X}} B(x,\alpha)$.

El teorema del nervio juega un papel fundamental en el ATD; proporciona una manera de codificar la topología de espacios continuos en estructuras combinatorias abstractas que se ajustan con facilidad al diseño de estructuras de datos y algoritmos efectivos.

Capítulo 2

Utilizando Cubiertas y Nervios para el Análisis de Datos Exploratorio y Visualización: El Algoritmo Mapper.

Usar el nervio de cubiertas como una manera de visualizar y explorar datos es una idea natural que fue propuesta para el ATD en el estudio por Singh et al. [1], dando lugar al algoritmo Mapper.

Definición 2.1. Sea $f : \mathbb{X} \to \mathbb{R}^d$, $d \geq 1$, una función continua y sea $\mathcal{U} = (U_i)_{i \in I}$ una cubierta de $\mathbb{R}^d$. laL cubierta pull-back de $\mathbb{X}$ inducida por $(f,\mathcal{U})$ es la colección de conjuntos abiertos $\left(f^{-1}(U_i)\right)_{i \in I}$. El pull-back refinado es una colección de componentes conexas de los abiertos $f^{-1}(U_i)$, $i \in I$.

La idea del algoritmo Mapper es, dado un conjunto de datos $\mathbb{X}$ y una función $f : \mathbb{X} \to \mathbb{R}^d$, sintetizar $\mathbb{X}$ a través del nervio del pull-back refinado de una cubierta $\mathcal{U}$ de $f(\mathbb{X})$. Para cubiertas bien escogidas $\mathcal{U}$, este nervio es una gráfica que encapsula de manera conveniente el detalle de los datos y los vuelve fáciles de visualizar (Ver Figura 2.1).

El algoritmo de Mapper es muy sencillo; pero este recalca las diferentes elecciones que son dejadas al usuario y que discutiremos a continuación.

- **Entrada:** Un conjunto de datos $\mathbb{X}$ con una métrica o medida de disimilaridad entre los puntos asociados a los datos, una función $f : \mathbb{X} \to \mathbb{R}$ (o bien, $f : \mathbb{X} \to \mathbb{R}^d$), y una cubierta $\mathcal{U}$ de $f(\mathbb{X})$. Para cada $U \in \mathcal{U}$, descomponer $f^{-1}(U)$ en agrupaciones $C_{U,1}, \ldots, C_{U,k_U}$. Calcular el nervio de la cubierta de X definido por los $C_{U,1}, \ldots, C_{U,k_U}$, $U \in \mathcal{U}$.

- **Salida:** Un complejo simplicial; el nervio que incluye un vértice $v_{U,i}$ por cada $C_{U,i}$ y una arista entre cada uno de los vértices $v_{U,i}$ y $v_{U',j}$ que cumplan $C_{U,i} \cap C_{U',j} \neq \varnothing$.

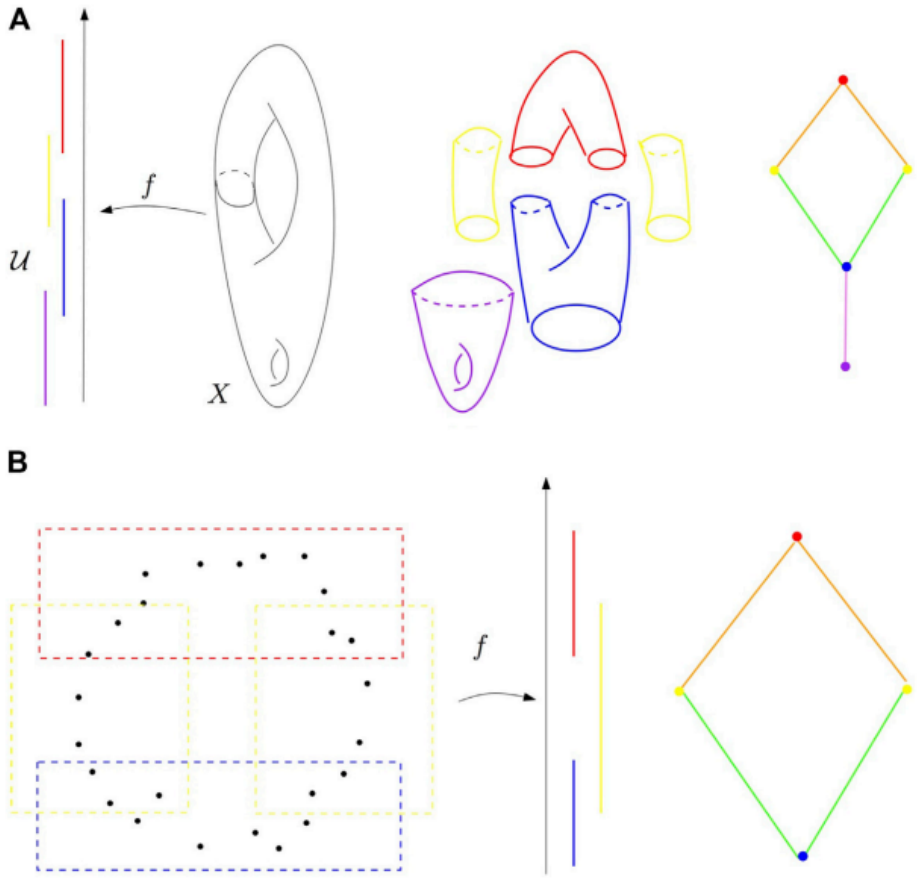


Figura 2.1: (A) Cubierta pull-back refinada de la función altura sobre una superficie en $\mathbb{R}^3$. (B) Algoritmo de Mapper en una nube de puntos muestreada alrededor de un círculo y la función altura.

La Elección de f

La elección de la función f, a veces llamada la función filtro o lente, depende fuertemente de las propiedades de los datos que uno pretende resaltar. Las siguientes son algunas de las más encontradas en la literatura:

- Estimadores de densidad: El complejo Mapper puede ser útil para entender la estructura y conexidad de áreas de alta densidad.
- Coordenadas de análisis de componentes principales (coordenadas PCA) o funciones coordenadas obtenidas de una técnica de reducción de dimensionalidad no lineal (NLDR), eigenfunciones de laplacianos de gráficas pueden ayudar a revelar y entender parte de la ambigüedad en el uso de reducciones de dimensionalidad no lineales.
- La función de centralidad $f(x) = \Sigma_{y\in\mathbb{X}} d(x,y)$ y la función de excentricidad $f(x) = \text{máx}_{y\in\mathbb{X}}\, d(x,y)$ a veces resultan ser buenas elecciones que no requieren de ningún

conocimiento específico acerca de los datos.

- Para datos muestreados sobre estructuras filamentarias de dimensión uno, la función distancia a un punto dado permite recuperar la topología subyacente de las estructuras filamentarias [44].

La Elección de la Cubierta $\mathcal{U}$

Cuando f es una función de valores reales, una elección estándar de $\mathcal{U}$ es un conjunto de intervalos espaciados regularmente y del mismo largo, $r > 0$, cubriendo al conjunto $f(\mathbb{X})$. El número real r es a veces llamado la resolución de la cubierta, y el porcentaje g de sobreposición entre dos intervalos consecutivos es llamado la ganancia de la cubierta. Nótese que si la ganancia g es escogida menor a $50\,\%$, entonces cada punto de la linea real es cubierto por, a lo más, 2 conjuntos abiertos de $\mathcal{U}$, y el nervio resultante es una gráfica. Es importante notar que la salida de Mapper es muy sensible a la elección de $\mathcal{U}$, y cambios pequeños en la resolución o ganancia puede afectar de manera significativa al resultado, volviendo el método muy inestable. Una estrategia clásica consiste en explorar un rango de parámetros y seleccionar aquellos que sean más informativos desde el punto de vista del usuario.

La Elección del Agrupamiento

El algoritmo Mapper requiere el agrupamiento de la preimagen de conjuntos abiertos $U \in \mathcal{U}$. Existen dos estrategias para realizar este agrupamiento. La primera consiste en aplicar, a cada $U \in \mathcal{U}$, un algoritmo de agrupamiento, escogido por el usuario, a la preimagen de $f^{-1}(U)$. La segunda, más global, consiste en construir una gráfica sobre el conjunto de datos $\mathbb{X}$, por ejemplo, una gráfica k-NN o una ϵ-gráfica y, para cada $U \in \mathcal{U}$, tomar las componentes conexas de la subgráfica con el conjunto de vértices $f^{-1}(U)$.

Aspectos Teoréticos y Estadísticos del Algoritmo Mapper

Basados en los resultados de estabilidad y la estructura de Mapper propuestos en el estudio por Carrière y Oudot (2017) [25], se han realizado avances en dirección a una versión de Mapper estadísticamente bien fundamentada en el estudio por Carrière et al. (2018) [27]. De aquí destaca que la convergencia de Mapper depende tanto del muestreo de los datos como de la regularidad de la función filtro. Más aun, estrategias de submuestreo pueden ser usadas para seleccionar un complejo en una filtración de Rips a una escala conveniente, así como la resolución y la ganancia para definir la gráfica Mapper. El caso para filtros estocásticos y multivariados también ha sido estudiado por Carrière y Michel (2019) [26]. Una descripción alternativa de la convergencia probabilística de Mapper, en términos de la categorificación, fue propuesta en el estudio por Brown et al. (2020) [15]. Otros acercamientos también fueron propuestos para estudiar y lidiar con la inestabilidad del algoritmo Mapper en los trabajos de Dey et al. (2016) [60], Dey et al. (2017) [59].

Análisis de Datos con Mapper

Como una herramienta del análisis de datos, Mapper se ha utilizado con éxito para tareas de agrupamiento y selección de atributos. La idea es identificar estructuras específicas en la gráfica (o complejo) Mapper, en particular, lazos. Estas estructuras son usadas para identificar cúmulos interesantes o seleccionar atributos que puedan diferenciar los datos en estas estructuras de manera apropiada. Aplicaciones en datos reales ilustrando estas técnicas pueden ser encontradas en, por ejemplo, los estudios por Carrière y Rabadán (2020) [28], Lum et al. (2013) [90], Yao et al. (2009) [129].

Capítulo 3

Reconstrucción Geométrica e Inferencia Homológica

Otra forma de construir cubiertas y usar sus nervios para exhibir la estructura topológica de los datos es considerar la unión de bolas centradas en los puntos de los datos. En esta sección suponemos que $\mathbb{X}_n = \{x_0, \ldots, x_n\}$ es un subconjunto de $\mathbb{R}^d$, muestrado de manera i. i. d. de acuerdo con la medida de probabilidad μ con soporte compacto $M \subset \mathbb{R}^d$. La estrategia general para inferir información topológica acerca de M a través de μ consiste en dos pasos:

1. Se cubre $\mathbb{X}_n$ con una unión de bolas de radio fijo con centros en las x_i's. Bajo algunas condiciones de regularidad en M, se puede relacionar la topología de esta unión de bolas con la de M.

2. Desde un una perspectiva práctica y algorítmica, las cualidades topológicas de M son inferidas del nervio de la unión de las bolas, utilizando el teorema del nervio.

De esta manera, es posible comparar espacios a través de equivalencias isotópicas, una noción más fuerte que la de homeomorfismo: $X \subseteq \mathbb{X}^d$ y $Y \subseteq \mathbb{X}^d$ se dicen ser (ambientalmente) isotópicos si existe una familia continua de homeomorfismos $H : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, H continua, tal que, para cualquier $t \in [0,1]$, $H_t = H(t, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ es un homeomorfismo, H_0 es el mapeo identidad en $\mathbb{R}^d$, y $H_1(X) = Y$. Es claro que, si X y Y son isotópicos, entonces son homeomorfos. El recíproco no es cierto: un círculo anudado y uno desanudado en $\mathbb{R}^3$ son homeomorfos pero no isotópicos.

Funciones DL y Reconstrucción

Dado un suconjunto compacto $K \subset \mathbb{R}^d$ y un número real no negativo r, la unión de bolas de radio r centradas en K, $K^r = \cup_{x \in K} B(x, r)$, llamado el r-cubrimiento de K, es el conjunto de r-subnivel de la distancia $d_K : \mathbb{R}^d \to \mathbb{R}$ definida por $d_K(x) = \text{ínf}_{y \in K} \|x - y\|$; es decir, $K^r = d_k^{-1}([0, r])$. Esto nos permite utilizar propiedades diferenciales de funciones distancia y nos ayuda a comparar la topología de los cubrimientos de conjuntos compactos que estén cercas el uno del otro con respecto a la distancia de Hausdorff.

Definición 3.1. (Distancia de Hausdorff en $\mathbb{R}^3$). La distancia de Hausdorff entre dos subconjuntos compactos K, K' de $\mathbb{R}^d$ esta definida como

$$d_H\left(K, K'\right) = \left\|d_K - d_{K'}\right\|_\infty = \inf_{x \in \mathbb{R}^d} |d_K(x) - d_{K'}(x)|$$

Aquí, los conjuntos compactos son el conjunto de datos $\mathbb{X}_n$ y el soporte M de la medida μ. Cuando M es una subvariedad compacta suave, bajo ciertas condiciones sobre $d_H(\mathbb{X}_n, M)$, para algún r bien escogido, las coberturas de $\mathbb{X}_n$ son homotópicamente equivalentes a M, Chazal y Lieutier (2008) [47], Niyogi et al. (2008) [97] (Ver Figura 3.1). Estos resultados se extienden a clases más grandes de conjuntos compactos y llevan a resultados fuertes sobre inferencia de los tipos de isotopías de las coberturas de M, Chazal et al. (2009c) [33], Chazal et al. (2009d)[34]. También llevan a resultados en la estimación de otras cantidades geométricas y diferenciales tales como normales, Chazal et al. (2009c) [33], curvaturas Chazal et al. (2009e) [35], o medidas de frontera, Chazal et al. (2010) [36] bajo ciertas condiciones en la distancia de Hausdorff entre la forma subyacente y los datos muestrales.

Estos resultados dependen de la 1-semiconcavidad del cuadrado de la función distancia d_K^2, esto es, la convexidad de la función $x \to \|x\|^2 - d_K^2(x)$, definida de a continuación.

Definición 3.2. Una función $\phi : \mathbb{R}^d \to \mathbb{R}_+$ es DL (distance-like) si es propia (la preimagen de cualquier conjunto compacto en $\mathbb{R}$ bajo ϕ es un compacto en $\mathbb{R}^d$) y $x \to \|x\|^2 - \phi^2(x)$ es convexa.

Gracias a su semiconcavidad, una función DL ϕ tiene un gradiente $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ bien definido, pero no continuo, que puede ser integrado en un flujo continuo (Petrunin, 2007 [101]) que permite rastrear la evolución de la topología de sus subniveles y compararla a una de los subniveles de funciones DL cercanas.

Definición 3.3. Sea ϕ una función DL y sea $\phi^r = \phi^{-1}([0, r])$ el r-subnivel de ϕ.

- Un punto $x \in \mathbb{R}^d$ es llamado α-crítico si $\|\nabla_x \phi\| \leq \alpha$. El valor $r = \phi(x)$ correspondiente, también es llamado α-crítico.
- El tamaño del atributo débil de ϕ en r es el mínimo $r > 0$ tal que ϕ no tiene ningún valor crítico entre r y $r + r'$. Lo denotamos por $\text{wfs}_\phi(r)$ (weak feature size). Para cualquier $0 < \alpha < 1$, el α-alcance de ϕ es el máximo r tal que $\phi^{-1}((0, r])$ no contiene ningún punto α-crítico.

El tamaño del atributo débil $\text{wfs}_\phi(r)$ (respecto al α-alcance) mide la regularidad de ϕ sobre sus r-niveles (respecto al O-nivel). Cuando $\phi = d_K$ es la función distancia a un conjunto compacto $K \subset \mathbb{R}^d$, el 1-alcance coincide con el alcance clásico de la teoría de la medida gemétrica, Federer (1959) [67]. Su estimación desde muestras aleatorias fue estudiada en Aamari et al. (2019) [2]. Una propiedad importante de una función DL ϕ es que la topología de sus subniveles ϕ^r sólo puede cambiar cuando r cruza un valor 0-crítico.

Lema 3.4. (Lema de isotopía). Sea ϕ una función DL y $r_1 < r_2$ dos números positivos tales que ϕ no tiene puntos 0-críticos, esto es, puntos x tales que $\nabla\phi(x) = 0$, en el subconjunto $\phi^{-1}([r_1, r_2])$. Entonces todos los subniveles $\phi^{-1}([0, r])$ son isotópicos para $r \in [r_1, r_2]$.

Como consecuencia inmediata del lema de isotopía, todos los subniveles de ϕ entre r y $r + \text{wfs}_\phi(r)$ tienen la misma topología. Ahora, el siguiente teorema de Chazal et al. [37], proporciona una conexión entre la topología de los subniveles de funciones DL cercanas.

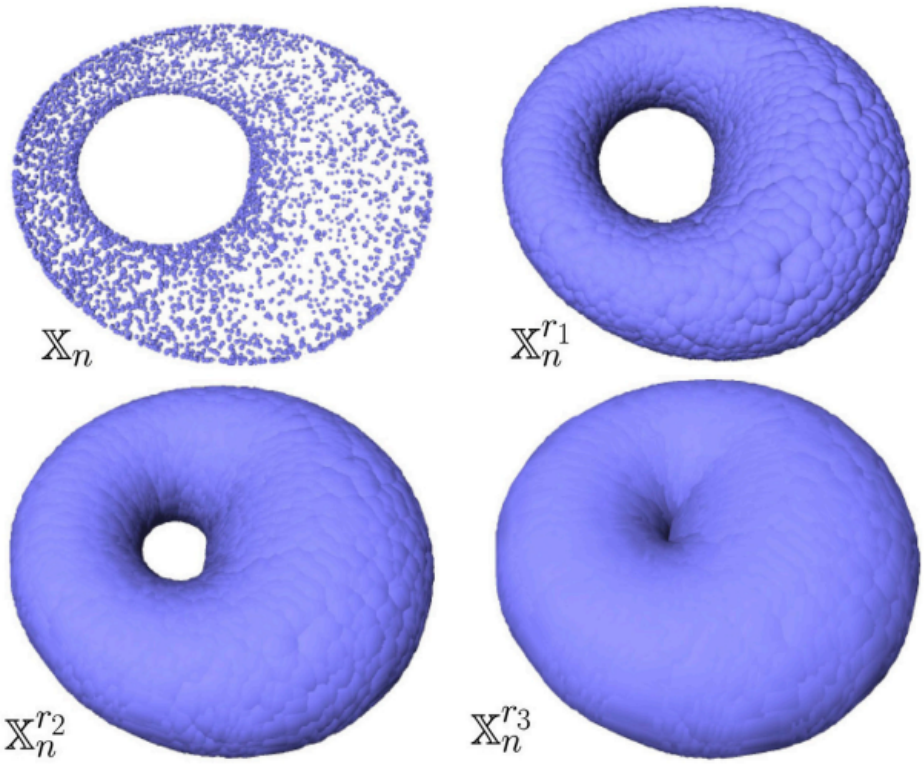


Figura 3.1: Ejemplo de una nube de puntos $\mathbb{X}_n$ muestreada en la superficie de un toro en $\mathbb{R}^3$ y sus coberturas para diferentes valores del radio $r_1 < r_2 < r_3$. Para valores bien escogidos del radio (por ejemplo r_1 y r_2), las coberturas son homotópicamente equivalentes al toro.

Teorema 3.5 (Teorema de reconstrucción)**.** Sean ϕ, ψ dos funciones DL, tales que $\|\phi - \psi\|_\infty \leq \epsilon$, con α-alcance $\text{reach}_\alpha(\phi) \geq R$ para algunos ϵ y α positivos. Entonces, para todo $r \in \left[4\epsilon/\alpha^2, R - 3\epsilon\right]$ y cada $\eta \in (0, R)$ los subniveles ψ^r y ϕ^η son homotópicamente equivalentes si:

$$\epsilon \leq \frac{R}{5 + 4/\alpha^2}$$

Bajo condiciones similares pero ligeramente más técnicas, el teorema de reconstrucción puede ser extendido para probar que los subniveles son homeomorfos e incluso isotópicos (Chazal et al., 2009 [33]; Chazal et al., 2008 [47]).

Consideremos una vez más $\phi = d_M$ y $\psi = d_{\mathbb{X}_n}$ las funciones distancia al soporte de M de la medida μ y al conjunto de puntos asociados a los datos $\mathbb{X}_n$, la condición $\text{wfs}_\alpha(d_M) \geq R$ puede ser interpretada como una condición de regularidad sobre M[1]. El teorema de reconstrucción junto con el teorema del nervio nos indican que para ciertos valores de r, η y los η-cobertura son homotópicamente equivalentes al nervio de la unión de las bolas de radio r centradas en $\mathbb{X}_n$, es decir, el complejo de Čech $Cech_r(\mathbb{X}_n)$.

[1]Por ejemplo, si M es una subvariedad compacta suave, el 0-alcance $\text{reach}_0(\phi)$ siempre es positivo se le llama el alcance de M Federer (1959) [67]

Desde un punto de vista estadístico, la principal ventaja de estos resultados sobre la distancia de Hausdorff es que el problema de estimación de cantidades topológicas se transforma en una serie de preguntas acerca de el soporte de ciertas medidas, las cuales han sido ampliamente estudiadas.

3.1. Inferencia Homológica

Los resultados anteriores proporcionan una estructura matemática bien fundamentada para inferir la topología de las formas de un complejo simplicial construido sobre una muestra finita que sirve como aproximación. Sin embargo, desde una perspectiva más práctica, aparecen dos problemas. Primero, el teorema de reconstrucción requiere de regularidad a través de la condición del α-alcance que a veces no puede ser garantizada, además de la elección del radio r que se debe realizar para construir el complejo de Čech $Cech_r(\mathbb{X}_n)$. Segundo, $Cech_r(\mathbb{X}_n)$ brinda una fiel descripción topológica de los datos a través de un complejo simplicial que normalmente no es adecuado para un procesamiento de datos adicional. Es conveniente tener descriptores topológicos que sean fáciles de manejar, en particular descriptores numéricos, que pueden ser calculados desde el complejo simplicial de manera sencilla. Este segundo problema se resuelve al considerar la homolgía del complejo simplicial en cuestión, tema que se desarrollara a continuación, por otra parte, el primer problema sera resuelto en la siguiente capítulo con la introducción a la homología persistente.

Homología

La homología es un concepto clásico en la topología algebraica, brinda una herramienta poderosa para formalizar y manejar la noción de características topológicas de un espacio topológico o un complejo simplicial de manera algebraica. Para cualquier dimensión k, los "hoyos" k-dimensionales son representados por un espacio vectorial H_k, cuya dimenión es el número de dichas propiedades. Por ejemplo, el grupo de homología 0-dimensional H_0 representa las componentes conexas del complejo, el grupo de homología 1-dimensional H_1 representa los lazos de dimensión uno, el grupo de homología 2-dimensional H_2 representa las cavidades de de dimensión dos, y así sucesivamente.

Para evitar dificultades y sutilezas técnicas, restringimos esta introducción a la homología al mínimo necesario para continuar con nuestro programa. En particular, nos restringimos al caso donde la homología tiene coeficientes en $\mathbb{Z}_2$, esto es, el campo con dos elementos, 0 y 1, tales que $1+1=0$, que tiene una interpretación geométrica más intuitiva. No obstante, todas las nociones y resultados presentados aquí se extienden de manera natural a la homología con coeficientes en cualquier campo. Referimos al lector al estudio por Hatcher (2001) [72] para una introducción completa a la homología y al estudio por Ghrist (2017) [70] para una introducción concisa y reciente a la topología algebraica aplicada y sus conexiones con el análisis de datos.

Sea K un complejo simplicial (finito) y k un entero no-negativo. El espacio de las k-cadenas en K, $C_k(K)$ es el conjunto cuyos elementos son las sumas formales (finitas) de los k-simplices de K. Más precisamente, si $\{\sigma_1, \ldots, \sigma_p\}$ es el conjunto de los k-simplices de K, entonces cualquier k-cadena puede ser escrita como:

$$c = \sum_{i=0}^{p} \epsilon_i \sigma_i \text{ con } \epsilon_i \in \mathbb{Z}_2$$

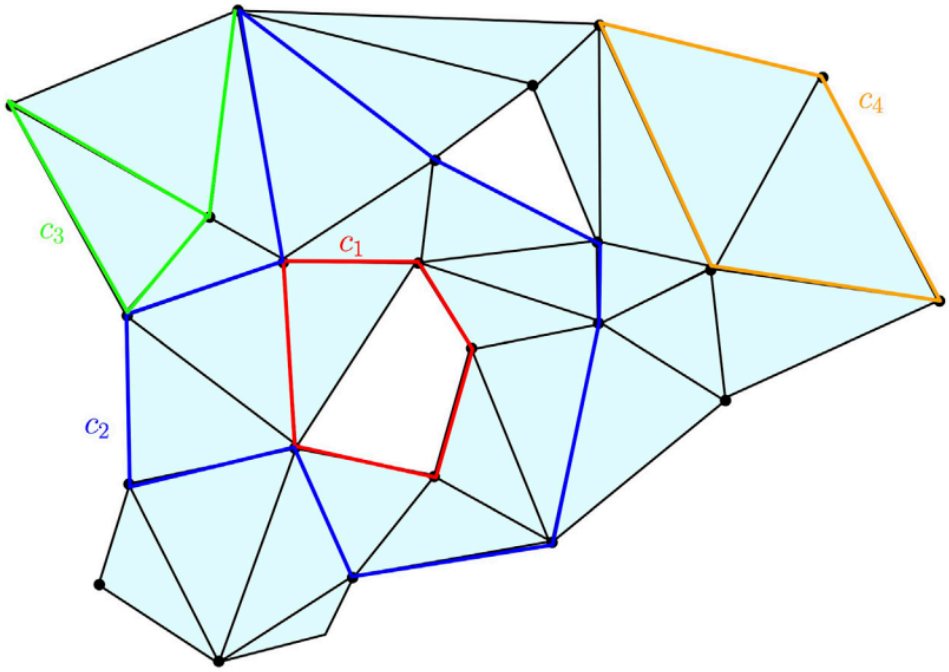


Figura 3.2: Algunos ejemplos de cadenas, ciclos y fronteras en un complejo K de dos dimensiones: c_1, c_2 y c_4 son 1-ciclos; c_3 es una 1-cadena pero no un 1-ciclo; c_4 es una 1-frontera, la frontera de la 2-cadena obtenida de la suma de los triángulos rodeados por c_4. Los ciclos c_1 y c_2 generan el mismo elemento en $H_1(K)$ ya que su diferencia es la 2-cadena representada por la unión de triángulos que rodean la unión de c_1 y c_2.

Si $c' = \sum_{i=1}^{p} \epsilon_i' \sigma_i$ es otra k-cadena y $\lambda \in \mathbb{Z}_2$, la suma $c + c'$ esta definida como $c + c' = \sum_{i=1}^{p} (\epsilon_i + \epsilon_i') \sigma_i$ y el producto $\lambda \cdot c$ esta definido como $\lambda \cdot c = \sum_{i=1}^{p} (\lambda \cdot \epsilon_i) \sigma_i$, convirtiendo a $C_k(K)$ en un espacio vectorial con coeficientes en $\mathbb{Z}_2$. Ya que estamos considerando los coeficientes en $\mathbb{Z}_2$, geométricamente, una k-cadena puede ser vista como una colección finita de k-simplices y la sumas de dos k-cadenas como la diferencia simétrica de las colecciones correspondientes[2].

La frontera de un k-simplejo $\sigma = [v_k, \ldots, v_k]$ es la $(k-1)$-cadena

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1) [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

donde $[v_0, \ldots, \hat{v}_i, \ldots, v_k]$ es el $(k-1)$-simplejo generado por todos los vértices a excepción de v_i[3]. Dado que los k-simplejos forman una base de $C_k(K)$, ∂_k se extiende como una función lineal de $C_k(K)$ a $C_{k-1}(K)$ llamado el operador frontera. El kernel de ∂_k denotado por: $Z_k(K) = \{c \in C_k(K) : \partial_k(c) = 0\}$ es llamado el espacio de k-ciclos de K, y la imagen de ∂_{k+1} denotada por: $B_k(K) = \{c \in C_k(K) : \exists c' \in C_{k+1}(K), \partial_{k+1}(c') = c\}$ es llamada el espacio de k-fronteras de K.

[2]Recordemos que la diferencia simétrica entre dos conjuntos A y B es el conjunto $A\Delta B = (AB) \cup (BA)$

[3]Ya que estamos considerando los coeficientes en $\mathbb{Z}_2$, se tiene que $-1 = 1$ y por lo tanto $(-1)^i = 1$ para cualquier i.

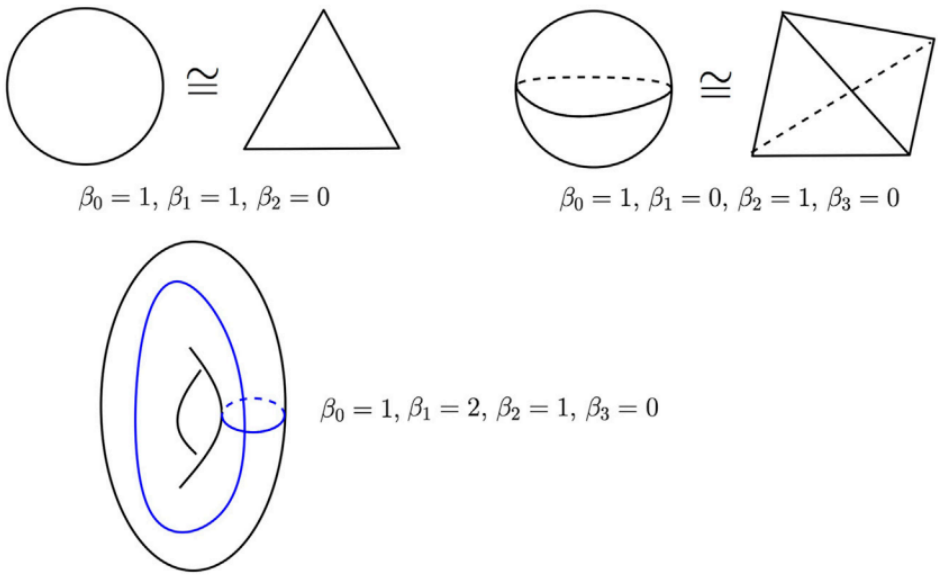


Figura 3.3: Números de Betti en el círculo, la esfera de dimensión dos y el toro de dimensión dos. Las curvas azules en el toro representan dos ciclos independientes cuya clase de homología es una base para el grupo de homología de dimensión uno.

El operador frontera satisface la siguiente propiedad fundamental:

$$\partial_{k+1} \circ \partial_{k+1} \equiv 0 \text{ para cualquier } k \geq 1.$$

en otras palabras, cualquier k-frontera es un k-ciclo, esto es, $B_k(K) \subseteq Z_k(K) \subseteq C_k(K)$. Estas nociones son ilustradas en la Figura 3.2.

Definición 3.6. (grupo de homología simplicial y números de Betti). El k-ésimo grupo de homología (simplicial) de K es el espacio cociente

$$H_k(K) = Z_k(K)/B_k(K).$$

El k-ésimo número de Betti de K es la dimensión $\beta_k(K) = \dim H_k(K)$ del espacio $H_k(K)$.

La Figura 3.3 muestra los números de Betti de algunos espacios sencillos. Dos ciclos, $c, c' \in Z_k(K)$, se dicen ser homólogos si difieren por una frontera, esto es, si existe una $(k+1)$-cadena d tal que $c' = c + \partial_{k+1}(d)$. Dichos ciclos dan lugar al mismo elemento de H_k. En otras palabras, los elementos de $H_k(K)$ son clases de equivalencia de ciclos homólogos.

Los grupos de homología simplicial y los números de Betti son invariantes topologicas; si K, K' son dos complejos simpliciales tales que sus realizaciones geométricas son homotópicamente equivalentes, entonces sus grupos de homología son isomorfos y sus números de Betti son iguales.

La homología singular es otra noción de homología que nos permite considerar una mayor variedad de espacios topológicos. Esta definida para cualquier espacio topológico X de manera similar a la homología simplicial, excepto que el concepto de simplejo, es reemplazado por por el de simplejo singular, que consiste en una función continua $\sigma : \Delta_k \to X$ donde Δ_k es el simplejo estándar de dimensión k. El espacio de las k-cadenas es el

espacio vectorial generado por los simplejos singulares k-dimensionales, y la frontera de un simplejo σ esta definida como la suma (alternante) de la restricción de σ a las caras $(k-1)$-dimensionales de Δ_k. Algo importante acerca de la homología singular es hecho de que esta coincide con la homología simplicial cuando X es homeomorfo a la realización gemétrica de un complejo simplicial. Esto nos permite hablar acerca de la homología de un espacio topológico o un complejo simplical, sin tener que especificar si nos referimos a la homología singular o simplicial.

Observemos que si $f : X \to Y$ es una función continua, entonces para cualquier simplejo singular $\sigma : \Delta_k \to X$ en X, se tiene que $f \circ \sigma : \Delta_k \to Y$ es un simplejo singular en Y, de aquí, deducimos que funciones continuas entre espacios topológicos inducen homomorfismos entre sus grupos de homología. En particular, si f es una equivalencia homotópica, entonces se induce un isomorfismo entre $H_k(X)$ y $H_k(Y)$ para cualquier k entero no-negativo. Por ejemplo, sea $X \subset \mathbb{R}^d$ cualquier conjunto de puntos y $r > 0$, se sigue del teorema del nervio que la r-cobertura X^r y el complejo de Čech $Cech_r(X)$ tienen grupos de homología isomorfos y los mismos números de Betti.

Además de esto, tenemos como consecuencia del teorema de reconstrucción 3.5 el siguiente resultado que nos auxilia en la estimación de números de Betti.

Teorema 3.7. Sea $M \subset \mathbb{R}^d$ un conjunto compacto con alcance, $\text{reach}_\alpha(d_M) \geq R > 0$ para algún $\alpha \in (0,1)$ y sea $\mathbb{X}$ un conjunto finito de puntos tales que:

$$d_H(M, \mathbb{R}) = \epsilon < \frac{R}{5 + 4/\alpha^2}.$$

Entonces, para cada $r \in \left[4\epsilon/\alpha^2, R - 3\epsilon\right]$ y cada $\eta \in (0, R)$, los números de Betti de $Cech_r(\mathbb{X})$ y M^η son iguales.

En particular, si M es una subvariedad suave de $\mathbb{R}^d$ de dimensión $m \in \mathbb{Z}$, entonces $\beta_k(Cech_r(\mathbb{R})) = \beta_k(M)$ para cualquier $k = 0, \ldots, m$.

Desde una perspectiva más pragmática, este resultado nos genera tres problemas: primero, la suposición de regularidad acerca del α-alcance de M puede ser demasiado restrictiva; segundo, el cálculo del nervio de la unión de bolas requiere de métodos para probar que la unión finita de bolas sea no-vacía; tercero, la estimación de números de Betti recae en la elección del parámetro r.

Para solucionar los problemas anteriores, Chazal y Oudot (2008) [50] establecieron el siguiente resultado que ofrece la solución a los primeros dos problemas.

Teorema 3.8. Sea $\mathbb{M} \subseteq \mathbb{R}^d$ un conjunto compacto tal que $\text{wfs}(M) = \text{wfs}_{d_M}(0) \geq R > 0$ y sea $\mathbb{X}$ un conjunto de puntos finito tal que $d_H(M, \mathbb{X}) = \epsilon < \frac{1}{9}\text{wfs}(M)$. Entonces para cualquier $r \in \left[2\epsilon, \frac{1}{4}(\text{wfs}(M) - \epsilon)\right]$ y cualquier $\eta \in (0, R)$,

$$\beta_k(X^\eta) = \text{rk}(H_k(Rips_r(\mathbb{X}))) \to H_k(Rips_{4r}(\mathbb{X}))$$

donde $\text{rk}(H_k(Rips_r(\mathbb{X}))) \to H_k(Rips_{4r}(\mathbb{X}))$ denota el rango de del homomorfismo inducido por la inclusión canónica (continua) $Rips_r(\mathbb{X}) \hookrightarrow Rips_{4r}(\mathbb{X})$.

Aunque este resultado deja abierta la elección del parámetro r, en el estudio realizado por Chazal y Oudot (2008) [50] se provee una descripción de una estrategia multiescala que ayuda a identificar las escalas relevantes en las cuales se puede aplicar el teorema anterior.

3.2. Aspectos estadísticos de la inferencia homológica

De acuerdo a los resultados de estabilidad presentados en la sección anterior, un acercamiento estadístico a la inferencia topológica se relaciona fuertemente al problema de estimación de soportes de distribuciones y estimaciones de conjuntos nivel bajo la métrica de Hausdorff. Afortunadamente se cuenta con una variedad de metodos y resultados que nos atudan a estimar el soporte de una distribución. Por ejemplo, el estimador de Devroye y Wise (Devroye y Wise 1980 [58]) definido en una muestra $\mathbb{X}_n$ es también una cobertura particular de $\mathbb{X}_n$. La tasa de convergencia de $\mathbb{X}_n$ y el estimador de Devroye y Wise al soporte de la distribución para la distancia de Hausdorff fueron estudiados por Cuevas y Rodriguez-Casal (2004) [55] en $\mathbb{R}^d$. Recientemente, las tasas de convergencia minimax de estimación de variedades bajo la metrica de Hausdorff, particularmente relevantes para la inferencia topológica, fueron estudiadas por Genovese et al. (2012) [69]. También existe literatura acerca de la estimacion de los conjuntos de nivel en varias métricas (vease, por ejemplo, Cadre, 2006 [21]; Polonik, 1995 [104]; Tsybakov, 1997 [121]) y, particularmente, para la métrica de Hausdorff Chen et al. (2017) [52]. Todos estos trabajos acerca de la estimación de soportes y conjuntos nivel, dan lugar al análisis estadístico de procesos de inferecia topológica.

En el estudio por Nigoyi et al (2008) [97], se muestra que el tipo de homotopía de variedades Riemannianas con alcance mayor que cierta constante puede ser recuperado con una alta probabilidad de las coberturas de una muestra en (o bien, cerca) de la variedad en cuestión. Este articulo fue probablemente el primer intento de considerar la inferencia topológica en términos de probabilidad. El estudio por Nigoyi et al.[97] derivo de un argumento de contracción de retracto y utilizó cotas estrechas sobre el número de cobertura de la variedad para controlar la distancia de Hausdorff entre la variedad y la nube de puntos observada. La inferencia homológica en el caso de ruido presente, esto es, en el sentido de que la distribución de la observación se concentra alrededor de la variedad, también fue estudiado por Nigoyi et al. (2008) [97], Nigoyi et al. (2011) [97]. La suposición de que el objeto geométrico es una variedad Riemanniana suave solo es usada en el artículo para controlar la distancia de Hausdorff entre la muestra y la variedad y no es realmente necesaria para la "parte topológica" del resultado, el cual es similar a aquellos en los estudios por Chazal et al. (2009) [34], Chazal y Lieutier (2008) [47] en el entorno particular de las variedades Riemannianas. Empezando por el resultado del estudio por Nigoyi et al. (2008) [97], las tasas de convergencia minimax del tipo de homología han sido estudiadas por Balakrishnan et al, (2012) [5] bajo varios modelos de variedades Riemannianas con alcance más grande que cierta constante. En contraste, no se ha propuesto una versión estadística del trabajo por Chazal et al. (2009) [34].

Más recientemente, siguiendo las ideas encontradas en Nigoyi et al. (2008) [97], Bobrowski et al (2014) [8] se ha propuesto un robusto estimador homológico para los conjuntos de nivel de funciones de densidad y regresión, por medio de considerar la inclusión entre pares anidados de conjuntos de nivel estimados obtenidos mediante un estimador del kernel.

3.3. Más allá de la Distancia de Hausdorff: Distancia a una Medida

Es bien sabido que los métodos del ATD fallan rotundamente en presencia de puntos aislados, añadir un solo punto asilado al conjunto de datos puede alterar la función distacia de manera dramática (ver Figura 3.4). Como respuesta a esto, Chazal et al. (2011) [37] introdujeron una función distancia alternativa la cual es resistente ante el ruido, la distancia a una medida.

Dada una distribución de probabilidad P en $\mathbb{R}^d$ y un parámetro real $0 \leq U \leq 1$, la noción de distancia al soporte de P puede ser generalizada como la función

$$\delta_{P,u} : x \in \mathbb{R}^d \mapsto \text{ínf}\, t > 0 : P\left(B\left(x,t\right)\right) \geq u$$

donde $B\left(x,t\right)$ es la bola cerrada (Euclideana) con centro en x y radio t. Para evitar problemas de discontinuidad con la función $P \to \delta_{P,u}$, la función distancia a la medida (DAM) con parámetro $m \in [0,1]$ y potencia $r \geq 1$ esta definida como

$$d_{P,m,r}\left(x\right) : x \in \mathbb{R}^d \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^r\left(x\right)\, du\right)^{\frac{1}{r}} \tag{3.1}$$

Una propiedad deseable de las DAM demostrada por Chazal et al. (2011)[37] es la estabilidad con respecto a las perturbaciones de P en la métrica de Wasserstein, más precisamente, la función $P \to d_{P,m,r}$ es $m^{-\frac{1}{r}}$-Lipschitz, esto es, si P y $\tilde{P}$ son dos distribuciones de probabilidad en $\mathbb{R}^d$, entonces

$$\left\| d_{P,m,r} - d_{\tilde{P},m,r} \right\|_\infty \leq m^{-\frac{1}{r}} W_r\left(P,\tilde{P}\right) \tag{3.2}$$

donde W_r es la distancia de Wasserstein para la métrica Euclidiana en $\mathbb{R}^d$, con exponente r[4]. Esta propiedad implica que la DAM asociada con distribuciones cercanas en la métrica de Wasserstein tienen conjuntos subnivel cercanos. Más aún, cuando $r = 2$, la función $d_{P,m,2}^2$ es semiconcava, lo cual asegura fuertes propiedades de regularidad en la geometría de sus subniveles. Usando estas propiedades, Chazal et al. (2011) [37] mostró que bajo suposiciones generales, si $\tilde{P}$ es una distribución de probabilidad que aproxima a P, así los conjuntos subnivel de $d_{\tilde{P},m,2}$ proveen una aproximación topológicamente correcta al soporte de P.

En la práctica, la medida P usualmente solo es conocida a través de un conjunto finito de observaciones $\mathbb{X}_n = \{X_1, \ldots, x_n\}$ muestreada desde P, dando lugar a la pregunta de una aproximación a la DAM. Una idea natural para estimar la DAM desde $\mathbb{X}_n$ es utilizar la medida empírica P_n en lugar de P en la definición de la DAM. Esto corresponde al computo de la distancia Ã la medida empírica (DAME). Para $m = \frac{k}{n}$, la DAME satisface

$$d_{P_n,k/n,r}^r\left(x\right) \coloneqq \frac{1}{k}\sum_{j=1}^{k} \left\|x - \mathbb{X}_n\right\|_{(j)}^r$$

donde $\left\|x - \mathbb{X}_n\right\|_{(j)}$ denota la distancia entre x y su j-ésima vecindad en $\{X_1, \ldots, X_n\}$. Esta cantidad es fácil de calcular en la práctica ya que solo requiere de distancias entre x y los puntos de la muestra. La convergencia de las DAME a las DAM ha sido estudiada por Chazal et al. (2017)[30] y Chazal et al (2016)[48].

[4] Ver Villiani (2003)[20] para la definición de distancia de Wasserstein

La introducción de las DAM a motivado trabajos y aplicaciones en diferentes direcciones tales como el anĺisis topológico de datos (Buchet et al., 2015[18]), análisis de trazas GPS (Chazal et al., 2011 [31]), estimación de densidad (Biau et al., 2011 [7]), pruebas de hipótesis (Brécheteau, 2019 [11]), y agrupamiento (Chazal et al., 2013 [43]), solo para nombrar algunos. También se han tomado en consideración, aproximaciones, generalizaciones, y variantes de las DAM (Guibas et al., 2013 [71]; Phillips et al., 2014 [102]; Buchet et al., 2015 [19]; Brécheteau y Levrard, 2020 [12]).

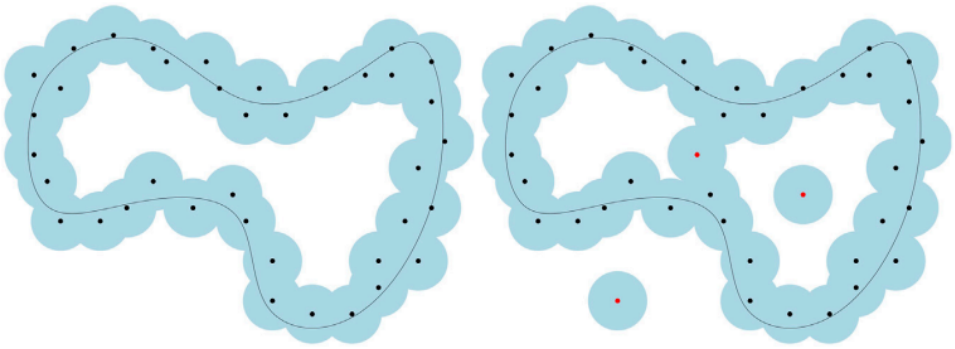


Figura 3.4: Efectos de los puntos aislados en los conjuntos subnivel de las funciones distancia. Añadir unos pocos puntos aislados a la nube puede alterar dramáticamente la función distancia y la topología de sus coberturas.

Capítulo 4

Homología Persistente

La homología persistente es una herramienta poderosa que es usada para el computo, estudio y codificación multiescala de propiedades topológicas de familias anidadas de complejos simpliciales y espacios topológicos. No solo provee algoritmos eficientes para calcular los números de Betti de cada complejo en las familias consideradas, como se requiere para la inferencia homológica cubierta en la sección anterior, sino que también codifica la evolución de los grupos de homología de los complejos anidados a través de las escalas. Ideas y resultados preliminares que culminan en la teoría de la homología persistente pueden ser encontrados desde antes del siglo XXI, en particular en los trabajos de Barannikov (1994) [13], Frosini (1992) [68], Robins (1999) [111]; pero su desarrollo en su forma moderna se concreto en los trabajos de Edelsbrunner et al. (2002) [64] y Zomorodian y Carlsson (2005) [131].

4.1. Filtraciones

Una filtración de un complejo simplicial K es una familia anidada de subcomplejos $(K_r)_{r\epsilon t}$, donde $T \subseteq \mathbb{R}$, tal que para cualquier $r, r' \in T$, si $r \leq r'$ entonces $K_r \subseteq K_{r'}$ y $K = \bigcup_{r\epsilon T} K_r$. El subconjunto T puede ser finito o infinito. En general, una filtración de un espacio topológico $\mathbb{M}$ es una familia anidada de subespacios $(M_r)_{r\epsilon T}$, donde $T \subseteq \mathbb{R}$, tal que para cualquier $r, r' \in T$, si $r \leq r'$ entonces $M_r \subseteq M_{r'}$ y $M = \bigcup_{r\epsilon T} M_r$. Por ejemplo, si $f : \mathbb{M} \to \mathbb{R}$ es una función, entonces la familia $M_r = f^{-1}((-\infty, r])$, $r \in \mathbb{R}$ define una filtración llamada la filtración del conjunto subnivel de f.

En la práctica, el parámetro $r \in T$ suele ser interpretado como un parámetro de escala, y las filtraciones comúnmente usadas en el ATD suelen pertenecer a uno de los siguientes dos tipos.

Filtraciones Sobre Datos

Dado un subconjunto $\mathbb{X}$ de un espacio métrico compacto (M, ρ), las familias de complejos de Vietoris-Rips $(Rips_r(\mathbb{X}))_{r\in\mathbb{R}}$ y los complejos de Čech $(Cech_r(\mathbb{X}))_{r\in\mathbb{R}}$ son filtraciones[1]. Aquí, el parámetro r puede ser interpretado como la resolución con la que se considera el conjunto de datos $\mathbb{X}$. Por ejemplo, si $\mathbb{X}$, es una nube de puntos en $\mathbb{R}^d$, gracias al teorema del nervio, la filtración $(Cech_r(\mathbb{X}))_{r\in\mathbb{X}}$ codifica la topología de todo la familia de

[1] Aquí consideramos $Rips_r(\mathbb{X}) = Cech_r(\mathbb{X}) = \varnothing$, si $r < 0$

uniones de bolas $\mathbb{X}^r = \cup_{x\in\mathbb{X}} B(x,r)$, cuando $0 < r < \infty$. Como la noción de filtración es algo flexible, se han considerado muchas otras filtraciones en la literatura para ser construidas sobre los datos, como el complejo testigo popularizado en el ATD por De Silva y Carlsson (2004)[116], las filtraciones de Rips con peso Buchet et al. (2015)[19], o las filtraciones DTM Anai et al. (2019)[4] que nos permiten trabajar con conjuntos de datos con ruido o con datos atípicos.

Filtraciones de Conjuntos Subnivel

Definir funciones en los vértices de un complejo simplicial da lugar a otro importante ejemplo de filtración: sea K el complejo simplicial con el conjunto de vértices V y $f : V \to \mathbb{R}$. Entonces f puede ser extendida a todos los simplices de K definiendo $f([v_0,\ldots,v_k]) = \max\{f(v_i) : i = 1,\ldots,k\}$ para cualquier simplejo $\sigma = [v_0,\ldots,v_k] \in K$ y la familia de subcomplejos, $K_r = \{\sigma \in K : f(\sigma) \leq r\}$, define una filtración llamada la filtraci'on del conjunto subnivel de f. La filtración del conjunto sobre-nivel de f se define de manera similar.

En la práctica, incluso si el índice del conjunto es infinito, todas las filtraciones consideradas son construidas en conjuntos finitos y son, en si, finitas. Por ejemplo, cuando $\mathbb{X}$ es finito, el complejo de Vietoris-Rips $Rips_r(\mathbb{X})$ cambia solo en un numero finito de índices, r. Esto nos permite manejarlos de manera sencilla desde una perspectiva algebraica.

4.2. Algunos Ejemplos

Dada una filtración $Filt = (F_r)_{r\in T}$ de un complejo simplicial o un espacio topológico, la homología de F_r cambia cuando r incrementa; pueden aparecen nuevos componentes conexos y algunos ya existentes pueden unirse, aros y cavidades pueden formarse o llenarse, etc. La homología persistente registra estos cambios, identifica las propiedades que aparecen y asocia un tiempo de vida a cada una. La información resultante se codifica como un conjunto de intervalos llamado código de barras, o bien, como un conjunto de puntos en $\mathbb{R}^2$ donde la coordenada de cada punto es el punto de inicio y final de cada intervalo correspondiente.

Antes de dar una definición formal, ilustraremos el concepto de homología persistente con unos ejemplos.

Ejemplo 1

Sea $f : [0,1] \to \mathbb{R}$ la función de la Figura 4.1, y sea $F_r = f^{-1}((-\infty, r))_{r\in\mathbb{R}}$ la filtración del conjunto subnivel de f. Todos los conjuntos subnivel de f son o bien vacios o la unión de intervalos, así que la única información topológica no-trivial que brindan es su homología cero dimensional, esto es, su número de componenetes conexas. Para $r < a_1$, F_r es vacio, pero para $r = a_1$, aparecen un primer componente conexo en F_{a_1}. La homología persistente registra a_1 como la "fecha de nacimiento" de una componente conexa la codifica como un intervalo que comienza en a_1. Luego, F_r permanece conexo hasta que r toma el valor de a_2 donde una segunda componente conexa aparece. La homología persitente registra esta nueva componente conexa creando un segundo intervalo que comienza en a_2. De manera similar, cuando r alcanza el valor de a_3, una nueva componente conexa aparece

y la homología persistente crea otro intervalo comenzando en a_3. Cuando r alcanza a_4, las dos componentes creadas en a_1 y a_3 se juntan para crear una sola componente conexa. En este paso, la homología persistente sigue la regla de que la componente que muere es la más reciente que ha aparecido en la filtración; Así, el intervalo que comenzó en a_3 termina en a_4, y el intervalo de persistencia que codifica el tiempo de vida de la componente nacida en a_3 es creado. Cuando r alcanza a_5, como en el caso previo la componente nacida en a_2 muere y se crea el intervalo (a_2, a_5). El intervalo creado en a_1 permanece hasta el final de la filtración, dando lugar al intervalo (a_1, a_6) si la filtración se detiene en a_6, o bien, (a_1, ∞) si r tiende a $+\infty$ (en cuyo caso la filtración se mantiene constante para $r > a_6$). El conjunto de intervalos obtenidos que codifican el tiempo de vida de diferentes características homológicas a lo largo de la filtracón es llamado el código de barras de persistencia de f. Cada intervalo (a, a') puede ser representado en el plano $\mathbb{R}^2$ por el punto (a, a'). El conjunto de puntos resultante es llamado el diagrama de persistencia de f. Es de notar que la función puede tener multiples copias del mismo intervalo en su código de barras de persitencia. Como consecuencia, el diagrama de persistencia de f es un multiconjunto donde cada punto tiene una multiplicidad entera asociada. Finalmente, por razones técnicas que seran claras más adelante, se añaden al diagrama de persistencia todos los puntos de la diagonal $\Delta = \{(b, d) : b = d\}$ con multiplicidad infinita.

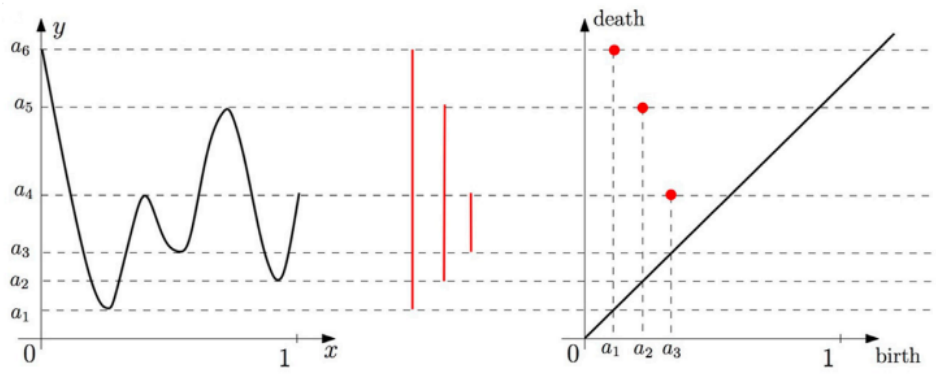


Figura 4.1: El código de barras de persistencia y el diagrama de persistencia de la función $f : [0, 1] \to \mathbb{R}$.

Ejemplo 2

Sea $f : M \to \mathbb{R}$ la función en la Figura 4.2, donde M es una superficie de dos dimensiones homeomorfa a un toro, y sea $F_r = f^{-1}\left((-\infty, r)\right)_{r \in \mathbb{R}}$ la filtración del conjunto subnivel de f. La homología persistente cero dimensional se calcula como en el ejemplo anterior, lo cual genera las barras rojas en el código de barras de persistencia. En este caso los subniveles también almacenan información acerca de características homológicas uno dimensionales. Cuando r alcanza la altura a_1, los conjuntos subnivel F_r que eran homeomorfos a dos discos se vuelven homeomorfos a la unión disjunta de un disco y un ánulo, creando un primer ciclo homólogo a σ_1 en la Figura 4.2. El nacimiento de este uno-ciclo es representado por un intervalo (en azul) que comienza en a_1. Similarmente, cuando r alcanza a_2, un

segundo ciclo, homólogo a σ_2, es creado, dando lugar al comienzo de un nuevo intervalo de persistencia. Estos dos ciclos nunca son rellenos (abarcan $H_1(M)$) de manera que los intervalos que les corresponden continuan por el resto de la filtración. Cuando r alcanza a_3, un nuevo ciclo es creado, el cual se rellena en a_4, lo cual genera el intervalo de persistencia (a_3, a_4). Esta vez, la filtración del conjunto subnivel da lugar a dos códigos de barras, uno par la homología cero dimensional (mostrado en rojo) y otro para la homología uno dimensional (mostrado en azul). Estos dos códigos de barras pueden ser representados de manera equivalente como diagramas en el plano.

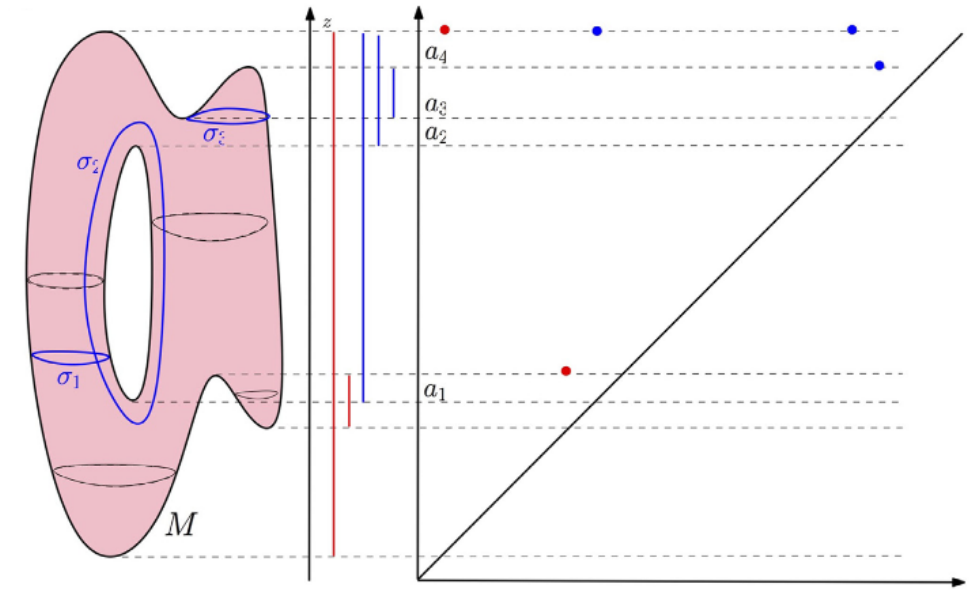


Figura 4.2: El código de barras de persistencia y el diagrama de persistencia de la función altura (proyección en el eje z) definida en una superficie en $\mathbb{R}^3$.

Ejemplo 3

En este último, consideramos la filtración dada por la unión de bolas (que crecen linealmente) centradas en el conjunto de puntos finitos C en la Figura 4.2. Notese que esta es la filtración del conjunto subnivel de la función distancia a C, y gracias al teorema del nervio, esta filtración es homotópicamente equivalente a la filtración de Čech construida sobre C. La Figura 4.2 muestra varios conjuntos subnivel de la filtración de la siguiente manera:

a) Para radio $r = 0$, la unión de bolas se reduce al conjunto de puntos finito inicial, cada uno de ellos correspondiendo a una componente conexa; se comienza un intervalo por cada una de estas componentes en $r = 0$.

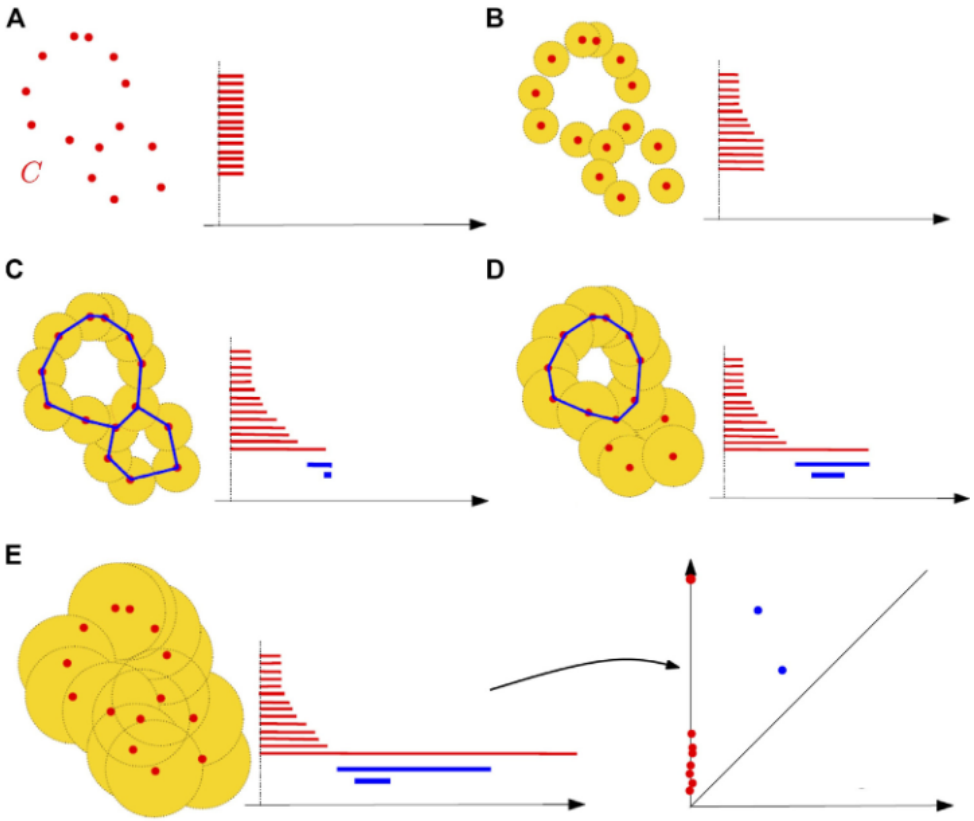
b) Algunas de las bolas comienzan a superponerse, resultando en la muerte de algunas de las componentes conexas que se han juntando entre sí; el diagrama de persistencia registra estas muertes, poniendo fin a los intervalos correspondientes.

c) Más componentes conexas se han juntado, dejando una sola componente conexa, y así, todos los intervalos asociados a características cero dimensionales terminan, con la excepción de los que corresponden a las componentes restantes; dos nuevas características uno dimensionales han aparecido, lo cual resulta en dos nuevos intervalos (en azul) que comienzan en ese valor de r.

d) Una de los dos ciclos uno dimensionales se ha rellenado, resultando en su muerte en la filtración y en el fin del intervalo correspondiente.

e) Todas la componenetes uno dimensionales han muerto, dejando un único intervalo rojo en el código de barras. como en ejemplos anteriores, el código de barras puede ser representado como un diagrama de persistencia, donde cada intervalo (a, b) es representa por un punto en $\mathbb{R}^2$ de coordenadas correspondientes.

Intuitivamente afirmamos que entre más largo sea un intervalo en el código de barras, o bien, equivalentemente, entre más alejado esté un punto de la diagonal en el diagrama correspondiente, más persistente, y por tanto, relevante, es la propiedad homológica que le corresponde a través de la filtración. Es de notar también que para un radio r dado, el k.-ésimo número de Betti de la unión de bolas en cuestión, es igual al número de intervalos de persistencia correspondiendo a características homológicas k dimensionales que contienen a r. Así, el diagrama de persistencia puede ser visto como una firma topológica que codifica la homologia de la unión de bolas abiertas, para todos los radios, así como su evolución a través de los valores que toma r.



4.3. Módulos y Diagramas de Persistencia

Los diagramas de persistencia pueden ser formalmente definidos de manera puramente algebraica. No obstante, lo anterior requiere de cuidado, nos limitaremos a dar nociones básicas dejando de lado sutilesas y dificultades técnicas. Una exposición detallada puede encontrarse en el trabajo por Chazal eta al. (2016)[38].

Sea $Filt = (F_r)_{r\in T}$ una filtración de un complejo simplicial o un espacio topológico. Dado K entero no negativo y considerando los grupos de homología $H_k(F_r)$, obtenemos una secuencia de espacios vectoriales donde las inclusiones $F_r \subset F_{r'}$, $r \leq r'$ inducen funciones lineales entre $H_k(F_r)$ y $H_k(F_{r'})$. Dicha secuencia de espacios vectoriales junto con las funciones lineales que los conectan es llamado un módulo de persistencia.

Definición 4.1. Un módulo de persistencia $\mathbb{V}$ sobre un subconjunto $T \subset \mathbb{R}$ es una familia indexada de espacios vectoriales $(V_r | r \in T)$ y una familia doblemente indexada de funciones lineales $(v_s^r : V_r \to V_s | r \leq s)$ la cual satisface la ley de composición $v_t^s \circ v_s^r = v_t^r$ donde $r \leq s \leq t$, y donde v_r^r es la identidad en V_r.

En ocasiones, es posible descomponer un módulo de persistencia en una suma directa de módulos de intervalos $\mathbb{I}_{(b,d)}$ de la forma

$$\ldots, \to 0 \to \ldots, \to 0 \to \mathbb{Z}_2 \to \ldots, \to \mathbb{Z}_2 \to 0 \to \ldots$$

donde las funciones $\mathbb{Z}_2 \to \mathbb{Z}_2$ son la identidad y las demas son la función 0. Denotando b y d respectivamente como el ínfimo y el supremo del intervalo de índices que corresponden a espacios vectoriales no cero; dicho módulo puede ser interpretado como una característica que aparece en la filtración en el índice b y desaparece en el índice d. Cuando un módulo de persistencia $\mathbb{V}$ puede ser descompuesto como una suma directa de módulos de intervalos, se puede mostrar que esta descomposición es única hasta un reordenamiento de los intervalos (ver Chazal et al., 2016[38], Teorema 2.7). Como consecuencia, el conjunto de intervalos resultantes es independiente de la descomposición de $\mathbb{V}$ y es llamado el código de barras de persistencia de V. Como en los ejemplos anteriores, cada intervalo (b, d) en el codigo de barras puede ser interpretado como un punto de coordenadas (b, d) en el plano $\mathbb{R}^2$. La unión disjunta de estos puntos, junto con la diagonal $\Delta = \{x = y\}$, es un multiconjunto llamado el diagrama de persistencia de $\mathbb{V}$.

El siguiente resultado, de (Chazal et al., 2016[38], Teorema 2.8), da algunas condiciones necesareas para que sea posible descomponer un módulo de persistencia en la suma directa de módulos de intervalos.

Teorema 4.2. Sea $\mathbb{V}$ un módulo de persistencia con índices en $T \subset \mathbb{R}$. Si T es un conjunto finito o si todos los espacios vectoriales V_r son de dimensión finita, entonces $\mathbb{V}$ se puede descomponer en una suma directa de módulos de intervalos. Más aún, para cualquier s, $t \in T$, $s \leq t$, el número β_t^s de intervalos que inician antes antes que s y finalizan despues que t es igual al rango de la función lineal v_t^s y es llamado el número de Betti (s, t)-persistente de la filtración.

Debido a que se satisfacen ambas de las condiciones anteriores para la homología persistente de filtraciones de complejos simpliciales finitos, una consecuencia inmediata de este resultado es que los diagramas de persistencia de dichas filtraciones siempre están bien definidos.

Así, es posible mostrar que los diagramas de persistencia pueden ser definidos tan pronto como la siguiente condicion se satisfaga.

Definición 4.3. Un módulo de $\mathbb{V}$ con índices en $T \subset \mathbb{R}$ es q-dócil si para todo $r < s$ en T, el rango de la función lineal $v_s^r : V_r \to V_s$ es finito.

Teorema 4.4. Chazal et al. (2009)[49], Chazal et al. (2016)[38]. Si $\mathbb{V}$ es un módulo de persistencia q-dócil, entonces tiene un diagrama de persistencia bien definido. Dicho diagrama de persistencia $\mathrm{dgm}(\mathbb{V})$ es la unión de los puntos de la dioagonal Δ de $\mathbb{R}^2$, contados con multiplicidad infinita, y un multiconjunto sobre la diagonal en $\mathbb{R}^2$ que es localmente finito. Aquí, con localmente finito nos referimos a que para cualquier rectángulo R con lados paralelos a los ejes coordenados que no intersecan a δ, el número de puntos de $\mathrm{dgm}(\mathbb{V})$, contados con multiplicidad, contenidos en R es finito. Además, a la parte del diagrama hecha de los puntos con segunda coordenada infinita es llamada la parte escencial del diagrama.

La construcción de diagramas de persitencia para módulos q-dóciles esta fuera de nuestro margen de estudio, pero da lugar a las mismas nociones que en el caso de módulos descomponibles. Puede ser creados siguiendo el acercamiento algebraico basado en las propiedades de descomposibilidad de los módulos o adoptando un acercamento por el lado de la teoría de la medida, la cual nos permite definir diagramas como medidas con valores enteros en un espacio de rectángulos en el plano. Vease Chazal et al. (2016)[38] para más información.

Aunque los módulos de persistencia encontrados en la práctica son descomponibles, la estructura general de los módulos de persistencia q-dóciles juega un papel fundamental en el análisis matemático y estadístico de la homología persistente. En especial al momento de asegurar la existencia de los diagramas límite cuando se estudian propiedades de convergencia (Ver Capítulo 5).

Una filtración $\mathrm{Filt} = (F_r)_{r \in T}$ de un complejo simplicial o un espacio topológico es llamado dócil si para cualquier entero k, el módulo de persistencia $(H_k(F_r) \,|\, r \in T)$ es q-dócil. Es de notar que las filtraciones de complejos simpliciales finitos siempre son dociles. Como consecuencia de lo anterior, para cualquier entero k, el diagrama de persistencia denotado $\mathrm{dgm}_k(\mathrm{Filt})$ es asociado con la filtración Filt. Cuando k no se especifica de manera explícita y no hay ambigüedadm es común dejar fuera de la notación el índice k y hablar de "el" diagrama de persistencia $\mathrm{dgm}(\mathrm{Filt})$ de la filtración Filt. Esta notación se entiende como "$\mathrm{dgm}_k(\mathrm{Filt})$ para alguna k."

4.4. Paisajes de persistencia

Los paisajes de persistencia introducidos en Bubenik (2015) [17] es una representación alternativa de los diagrama de persistencia. Este acercamiento se enfoca en representa la información topológica codificada en los diagramas de persitencia como elementos de un espacio de Hilbert, en el cual métodos de aprendizaje estadístico pueden ser directamente aplicados. Los paisajes de persistencia son una colección de funciones continuas y lineales por pedazos $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ las cuales resumen un diagrama de persistencia dgm.

Un par $p = (b, d) \in \mathrm{dgm}$ es transformado en el punto $\left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ (Ver Figura 4.3). En esta definición descartamos los puntos con persistencia infinita. El paisaje se define entonces considerando el conjunto de funciones creadas cubriendo las características del diagrama de persistencia rotado como sigue:

$$\Lambda_p(t) = \begin{cases} t-b & t \in \left[b, \frac{b+d}{2}\right] \\ d-t & t \in \left(\frac{b+d}{2}, d\right] \\ 0 & \text{en otro caso.} \end{cases}$$

El paisaje de persistencia λ_{dgm} de dgm es un resumen del arreglo de curvas lineales por pedazos que se obtiene al sobreponer las gráficas de las funciones $\{\Lambda_p\}_{p \in \text{dgm}}$. Formalmente, el paisaje de persistencia de dgm es una colección de funciones

$$\lambda_{\text{dgm}}(k,t) = \underset{r \in \text{dgm}}{\text{kmax}} \Lambda_r(t), \ t \in [0,T], k \in \mathbb{N},$$

donde kmax es el k-ésimo valor más grande del conjunto. Dado $k \in \mathbb{N}$, la función $\lambda_{dgm}(k,.) : \mathbb{R} \to \mathbb{R}$ es llamado el k-ésimo paisaje de persistencia de dgm. Más aún, una función que asocia cada diagrama de persistencia con su paisaje correspondiente es inyectiva. En otras palabras, no se pierde inforamción cuando se utiliza un paisaje de persistencia para representar un diagrama.

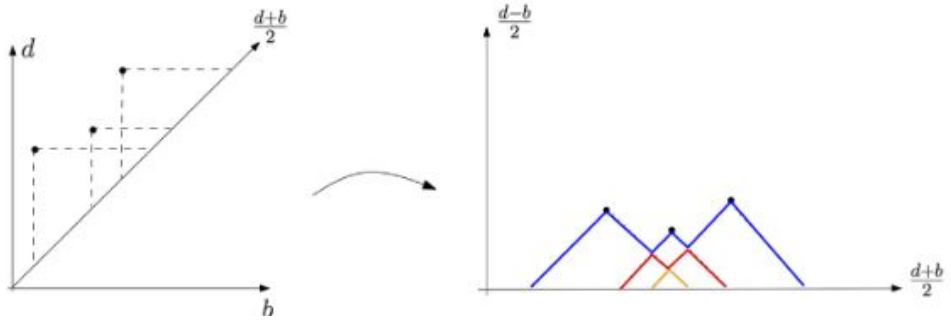


Figura 4.3: Ejemplo de un paisaje de persitencia (derecha) asociado con un diagrama de persistencia (izquierda).

La ventajas de la representación por paisajes de persistencia son. Primero, los diagramas de persistencia se convierten en elementos de un espacio funcional, permitiendo el uso de una variedad más amplia de herramientas de la estadística y el análisis de datos para el procesamiento de características topológicas, vease Bubenik (2015)[17], Chazal et al. (2015)[42]. Segundo, y fundamental para la perspectiva teórica, los paisajes de persistencia comparten las mismas propiedades de estabilidad que los diagramas de persistencia (Ver Sección 4.7).

4.5. Representaciones Lineales de la Persistencia Homológica

Un diagrama de persistencia sin su parte escencial puede representarse como una medida discreta $\delta^+ = \{p = (b,d), b < d < \infty\}$. Con un ligero abuso de la notación, podemos escribir lo siguiente:

$$\text{dgm} = \sum_{p \in \text{dgm}} \delta_p,$$

donde las características son contadas con multiplicidad y $\delta_{(b,d)}$ denota la medida de Dirac en $p = (b, d)$. La mayoria de los descriptores propuestos para analizar la persistencia pueden ser expresados como transformaciones lineales de un diagrama de persistencia, vistos como:

$$\Psi(\mathrm{dgm}) = \sum_{p \in \mathrm{dgm}} f(p),$$

para alguna función f definida en Δ que toma valores en un espacio de Banach.

En la mayoria de los casos, queremos que estas transformaciones se apliquen de manera independiente a cada dimensión homológica. Para $k \in \mathbb{N}$ alguna dimensión homológica dada, consideramos alguna transformación lineal del diagrama de persistencia restringida a las características topológicas de la dimensión k como sigue:

$$\Psi_k(\mathrm{dgm}_k) = \sum_{p \in \mathrm{dgm}_k} f_k(p), \tag{4.1}$$

donde dgm_k es el diagrama de persistencia de las características topológicas de la dimensión k donde f_k está definido en Δ y toma valores en un espacio de Banach.

Curvas de Betti

La manera más sencilla de representar la persistencia homológica es la funciń de Betti o curva de Betti. La curva de Betti de una dimensión homológica k esta definida como:

$$\beta_k(t) = \sum_{(b,d) \in \mathrm{dgm}} w(b,d)\, \mathbb{1}_{t \in [b,d]}$$

donde w es una función peso definida en Δ. En otras palabras, la curva de Betti es el número de cadigos de barra en el tiempo m. Este descriptor es una reapresentación linear de la homología persistente, tomando f como en (4.1) de forma que $f(b,d)(t) = w(b,d)\, \mathbb{1}_{t \in [bd]}$. Una elección típica para la función peso es una función creciente de la persistencia $w(b,d) = \tilde{w}(d-b)$ donde $\tilde{w}$ es una función creciente definida en $\mathbb{R}^+$. Una de las primeras aplicaciones de las curvas de Betti se puede encontrar en Umeda (2017)[124].

Superficies de Persistencia

Una superficie de persistencia (también llamadas imágenes de persistencia) se obtiene mediante la convolución de un diagrama de persistencia con un kernel. Introducidas en Adams et al. (2017)[3]. Sea $K : \mathbb{R}^2 \to \mathbb{R}$, un kernel, y H una matriz 2×2 simétrica y definida positiva, y dada $u \in \mathbb{R}^2$ definimos

$$K_H(u) = \det(H)^{-\frac{1}{2}} K\left(H^{-\frac{1}{2}} u\right).$$

Sea $w : \mathbb{R}^2 \to \mathbb{R}_+$ una función peso definida en Δ. Se define la superficie de persistencia homológica de dimensión k asociada con el diagrama dgm, con kernel K y matriz H como sigue:

$$\forall u \in \mathbb{R}^2, \rho_k(\mathrm{dgm})(u) = \sum_{p \in \mathrm{dgm}_k} w(r)\, K_H(u-p)$$

La superficie de persistencia es entonces una represetación lineal de la persistencia homológica. Típicamente las funciones peso son funciones crecientes de la persistencia.

4.6. Métricas en el Espacio de Diagramas de Persistencia

Para aprovechar la información y características topológicas inferidas por la homologiía persistente, necesitamos alguna manera de comparar diagramas de persistencia, esto es, otorgar al espacio de los diagramas de persistencia una estructura de espacio métrico. Aunque se han considerado una variedad de métricas, la más fundamental de ellas se le conoce como la distancia de cuello de botella.

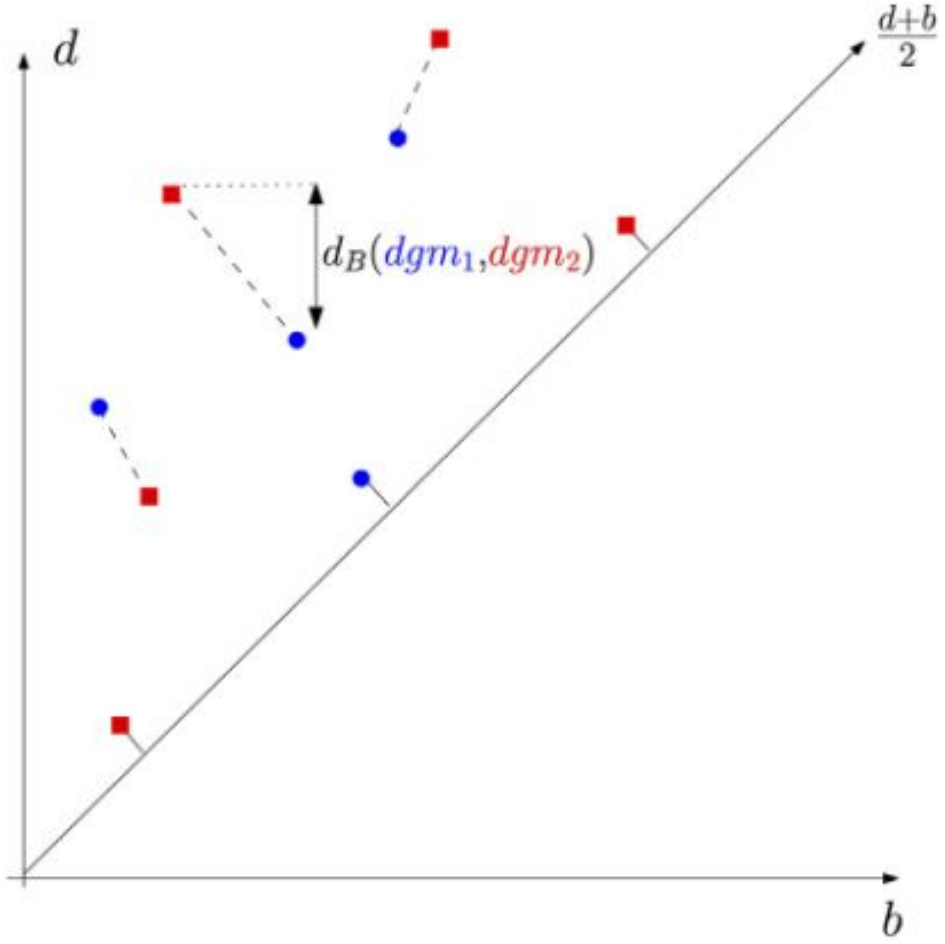


Figura 4.4: Emparejamiento perfecto y la distancia de cuello de botella entre un diagrama en rojo y un diagrama en azul. Nótese que algunos puntos en ambos diagramas son emparejados a puntos en la diagonal.

Recordemos que un diagrama de persistencia es la unión de un multiconjunto discreto en la parte del plano por encima de la diagonal Δ y, por razones técnicas que veremos adelante, Δ, donde el punto de Δ se cuenta con multiplicidad infinita. Un emparejamiento (ver Figura 4.4) entre dos diagramas, dgm_1 y dgm_2, es un subconjunto $m \subseteq \mathrm{dgm}_1 \times \mathrm{dgm}_2$ tal que cada punto en $\mathrm{dgm}_1 - \Delta$ y $\mathrm{dgm}_2 - \Delta$ aparece exactamente una vez en m. En otras palabras, para cada $p \in \mathrm{dgm}_1 - \Delta$ y para cualquier $q \in \mathrm{dgm}_2 - \Delta$, $(\{p\} \times \mathrm{dgm}_2) \cap m$ y $(\mathrm{dgm}_1 \times \{q\}) \cap m$ contiene cada uno un único par. La distancia de cuello de botella entre

dgm_1 y dgm_2 esta definida como

$$\mathrm{d_b}\left(\mathrm{dgm}_1, \mathrm{dgm}_2\right) = \inf_{\text{empar. } m} \max_{(p,q)\in m} \|p-q\|_\infty .$$

El cálculo práctico de la distancia de cuella de botella se resume en encontrar un cierto emparejamiento perfecto en gráficas bipartitas, y para esto podemos utilizar algoritmos clásicos.

La distancia de cuello de botella es una métrica parecida a una métrica L_∞. Resulta ser una métrica natural para expresar las propiedades de estabilidad de los diagramas de persistencia presentados en la sección 4.7, pero sufre de las mismas desventajas que las métricas usuales en L_∞, esta completamente determinada por la mayor distancia entre los pares y no toma en cuenta la cercania de los pares de puntos restantes. Una variante para superar este problema, es la distancia de Wasserstein entre diagramas. Dada $p \geq 1$, se define como

$$\mathrm{W}_p\left(\mathrm{dgm}_1, \mathrm{dgm}_2\right)^p = \inf_{\text{empar. } m} \sum_{(p,q)\in m} \|p-q\|_\infty^p$$

Se han presentado resultados de utilidad acerca de la persistencia en la métrica W_p, particularmente en el estudio por Cohen-Steiner et al. (2010)[54], pero dependen de supuestos que los vuelve consecuencias de los resultados de estabilidad en la distancia de cuello de botella. Un estudia general del espacio de diagramas de persistencia dotado con métricas W_p se ha considerada en Divol y Lacombe (2020)[63], donde so propone un marco general basado en transportes parciales óptimos, donde muchas propiedades importantes de los diagramas de persistencia pueden probarse de manera natural.

4.7. Propiedades de Estabilidad en los Diagramas de Persistencia

Una propiedad fundamental de la homología persistente es que los diagramas de persistencia de filtraciones construidas sobre conjuntos de datos resultan ser muy estables con respecto a ciertas perturbaciones de los datos. Para formalizar y cuantificar dichas propiedades, primero necesitamos ser precisos con respecto a que perturbaciones son permitidas.

En lugar de trabajar directamente con filtraciones sobre conjuntos de datos, resulta ser más conveniente definir una noción de proximidad entre módulos de persistencia, de donde derivaremos un resultado de estabilidad general para la homología persistente. Entonces, la mayoria de los resultados de estabilidad para filtraciones especificas serán consecuencia de este teorema general. Para evitar discuciones técnicas y sin pérdida de generalidad, suponemos que los módulos de persistencia considerados están indexados por $\mathbb{R}$.

Definición 4.5. Sean $\mathbb{B}$, $\mathbb{W}$ dos módulos de persistencia indexados por $\mathbb{R}$. Dado $\delta \in \mathbb{R}$, un homomorfismo de grado δ entre $\mathbb{V}$ y $\mathbb{W}$ es una colección Φ de funciones lineales $\phi_r : V_r \to W_{r+\delta}$, para todo $r \in \mathbb{R}$ tal que para cualquier $r \leq s$, $\phi_r \circ v_s^r = w_{s+\delta}^{r+\delta} \circ \phi_r$.

Un ejemplo importante de un homomorfismo de grado δ es el endomorfismo de cambio $1_{\mathbb{V}}^{\delta}$ el cual consiste de las familias de funciones lineales $\left(v_{r+\delta}^r\right)$. Nótese que los homomorfismos de módulos pueden ser naturalmente compuestos; la compisición de un homomorfismo Ψ de grado δ entre $\mathbb{U}$ y $\mathbb{V}$ y un homomorfismo Φ de grado δ' entre $\mathbb{V}$ y $\mathbb{W}$ naturalmente da lugar a un homomorfismo $\Phi\Psi$ de grado $\delta + \delta'$ entre $\mathbb{U}$ y $\mathbb{W}$.

Definición 4.6. Sea $\delta \geq 0$. Dos módulos de persistencia $\mathbb{V}$, $\mathbb{W}$ son *delta*-intercalados si existen dos homomorfismos de grado δ, Φ, de $\mathbb{V}$ a $\mathbb{W}$ y Ψ de $\mathbb{W}$ a $\mathbb{V}$ tales que $\Psi\Phi = 1_{\mathbb{V}}^{2\delta}$ y $\Phi\Psi = 1_{\mathbb{W}}^{2\delta}$.

Aunque no define un amétrica en el espacio de los módulos de persistencia, la noción de cercania entre dos módulos de persistencia puede ser definida como el más pequeño δ no negativo tal que los módulos sean δ-intercalados. Más aún, nos ayuda a formalizar el siguiente teorema fundamental (Chazal et al., 2009[49]; Chazal et al., 2016[38]).

Teorema 4.7 (Estabilidad de la persistencia)**.** Sean $\mathbb{V}$ y $\mathbb{W}$ dos módulos de persistencia q-dóciles. Si $\mathbb{V}$ y $\mathbb{W}$ son δ-intercalados para algún $\delta > 0$, entonces

$$\mathrm{d_b}\left(\mathrm{dgm}\left(\mathbb{V}\right), \mathrm{dgm}\left(\mathbb{W}\right)\right) \leq \delta.$$

Con este resultado obtenemos una herramienta eficiente para establecer resultados de estabilidad concretos en el ATD. Por ejemplo, podemos recuperar el primer resultado de estabilidad de la persistencia que aparece en la literatura (Cohen-Steiner et al., 2005)[53].

Teorema 4.8. Sean $f, g : M \to \mathbb{R}$ dos funciones de valores reales definidas en un espacio topológico M que son q-dóciles, esto es, que sus filtraciones de los conjuntos subnivel de f y g inducen módulos q-dóciles en el nivel de la homología. Entonces, para cualquier entero k,

$$\mathrm{d_b}\left(\mathrm{dgm}_k\left(f\right), \mathrm{dgm}_k\left(g\right)\right) \leq \left\|f - g\right\|_\infty = \sup_{x \in M} \left|f\left(x\right) - f\left(x\right)\right|$$

donde $\mathrm{dgm}_k\left(f\right)$ es el diagrama de persistencia de el módulo de persistencia $\left(H_k\right)\left(f^{-1}\left(-\infty, r\right)\right) | r \in \mathbb{R}$, y respectivamente para $\mathrm{dgm}_k\left(f\right)$, donde las funciones lineales son las inducidas por las inclusiones canónicas entre los conjuntos subnivel.

Prueba: Denotamos $\delta = \left\|f - g_\infty\right\|$, tenemos que para cualquier $r \in \mathbb{R}$, $f^{-1}\left(-\infty, r\right) \subseteq g^{-1}\left(-\infty, r + \delta\right)$ y $g^{-1}\left(-\infty, r\right) \subseteq f^{-1}\left(-\infty, r + \delta\right)$. este intercalado entre los conjuntos subnivel de f induce un δ-intercalamiento entre los módulos de persistencia en el nivel de la homología, y así, el resultado se sigue de una aplicación directa del teorema anterior. □

El *Teorema 4.7* también implica el siguiente resultado de estabilidad para los diagramas de persistencia de filtraciones construidas sobre conjuntos de datos.

Teorema 4.9. Sean $\mathbb{X}$ y $\mathbb{Y}$ dos espacios métricos compactos y sean $\mathrm{Filt}\left(\mathbb{X}\right)$ y $\mathrm{Filt}\left(\mathbb{Y}\right)$ las filtraciones de Vietoris-Rips o de Čech contruidos sobre $\mathbb{X}$ y $\mathbb{Y}$. Entonces

$$\mathrm{d_b}\left(\mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{X}\right)\right), \mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{Y}\right)\right)\right) \leq 2\mathrm{d_{GH}}\left(\mathbb{X}, \mathbb{Y}\right)$$

donde $\mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{X}\right)\right)$ y $\mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{Y}\right)\right)$ denotan los diagramas de persistencia de las filtraciones $\mathrm{Filt}\left(\mathbb{X}\right)$ y $\mathrm{Filt}\left(\mathbb{Y}\right)$.

Como notamos en el Ejemplo 3 de la Sección 4.2, los diagramas de persistencia pueden ser interpretados como característicos topológicos multiescala de $\mathbb{X}$ y $\mathbb{Y}$. Además, el Teorema 4.9 nos dice que estos característicos son resilientes con respecto a perturbaciones de los datos en la métrica de Gromov-Hasudorff. Así, pueden ser usados como propiedades discriminatorias para tareas de clasificación entre otras (vease, por ejemplo, Chazal et al. (2009)[32] para un aplicación a la clasificación no rígida de figuras 3D).

Ahora damos resultados similares para la representaciones alternativas de la homología persistente que vimos con anterioridad. De la definición de paisajes de persistencia, obsevamos que $\lambda\left(k, \cdot\right)$ es 1-Lipschitz, y esto es suficiente para que propiedades de estabilidad similares a las de los diagramas de persistencia se satisfagan para los paisajes.

Proposición 4.10 (Estabilidad de los paisajes de persistencia; Bubenik (2015)[17])**.** Sean dgm y dgm$'$ dos diagramas de persistencia (sin sus partes escenciales). Para cualquier $t \in \mathbb{R}$ y cualquier $k \in \mathbb{N}$, tenemos lo siguiente:

(I) $\lambda(k,t) \geq \lambda(k+1,t) \geq 0$.

(II) $|\lambda(k,t) - \lambda'(k,t)| \geq \mathrm{d_b}\left(\mathrm{dgm}, \mathrm{dgm}'\right)$.

Una amplia gama de representaciones lineales es continua con respecto a la métrica de Wasserstein W_s en el espacio de los diagramas de persistencia y con respecto a la norma de Banach para las representaciones lineales de la persistencia. En general, no suempre es posible dar una cota superior para el módulo de continuidad del operador de representaciones lineales. Sin embargo, en el caso donde $s = 1$, es posible probar además un resultado de estabilidad si la función de peso toma valores pequeños para puntos cercanos a la diagonal. (vease Divol y Lacombe (2020)[63], Hofer et al. (2019)[75]).

La Estabilidad y la Capacidad Discriminativa de las Representaciones de Persistencia

Los resultados en el estudio por Divol y Lacombe (2020)[63] muestran que la continuidad y estabilidad solo son posible con funciones de pesos que toman valores pequeños para puntos cercanos a la diagonal. No obstante, en general, no hay razón específica para considerar que los puntos cercanos a la diagonal son menos importantes, dada alguna tarea de aprendizaje. Desde una perspectiva del aprendizaje automático, también es relevante diseñar representaciones lineales con funciones de peso generales, aunque seria más difícil probar la consistencia de los métodos correspondientes sin al menos la continuidad de la representación. Si bien la estabilidad estabilidad es importante, es quizás un requisito muy fuerte para muchos de los problemas en la ciencia de datos. Posiblemente, diseñar representaciones lineales que sean sensibles a partes específicas de los diagramas de persistencia, sea una mejor estrategia en la práctica que buscar la estabilidad global de la representación.

Capítulo 5

Aspectos Estadísticos de la Homología Persistente

cap 5

Por sí sola, la homología persistente no toma en cuenta la naturaleza estocástica de los datos y la variabilidad intrinseca de las cantidades topológicas que infieren. Buscamos ahora un acercamiento estadístico a la homología persistente, considerando que los datos son generados de alguna distribución desconocida. Comenzamos dando varos resultados de consistencia para la inferencia de la homología persistente.

5.1. Resultados de Consistencia para la Homología Persistente

Supóngase que observamos n puntos $(X_1, \ldots, X_n)$ en un espacio métrico (M, ρ) obtenidas i. i. d. de una medida de probabilidad desconocida μ con soporte compacto $\mathbb{X}_\mu$ la distancia de Gromov-Hausdorff nos permite comparar $\mathbb{X}_\mu$ con otros espacios métricos compactos no necesariamente encajados en M. Definimos a continuación, $\hat{\mathbb{X}}$ un estimador de $\mathbb{X}_\mu$ como una función de $X_1, \ldots, X_n$ que toma valores en el conjunto de espacios métricos compactos.

Sean $\text{Filt}(\mathbb{X}_\mu)$ y $\text{Filt}\left(\hat{\mathbb{X}}\right)$ dos filtraciones definidas en $\mathbb{X}_\mu$ y $\hat{\mathbb{X}}$. Con el teorema 4.9 hemos visto que una estrategia natural para estimar la homologiía persistente de $\text{Filt}(\mathbb{X}_\mu)$ consiste en estimar el soporte de $\hat{\mathbb{X}}$. Nótese que en algunos casos, el espacio M puede ser desconocido y las observaciones $X_1, \ldots, X_n$ solo se conocen mediante de sus distancias por pares $\rho(X_i, X_j)$, $i, j = 1, \ldots, n$. La distancia de Gromov-Hausforff nos permite considerar el conjunto de observaciones como un espacio métrico abstracto de cardinalidad n, independietemente de la manera en la que esta encajado en M. Esta estructura general incluye el acercamiento más estándar que consiste en estamar el soporte con respecto a la distancia de Hausdorff restringiendo los valores de $\hat{\mathbb{X}}$ a los conjuntos compactos en M.

El conjunto finito $\mathbb{X}_n \coloneqq \{X_1, \ldots, X_2\}$ es un estimador natural para el soporte de $\mathbb{X}_\mu$. En muchos de los contextos que veremos a continuación, $\mathbb{X}_\mu$ muestra tasas de convergencia óptimas con respecto a la ditancia de Hausdorff. Para algunas constantes $a, b > 0$, decimos que μ satisface el supuesto (a, b)-estándar si para cualquier $x \in \mathbb{X}_\mu$ y cualquer $r > 0$,

$$\mu(B(x, r)) \geq \min\left(ar^b, 1\right). \tag{5.1}$$

Este supuesto es ampliamente usado en la literatura de la estimación de conjuntos bajo la distancia de Hausdorff (Cuevas y Rodriguez-Casal, 2004[55]; Singh et al., 2009[117]). Bajo este supuesto, puede deducirse que la tasa de convergencia de $\mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{X}_n\right)\right)$ a $\mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{X}_\mu\right)\right)$ para la métrica de cuello de botella es acotada superiormente por $O\left(\frac{\log n}{n}\right)^{1/b}$. Más precisamente, esta tasa acota superiormente la tasa de convergencia minimax sobre el conjunta de medidas de probabilidad en el espacio métrico (M, ρ) satisfaciendo el supuesto (a, b)-estándar en M.

Teorema 5.1. Chazal et al. (2014)[42] para algunas constantes positivas, a y b, sea

$$\mathcal{P} := \left\{\mu \text{ en } M | \mathbb{X}_\mu \text{ es compacto y } \forall x \in \mathbb{X}_\mu, \forall r > 0,\ \mu\left(B\left(x, r\right)\right) \geq \min\left(1, ar^b\right)\right\}$$

Entonces, se tiene que

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d_b}\left(\mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{X}_\mu\right)\right), \mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{X}_n\right)\right)\right)\right] \leq C\left(\frac{\log n}{n}\right)^{1/b}$$

Donde la constante C depende solo de a y b.

Bajo algunos supuestos técnicos adicionales, se pueden evidenciar las cotas inferiores correspondientes (hasta un término logarítmico) (vease Chazal et al. (2014)[42]). Utilizando resultados de estabilidad, se pueden obtener resultados de consistencia similares bajo modelos generativos alternativos siempre que se conosca un estimador del soporte consistente bajo la métrica de Hausdorff. Por ejemplo, de los resultados del estudio por Genovese et al. (2012)[69] sobre la estimación del soporte Hausforff bajo ruido aditivo, se puede deducir que las tasas de convergencia minimax para la estimación de diagramas de persistecia son más rápidas que $(\log n)^{-1/2}$. Más aún, siempre que se disponga de un resultado de estabilidad para alguna representación de la persistencia dada, resultados de consistencia similares pueden ser directamente derivados de la consistencia de diagramas de persistencia.

Estimación de la Homología Persistente de Funciones

El Teorema 4.7 abre la puerta a la estimación de la homología persistente en funciones definidas en $\mathbb{R}^d$, en una subvariedad de $\mathbb{R}^d$ o, más generalmente, en un espacio métrico. La homología persistente de funciones de regresión a sido estudiada por Bubenik et al. (2010)[126]. El acercamiento alternativo de Bobrowski et al. (2014)[8], basado en la inclusión entre pares anidados de los conjuntos de nivel estimados, puede aplicarse con estimadores del kernel de regresión y de la densidad del kernel para estimar la homología persistente de funciones de densidad y regresión. Otra rama de investigación en este tema trata con varias versiones robustas de ATD. Una solución es estudiar la homología persistente de de los conjuntos super-nivel de estimadores de densidad (Fasy et al., 2014[66]). Otra alternativa, más estrechamente relacionada a la función distancia, pero robusta al ruido, consiste en estudiar la homología persistente de conjuntos subnivel de la distancia a una medida definida en la sección 3.3 (Chazal et al., 2017[30]).

5.2. Estadísticos de la Homología Persistente Calculados en una Nube de Puntos

Para muchas aplicaciones, en especial donde el soporte de la nube de puntos no esta dibujado sobre o es cercano a una figura geométrica, los diagramas de persistencia pueden ser difíciles de analizar. En particular, muchos característicos topológicos estan cerca de la diagonal. Como corresponden a estructuras topológicas que viven por un periodo muy corto de tiempo, estos puntos son generalmente considerados ruido (ver Figura 5.1). Las regiones de confianza para los diagramas de persistencia nos otorgan una respuesta de rigor al problema de distingir entre la señal y el ruido en estas representaciones.

Los resultados de estabilidad dados en la sección 4.7 motivan el uso de la distancia de cuello de botella para definir regiones de confianza. Sin embargo, distancias alternativas inspiradas en distancias de Wasserstein también pueden ser propuestas. Cuando se estima un diagrama de persistecia dgm con un estimador $\widehat{\text{dgm}}$, buscamos un valor η_α tal que

$$P\left(\mathrm{d_b}\left(\widehat{\text{dgm}}, \text{dgm}\right) \geq \eta_\alpha\right) \leq \alpha,$$

para $\alpha \in (0,1)$. Sea B_α la bola cerrada de radio α para la distancia de cuello de botella, centrada en $\widehat{\text{dgm}}$ en el espacio de los diagramas de persistencia. Siguiendo a Fasy et al. (2014)[66], podemos visualizar los puntos que pertenecen a esta bola de varias maneras. Una primera opción es centrar una caja con lados de largo 2α en cada punto del diagrama $\widehat{\text{dgm}}$. Una solución alternativa es visualizar el conjunto de confianza añadiendo una banda a una distancia $\eta_\alpha/2$ de la diagonal (siendo la distancia de cuello de botella definida para la norma ℓ_∞) (ver Figura 5.1 para una ilustración). Luego los puntos fuera de la banda son considerados, cualidades topológicas significativas (véase el estudio por Fasy et al. (2014)[66] para más detalles).

Se han propuesto una variedad de métodos para estimar η_α en el estudio por Fasy et al. (2014)[66]. Estos métodos recaen principalmente en los resultados de estabilidad para diagramas de persistencia; se pueden derivar conjuntos de confianza para diagramas de los conjuntos de confianza en el espacio muestral.

Acercamiento por submuestreo

Este método se basa en una región de confianza para el soporte K de la distribución de la muestra en la distancia de Hausdorff. Sea $\tilde{\mathbb{X}}_b$ un submuestreo de tamaño b de una muestra $\tilde{\mathbb{X}}_n$, donde $b = o(n/\log n)$. Sea $\mathrm{q}_b(1-\alpha)$ un cuantil de la distribución de $\mathrm{Haus}\left(\tilde{\mathbb{X}}_b, \mathbb{X}_n\right)$. Sea $\hat{\eta}_\alpha := 2\hat{q}_b(1-\alpha)$, donde $\hat{q}_b$ es una estimación $\mathrm{q}_b(1-\alpha)$ usando un procedimiento de Monte Carlo estándar. Bajo un supuesto (a,b) estándar y para una n suficientemente grande, el estudio por Fasy et al. (2014)[66] muestra que

$$P\left(\mathrm{d}_b\left(\mathrm{dgm}\left(\mathrm{Filt}\left(K\right)\right), \mathrm{dgm}\left(\mathrm{Filt}\left(\mathbb{X}_n\right)\right)\right) > \hat{\eta}_\alpha\right) \leq P\left(\mathrm{Haus}\left(K, \mathbb{X}_n\right) > \hat{\eta}_\alpha\right) \leq \alpha + O\left(\frac{b}{n}\right)^{\frac{1}{4}}.$$

Bootstrap de Cuello de Botella

Los resultados de estabilidad suelen llevar a conjuntos de confianza conservativos. Una estrategia alternativa es el bootstrap de ceullo de botella introducido en el estudio por Chazal et al. (2016)[38]. Consideramos el escenario general donde un diagrama de persistencia

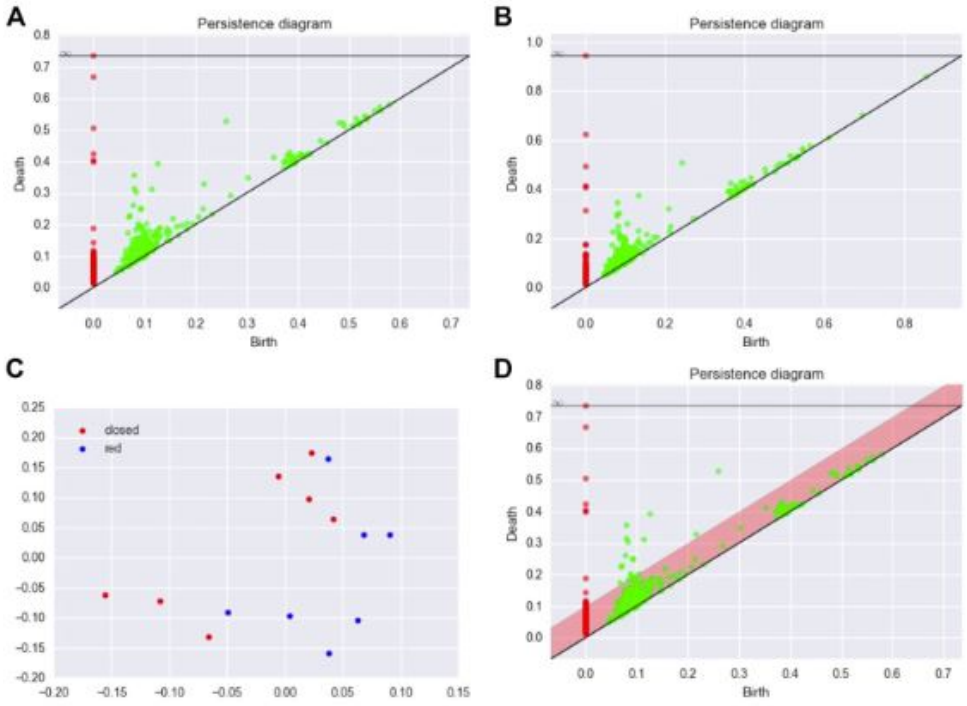


Figura 5.1: (A, B) Dos diagramas de persistencia para dos configuraciones de la MBP (Maltose-Binding Protein). (C) Configuración MDS (Multi-Dimensional Scaling) para la matriz de distancias de cuello de botella. (D) Diagrama de persistencia con región de confianza para la MBP.

$\hat{\text{dgm}}$ se define como la observación $(X_1, \ldots, X_n)$ en un espacio métrico. Este diagrama de persistencia corresponde a la estimación de un diagrama de persistencia subyacente dgm, que puede ser relacionado, por ejemplo, al soporte de la medida, o al los conjutos subnivel de la función relacionada a esta distribución (Por ejemplo, una función de densidad donde las X_i's estan en $\mathbb{R}^d$). Sea $(X_1^*, \ldots, X_n^*)$ una muestra de una medida empírica definida de las observaciones $(X_1, \ldots, X_n)$. Y sea $\widehat{\text{dgm}}^*$ el diagrama de persistencia derivado de esta muestra. Podemos tomar entonces para η_α la cantidad $\hat{\eta}_\alpha$ definida por:

$$P\left(\mathrm{d_b}\left(\widehat{\text{dgm}}^*, \widehat{\text{dgm}}\right) > \widehat{\eta}_\alpha | X_1, \ldots, X_n\right) = \alpha. \tag{5.2}$$

Es de notar que $\hat{\eta}_\alpha$ puede ser estimada con facilidad utilizando métodos de Monte Carlo. Se ha demostrado en el estudio por Chazal et al. (2016)[48] que el bootstrap de cuello de botella es una opción válida para calcular los conjuntos subnivel de un estimador de densidad.

Bootstrap para Números de Betti Persistentes

Como se ha mencionado, las regiones de confianza basadas en las propiedades de estabilidad de la persistencia suelen dar lugar a regiones de confianza muy conservadoras.

Basados en los conceptos de estadísticos estabilizantes, Penrose y Yukick (2001)[100], recientemente si ha mostrado normalidad asintótica para números de Betti persistentes en Krebs y Polonik (2019)[82], y en Roycraft et al. (2020)[113] bajo muy pocas condiciones en la filtración y distribución de la nube muestral. Además los procedimientos de bootstrapping también son válidos en este acercamiento. Más precisamente, un procedimiento bootstrap suavizado junto a un conveniente reescalado de la nube de puntos, parece ser un acercamiento prometedor para el bootstrapping de características ATD de la nube de puntos de datos.

5.3. Estadísticos para una Familia de Diagramas de Persistencia y Otras Representaciones

Hasta ahora hemos considerado estadísticos basados solamente un diagrama de persistencia. Dirigimos nuestra atención ahora a un nuevo esquema donde se encuentran a disposición una variedad de diagramas de persistencia (y otras representaciones), y estamos interesados en proveer una tendencia central, regiones de confianza y pruebas de hipótesis para descriptores topológicos construidos en esta familia.

5.3.1. Tendencia Central para la Homología Persistente

Media de Distribuciones de Diagramas

Dado que el espacio de diagramas de persistencia es un espacio métrico general pero no un espacio de Hilbert, la definición de media en diagramas de persistencia no es obvia ni única. Un primer acercamiento natural para definir una medida de tendencia central en este contexto es considerar la media de Fréchet de distribuciones de diagramas. Su existencia ha sido demostrada en el estudio por Mileyko et al. (2011)[93], y también han sido caracterizadas en el estudio por Turner er al. (2014)[122]. Sin embargo, pueden no ser únicas, y resultan complicadas de calcular en la práctica. Algunos acercamientos para tratar de solucionar estas problematicas que se han propuesto recientemente incluyen la propuesta basada en transporte óptimo númerico Lacombe et al. (2018)[85] así como las representaciones lineales y métodos basados en el kernel por Divol y Chazal (2020)[39].

Firmas Topológicas de Submuestras

También podemos usar propiedades de tendencia central de la homología persistente para calcular firmas topológicas para conjuntos de datos de gran tamaño, como alternativa al prohibitivo costo de los cálculos de persistencia. Dada una gran nube de puntos, la idea es extraer muchas submuestras, calcular el paisaje de persistencia de cada submuestra, y luego combinar la información.

Para cualquier entero positivo m, sea $X = \{X_1, \ldots, x_m\}$ un muestra de m puntos tomada de una medida μ en un espacio métrico M cuyo soporte es denotado por $\mathbb{X}_\mu$. Suponemos que el diametro de $\mathbb{X}_\mu$ es finito y acotado superiormente por $\frac{T}{2}$, donde T es la misma constante que en la definición de paisajes de persistencia en la Sección 4.4. Concentramos nuestra atención ahora en el caso de $k = 1$ y el conjunto $\lambda(t) = \lambda(1, t)$. Sin embargo, los resultados que presentaremos a continuación se sostienen para $k > 1$. El paisaje de persistencia correspondiente (asociado con el diagrama de persistencia de la filtración de Čech o Vietoris-Rips) es λ_X y denotamos por ψ_μ^m la medida inducida por $\mu^{\otimes m}$ en el espacio de paisajes de persistencia. Nótese que el paisaje de persistencia λ_X puede ser visto como

una única muestra de la medida λ_μ^m. La esperanza puntual del paisaje de persistencia (aleatorio) bajo esta medida se define por $\mathbb{E}_{psi_\mu^m}[\lambda_X(t)], t \in [0,T]$. El paisaje promedio $\mathbb{E}_{psi_\mu^m}[\lambda_X]$ tiene una contraparte empírica natural, que puede ser usada como un estimador insesgado. Sea $S_1^m, \ldots, S_\ell^m$, ℓ muestras independientes de tamaño m de $\lambda^{\otimes m}$. Definimos el paisaje empírico promedio como

$$\overline{\lambda_\ell^m}(t) = \frac{1}{b}\sum_{i=1}^{b} \lambda_{S_i^m}(t), \text{ para todo } t \in [0,T], \tag{5.3}$$

y proponemos usar $\overline{\lambda_\ell^m}$ para estimar $\lambda_{\mathbb{X}_\mu}$. Nótese que calcular la homología persistente de $\mathbb{X}_n$ es $O(\exp(n))$, mientras que calcular la del paisaje promedio es $O(b\exp(n))$.

Otra motivación para este acercamiento por submuestreos es que tambíen puede ser aplicado cuando μ es una medida discreta con el soporte $\mathbb{X}_N = \{x_1, \ldots, x_N\}$ en el espacio métrico M. Lo anterior es muy común en la práctica, cuando una medida continua (pero desconocida) es aproximada par una medida uniforme discreta μ_N en $\mathbb{X}_N$.

El paisaje promedio $\mathbb{E}_{\psi\mu^m}[\lambda_X]$ es interesante por si mismo, ya que trae contiene cierta información topológica estable acerca de la medida subyacente μ, de la cual se generan los datos.

Teorema 5.2 (Chazal et al. (2015a)[40])**.** Sea $X \sim \mu^{\otimes m}$ y $Y \sim \nu^{\otimes m}$, donde μ y ν son dos medidas de probabilidad en M. Para cualquier $p \geq 1$, tenemos

$$\left\| \mathbb{E}_{\psi_\mu^m}[\lambda_X] - \mathbb{E}_{\psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2m^{\frac{1}{p}} \mathrm{W}_p(\mu,\nu),$$

donde W_p es la p-ésima distancia de Wasserstein en M.

Lo anterior nos resulta útil por dos razones. Primero, nos dice que para una m fija, el "comportamiento topológico" del conjunto de m puntos trae consigo cierta información estable sobre la medida subyacente de la cual se generan los datos.

5.3.2. Normalidad Asintótica

Como en la sección anterior, consideramos una multitud de diagramas de persistencia (o alguna otra representación). El siguiente paso despues de dar descriptores de tendencia central para la persistencia homológica es proveer resultados de normalidad asintótica, así como procesos bootstrap para ser capaces de derivar intervalos de confianza. Por supuesto, resulta más sencillo otorgar resultados para representaciones funcionales de la persistencia. En los estudios por Chazal et al. (2015a)[41], Chazal et al. (2015c)[45], siguiendo esta estrategia, se proponen bandas de confianza para paisajes a partir de observaciones de paisajes $\lambda_1, \ldots, \lambda_N$ obtenidas i. i. d. de una distribución aleatoria en el espacio de paisajes. La validez asintótica y la convergencia uniforme del multiplicador bootstrap se demuestra en este marco. Nótese que resultados similares también se han propuesto para muchas otras representaciones de la persistencia, en particular provando que los espacios funcionales correspondientes son espacios de Donsker.

5.4. Otros Acercamientos Estadísticos al Análisis Topológico de Datos

Se ha mostrado un aumento en el interés por acercamientos estadísticos al ATD y se han propuesto y estudiado varios en los años recientes, a continuación una lista no exhaustiva de ejemplos.

Prueba de Hipótesis

Se han propuesto varios métodos para procedimientos de prueba de hipótesis para la persistencia homológica, en su mayoría para pruebas de dos muestras y basados en estrategias de permutación. Robinson y Turner (2017)[112] se concentran en distancias por pares de diagramas de persistencia, mientras que Berry et al. [6] estudian acercamientos más generales. Tambíen se han propuesto pruebas de hipótesis basadas en acercamientos enfocados en el kernel en el estudio por Kusano (2019)[83]. Además una prueba de hipótesis de dos etapas de filtración y prueba para imagenes de persistencia fue presentada en el estudio por Moon y Lazar (2020)[94].

Transformada de la Homología Persistente

Las representaciones introducidas anteriormente son todas transformaciones derivadas del diagrama de persistencia calculado de una filtración fija construida sobre un conjunto de datos. La transformada de la homlogía persistente introducida en los estudios por Curry et al. (2018)[56], Turner et al. (2014b)[123] para estudiar figuras en $\mathbb{R}^d$ toma un camino diferente enfocandose en la homología persistente del conjunto subnivel de la filtración inducida por la proyección de la figura considerada en cada dirección de $\mathbb{R}^d$. Tiene varias propiedades interesantes; en particular, la transformada de la homología persistente es un estadístico suficiente para distribuciones definidas en el conjunto de complejos simpliciales geométricos y finitos, encajados en $\mathbb{R}^d$.

Estadística Bayesiana para el Análisis Topológico de Datos

Se ha propuesto un acercamiento Bayesiano a la inferencia de diagramas de persistencia en el estudio por Maroulas et al. (2020)[92] que consiste en entender un diagrama de persistencia como una muestra de un proceso puntual. Este método Bayesiona calcula la intensidad posterior del proceso puntual basado en una intensidad de mezcla Gaussiana para el previo.

5.5. Homología Persistente y el Aprendizaje Autoḿatico

Usando el ATD y, más especificamente, la homología persistente para el apredizaje automático ha sido un tema de alto muy alto interés que ha generado amplia e intensa investigación. Aunque los avances recientes se escapan el alcance de este trabajo, introducimos a continuación las pricipales direcciones que la investigación ha tomado con algunas referencias de interés particular.

Análisis Topológico de Datos para Exploración de Datos y Estadísticos Descriptivos

In algunos dominios, el ATD puede ser usado como una herramienta para el análisis exporatorio y la visualización. Por ejemplo, el algoritmo Mapper provee un acercamiento poderoso para explorar y visualizar la estructura topológica global de conjuntos de datos intricados. En algunos casos, diagramas de persistencia obtenidos de datos pueden ser directamente interpretados y explotados para el mejor entendimiento del fenómeno del cual los datos han sido generados. Este es el caso en el estudio sobre campos de fuerza en medios granulares (Kramar et al., 2013[80]) o en el de las estructuras atómicas en vidrio (Nakamura et al., 2015[96]) en la ciencia de materiales, en el estudio de la evolución de patrones de convección en la dinámica de fluidos (Kramár et al., 2016[81]), y en el monitoreo

maquinista (Khasawneh y Munch, 2016[77]) o en el análisis de estructuras nanoporosas en química (Lee et al., 2017[86]) donde características topológicas pueden ser relacionadas con estructuras geométricas topológicas y patrones en los datos considerados.

Homología Persistente para la Ingenieria de Características

Hay muchas otros casos de las carácteristicas de la persistencia no pueden ser directamente interpretadas de manera sencilla, sin embargo, existe información de valor en ellas que requiere un procesamiento adicional. No obstante, la naturaleza altamente no lineal de los diagramas de persistencia nos impide usarlos de manera inmediata como características estándar en algoritmos de aprendizaje automático.

Los paisajes de persistencia y sus representaciones lineales ofrecen una primera opción para convertir diagramas de persistencia en elementos de un espacio vectorial que pueden ser directamente usados como características en los esquemas clásicos de aprendizaje automático. Ese acercamiento ha sido utilizado, por ejemplo, en la unión de proteinas (Kovacev-Nikolic et al., 2016[79]), reconocimiento de objetos (Li et al., 2014[88]), o en el análisis de series de tiempo. De manera similar, la construcción de kernels para diagramas de persistencia que preservan las propiedades de estabilidad a sido objeto de atención recientemente. La mayoria de ellos han sido obtenidos considerando diagrams como medidas discretas en $\mathbb{R}^2$. Convolucionando una versión simetrizada (con respecto a la diagonal) de los diagramas de persistencia con una distrivución Gaussiana 2-dimensional, Reininghaus et al. (2015)[108] introdujo un kernel multiescala y lo aplicó a la clasificación de figuras y problemas de reconocimiento de texturas. Considerando la distancia de Wasserstein entre proyecciones de diagramas de persistencia sobre rectas, Carriere et al. (2017)[25] contruyó otro kernel y probó su desempeño en varias pruebas de referencia. También se han propuesto otros kernels en estudio por Kusano et al.(2017)[84] obtenidos de la misma manera consideranda diagramas de persistencia como medidas.

Se han propuesto otros resumenes vectoriales para diagramas de persistencia con la intención de ser usados como características para diversos problemas. Por ejemplo, resumenes básicos fueron considerados en el estudio por Bonis et al. (2016)[10] y combinados con métodos de cuantización y agrupamiento para abordar problemas de análisis de formas no rígidas; curvas de Betti extraidas de diagramas de persistencia fueron utilizadas con redes neuronales convolucionales de una dimensión para analizar datos dependientes del tiempo y reconocer actividades humanas con medidas de sensores inerciales en los estudios por Dindin et al. (2020)[62], Umeda (2017)[124]; imagenes de persistencia introducidas en el estudio por Adams et al. (2017)[3] y se consideraron para afrontar problemas inversos utilizando modelos lineales de aprendizaje automático en el estudio por Obayashi et al. (2018)[98].

Los kernels y resumenes vectoriales de los diagramas de persistencia mencionados anteriormente fueron construidos independientemente de los análisis o tareas en los que se utilizaron. Más aún en muchos casos, la información topológica relevante no es transportada por todo el diagrama de persistencia, sino que se encuentra concentrada en regionas localizadas que pueden no ser sencillas de identificar. Esto hace de la elección del kernel o resumen vectorial a considerar una muy difícil para el usuario. Para superar este problema, varios autores han propuesto acercamientos de aprendizaje que nos permitan identifical las características topológicas relevantes para una tarea dada. Con esta intención, Hofer et al. (2017)[76] propuso un acercamiento de aprendizaje profundo para detectar los parámetros de imagenes de persistencia, mientras que Kim et al. (2020)[78] introdujo una capa de red neuronal para paisajes de persistencia. En el estudio por Carrière et al. (2020a)[29], los

autores introdujeron una capa general de red neuronal para diagramas de persistencia que puede ser usada para aprender vectorizaciónes adecuadas o directamente integrada en una arquitectura de red neuronal profunda. Otros métodos, inspirados en k-medias, pronen la utilizacíon de aprendizaje no supervisado para vectorizar diagramas de persistencia (Royer et al., 2021[114]; Zieliński et al., 2010[130], algunos de ellos incluso proveen garantias teóricas (Chazal et al., 2020[46]).

Homología Persistente para Optimización de Arquitecturas de Aprendizaje Automático y Selección de Modelos

Recientemente, el ATD ha visto nuevos desarrollos en el área de aprendizaje automático donde la homología persistente es usada no en para la ingenieria de características, sino como una herramienta para diseñar, mejorar y seleccionar modelos (Véase kCarlsson y Gabrielsson (2020)[23], Chen et al. (2019)[51], Gabrielsson y Carlsson (2019)[16], Hofer et al. (2019a)[74], Moor et al. (2020)[95], Ramamurthy et al. (2019)[107], Rieck et al(2019)[109]). Muchas de estas herramientas en el uso de funciones de pérdida o regularización dependientes de características de la homología persistente, lo que convierte su optimización en un problema más a resolver. En los estudios por (Poulenard et al., 2018[105]; Gabrielsson et al., 2019[16]) se ha propuesto y experimentado con métodos prácticos, construidos utilizando las herramientas proporcionadas por las librerias de PyTorch o TensorFlow, los cuales nos permiten codificar y optimizar una extensa familia de funciones basadas en la persistencia. Un esquema general para la optimización de funciones basadas en la persistencia que hace uso de algoritmos de descenso de subgradiente estocástico con garantias de convergencia ha sido recientemente propuesto e implementado en una herramienta de software de fácil utilización (Carrière et al., 2020b[24]). Con otro enfoque, un esquema teórico distinto para el estudio de la estructura diferenciable de funciones de diagramas de persistencia ha sido propuesta en el estudio por Leygonie et al. (2021)[87].

Capítulo 6

Análisis Topológico de datos para Ciencia de Datos con la Libreria GUDHI

En esta sección, ilustraremos métodos del ATD usando la libreria de Python GUDHI[1] (Maria et al., 2014[91]) junto con otras librerias populares como Numpy (Walt et al., 2011[125]), scikit-learn (Pedregosa et al., 2011)[99], y pandas (McKinney, 2010[128]). Esta sección se enfoca en demostrar que las firmas topológicas del ATD pueden ser facilmente calculadas y explotadas usando GUDHI. Se pueden encontrar más ejemplos en el tutorial de GUDHI en GitHub[2].

6.1. Bootstrap y Comparasión de Configuraciones de Unión de Proteínas

Este ejemplo lo tomamos prestado de Kovacev-Nikolic et al (2016)[79]. En este artículo, la homología persistente es usada para analizar la unión de proteínas, y más precisamente, compara las formas abiertas y cerradas de la proteina de unión a la maltosa (MBP), una biomolécula que consiste de 370 residuos de aminoácidos. El anánlisis no se basa en distancias geométricas en $\mathbb{R}^3$ sino en una métrica de distancias dinámicas definida por

$$D_{ij} = 1 - |C_{ij}|,$$

donde C son las matrices de correlación entre los residuos. Los datos pueden descargarse en el siguiente link[3].

```
import numpy as np
import gudhi as gd
import pandas as pd
import seaborn as sns

corr_protein = pd.read_csv("mypath/1anf.corr_1.txt", header=None,
    delim_whitespace = True)
dist_protein_1 = 1 - np.abs(corr_protein_1.values)
```

[1] https://gudhi.inria.fr/python/latest
[2] https://github.com/GUDHI/TDA-tutorial
[3] https://www.researchgate.net/publication/301543862_corr

```
rips_complex_1 = gd.RipsComplex(distance_matrix = dist_protein_1,
    max_edge_length = 1.1)
simplex_tree_1 = rips_complex_1.create_simplex_tree(max_dimension = 2)
diag_1 = simplex_tree_1.persistence()

gd.plot_persistence_diagram(diag_1)
```

Para comparar diagramas de persistencia, usamos la distancia de cuello de botella. El bloque a continuación calcula los intervalos de persistencia y la distancia de cuello de botella para la 0-homología y 1-homología

```
interv0_1 = simplex_tree_1.persistence_intervals_in_dimension(0)
interv0_2 = simplex_tree_2.persistence_intervals_in_dimension(0)
bot0 = gd.bottleneck_distance(interv0_1, interv0_2)

interv1_1 = simplex_tree_1.persistence_intervals_in_dimension(1)
interv1_2 = simplex_tree_2.persistence_intervals_in_dimension(1)
bot1 = gd.bottleneck_distance(interv1_1, interv1_2)
```

De esta manera, podemos calcular la matriz de distancias de cuello de botella entre las catorce MBP. Finalmente, aplicamos metodos de reescalado multidimensional (MDS) para encontrar la configuración en $\mathbb{R}^2$ que se asimile a las distancias de cuello de botella (ver Figura 5.1C). Hacemos uso de la libreria scikit-learn para el MDS como sigue:

```
import matplotlib.pyplot as plt
from sklearn import manifold

mds = manifold.MDS(n_components2, dissimilarity = "precomputed")
config = mds.fit(M).embedding_

plt.scatter(cong[0:7,0], config[0:7, 1], color = "red", label="closed")
plt.scatter(config[7:1,0], config[7:1, 1], color = "blue", label="red")
plt.legend(loc = 1)
```

Ahora, defnimos una banda de confianza para un diagrama usando el acercamiento de bootstrap de cuello de botella. Remuestreamos sobre las lineas (y columnas) de la matriz de distancias, y calculamos la distancia de cuello de botella entre el diagrama de persistencia original y el diagrama de persistencia bootstrap. Repetimos el procedimiento las veces que sean necesareas, y finalmente, estimamos el cuantil 95 % de esta colección de distancias de cuello de botella. Tomamos el valor del cuantil para definir la banda de confianza en el diagrama original (ver Figura 5.1D). Sin embargo, se debe ser cuidadoso a la hora de considerar este procedimiento ya que, hasta donde sabemos, la valides de bootstrap de cuello0 de botella no ha sido probada en este contexto.

6.2. Clasificación de Datos de Sensores

En este experimento, la aceleración (en 3D) de 3 sujetos caminando (A, B, y C) ha side monitoreada utilizando el sensor de un smartphone[4]. La homología persistente no es sensible a la elección de ejes, asÃ que no es necesario ningun preprocesamiento de los datos para alinear las 3 series de tiempo a los mismos ejes. De estas series, se escogieron aleatoriamente extractos de 8 segundos de la serie de tiempo completa, esto es, 200 puntos consecutivos de aceleración en $\mathbb{R}^3$. Por cada sujeto, se extrajeron 100 series de tiempo de

[4]Los datos pueden ser descargados en: http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/data_acc

esta manera. El siguiente cloque calcula la persistencia para las filtraciones de complejos alpha para una de las 100 series de tiempo de la aceleración del sujeto A.

```
alpha_complex_sample = gd.AlphaComplex(points = data_A_sample)
simplex_tree_sample = alpha_complex_sample.create_simplex_tree(
    max_alpha_square = 0.3)
diag_Alpha = simplex_tree_sample.persistence()
```

Con `diag_Alpha`, podemos calcular y graficar los paisajes de persistencia con facilidad (ver Figura 6.1A). Para todas las 300 series de tiempo, calculamos los paisajes de persistencia de dimensión 0 y 1, y calcylamos los primeros tres paisajes de dimensión 2. Más aún, cada paisaje de persistencia es discretizado en 1000 puntos. Cada serie de tiempo es entonces descrita por 6000 variables topológicas. Para predecir al sujeto con estas características, usamos un bosque aleatorio (Breiman, 2001[14]), ya que es eficiente cuando se maneja una alta dimensionalidad en los datos. Separamos nuestros datos en conjuntos de entrenamiento y prueba de manera aleatoria varias veces. Con esto obtenemos un error de clasificación promedio alrededor de 0,95. Al utilizar un modelo de bosque aleatorio, tenemos acceso a una visualización de las variables más importantes (ver Figura6.1B).

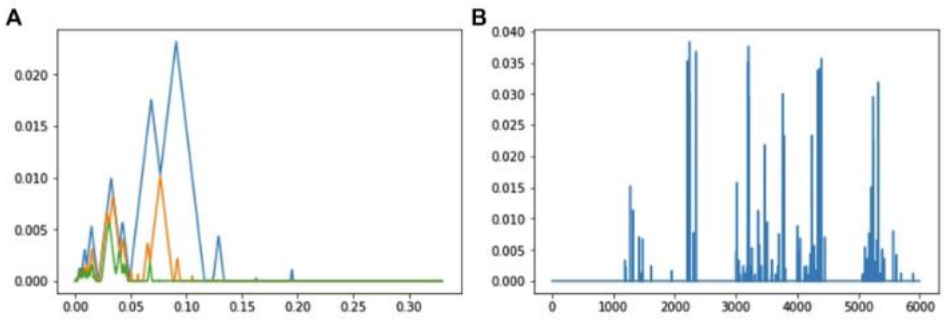


Figura 6.1: (A) Los primeros tres paisajes para la 0-homologiá de la filtración alpha definida para una de las series de tiempo de aceleración del sujeto A. (B) Coeficientes de significancia de las variables del paisaje para la clasificación de los sujetos. Los primeros 3000 coeficientes corresponden a los tres paisajes de dimensión 0 y los últimos 3000 a los paisajes de dimensión 1. Hay 1000 coeficientes por paisaje. Nótese que el primer paisaje de dimensión 0 es siempre el mismo utilizando la filtración del complejo de Rips (un paisaje trivial), por tanto, loes coeficientes que le corresponden tiene un valor 0 de significancia.

Capítulo 7

Discusión

En este documento, porponemos un marco general de los métodos más estándar en el área del análisis topológico de datos. También damos una presentación de los fundamentos matemáticos del ATD, en aspectos topológicos, algebraicos, geométricos y estadísticos. La robustez de los métodos del ATD (debido a su invariancia ante la deformación e independencia de las coordenadas) y la representación comprimida de los datos que ofrece hacen muy interesante su uso para el análisis de datos, apredizaje automático y la inteligencia artificial transparente. Se han propuesto muchas aplicaciones en esta dirección durante los ultimos años. Así, el ATD forma parte de la caja de herramientas del científico de datos.

Por supuesto, aunque el ATD este equipado para afrontar todo tipo de problemas, aquellos que se aventuren en el uso de estos métodos pueden encontrarse con una variedad de problemas. En los aspectos algoritmicos, calcular la homología persistente puede ser costoso en tiempo y recursos. Aunque aún hay posibilidades de mejora, los recientes avances computacionales han permitido que el ATD se convierta en un método efectivo para el análisis de datos, gracias a librerias como GUDHI, por ejemplo. Además, combinar el ATD usando métodos de cuantización, simplificación de graficas, o reducción de dimensionalidad, puede reducir el coste computacional de los algoritmos del ATD de manera significativa. Otro potencial problema que podemos encontrar es que volver a los datos para interpretar las firmas topológicas puede llegar a ser complicado ya que estas firmas corresponden a clases de equivalencia de ciclos. Esto puede ser un problema al momento de identificar que parte de la nube de puntos "A creado" una cierta firma topológica. Finalmente, la información topológica y geométrica que puede ser extraida de los datos no es simepre eficaz para resolver cualquier problema dado en la ciencia de datos. Combinar características topológicas con otro tipo de descriptores es generalmente el acercamiento adecuado.

Hoy en día, el ATD es un campo de investigación activo, relevante en muchos campos científicos. En particular, actualmente existe un interés intenso por combinar de manera efectiva el aprendizaje automático, la estadística y el ATD. En esta perspectiva, creemos que aún existe una necesidad de resultados estadísticos que demuestren y cuantifiquen el interés de estos acercamientos basados en el ATD.

Apéndices

Apéndice A

Cosas que no deberían ir en el texto principal

Un apéndice, por qué no?!

Bibliografía

[1] *8. Tensor Decomposition*, pages 91–100. 2007.

[2] E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359 – 1399, 2019.

[3] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[4] H. Anai, F. Chazal, M. Glisse, Y. Ike, H. Inakoshi, R. Tinarrage, and Y. Umeda. Dtm-based filtrations. In N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, editors, *Topological Data Analysis*, pages 33–66, Cham, 2020. Springer International Publishing.

[5] S. Balakrishnan, A. Rinaldo, D. Sheehy, A. Singh, and L. Wasserman. Minimax rates for homology inference. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 64–72, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[6] E. Berry, Y.-C. Chen, J. Cisewski-Kehe, and B. T. Fasy. Functional summaries of persistence diagrams. *Journal of Applied and Computational Topology*, 4(2):211–262, Jun 2020.

[7] G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5(none):204 – 237, 2011.

[8] O. Bobrowski, S. Mukherjee, and J. E. Taylor. Topological consistency via kernel estimation. *Bernoulli*, 23(1), feb.

[9] J.-D. Boissonnat, F. Chazal, and M. Yvinec. *Geometric and Topological Inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2018.

[10] T. Bonis, M. Ovsjanikov, S. Oudot, and F. Chazal. Persistence-based pooling for shape pose recognition. In A. Bac and J.-L. Mari, editors, *Computational Topology in Image Context*, pages 19–29, Cham, 2016. Springer International Publishing.

[11] C. Brécheteau. A statistical test of isomorphism between metric-measure spaces using the distance-to-a-measure signature. *Electronic Journal of Statistics*, 13(1):795–849, 2019.

[12] C. Brécheteau and C. Levrard. A k-points-based distance for robust geometric inference. *Bernoulli*, 26(4):3017 – 3050, 2020.

[13] C. Brécheteau and C. Levrard. A k-points-based distance for robust geometric inference. *Bernoulli*, 26(4):3017 – 3050, 2020.

[14] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[15] A. Brown, O. Bobrowski, E. Munch, and B. Wang. Probabilistic convergence and stability of random mapper graphs. *Journal of Applied and Computational Topology*, 5(1):99–140, Mar 2021.

[16] R. BrÃ¼el Gabrielsson and G. Carlsson. Exposition and interpretation of the topology of neural networks. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1069–1076, 2019.

[17] P. Bubenik. Statistical topological data analysis using persistence landscapes, 2015.

[18] M. Buchet, F. Chazal, T. K. Dey, F. Fan, S. Y. Oudot, and Y. Wang. Topological analysis of scalar fields with outliers. In *Proc. Sympos. on Computational Geometry*, 2015.

[19] M. Buchet, F. Chazal, S. Y. Oudot, and D. R. Sheehy. *Efficient and Robust Persistent Homology for Measures*, pages 168–180.

[20] G. BURTON. Topics in optimal transportation (graduate studies in mathematics 58) by cdric villani: 370 pp., us$59.00, isbn 0-8218-3312-x (american mathematical society, providence, ri, 2003). *Bulletin of the London Mathematical Society*, 36:285 – 286, 03 2004.

[21] B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4):999–1023, 2006.

[22] G. Carlsson and R. B. Gabrielsson. Topological approaches to deep learning. In N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, editors, *Topological Data Analysis*, pages 119–146, Cham, 2020. Springer International Publishing.

[23] G. Carlsson and R. B. Gabrielsson. Topological approaches to deep learning. In N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, editors, *Topological Data Analysis*, pages 119–146, Cham, 2020. Springer International Publishing.

[24] M. Carriere, F. Chazal, Y. Ike, T. Lacombe, M. Royer, and Y. Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2786–2796. PMLR, 26–28 Aug 2020.

[25] M. Carrière, M. Cuturi, and S. Oudot. Sliced Wasserstein kernel for persistence diagrams. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 664–673. PMLR, 06–11 Aug 2017.

[26] M. Carrière and B. Michel. Statistical analysis of mapper for stochastic and multivariate filters, 2021.

[27] M. Carrière, B. Michel, and S. Oudot. Statistical analysis and parameter selection for mapper. *Journal of Machine Learning Research*, 19(12):1–39, 2018.

[28] M. Carrière and R. Rabadán. Topological data analysis of single-cell hi-c contact maps. In N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, editors, *Topological Data Analysis*, pages 147–162, Cham, 2020. Springer International Publishing.

[29] M. CarriÃ¨re, F. Chazal, M. Glisse, Y. Ike, and H. Kannan. A note on stochastic subgradient descent for persistence-based functionals: convergence and practical aspects. 10 2020.

[30] F. Chazal. High-Dimensional Topological Data Analysis. In *3rd Handbook of Discrete and Computational Geometry*. CRC Press, 2016.

[31] F. Chazal, D. Chen, L. Guibas, X. Jiang, and C. Sommer. Data-driven trajectory smoothing. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 251–260, New York, NY, USA, 2011. Association for Computing Machinery.

[32] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. MÃ©moli, and S. Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, 28(5):1393–1403, 2009.

[33] F. Chazal, D. Cohen-Steiner, and A. Lieutier. Normal cone approximation and offset shape isotopy. *Computational Geometry*, 42(6):566–581, 2009.

[34] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean space. *Discrete & Computational Geometry*, 41(3):461–479, Apr 2009.

[35] F. Chazal, D. Cohen-Steiner, A. Lieutier, and B. Thibert. Stability of Curvature Measures. *Computer Graphics Forum*, 2009.

[36] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Boundary measures for geometric inference. *Foundations of Computational Mathematics*, 10(2):221–240, Apr 2010.

[37] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, Dec 2011.

[38] F. Chazal, V. de Silva, M. Glisse, and S. Y. Oudot. *The Structure and Stability of Persistence Modules*. Springer Briefs in Mathematics. Springer, 2016.

[39] F. Chazal and V. Divol. The density of expected persistence diagrams and its kernel based estimation. *CoRR*, abs/1802.10457, 2018.

[40] F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Subsampling methods for persistent homology. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2143–2151, Lille, France, 07–09 Jul 2015. PMLR.

[41] F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman. Stochastic convergence of persistence landscapes and silhouettes, 2013.

[42] F. Chazal, M. Glisse, C. LabruÃ¨re, and B. Michel. Convergence rates for persistence diagram estimation in topological data analysis. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 163–171, Bejing, China, 22–24 Jun 2014. PMLR.

[43] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6), nov 2013.

[44] F. Chazal, R. Huang, and J. Sun. Gromov–hausdorff approximation of filamentary structures using reeb-type graphs. *Discrete & Computational Geometry*, 53(3):621–649, Apr 2015.

[45] F. Chazal, R. Huang, and J. Sun. Gromov–hausdorff approximation of filamentary structures using reeb-type graphs. *Discrete & Computational Geometry*, 53(3):621–649, Apr 2015.

[46] F. Chazal, C. Levrard, and M. Royer. Optimal quantization of the mean measure and applications to statistical learning, 2021.

[47] F. Chazal and A. Lieutier. Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees. *Computational Geometry*, 40(2):156–170, 2008.

[48] F. Chazal, P. Massart, and B. Michel. Rates of convergence for robust geometric inference. *Electronic Journal of Statistics*, 10(2):2243–2286, 2016.

[49] F. Chazal and B. Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4, 2021.

[50] F. Chazal and S. Y. Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry*, SCG '08, pages 232–241, New York, NY, USA, 2008. Association for Computing Machinery.

[51] C. Chen, X. Ni, Q. Bai, and Y. Wang. A topological regularizer for classifiers via persistent homology. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2573–2582. PMLR, 16–18 Apr 2019.

[52] Y.-C. Chen, C. R. Genovese, and L. Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–1696, 2017.

[53] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proceedings of the Twenty-First Annual Symposium on Computational Geometry*, SCG '05, pages 263–271, New York, NY, USA, 2005. Association for Computing Machinery.

[54] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have lp-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, Apr 2010.

[55] A. Cuevas and A. RodrÃguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36(2):340–354, 2004.

[56] J. Curry, S. Mukherjee, and K. Turner. How many directions determine a shape and other sufficiency results for two topological transforms, 2021.

[57] V. de Silva and R. Ghrist. Homological sensor networks. 2007.

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

[59] T. K. Dey, F. Memoli, and Y. Wang. Topological analysis of nerves, reeb spaces, mappers, and multiscale mappers, 2017.

[60] T. K. Dey, F. MÃ©moli, and Y. Wang. *Multiscale Mapper: Topological Summarization via Codomain Covers*, pages 997–1013.

[61] M. Dindin, Y. Umeda, and F. Chazal. Topological data analysis for arrhythmia detection through modular neural networks. In C. Goutte and X. Zhu, editors, *Advances in Artificial Intelligence*, pages 177–188, Cham, 2020. Springer International Publishing.

[62] M. Dindin, Y. Umeda, and F. Chazal. Topological data analysis for arrhythmia detection through modular neural networks. In C. Goutte and X. Zhu, editors, *Advances in Artificial Intelligence*, pages 177–188, Cham, 2020. Springer International Publishing.

[63] V. Divol and T. Lacombe. Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport. *Journal of Applied and Computational Topology*, 5(1):1–53, Mar 2021.

[64] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002.

[65] B. T. Fasy, J. Kim, F. Lecci, and C. Maria. Introduction to the R package TDA. *CoRR*, abs/1411.1830, 2014.

[66] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6), dec 2014.

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

[68] P. Frosini. Measuring shapes by size functions. In D. P. Casasent, editor, *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, volume 1607, pages 122 – 133. International Society for Optics and Photonics, SPIE, 1992.

[69] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941 – 963, 2012.

[70] R. Ghrist. Homological algebra and data. *IAS/Park City Mathematics Series*, 2018.

[71] L. Guibas, D. Morozov, and Q. Mérigot. Witnessed k-distance. *Discrete & Computational Geometry*, 49(1):22–45, Jan 2013.

[72] A. Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2001.

[73] F. Hensel, M. Moor, and B. Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4, 2021.

[74] C. Hofer, R. Kwitt, M. Niethammer, and M. Dixit. Connectivity-optimized representation learning via persistent homology. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2751–2760. PMLR, 09–15 Jun 2019.

[75] C. D. Hofer, R. Kwitt, and M. Niethammer. Learning representations of persistence barcodes. *Journal of Machine Learning Research*, 20(126):1–45, 2019.

[76] C. D. Hofer, R. Kwitt, M. Niethammer, and A. Uhl. Deep learning with topological signatures. *CoRR*, abs/1707.04041, 2017.

[77] F. A. Khasawneh and E. Munch. Chatter detection in turning using persistent homology. *Mechanical Systems and Signal Processing*, 70-71:527–541, 2016.

[78] K. Kim, J. Kim, M. Zaheer, J. Kim, F. Chazal, and L. Wasserman. Pllay: Efficient topological layer based on persistent landscapes. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15965–15977. Curran Associates, Inc., 2020.

[79] V. Kovacev-Nikolic, P. Bubenik, D. Nikolić, and G. Heo. Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology*, 15(1):19–38, 2016.

[80] M. Kramar, A. Goullet, L. Kondic, and K. Mischaikow. Persistence of force networks in compressed granular media. *Phys. Rev. E*, 87:042207, Apr 2013.

[81] M. Kramár, R. Levanger, J. Tithof, B. Suri, M. Xu, M. Paul, M. F. Schatz, and K. Mischaikow. Analysis of kolmogorov flow and rayleigh-bénard convection using persistent homology. *Physica D: Nonlinear Phenomena*, 334:82–98, 2016. Topology in Dynamics, Differential Equations, and Data.

[82] J. Krebs and W. Polonik. On the asymptotic normality of persistent betti numbers, 2023.

[83] G. Kusano. On the expectation of a persistence diagram by the persistence weighted kernel. *Japan Journal of Industrial and Applied Mathematics*, 36(3):861–892, Sep 2019.

[84] G. Kusano, K. Fukumizu, and Y. Hiraoka. Kernel method for persistence diagrams via kernel embedding and weight factor, 2017.

[85] T. Lacombe, M. Cuturi, and S. Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport, 2018.

[86] Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, and B. Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8(1):15396, May 2017.

[87] J. Leygonie, S. Oudot, and U. Tillmann. A framework for differential calculus on persistence barcodes, 2021.

[88] C. Li, M. Ovsjanikov, and F. Chazal. Persistence-based structural recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2003–2010, 2014.

[89] M. Z. Li, M. S. Ryerson, and H. Balakrishnan. Topological data analysis for aviation applications. *Transportation Research Part E: Logistics and Transportation Review*, 128:149–174, 2019.

[90] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3(1):1236, Feb 2013.

[91] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec. The gudhi library: Simplicial complexes and persistent homology. In H. Hong and C. Yap, editors, *Mathematical Software – ICMS 2014*, pages 167–174, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[92] V. Maroulas, F. Nasrin, and C. Oballe. A bayesian framework for persistent homology. *SIAM Journal on Mathematics of Data Science*, 2(1):48–74, 2020.

[93] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, nov 2011.

[94] C. Moon and N. A. Lazar. Hypothesis testing for shapes using vectorized persistence diagrams, 2023.

[95] M. Moor, M. Horn, B. Rieck, and K. Borgwardt. Topological autoencoders. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7045–7054. PMLR, 13–18 Jul 2020.

[96] T. Nakamura, Y. Hiraoka, A. Hirata, E. G. Escolar, and Y. Nishiura. Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26(30):304001, jul year.

[97] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, Mar 2008.

[98] I. Obayashi and Y. Hiraoka. Persistence diagrams with linear machine learning models, 2017.

[99] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012.

[100] M. D. Penrose and J. Yukich. Central Limit Theorems for Some Graphs in Computational Geometry. *The Annals of Applied Probability*, 11(4):1005 – 1041, 2001.

[101] A. Petrunin. *Applied Manifold Geometry*, pages 137–483. World Scientific, 2007.

[102] J. M. Phillips, B. Wang, and Y. Zheng. Geometric inference on kernel density estimates, 2013.

[103] J. A. Pike, A. O. Khan, C. Pallini, S. G. Thomas, M. Mund, J. Ries, N. S. Poulter, and I. B. Styles. Topological data analysis quantifies biological nano-structure from single molecule localization microscopy. *Bioinformatics*, 36(5):1614–1621, 10 2019.

[104] W. Polonik. Measuring Mass Concentrations and Estimating Density Contour Clusters-An Excess Mass Approach. *The Annals of Statistics*, 23(3):855 – 881, 1995.

[105] A. Poulenard, P. Skraba, and M. Ovsjanikov. Topological function optimization for continuous shape matching. *Computer Graphics Forum*, 37(5):13–25, 2018.

[106] T. Qaiser, Y.-W. Tsang, D. Taniyama, N. Sakamoto, K. Nakane, D. Epstein, and N. Rajpoot. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Medical Image Analysis*, 55:1–14, 2019.

[107] K. N. Ramamurthy, K. Varshney, and K. Mody. Topological data analysis of decision boundaries with application to model selection. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5351–5360. PMLR, 09–15 Jun 2019.

[108] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A stable multi-scale kernel for topological machine learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4741–4748, 2015.

[109] B. Rieck, M. Togninalli, C. Bock, M. Moor, M. Horn, T. Gumbsch, and K. Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology, 12 2018.

[110] B. Rieck, T. Yates, C. Bock, K. Borgwardt, G. Wolf, N. Turk-Browne, and S. Krishnaswamy. Uncovering the topology of time-varying fmri data using cubical persistence. *Advances in neural information processing systems*, 33, 2020.

[111] V. Robins. Towards computing hology from finite approximations. *Topology Proceedings*, 24:503–532, 1999.

[112] A. Robinson and K. Turner. Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology*, 1(2):241–261, Dec 2017.

[113] B. Roycraft, J. Krebs, and W. Polonik. Bootstrapping persistent betti numbers and other stabilizing statistics. *The Annals of Statistics*, 51(4), Aug. 2023.

[114] M. Royer, F. Chazal, C. Levrard, U. Yuhei, and I. Yuichi. Atol: Measure vectorization for automatic topologically-oriented learning, 2020.

[115] L. M. Seversky, S. Davis, and M. Berger. On time-series topological data analysis: New data and opportunities. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1014–1022, 2016.

[116] V. d. Silva and G. Carlsson. Topological estimation using witness complexes. In M. Gross, H. Pfister, M. Alexa, and S. Rusinkiewicz, editors, *SPBG'04 Symposium on Point - Based Graphics 2004*. The Eurographics Association, 2004.

[117] A. Singh, C. Scott, and R. Nowak. Adaptive Hausdorff estimation of density level sets. *The Annals of Statistics*, 37(5B):2760 – 2782, 2009.

[118] A. E. Sizemore, J. E. Phillips-Cremins, R. Ghrist, and D. S. Bassett. The importance of the whole: Topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656–673, 07 2019.

[119] P. Skraba, M. Ovsjanikov, F. Chazal, and L. Guibas. Persistence-based segmentation of deformable shapes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 45–52, 2010.

[120] A. D. Smith, P. Dlotko, and V. M. Zavala. Topological data analysis: Concepts, computation, and applications in chemical engineering. *Computers & Chemical Engineering*, 146:107202, 2021.

[121] A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948 – 969, 1997.

[122] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, Jul 2014.

[123] K. Turner, S. Mukherjee, and D. M. Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344, 12 2014.

[124] Y. Umeda. Time series classification via topological data analysis. *Transactions of the Japanese Society for Artificial Intelligence*, 32(3):D–G72_1–12, 2017.

[125] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30, 2011.

[126] M. A. G. Viana and H. P. Wynn, editors. *Algebraic Methods in Statistics and Probability II*. American Mathematical Society, 2010.

[127] L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018.

[128] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[129] Y. Yao, J. Sun, X. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande, and G. Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of Chemical Physics*, 130(14):144115, 2009.

[130] B. Zieliński, M. Lipiński, M. Juda, M. Zeppelzauer, and P. Dłotko. Persistence bag-of-words for topological data analysis, 2019.

[131] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, Feb 2005.