RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Análisis Topológico de Datos: una mirada estadística

Guillermo Aguirre Carrazana y Edel García Reyes

RT_088 **enero 2017**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

Análisis Topológico de Datos: una mirada estadítica

Guillermo Aguirre Carrazana y Edel García Reyes

RT_088 **enero 2017**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

Lista de figuras

Análisis Topológico de Datos: una mirada estadística

Guillermo Aguirre Carrazana y Edel García Reyes

Equipo de Investigaciones de Reconocimiento de Patrones, CENATAV - DATYS, La Habana, Cuba
{gaguirre, egarcia}@cenatav.co.cu

RT_088, Serie Azul, CENATAV - DATYS
Aceptado: 16 de diciembre de 2016

Resumen. El Análisis Topológico de Datos (TDA, *por sus siglas en inglés*), es un área emergente de las matemáticas aplicadas que se ha ganado todo tipo de atención en el mundo de la analítica. Este trabajo tiene como objetivo fundamental brindar una introducción a los principales resultados alcanzados dentro del área desde un enfoque estadístico. Se presentan algunos métodos para transformar nubes de puntos en complejos y obtener una representación reducida de los espacios topológicos conocida como complejos simpliciales. Se muestran aspectos y propiedades importantes de los descriptores topológicos introducidos al aplicar el nuevo enfoque sobre TDA. Se expone la técnica bootstrap de gran utilidad para obtener estimaciones del error estadístico y calcular intervalos de confianza sobre los resúmenes. Se brindan resultados en el área de prueba de hipótesis sobre Análisis Topológico de Datos. Se abordan representaciones funcionales de los descriptores topológicos estándar; esto es, la persistencia landscape y la función rango y así solucionar algunos problemas que presentan los descriptores clásicos expuestos en el trabajo. Se introduce la teoría de categorías sobre la homología persistente y se abordan ideas relacionadas con el método de submuestreo para homología persistente. Se introducen generalizaciones para la teoría de persistencia, se expone la teoría de homología para persistente zigzag y son comentados los vineyards, nueva teoría para el estudio de los diagramas de persistencia variables en el tiempo. Se detectaron problemáticas no abordadas hasta el momento, las cuales se resumen en las conclusiones del trabajo.

Palabras clave: análisis topológico de datos, homología persistente, persistencia landscape, persistencia zigzag, función rango.

Abstract. Topological Data Analysis (TDA) is an area of applied mathematics currently garnering all sorts of attention in the world of analytics. The main goal of this work is to provide an introduction to the main contributions achieved in the field with an underlying statistical approach. Some methods are outlined for transforming clouds of points in complexes and obtaining a reduced representation of the topological space known as simplicial complexes. Important aspects and properties of the topological descriptors introduced while applying the new approach about TDA are shown. The bootstrap technique, which is of great importance for obtaining statistical error estimations and to determine confidence intervals for the descriptors is also analyzed. Some results in the area of hypothesis testing in TDA are boarded. Functional representations of the standard topological descriptors are exposed such as the persistence landscape and the range function, which allow to solve some issues of the classic descriptors addressed in this paper. The category theory in persistent homology is introduced and some ideas related with the sub-sampling method for persistent homology are discussed. Some generalizations of the theory of persistence are introduced: homology theory for zigzag persistence and the vineyards are briefly analyzed, which is the new theory for the study of the time-variable persistence diagrams. All unaddressed issues so far in the literature, are summarized in the conclusions of this work.

Keywords: TDA, persistent homology, landscape persistence, zigzag persistence, vineyards, rank function.

1. Introducción

En la actualidad el mundo se encuentra generando grandes volúmenes de datos, que se recogen de diferentes formas y a una gran velocidad, por lo cual surge el problema de cómo manejar esta información para obtener inferencia y deducir patrones. Su investigación requiere de nuevas y complejas técnicas que han sido objeto de estudio en los últimos años. Los datos provienen de nubes de puntos en espacios de alta dimensión que por lo general no distribuyen de manera uniforme; a pesar de que provienen de espacios abstractos contienen a menudo estructuras geométricas y topológicas específicas. Para comprender estas complejas estructuras se requiere de nuevas técnicas diseñadas para encontrar y describir la forma de los datos.

Henri Poincaré (1854-1912), fue el fundador de la **Topología**, rama de la matemática que estudia las deformaciones continuas. Una *deformación continua* permite transformar una superficie en otra; por ejemplo, la taza y el toro son objetos diferentes, pero se puede pasar del uno al otro mediante una deformación continua que no introduce ninguna rotura. Para poder decidir si dos superficies son topológicamente diferentes, se hace necesario clasificarlas atendiendo al número de agujeros que presentan. Fue Poincaré quién ideó las teorías necesarias para abordar esta cuestión, definiendo los *grupos fundamentales* y de *homología*, conceptos que impulsaron el desarrollo de la **Topología Algebraica**, rama de la matemática que proporciona técnicas capaces para analizar en profundidad los espacios topológicos usando herramientas del álgebra abstracta, por lo que sus resultados se consideran puramente algebraicos, con aplicaciones en la teoría de grupos y la teoría de números. La **Topología Algebraica**, por ejemplo, permite contar el número de agujeros que se encuentran en una estructura, pero no permite medir el tamaño ni puede ver otras formas anómalas; para hacer frente a esto surge la persistencia.

La homología persistente es una de las principales herramientas aplicables en el campo emergente de la topología computacional introducida por Edelsbunner [1]. Intuitivamente realiza un seguimiento de las características topológicas en una secuencia creciente de formas, es decir para su cálculo se requiere de la construcción de un complejo de celdas filtradas. Su aplicación permite desarrollar herramientas con el fin de estudiar estructuras topológicas relevantes cualitativas y cuantitativamente de los datos. Algunas de estas pueden ser: agrupamientos, ciclos, *tendrils*, estructuras gráficas, etc.

La teoría clásica de la homología persistente restringe su trabajo a funciones y filtraciones de complejos simpliciales finitos y aborda ciertas cuestiones necesarias, como por ejemplo la inferencia de homología multiescala para espacios métricos o análisis de campos escalares en datos discretos, reducción de dimensionalidad manteniendo o mejorando la capacidad de hacer inferencia geométrica y estadística entre otros. El rápido crecimiento en el rango de las aplicaciones de la topología algebraica sugiere la necesidad de algoritmos eficientes para el cálculo de los grupos de homología, la homología persistente y mapas inducidos en la homología; para ello se adoptan diferentes estrategias. Es así como surge el Análisis Topológico de Datos; al cual nos referiremos en lo adelante como TDA, *por sus siglas en inglés* fundamentado sobre las bases de la Topología Algebraica y la Inferencia Estadística; donde la homología persistente es considerada una herramienta fundamental para aplicar TDA.

El campo del TDA se refiere a varios enfoques y métodos para la exploración de los datos. Los dos enfoques más populares son el algoritmo Mapper [2] y la homología persistente [3]. Mapper es un algoritmo para describir conjuntos de datos en alta dimensión en términos de objetos geométricos simples. Se basa en la idea de agrupación parcial de los datos guiados por un conjunto de funciones definidas en los datos

denominadas filtros. El objetivo sigue siendo recuperar una representación de la nube de puntos, para crear un buen descriptor. Es un método de visualización que conserva la estructura topológica; mientras que la homología persistente ofrece un marco y algoritmos eficientes para codificar la evolución de la topología de la forma, de pequeña a gran escala.

La idea central para aplicar TDA es comenzar con una nube de puntos y calcular resúmenes topológicos de los datos: diagramas de persistencia, códigos de barra y persistencia landscape entre otros. Dada una función de valores reales f, la homología persistente describe como la topología de los conjuntos de nivel inferior $\{x : f(x) \leq t\}$ (o conjuntos de nivel superior $\{x : f(x) \geq t\}$) cambia cuando t aumenta de $-\infty$ a ∞ (o decrece ∞ a $-\infty$). Esta información es codificada en los resúmenes topológicos por ejemplo los diagramas de persistencia. Estos resúmenes proporcionan informaciones útiles acerca de la estructura y geometría de los datos, su objetivo fundamental: cuantificar la incertidumbre, ruido y la reproducibilidad de los descriptores. La principal premisa en el marco de TDA es poder definir objetos estadísticos o estadígrafos sobre estos resúmenes: media, mediana, varianzas y esperanzas condicionales.

De manera general la teoría TDA se basa principalmente en enfoques deterministas que no tienen en cuenta la naturaleza aleatoria de los datos y la variabilidad intrínseca de las cantidades topológicas que infieren. En consecuencia la mayoría de los métodos correspondientes permanecen en exploración, sin ser capaz de distinguir de manera eficiente entre la información y el ruido topológico. Un enfoque estadístico para TDA significa considerar que los datos se generaron a partir de una distribución desconocida, pero que las características topológicas inferidas por sus métodos son vistos como estimadores de cantidades topológicas que describen un objeto subyacente. Los objetivos principales de este nuevo enfoque siguen las siguientes líneas:

- Estudiar razones de convergencia para métodos TDA
- Proporcionar regiones de confianza para características topológicas y discutir las importancias de la cantidades topológicas estimadas.
- Seleccionar escalas relevantes en la que se consideran los fenómenos topológicos como una función de los datos observados.
- Trabajar con valores extremos y proporcionar métodos robustos para TDA
- Proporcionar enfoques funcionales para poder inferir con mayor facilidad estadísticos sobre los descriptores.

El presente reporte continúa de la siguiente forma: en la sección 2 se introducen algunos trabajos en los que ha sido aplicado TDA como herramienta en diferentes campos de manera exitosa. Luego en la sección 3 presenta algunos métodos para transformar nubes de puntos en complejos, es decir obtener una representación discreta de espacios topológicos conocida como *complejos simpliciales*. En la sección 4 son estudiados aspectos estadísticos relacionados con la homología persistente, donde se expone un nuevo enfoque para este campo emergente. Luego, las secciones 5 y 6 muestran aspectos y propiedades importantes de los descriptores topológicos introducidos en la sección 4 para su trabajo desde un punto de vista probabilístico. Más tarde en la sección 7 se expone la técnica bootstrap de gran utilidad para obtener estimaciones del error estadístico, y calcular intervalos de confianza sobre los resúmenes.

La sección 8 muestra algunos resultados en el área de prueba de hipótesis sobre TDA. En los capítulos 9 y 10 se abordan representaciones funcionales de los descriptores topológicos estándar con el objetivo de atacar algunos problemas que se presentan cuando son combinados con herramientas de la inferencia estadística; los enfoques expuestos son la persistencia landscape y la función rango. Luego en la sección

11 como parte de la búsqueda por axiomatizar de forma abstracta diversas estructuras matemáticas como una sola, mediante el uso de objetos y morfismos se introduce la *teoría de categorías* sobre la homología persistente. Para enfrentar algunos obstáculos que surgen al aplicar técnicas de TDA sobre problemas de alta dimensión, la sección 12 aborda ideas relacionadas con el método de submuestreo para persistencia homología persistente. Para finalizar en los últimos capítulos se introducen generalizaciones para la teoría de persistencia, en el capítulo 13 se expone la teoría de homología persistente *zigzag* y por último son comentados los *vineyards* en el capítulo 14, nueva teoría para el estudio de los diagramas de persistencia variables en el tiempo. Al final del reporte se encuentran en forma de anexos algunos aspectos básicos de la inferencia estadística, en donde se producen la bases teóricas de esta nueva rama; el lector puede consultar las definiciones de los conceptos utilizados en caso de ser necesario.

2. Análisis Topológico de Datos en datos reales

El Análisis Topológico de Datos ha sido aplicado sobre datos científicos y provenientes de contextos muy variados. Algunos de los ámbitos en los que ya se ha usado son:

- **Biología**: Actualmente la Biología es probablemente el mayor campo de aplicación de TDA. Existe una amplia literatura que utiliza homología persistente y el Algoritmo Mapper para analizar diferentes tipos de datos biológicos.
- **Neurociencia**: TDA posee un potencial significativo en el estudio de datos complejos que surgen de los laboratorios de neurociencia. Una parte importante de las investigaciones en este campo consiste en estudiar las redes, y las redes son particularmente susceptibles a las herramientas topológicas.
- **Química**: En los laboratorios de Química Analítica o Química Física se generan grandes volúmenes de datos. En este sentido se desarrollan nuevas herramientas para explorar y valorizar tales conjuntos de datos y algunos trabajos muestran que el concepto topológico es útil para los diferentes análisis necesarios.
- **Ciencia de los materiales**: La homología persistente ha encontrado recientemente algunas aplicaciones en el estudio de las estructuras de los materiales; (por ejemplo en materiales amorfos [4])
- **Minería de datos**: El análisis de datos espacio-temporales y la minería de datos experimenta un crecimiento debido a la creciente disponibilidad y conocimiento de una gran cantidad de conjuntos de datos. En la actualidad se realizan avances para modelar y representar los fenómenos, por ejemplo análisis de redes sociales, utilizando técnicas que ofrece TDA. A pesar de que existen resultados experimentales sigue siendo una dirección de investigación ampliamente inexplorada.
- **Visión por computadora**: Se trabaja sobre posibles enfoques que combinan las herramientas de TDA con *machine learning* para problemas prácticos en la visión por computadora, estos análisis introducen una nueva área de investigación dentro del Análisis Topológico de Datos conocida como *Topological Computer Vision*.

A continuación se exponen algunas aplicaciones exitosas en diferentes campos:

2.1. Obstrucción topológica de Filogenia

En [5] Chan y Carlsson realizan una aplicación de TDA en filogenética. La filogenia es la historia del desarrollo evolutivo de un grupo de organismos y la filogenética molecular intenta reconstruir este proceso evolutivo a partir de las secuencias de ADN de los individuos. El método que proponen es el uso de TDA para distinguir que tan cercano está un espacio métrico de ser aditivo. No proponen un árbol o

una red, sino que describen las propiedades topológicas del proceso evolutivo y exponen un estimador para la taza de eventos horizontales, logrando capturar eventos horizontales complejos (aquellos en los que intervienen más de dos especies). La hipótesis principal del método es que encontrar agujeros en la estructura topológica asociada a los datos implica la existencia de eventos reticulares.

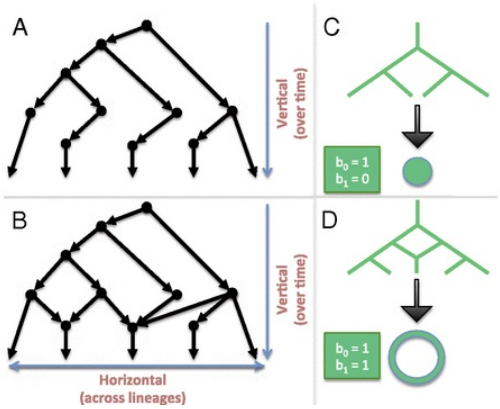

Fig. 1. Vinculación de la topología algebraica a la evolución. (A) un árbol que representa la evolución vertical, (B) una estructura reticulada captura la evolución horizontal, (C) un árbol se puede comprimir en un punto, (D) lo mismo no se puede hacer para un estructura reticulada sin destruir el agujero en el centro.

2.2. Análisis de superficies 3D

En [6] se presenta una investigación sobre descriptores topológicos para el análisis de superficies 3D con el objetivo de su descripción y clasificación de acuerdo con su micro-estructura geométrica. Se investigan diferentes descriptores topológicos y se analiza su capacidad para discriminar de forma estructural diferentes parches de superficies 3D.

Los espacios topológicos que aparecen en el análisis de datos son construidos de pequeñas piezas. Una herramienta natural en el estudio de imágenes multidimensionales con métodos topológicos son los hiper-cubos(puntos, bordes, cuadrados, cubos, etc...), por ejemplo un pixel en un imagen 2-dimensional es equivalente a un cuadrado y un voxel en un volumen 3-dimensional es equivalente a un cubo. Tales representaciones se convierten en una herramienta natural en el estudio de conjuntos de datos multidimensionales.

En el artículo se realizan una serie de experimentos donde se investiga sobre la robustez de los descriptores topológicos para el análisis de superficies 3-D y se compara y combina con descriptores tradicionales no topológicos. Se muestra que los descriptores topológicos contribuyen con información adicional valiosa a los descriptores anteriores y mejoran la precisión de clasificación cuando se combina con descriptores no topológicos. Se llega a la conclusión de que los descriptores topológicos son complementarios a los descriptores de imágenes tradicionales y representan la información necesaria para obtener el máximo rendimiento en clasificación que lo esperado.

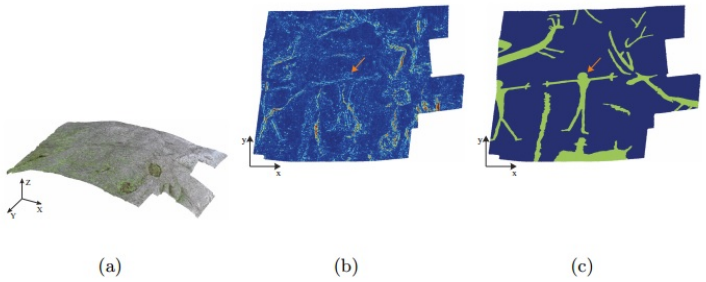

Fig. 2. (a) Nube de puntos 3D de la superficie, (b) La proyección profundidad de la superficie con curvatura global, (c) etiquetado real que especifica las zonas con diferente topografía, tales como la figura con forma humana en el centro cuya cabeza está marcada con una flecha.

2.3. Aplicación sobre diagramas de fases en aleaciones metálicas

En el artículo [7] se estudia la dinámica de separación de fases en aleaciones metálicas binarias como se describe en el modelo de Cahn-Hilliard-Cook estocástico. Los autores proponen la persistencia landscape (representación funcional de los diagramas de persistencia) como una métrica topológica para analizar la información de conectividad en micro-estructuras. Utilizando la persistencia landscape se puede obtener una secuencia de objetos discretos que caracteriza la evolución de la topología y se demuestra que el promedio landscapes puede ser utilizado para recuperar información en la teoría de Cahn-Hilliard de separación de fases. En el artículo se demuestra que cuando se trabaja en un marco estocástico y evolucionando el tiempo, la información topológica codifica mucho más de lo previsto. Debido a que el modelo es de naturaleza estocástica, se captura el comportamiento típico realizando un promedio de la persistencia landscape, donde se demuestra que este codifica información suficiente para tomar las decisiones.

La información topológica de la evolución de las micro-estructuras sólo es suficiente para detectar la información de concentración y la etapa de descomposición real de los datos. Los resultados indicaron que los parámetros del sistema en un proceso de separación de fases afectan a la topología considerablemente más de lo previsto. La Homología como conocemos es un camino para cuantificar los espacios topológicos a través de una secuencia de números enteros en su forma más reducida. Este estudio es el primero en utilizar la información homología en el contexto de la validación del modelo. Basado en los datos experimentales se demostró que si el ruido en el sistema es demasiado bajo, los números de Betti observados en la evolución son cualitativamente diferente de los experimentales. Además se demuestra que mientras los números de Betti se pueden utilizar para separar grandes cantidades en un comportamiento límite, la característica de Euler promediada sólo puede describir los efectos de contorno.

2.4. Desarrollo de un marco estadístico en Análisis Topológico de Datos para ser aplicado en datos reales

En [8] se propone un método núcleo en los diagramas de persistencia para desarrollar un marco estadístico en TDA (ver fig:3), el núcleo propuesto satisface las propiedades de estabilidad con respecto a la distancia. El método se aplica en datos prácticos sobre proteínas y vidrio de óxidos, los resultados muestran ventaja en comparación con otros métodos existentes.

Dado que un diagrama de persistencia es un conjunto de puntos de tamaño variable, no es sencillo aplicar métodos de análisis de datos estadísticos, que normalmente asumen los datos vectoriales. Los autores proponen un núcleo para diagramas de persistencia, llamado *Persistence Weighted Gaussian Kernel* (PWGK), la vectorización de los diagramas de persistencia permite aplicar cualquier método núcleo para diagramas de persistencia. El núcleo utilizado permite controlar el efecto de la persistencia en el análisis de datos y desde el punto de vista de los cálculos proporciona una aproximación precisa y eficiente para calcular la matriz de Gram, adecuado para aplicaciones prácticas en TDA.

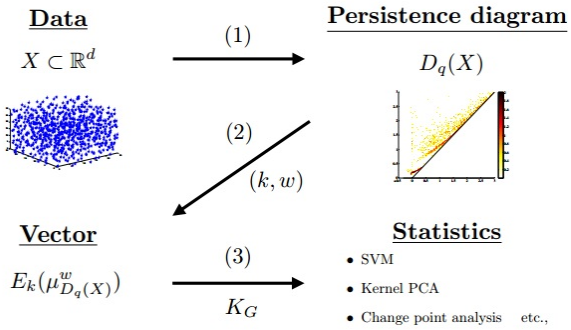

Fig. 3. Una data X se transforma en un diagrama de persistencia $D_q(X)$, (2) $D_q(X)$se asigna a un vector $E_{k_G}(\mu^{w_{arc}}_{D_q(X)})$, donde k es el núcleo y w es el peso controlando el efecto de persistencia. Este vector proporciona método estadístico para diagramas de persistencia.

2.5. Reconocimiento de acciones humanas en video

Los autores en [9] proponen un nuevo marco para el análisis dinámico de las acciones humanas a partir de los datos de captura de movimiento en 3D usando TDA. La tarea de reconocer las actividades humanas tiene una amplia gama de aplicaciones tales como vigilancia, seguimiento de la salud y la animación. El modelado de la evolución espacio-temporal de las articulaciones del cuerpo humano se realiza tradicionalmente mediante la definición de un espacio estado y el aprendizaje de una función que transforma el estado actual al estado siguiente. En el artículo se trabaja con un espacio de fases reconstruido a partir de los datos de series temporales, que preserva las propiedades topológicas del sistema dinámico subyacente de una acción determinada; tratan el atractor reconstruido como una nube de puntos y extraen las características topológicas de la nube de puntos basado en homología persistente. Además incorporan relaciones entre los puntos de tiempo adyacentes en la construcción del complejo simplicial de la nube de puntos.

El enfoque propuesto aborda los inconvenientes de los métodos tradicionales, mediante la combinación de los principios del análisis de series de tiempo no lineales y TDA, para extraer características robustas y discriminativas del espacio de fase reconstruido.

2.6. Algoritmo Mapper para identificar un subgrupo del cáncer de seno

En [10] se introduce un método que extrae información de los datos microarrays de alto rendimiento y mediante el uso de la topología proporciona un mayor detalle de la información que las técnicas analíticas actuales. El método denominado Análisis de Progresión de la Enfermedad (PAD), identifica aspectos robustos del análisis de cluster, luego encuentra características biológicamente significativas en estos datos

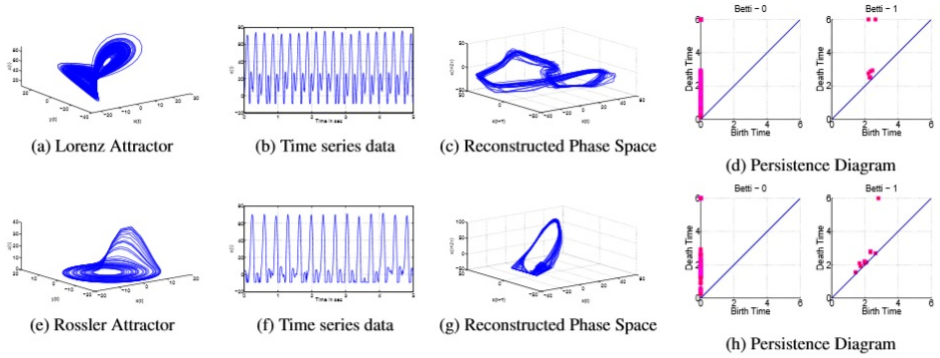

Fig. 4. Reconstrucción del espacio fase de atractores dinámicos mediante la incorporación de retraso. a) y e) muestra la vista 3D de las trayectorias de los atractores Lorenz y Rossler. Este ejemplo muestra que la reconstrucción del espacio fase preserva ciertas propiedades topológicas del atractor original.

y aporta una imagen sencilla o gráfico para explorar aún más estos datos. En el artículo se utiliza como un ejemplo para analizar la progresión del cáncer de mama. Mediante la preservación de la geometría de los datos, PAD identifica un subconjunto único de los cánceres de mama que presenta características clínicas claras y coherentes. El método tiene la capacidad de capturar los detalles, incluso, en un gran conjunto de datos, en situaciones en la que los métodos tradicionales tienden a desaparecer esos detalles en cuestión y se puede aplicar a cualquier situación en la que exista una noción de similitud o proximidad, no sólo en los datos euclidianos.

El método es una aplicación del Algoritmo Mapper [2]; herramienta matemática que identifica la forma de un conjunto de datos a lo largo de una función filtro preasignada. La idea principal es identificar las agrupaciones locales dentro de los datos y luego comprender la interacción entre estos pequeños grupos al conectarlos donde se forma un gráfico cuya forma captura aspectos de la topología del conjunto de datos. La figura:5, ilustra como la construcción resulta un conjunto de puntos con una forma más o menos circular en un gráfico. Claramente formas similares tienen gráficos similares, incluso cuando la forma es un tanto distorsionada.

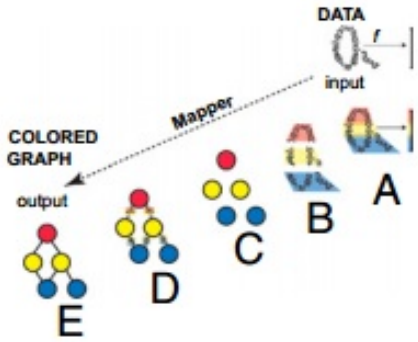

Fig. 5. Mapper comienza con un conjunto de puntos de datos y una función filtro f, produce un gráfico de color que captura la forma de los datos. (A) La imagen de la función f se subdivide en intervalos superpuestos. (B) cada pieza es agrupada por separado, (C) cada agrupación es representada por un disco coloreado, un bin de puntos, (D) identifica pares de bins que tiene puntos en común, (E) conecta pares de bins que tienen puntos en común por un borde.

2.7. Homología Persistente en redes ponderadas

Dentro del campo de la neurociencia en [11] se describe por primera vez una variación de la homología persistente que permite tratar con redes ponderadas, se conoce como la homología *scaffolds* que ofrece una nueva medida de la importancia topológica de bordes en el sistema original en cuanto a la frecuencia con que son parte de los generadores de los grupos de homología y como la persistencia son los generadores a la cual pertenecen. Aplican el método a un conjunto de datos de resonancia magnética funcional que comprende un grupo de sujetos inyectados con una placebo y otro inyectado con psilocibina. El análisis de la homología *scaffolds* revela la existencia de un conjunto de bordes que son predominantes en términos de su persistencia a pesar de que son estadísticamente parte del mismo número de ciclos en las dos condiciones.

2.8. Estudio de imágenes cerebrales y autismo

Chung y colaboradores presentan en [12] un nuevo enfoque para caracterizar señales en imágenes, utilizando técnicas de la topología algebraica computacional.
El método que se propone, utiliza todos los valores críticos locales en la caracterización de la señal y al hacerlo ofrece un nuevo marco de reducción y análisis de datos para la cuantificación de la señal. Se aplica el método para señales simuladas unidimensionales y datos de grosor cortical 2D. Dado que el grosor cortical es muy ruidoso se aplica un núcleo de suavizado para eliminar el ruido de alta frecuencia espacial antes de la filtración. Este es el primer trabajo que aplica el concepto de homología persistente para datos de imágenes médicas.

El método es aplicado tanto en datos de neuroimágenes reales como en simuladas, donde para la simulación utilizan el ruido gaussiano 1D. Los datos de neuroimagen 2D proviene de un estudio de resonancia magnética, donde el interés se centra en la cuantificación del patrón de grosor cortical anómalo en sujetos autistas, si existe. Se demuestra que existen patrones de homología persistente únicos para el grupo de autismo.

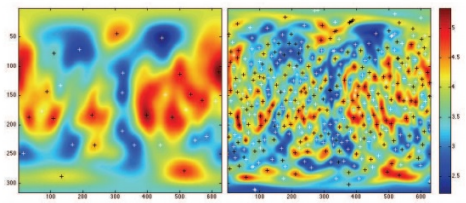

Fig. 6. Mapas planos de grosor cortical a diferentes escalas de suavizado. Los máximos y mínimos se denotan con cruces negras y blancas respectivamente. El suavizado se realiza a lo largo de la esfera unidad usando los ángulos asociados a una 2-esfera, lo que produce menos cantidad de puntos críticos y a su vez menor número de emparejamientos.

2.9. Proteínas

En [13] utilizan TDA para el estudio de la proteína *maltose-binding*(MBP) la cual se encuentra en el Escherichia coli. Un ejemplo de la proteína lo podemos ver en la figura:7. La proteína es una estructura dinámica y los cambios en su estructura son de relevancia biológica; puede estar en una conformación abierta o cerrada. El objetivo de los autores en el artículo es clasificar el estado de la proteína.

Cada proteína es representada por 370 puntos (correspondiente a los aminoácidos) en un espacio 3D. Los autores construyen un modelo dinámico de la estructura de la proteína (ya que la estructura cambia en el tiempo) a partir de la que definen distancias dinámicas entre los 370 puntos. Entonces una proteína es representada por una matriz distancia ($370x370$). A partir de la matriz distancia se construye un diagrama de persistencia. En el siguiente paso se convierte el diagrama de persistencia en un conjunto de funciones llamadas *landscapes* como se define en [14]. Al convertir el diagrama en un conjunto de funciones unidimensionales, permite que sea más fácil utilizar herramientas estadísticas. En el trabajo se realiza una prueba de permutación de dos muestras usando las distancias integradas entre las funciones *landscapes* como una prueba estadística. El p-valor es $5{,}83x10^{-4}$, lo que sugiere una diferencia entre las conformaciones abiertas y cerradas. Esto sugiere que los *landscapes* pueden ser utilizados para clasificar proteínas como abiertas o cerradas. También muestran que ciertos sitios en la proteína, conocidos como sitios activos, están asociados con bucles en la proteína.

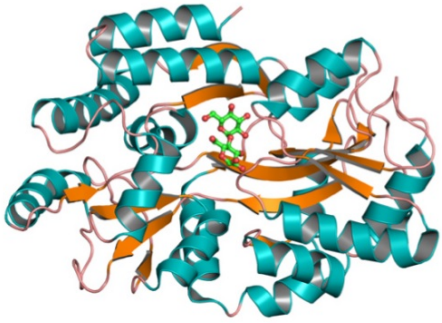

Fig. 7. Una proteína *maltose binding*(MBP).

2.10. Aplicación sobre seguimiento de objetos en video

El trabajo [15] presenta una teoría unificada para el seguimiento de objetos: utilizando Seguimiento de múltiples hipótesis (MHT), Análisis Topológico de Datos (TDA) y *machine learning*. Realizan una serie de innovaciones como son el uso de características topológicas robustas para codificar la información del comportamiento donde modelos estadísticos se ajustan a distribuciones sobre estas características. La idea principal es usar medidas topológicas del comportamieno para reducir *tracklets* improbables en el algoritmo MHT. TDA no sólo nos ofrece la clasificación del comportamiento objetivo, sino también ayuda a resolver un problema más difícil, el seguimiento de objetivos: conectar los puntos mediante la asociación, con los datos objetivos.

Para cada *tracklet* asocian funciones que describan el comportamiento. En el artículo se centran en

la velocidad, aceleración y giro, pero en un marco general trabajan para cualquiera de las funciones. Dado dos *tracklets* T_1 y T_2 utilizan estas funciones para conocer si los agentes asociados a T_1 y T_2 son los mismos.

Por supuesto, existen diferentes maneras para resumir los datos funcionales: valores críticos, variación total entre otros. La solución propuesta consiste en considerar un diagrama de persistencia (PD) que proporciona una imagen de los datos funcionales estables al ruido, fácil de calcular, y captura las características importantes de cada función sin necesidad de alineaciones. Como conocemos los diagramas de persistencia son una de las principales herramientas de TDA, adapta métodos de topología algebraica para encontrar la estructura de los conjuntos de datos complejos. Una vez que los diagramas han sido calculados, la pregunta que nos surge es ¿cómo interpretarlo?. En el trabajo se propone un método de *machine-learning* aplicado para interpretar los PD en un contexto estadísticon y clasificarlos en tipos de comportamiento.

En el trabajo se demuestra la utilidad de los métodos de TDA mediante su integración con un programa de seguimiento MHT existente, donde se puede notar la mejora apreciable de los resultados.

En la figura podemos notar que MHT con características topológicas es capaz de corregir el error asociado después de la intersección. Se estima el comportamiento del conductor antes de la intersección y que el comportamiento que surgió después de la intersección se puede utilizar la información del comportamiento para asociar correctamente los datos. Sin las características topológicas el *tracklet* no podría corregirse a sí misma.

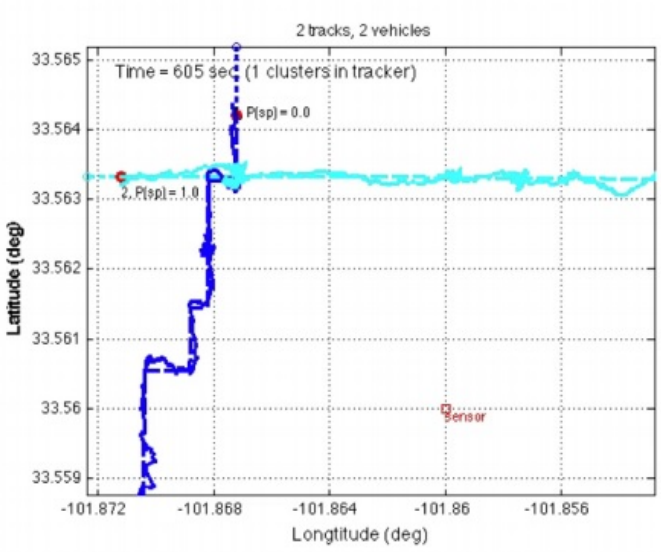

Fig. 8. Seguimiento luego de interceptarse 2 vehículos.

2.11. Datos financieros

En [16] aplican la teoría TDA para el análisis de datos financieros, usando los datos generados por un método general de matemática financiera. Se investigó la relación entre los descriptores topológicos de

TDA y las medidas de riesgo financieras tradicionales, como las tasas de crecimiento, la volatilidad y los coeficientes de correlación. Se estudió el espacio utilizado en la fijación de precios de las opciones europeas donde los resultados apoyaron la eficacia de TDA como medida de riesgo. De manera particular los códigos de barra pueden detectar el cambio rápido en un corto período de tiempo. En este sentido TDA ofrece una nueva naturaleza que es diferente de las medidas de riesgo tradicionales.

2.12. Otras aplicaciones

En el reporte son mencionados brevemente algunos ejemplos de aplicaciones en diferentes áreas. La característica de Euler es una cantidad topológica que ha desempeñado un papel importante en varios aspectos de la probabilidad, así como para aplicaciones en astrofísica y neurociencia [17,18]. También se ha utilizado para la clasificación de formas [19]. En [20] utilizan métodos topológicos para estudiar las interacciones entre los sistemas de raíces de las plantas. Carstens en [21] utiliza la homología persistente para describir la estructura de las redes de colaboración. En el artículo [22] utilizan TDA en el análisis de biomoléculas. En [23] se introduce el rol que cumple TDA en la Quimiometría. En la actualidad existe una amplia literatura sobre las aplicaciones de TDA en la neurociencia incluidas [24,25,26,27,28]. El sitio web `http://www.chadgiusti.com/algtop-neuro-bibliography.html` mantiene una bibliografía de referencia en este campo.

Las matemáticas aplicadas enfrentan el problema de la complejidad computacional; un problema en relación con el cálculo de la homología persistente es la elección del complejo, donde se siguen estudiando formas eficientes para reducir el costo computacional. Debido a esto antes de indagar en el enfoque estadístico sobre TDA, se deben introducir los métodos que existen para obtener los *complejos simpliciales*.

3. De la nube de puntos a los complejos

En Análisis Topológico de Datos se asume que los datos son muestreados del espacio subyacente $\mathbb{X}$ y el objetivo principal es recuperar la topología de $\mathbb{X}$. El proceso generalmente sigue los dos pasos siguientes donde el segundo resulta ser el más complejo:

1. Aproximar $\mathbb{X}$ utilizando una estructura combinatorial por ejemplo: complejos simpliciales.
2. Utilizar técnicas de la topología algebraica para calcular invariantes topológicos de estas estructuras por ejemplo: homología persistente [3,29] .

Existen numerosos métodos para completar el primer paso del análisis; separados en geométricos y algebraicos. Los complejos *Cech* y *Vietoris-Rips* son los métodos algebraicos más comunes, sin embargo la complejidad del cálculo resulta igual a la multiplicación de matrices; el primero a menudo puede ser aproximado por el segundo, lo cual es relevante por su análisis en conjunto de datos de altas dimensiones. Se han estudiado formas eficientes para reducir el costo computacional introduciendo métodos geométricos: *los complejos alpha* [30], *los complejos flow* [31]. Tienen la ventaja de ser métodos rápidos y relativamente pequeños, pero desafortunadamente dependen de los complejos Delaunay [32]. Otro método popular es el trabajo con *witness complex* [33]. A continuación son abordados temas relacionados con los dos métodos algebraicos más utilizados, construidos sobre puntos aleatorios (i.i.d) en un espacio euclidiano $\mathbb{R}^d$. Luego son explicados los de carácter geométricos, menos usados en la actualidad.

Un problema principal al usar herramientas de homología simplicial para estudiar una base de datos $\mathbb{X} = \{x_i\}_{i=1}^m \subset \mathbb{R}^n$ es que no se dispone de una estructura simplicial a priori. Tratar de construir un complejo simplicial a partir de $\mathbb{X}$ puede resultar difícil. Una primera estrategia es la de considerar la homología de los espacios $\mathbb{X}_\varepsilon = \bigcup_{i=1}^m B(x_i, \varepsilon)$, donde una bola de radio ε es centrada alrededor de cada punto en $\mathbb{X}$. La unión de bolas $\mathbb{X}_\varepsilon$ constituye un buen descriptor combinatorial, el proceso se puede observar en 9 .

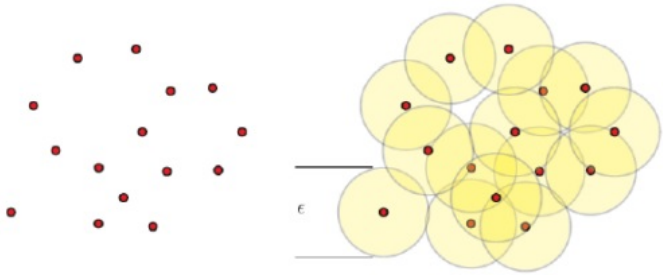

Fig. 9. Descriptor combinatorial.

En un nube de puntos $\mathbb{X}$, a pesar de que el parámetro ε es continuo se puede verificar que en realidad existe sólo un número finito de complejos simpliciales $K_1 \subset K_2 \subset \cdots K_r$(concepto de filtración)[34] que se puede construir a partir de $\{X_\varepsilon | \varepsilon > 0\}$ un ejemplo se puede ver en la figura: 10.

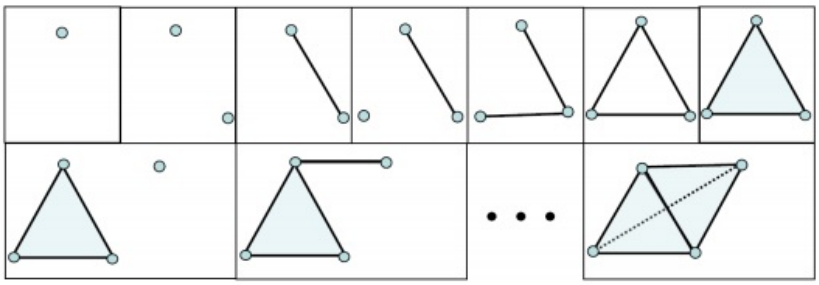

Fig. 10. Esquema de la filtración de un complejo.

Por otra parte una de las aplicaciones de la homología persistente es construir complejos simpliciales a partir de nubes de puntos en $\mathbb{R}^n$. Si $\mathbb{X} = \{x_i\}_{i=1}^m \subset \mathbb{R}^n$ para cada valor de $\varepsilon > 0$ son calculados $H(\mathbb{X}_\varepsilon)$, la homología del complejo simplicial resultante. Mientras incremente ε la unión de bolas crece, y las inclusiones resultantes inducen una correspondencia entre homología de grupos. Es decir si $\varepsilon \leq \delta$, la inclusión $i_\varepsilon^\delta : \mathbb{X}_\varepsilon \to \mathbb{X}_\delta$ induce un mapeo entre los grupos de homología: $H(i_\varepsilon^\delta) : H(\mathbb{X}_\varepsilon) \to H(\mathbb{X}_\delta)$ denominado homomorfismo de inclusión. Cada nivel de la filtración contiene su propia homología y las inclusiones relacionan las homologías de los distintos niveles. Los grupos de homología persistentes contienen clases homológicas que son estables en el intervalo de ε a δ: tales clases nacen en un tiempo no posterior al ε, y siguen vivas en δ. Sea $\delta = \varepsilon + p$; las clases de homología persistentes que permanecen vivas para grandes valores de p detectan características topológicas estables de $\mathbb{X}$, mientras que las clases que sobreviven sólo para pequeños valores de p son inestables ó componentes topológicas tipo ruido.

3.1. Métodos algebraicos

3.1.1. Complejos de Cech

El Complejo de Cech generado por un conjunto de puntos $\mathbb{X}$ es un complejo simplicial formado por vértices, bordes, triángulos y caras de altas dimensiones. La definición general es bastante amplia, en la mayoría de los artículos revisados se trabaja un caso especial expuesto en [35] que usa la intersección de bolas euclidianas suficientes para el posterior análisis.

***Definición* 1** *(**Complejos de Cech**) Sea $\mathbb{X} = \{x_1, x_2, \cdots, x_n\}$ un colección de puntos en $\mathbb{R}^d$ y sea $\varepsilon > 0$. El complejo $\hat{C}_\varepsilon(\mathbb{X})$ de un conjunto de bolas $\{B_{x_i}(\varepsilon)\}$ se construye como sigue:*

1. *Un 0-símplice (vértices) son los puntos en $\mathbb{X}$*
2. *Un k-símplice $[x_{i_0}, \cdots, x_{i_k}]$ es un $\hat{C}_\varepsilon(\mathbb{X})$ si $\bigcap_{n=0}^{k} B_\varepsilon(x_{i_n}) \neq \emptyset$ (los k-símplices corresponden a $k+1$ bolas con intersección no vacía)*

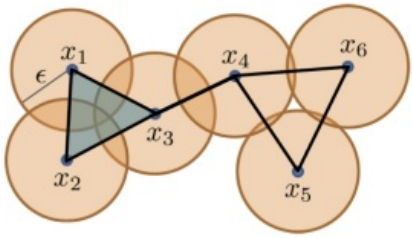

Fig. 11. $\hat{C}_\varepsilon(\mathbb{X})$ para $\mathbb{X} = \{x_1, x_2, \cdots, x_6\}$ y $\varepsilon > 0$. El complejo contiene 6 vértices, 2 bordes y un triángulo.

Un paradigma en TDA es crear $\hat{U}_\varepsilon(\mathbb{X})$(conjunto de vecindades) como una estimación de alguna subvariedad subyacente, $\mathcal{M} \subset \mathbb{R}^d$, a partir del cual $\mathbb{X}$ es la muestra y luego considerar su homología, normalmente a través de lo que se conoce como la homología persistente o por medio de números Betti como se aborda en [36]. Los lectores interesados en la teoría de homología pueden referirse a [37,38]. Un resultado importante en esta área es el Lema del Nervio [39], une el Complejo de Cech $\hat{C}_\varepsilon(X)$ y el conjunto de vecindades $\hat{U}_\varepsilon(X)$ y desde un punto de vista de TDA afirma que son equivalentes homotópicamente, y en particular tienen los mismos grupos de homología. Sin embargo dado que la definición de Complejo de Cech es esencialmente combinatoria, es computacionalmente más accesible y por lo tanto de mayor uso en las aplicaciones. El interés en trabajar con los complejos de Cech es debido a que primeramente es un complejo de alta dimensión análogo de un grafo geométrico; un estudio extenso sobre grafos geométricos aleatorios lo podemos ver en [40]. Además resulta menos complejo examinar el Complejo de Cech en lugar de la estructura geométrica $\hat{U}_\varepsilon(\mathbb{X})$. La topología de Cech esta estrechamente relacionada con la topología del *complejos alpha*. Sin embargo cuando la dimensión es mayor que 3 su cálculo se convierte en poco práctico.

Otro método utilizado, estrechamente vinculado con el anterior y más fácil para calcular la filtración es *Vietoris-Rips*.

3.1.2. Complejo Vietoris-Rips

En ocasiones no es factible calcular complejos Cech en la práctica y los *complejos alpha* sólo pueden calcularse de manera eficiente en la dimensión 3 o menor. Los complejos Vietoris-Rips fueron introducidos por Vietoris en [41] con el fin de extender la homología simplicial a una teoría de la homología de espacios

métricos más generales. Aunque generalmente no son tan rápidos como los *complejos alpha* en bajas dimensiones, su cálculo puede ser eficiente en altas dimensiones.

Definición* 2** ***(Complejos Vietoris-Rips) *Sea $X = \{x_1, x_2, \cdots, x_n\}$ un colección de puntos en $\mathbb{R}^d$ y sea $\varepsilon > 0$. El complejo $\hat{R}_\varepsilon(X)$ se construye como sigue:*

$$\sigma = \{x_1, x_2, \cdots, x_k\} \in \hat{R}_\varepsilon(X) \;\leftrightarrow\; ||x_i - x_j|| \leq \varepsilon \;\; \forall\, i, j \in \{1, \cdots, k\}\,. \tag{1}$$

La definición de los complejos Cech y Vietoris-Rips no se limita solo al caso de espacio euclidianos, se pueden definir para un conjunto de puntos en cualquier espacio métrico. De hecho en[42] se extiende para cualquier espacio métrico. Por otra parte satisface la propiedad siguiente que juega un papel importante en TDA; para mayor detalle se puede revisar [43].

***Lema* 1** *Sea $\mathbb{X}$ un conjunto finito de puntos en $\mathbb{R}^d$ y $\varepsilon \geq 0$. Entonces existe una cadena de mapas de inclusión[44].*

$$\hat{R}_\varepsilon(X) \to \hat{C}_{\sqrt{2}\varepsilon}(X) \to \hat{R}_{\sqrt{2}\varepsilon}(X). \tag{2}$$

Esto significa que una propiedad topológica que persiste bajo la inclusión $\hat{R}_\varepsilon(\mathbb{X}) \to \hat{R}_{\varepsilon'}(\mathbb{X})$ con $\varepsilon' \geq \sqrt{2}\varepsilon$ es una característica topológica de $\hat{C}_{\varepsilon'}(\mathbb{X})$. La idea principal es que la información sobre las características topológicas que persisten bajo la inclusión anterior revelan mayor información que las dos por separado. En [45] se analiza desde un punto de vista computacional, que el complejo de Rips es menos costoso que el correspondiente complejo de Cech, a pesar de tener más símplices.

Un aspecto insatisfactorio de los resultados anteriores es la dependencia de un conocimiento a priori de la escala característica ε. Una manera de manejar el hecho de poder realizar una buena elección del ε es considerar invariantes homológicos multiescala que codifican los cambios en forma de homología, cuando ε varía. En [45] realizan un estudio de como seleccionar el parámetro ε; para ε suficientemente pequeño, el complejo es un conjunto discreto y para ε suficientemente grande, el complejo es un símplice de alta dimensión.

Un problema que persiste es que los símplices pueden tener dimensiones muy altas. A continuación son abordados algunos complejos expuestos en [33,46,47] que surgen de las técnicas de la geometría computacional para atacar el problema.

3.2. Métodos geométricos

3.2.1. Diagrama de Voronoi y los Complejos Delaunay

Sea $\mathbb{X}$ un conjunto finito de puntos en $\mathbb{R}^d$. Como primer paso se define la celda Voronoi de un punto x en $\mathbb{X}$ siendo el conjunto de puntos $V_x \subseteq \mathbb{R}^d$ para x más cercana de los puntos en $\mathbb{X}$:

$$V_x = \left\{ u \in \mathbb{R}^d : ||u - x|| \leq ||u - x'|| \;\forall\, x' \in X \right\}. \tag{3}$$

Veamos que si x y x' son dos puntos en el plano entonces sus regiones Voronoi se intersectan a lo largo del punto medio de los dos puntos. Con n puntos diferentes la región de Voronoi de x se convierte en la intersección de las normales de punto medio de x y los otros puntos (ver fig: 12).

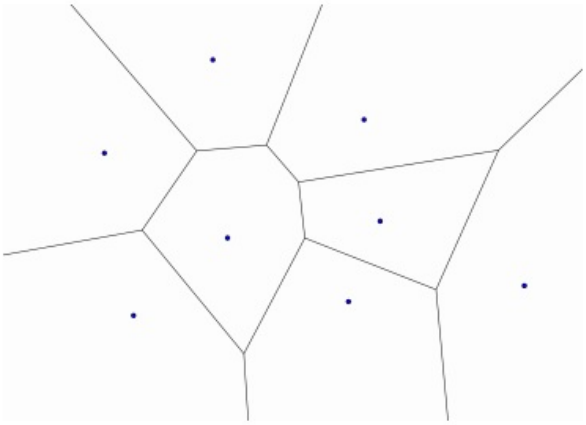

Fig. 12. Diagrama Voronoi de puntos en el plano.

Podemos ver que la unión de celdas Voronoi cubre $\mathbb{R}^d$ y se definen los diagramas Voronoi de $\mathbb{X}$ como una colección de celdas de sus puntos.

Definición* 3 *Complejo Delaunay
El Complejo Delaunay $\mathcal{D}(\mathbb{X})$ de un conjunto finito $\mathbb{X} \in \mathbb{R}^d$ se define como el nervio del diagrama de Voronoi.

A continuación introducimos una familia de subcomplejos de los Complejos Delaunay.

3.2.2. Los complejos Alpha

Estos complejos son similares a los Complejos Cech, pero difieren de ellos por tener realización geométrica natural. Sea $B(x,\varepsilon)$ bola cerrada en x con radio ε y se define $R(x,\varepsilon)$ como la intersección de las celdas Voronoi V_x con $B(x,\varepsilon)$

$$R(x,\varepsilon) = V_x \cap B(x,\varepsilon). \tag{4}$$

Sean x y $x' \in \mathbb{X}$ dos elementos cualquiera o bien tienen intersección disjunta o se solopan a lo largo de un pieza común de sus limítes. Por lo tanto veamos que la unión de $R(x,\varepsilon)$ $\forall x \in \mathbb{X}$ cubre la unión de las bolas cerradas.

Definición* 4 *Complejos Alpha

$$\mathcal{A}(X,\varepsilon) := \left\{ \sigma \in X \,\middle|\, \bigcap_{x\in\sigma} R(x,\varepsilon) \neq \emptyset \right\}. \tag{5}$$

Una explicación más detallada es dada en [47].

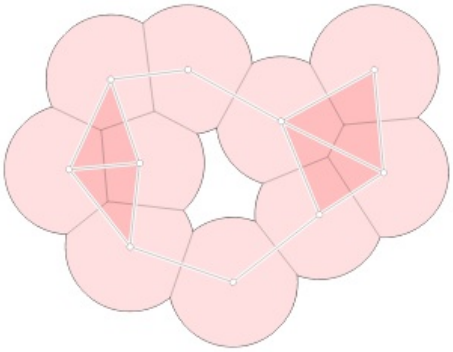

Fig. 13. La unión de bolas se descompone en regiones convexas por las celdas Voronoi.

3.2.3. Complejos Witness

Los *complejos alpha* requieren del cálculo del diagrama de Voronoi, esto significa que la complejidad depende del espacio ambiente, así como del número de puntos. Una posible solución a esto es utilizar los ***complejos witness*** donde para hacer frente al problema de la cantidad de puntos se introducen los puntos *landmark*

La idea es seleccionar un subconjunto de puntos $L \subset \mathbb{X}$ que capture con mayor precisión posible la topología de los datos originales. Existen dos enfoques para la selección de puntos *landmark*. El enfoque simple es elegir un subconjunto de $\mathbb{X}$ al azar y un enfoque más sofisticado es el método secuencial max-min. El algoritmo elige puntos tales que la distancia desde los datos originales a los puntos *landmark* es mínima.

Sea L_{i-1} el conjunto de los primeros $(i-1)$- puntos *landmark*, entonces el punto *landmark* i-ésimo es el $x \in \mathbb{X}$ que maximiza $||l-x||$ sobre L_{i-1}. Los puntos *landmark* son de manera general uniformemente distribuidos, pero el método también tiende a escoger puntos extremales. A continuación definimos los *complejos witness* y algunas ventajas, detalles específicos se pueden encontrar en [33].

Definición* 5 *Complejos witness

$W(\mathbb{X}, L, \varepsilon)$ se define de tal manera que su conjunto de vértices es L, y para $k > 0$ y vértices l_i el p-símplice formado por $\{l_0, \cdots, l_p\}$ pertenece al complejo si todas sus caras están y si existe un punto $x \in \mathbb{X}$ tal que:

$$max\{||l_0 - x||, \cdots, ||l_p - x||\} \leq \varepsilon + m_k(x), \tag{6}$$

$m_k(x)$ es la distancia de un punto $x \in \mathbb{X}$ a sus $k+1$-puntos landmark cercanos (tener en cuenta que el punto x puede ser un punto landmark).

De acuerdo a [33] estos complejos se pueden calcular fácilmente, son adaptables a métricas arbitrarias, utilizan un número pequeño de celdas y no sufren el problema de la dimensionalidad. Se muestran algunos ejemplos donde la combinación de estos complejos con homología persistente es muy eficaz en la práctica, incluso para datos ruidosos.

Expuestas las diferentes vías tanto de manera algebraica cómo geométricas para obtener los complejos simpliciales, como paso siguiente al análisis se explica la homología persistente que estudia la

forma en que varía la homología en una filtración. A continuación se expone un enfoque estadístico sobre la temática, principal para el estudio de TDA.

4. Enfoque estadístico para homología persistente

TDA se refiere a un conjunto de métodos para encontrar la estructura topológica de los datos. Un rama del TDA es la homología persistente que aplicada a una filtración de un complejo simplicial, se encarga de llevar la cuenta del momento i en el que nace una clase de homología y del momento j en el que desaparece la misma clase; lo que conduce el análisis hacia los descriptores topológicos, por ejemplo los diagramas de persistencia. En el capítulo se introducen aspectos estadísticos relacionados con el tema.

La Homología simplicial es un formalismo matemático utilizado para resumir la conectividad global de un espacio topológico; asocia un espacio topológico dado con una sucesión de grupos abelianos(o en contextos más generales módulos o cualquier elemento sobre una categoría abeliana). Su principal motivación es detectar las componentes conexas, túneles, agujeros, etc., de un espacio topológico $\mathbb{X}$ con este objetivo surgen los grupos de homología. El p-ésimo grupo de homología $H_p(\mathbb{X})$ es el conjunto de las clases de equivalencia de ciclos que encierra los agujeros p-dimensional. El rango β_p es llamado el p-ésimo número Betti, en el caso de las tres primeras dimensiones existe una interpretación intuitiva ya que β_0 concuerda con el número de componentes conexas de S, β_1 con el número de agujeros y β_2 con el número de cavidades [48].

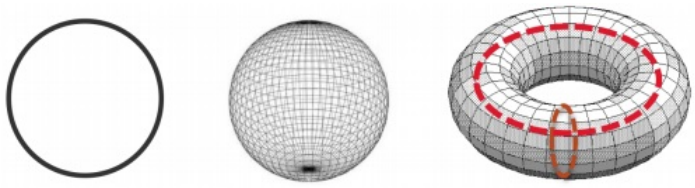

Fig. 14. El círculo tiene una componente conexa y un agujero ($\beta_0 = 1, \beta_1 = 1$). Una esfera en $\mathbb{R}^3$ tiene una componente conexa y una cavidad ($\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$). El toro tiene una componente conexa, dos agujeros(los dos círculos no equivalentes en rojo) y una cavidad encerrada ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$).

A continuación se expone una introducción a esta herramienta, más detalles son explicados en [48,49]. Dada una función de valores reales $f : \mathbb{X} \to \mathbb{R}$ definida para un subespacio triangulable de $\mathbb{R}^D$, la homología persistente describe los cambios en la topología de los conjuntos de nivel inferior(o superior) $f^{-1}(-\infty, t]$ cuando t aumenta de $-\infty$ a ∞ . Por ejemplo considerando los conjuntos de nivel inferior $L_t = \{x \in \mathbb{X} : f(x) \leq t\}$, el índice t se considera como un parámetro de escala que conduce la filtración de subespacios, tales que $L_t \subseteq L_s$, $\forall\, t \leq s$. Tal filtración induce una familia $\{H(L_t) : t \in \mathbb{R}\}$ de grupos de homología y la inclusión $L_t \to L_s$ induce una familia de homomorfismos $H(L_t) \to H(L_s)$.

La homología persistente describe f mediante los resúmenes topológicos que contienen la información topológica del nacimiento y muerte de las características homológicas que existieron durante algún intervalo de tiempo t. Las principales características de un conjunto incluyen las componentes conexas (homología de orden 0), los túneles (homología de orden 1), huecos (homología de orden 2), etc. Estas características aparecen y desaparecen cuando t aumenta. Por ejemplo las componentes conexas de L_t mueren cuando se combinan con otras componentes conexas.

Cada característica topológica tiene un tiempo de nacimiento b y un tiempo de muerte d. En general existe un conjunto de características con tiempos de vida y muerte asociados $(b_1,d_1),\cdots,(b_m,d_m)$. Estos puntos pueden ser ploteados en el plano resultando un diagrama de persistencia P. Alternativamente los pares (b_i,d_i) son presentados como intervalos $[b_i,d_i]$, donde el conjunto de intervalos ploteados se representa en un código de barra. Se consideran los diagramas de persistencia y los códigos de barra como resúmenes topológicos de la función de entrada o los datos. Puntos cercanos a la diagonal en el diagrama (es decir intervalos pequeños) tienen tiempo de vida cortos y se consideran ruidos topológicos. La mayoría de las aplicaciones se interesan en el características que podemos distinguir del ruido, es decir aquellas que persisten para un rango mayor de valores. Se han encontrado aplicaciones en diferentes campos, incluyendo neurociencia [50], bio-informática [51], clasificación de formas [52], agrupamiento [53], redes de sensores [54].

Entre los primeros resultados estadísticos sobre homología persistente en un entorno paramétrico lo encontramos en [55], donde suponen que los datos son muestreados aleatoriamente de una distribución de probabilidad desconocida, se construyen dos complejos de cadenas filtrados: el complejo Morse y el complejo Cech. Usando la homología persistente uno puede calcular los números Betti que proporciona una descripción topológica de la distribución de probabilidad. Se demuestra que con el uso de estimadores estadísticos para las muestras de ciertas familias de distribuciones se puede recuperar la homología persistente de la distribución subyacente. El enfoque dado en [56] desarrolla la teoría de probabilidad necesaria para definir objetos estadísticos básicos tales como medias, varianzas y probabilidades condicionales en el espacio de los diagramas de persistencia, donde demuestra que estos espacios con una métrica Wasserstein son completos y separables. Bubenick introduce en [14] una representación funcional de los diagramas de persistencia, denominado persistencia landscapes que veremos su desarrollo en 9.

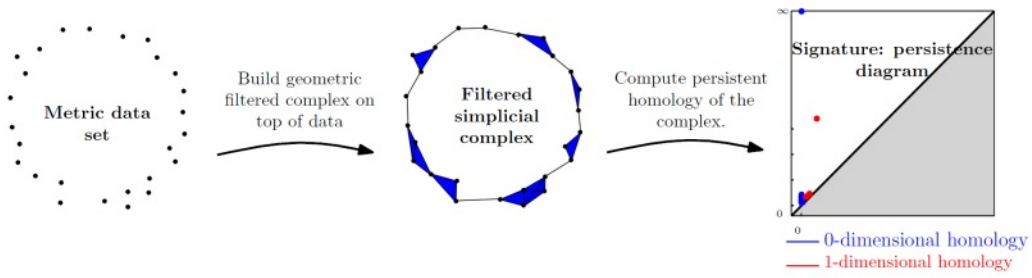

Fig. 15. Proceso clásico para persistencia en TDA.

A continuación son expuestos algunos detalles de la homología persistente para las funciones de distancia y funciones de densidad, que proporciona una mayor información para la construcción de los diagramas de persistencia [57].

4.1. Homología persistente de la función distancia

Primero consideramos el caso donde f es la función distancia. Dado los datos $S_n = \{X_1,\cdots,X_n\}$ interesa entender la homología del espacio topológico compacto d-dimensional $\mathbb{X} \subset \mathbb{R}^D$ de la que fueron muestreados los datos y sea $d_{\mathbb{X}} : \mathbb{R}^D \to \mathbb{R}$ la función distancia a $\mathbb{X}$, no negativa definida por:

$$d_{\mathbb{X}}(x) = inf_{y\in\mathbb{X}} d(x,y) \;\; \forall x \in \mathbb{R}^d. \tag{7}$$

Donde $d(x,y) = ||x-y||_2$ es la distancia euclidiana entre x e y en $\mathbb{R}^d$. La función de distancia a $\mathbb{X}$ es continua y en efecto 1-Lipstchiz: $\forall x,x' \in \mathbb{R}^d,\ \ |d_K(x) - d_K(x')| \leq ||x-x'||$. Además $\mathbb{X}$ es completamente caracterizada por $d_{\mathbb{X}}$ dado que $\mathbb{X} = d_{\mathbb{X}}^{-1}(0)$.

La homología persistente resume como las características topológicas de $L_t = \{x : d_{S_n} \leq t\}$ cambian en función de t. Cada característica tiene un tiempo de vida b y de muerte d. En general se tienen un conjunto de características con sus tiempos de vida y muerte asociados: $\{(b_i,d_i)\}_{i=1}^n$, estos puntos pueden ser ploteados en el plano y son obtenidos lo que se conoce como diagramas de persistencia P (resumen de la función de entrada o los datos).[48,58].

Si la muestra es lo suficientemente densa y el espacio topológico se encuentra en $\mathbb{R}^D$, entonces $F_p(\mathbb{X})$ es un subgrupo del grupo de homología p-ésimo del conjunto subnivel $\widehat{L}_t = \{x : D_{S_n}(x) \leq t\}$ para un intervalo de valores de tiempo t. Tarea difícil resulta seleccionar el correcto valor t: t pequeño tiene la homología de n puntos y un t mayor nos da la homología de un solo punto. Utilizando la homología persistente se evita la elección de un único t mediante la asignación de un valor de persistencia a cada generador de homología no trivial, que es realizado como $\widehat{L}_t$ para algún t no negativo. Como t varía, las características de nacimiento y muerte S_n son resumidas utilizando el diagrama de nacimiento empírico $\widehat{P}$. Donde $\widehat{P}$ es considerado como una estimación del diagrama de persistencia observado del espacio subyacente $\mathbb{S}$. Recordar que los puntos cerca de la diagonal tienen poco tiempo de vida y son considerados ruidos topológicos.

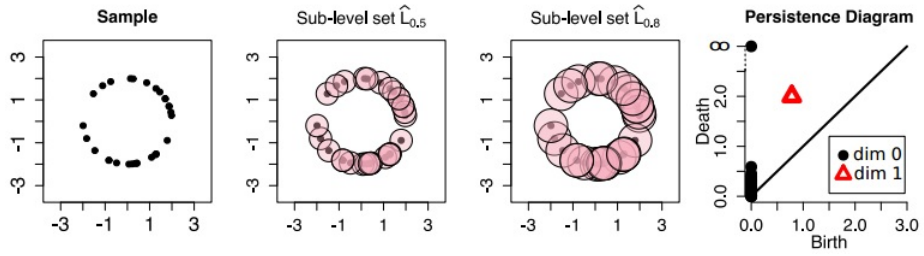

Fig. 16. Sample:S_{30} muestreado del círculo de radio 2; **Sub-level set** $\widehat{L}_{0,5}$: consta de 2 componentes conexas y 0 loops; **Sub-level set** $\widehat{L}_{0,8}$= $\{x : d_{S_{30}} \leq 0{,}8\}$, al aumentar t se observa el nacimiento y muerte de características topológica por ejemplo una de las componentes conectadas muere (se fusionó con otra) y se formo un agujero de dimensión 1, las bolas rosadas representan la función de distancia que se tocan entre sí en el centro del círculo. **Persistence Diagram** resume las características topológicas de los puntos muestreados. Los puntos negros representan las componentes conexas: 30 componentes conexas se presentan con $t = 0$ y mueren con el aumento de t, sólo una componente conexa persiste para grandes valores de t. El triángulo rojo representa el único agujero 1-dimensional que es formado con $t = 0{,}8$ y muere en $t = 2$.

4.1.1. Puntos críticos de la función distancia

Dado un conjunto compacto $\mathbb{X} \subset \mathbb{R}^d$, la función distancia es usualmente diferenciable. Por ejemplo si $\mathbb{X}$ es un cuadrado en el plano, $d_{\mathbb{X}}$ es no diferenciable a lo largo de la diagonal de $\mathbb{X}$. Sin embargo, es posible definir un campo de vector gradiente generalizado $\nabla_{\mathbb{X}} : \mathbb{R}^d \to \mathbb{R}^d$ para $d_{\mathbb{X}}$ que coincide con el gradiente clásico en los puntos donde $d_{\mathbb{X}}$ es diferenciable.

Sea $x \in \mathbb{R}^d$ y $\Gamma_{\mathbb{X}}(x)$ el conjunto de puntos en $\mathbb{X}$ más cercano a x:

$$\Gamma_{\mathbb{X}}(x) = \{y \in \mathbb{X} : d(x,y) = d_{\mathbb{X}}(x)\}, \tag{8}$$

este es un subconjunto compacto no vacío de $\mathbb{X}$.

Sea $\sigma_{\mathbb{X}}(x)$ bola cerrada más pequeña que encierra $\Gamma_{\mathbb{X}}(x)$ y sea $\theta_{\mathbb{X}}(x)$ el centro y de radio $F_{\mathbb{X}}(x)$. Para $x \in \mathbb{R}^d$, el gradiente generalizado se define:

$$\nabla_{\mathbb{X}}(x) = \frac{x - \theta_{\mathbb{X}}(x)}{R_{\mathbb{X}}(x)}, \tag{9}$$

y para $x \in \mathbb{X}$, $\nabla_{\mathbb{X}}(x) = 0$

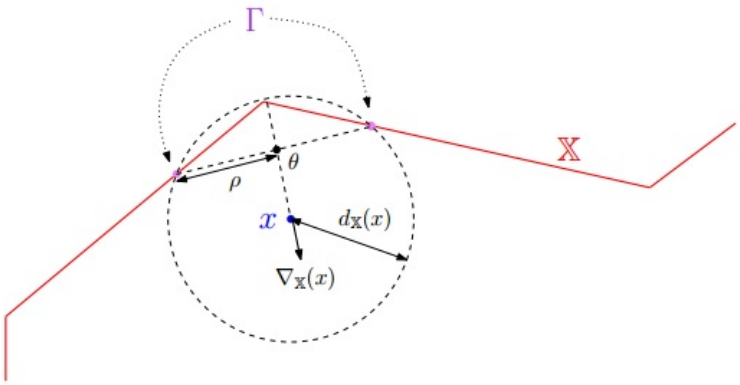

Fig. 17. Gradiente generalizado de la función distancia para un conjunto compacto.

La norma del gradiente está dada por:

$$||\nabla_{\mathbb{X}}(x)||^2 = 1 - \frac{F_{\mathbb{X}}(x)^2}{R_{\mathbb{X}}(x)^2}. \tag{10}$$

Equivalentemente la norma de $\nabla_{\mathbb{X}}(x)$ es el coseno del ángulo mitad del cono más pequeño que contiene $\Gamma_{\mathbb{X}}(x)$. Una explicación más detalla la podemos encontrar en [34].

En [59] realizan un estudio del número de puntos críticos de $d_{\mathbb{X}}$ cuando $\mathbb{X}$ es un proceso Poisson. La teoría de Morse no se aplica directamente a la función de distancia, principalmente porque no es diferenciable en todas partes. Si embargo se demuestra que se puede definir una noción de puntos críticos no degenerada para la función distancia, así como su índice de Morse. En particular analizan el comportamiento límite de N_k- número de puntos críticos de $d_{\mathbb{M}}$ con índice Morse k- como la densidad de puntos. Se analiza como los puntos críticos son por sí mismo intrínsecamente interesante, y el conocimiento de su comportamiento tiene implicaciones inmediatas a través de la teoría de Morse para el estudio de la topología de los complejos de Cech construidos sobre los conjuntos de puntos aleatorios.

4.2. Homología persistente de la función de densidad

La mayor parte de la literatura sobre la topología computacional se centra en la función de distancia. Pero si se quiere la homología del conjunto $\mathbb{X}$ y no se observa $\mathbb{X}$ directamente, más bien se observa una muestra $S_n = \{X_1, \cdots, X_n\}$ de una distribución P que se concentra en ó cerca $\mathbb{X} \subset \mathbb{R}^D$. Es decir, utilizar los datos

para construir un estimador de densidad suave y luego encontrar el diagrama de persistencia definido por una filtración de los conjuntos de nivel superior del estimador de densidad [48,49].

Por ejemplo, si se supone que $\mathbb{X}$ es un círculo. Entonces la homología del conjunto de datos S_n no es igual a la homología de $\mathbb{X}$; sin embargo el conjunto $\widehat{L}_\varepsilon = \{x : d_{S_n} \leq \varepsilon\} = \bigcup_{i=1}^n B(x_i, \varepsilon)$ captura a homología de $\mathbb{X}$ para un intervalo de valores ε.

Sea $X_1, \cdots, X_n$ una muestra de P, donde $X_i \in \mathbb{R}^D$, se utilizan los datos para construir un estimador de densidad suavizado. Un enfoque diferente de suavizado basado en distancias de difusión es discutido en [60].

Sea $\mathbb{X}$ el soporte d-dimensional de P. Se define la densidad de la medida de probabilidad $\mathcal{P}_h$ versión suavizada por la convolución con el Núcleo.

$$p_h(x) = \int_{\mathbb{X}} \frac{1}{h^D} K\left(\frac{||x-u||_2}{h}\right) dP(u), \tag{11}$$

$P_h = P * \mathbb{K}_h$, donde $\mathbb{K}_h(A) = h^{-D}\mathbb{K}(h^{-1}A)$ y $\mathbb{K} = \int_A K(t)dt$. El estimador estándar para p_h es el estimador de densidad del Núcleo:

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^D} K\left(\frac{||x-X_i||}{h}\right). \tag{12}$$

Donde el Núcleo K satisface $\int K(x)dx = 1$ y el parámetro de suavizado h es conocido como el ancho de banda. En la práctica el Núcleo K se elige generalmente para ser una función de densidad de probabilidad unimodal simétrica alrededor de cero. Una popular elección en los trabajos revisados para el Núcleo es el gaussiano $K(y) = \frac{1}{\sqrt{2\pi}} exp(\frac{-y^2}{2})$[61].

El objetivo es poder estimar P_h usando el diagrama $\widehat{P}_h$ de los conjuntos de nivel superior $\{x : \widehat{p}_h(x) \geq t\}$. p_h es de interés por varias razones, primero los conjuntos de nivel superior de una densidad son de interés intrínseco en estadística y machine learning y la homología de estos conjuntos proporciona información estructural acerca de la densidad. Las componentes conexas de estos conjuntos se usan para la agrupación. En segundo lugar bajo condiciones apropiadas se obtiene información topológica sobre un conjunto de interés $\mathbb{X}$. Es mostrado en la fig: 18, suponiendo que $\mathbb{X}$ es un compacto d-manifold suave, p es la densidad de P con respecto a la medida de Hausdorff en $\mathbb{X}$.

Luego de expuesto una panorámica acerca del enfoque estadístico para la homología persistente y algunos conceptos básicos, el siguiente capítulo ofrece una descripción de los resúmenes topológicos estándar que se obtienen y son analizados los resultados desde un punto de vista probabilístico.

5. Diagramas de persistencia

Utilizando la sección 4 se considera que la clase de homología α nació en L_t y muere al entrar en L_s, establecer $b(\alpha) = t$ y $d(\alpha) = s$, representar cada clase (α) por un punto $(b(\alpha), d(\alpha))$, resultando entonces un multiconjunto de puntos en $\mathbb{R}^2$ con el eje horizontal correspondiente a el nacimiento de la clase y el vertical a la muerte. La persistencia de α es la diferencia $\text{pers}(\alpha) = d(\alpha) - b(\alpha)$, donde en un contexto general se obtienen puntos con infinita persistencia que corresponden a los puntos de la forma (∞, s) o (t, ∞).

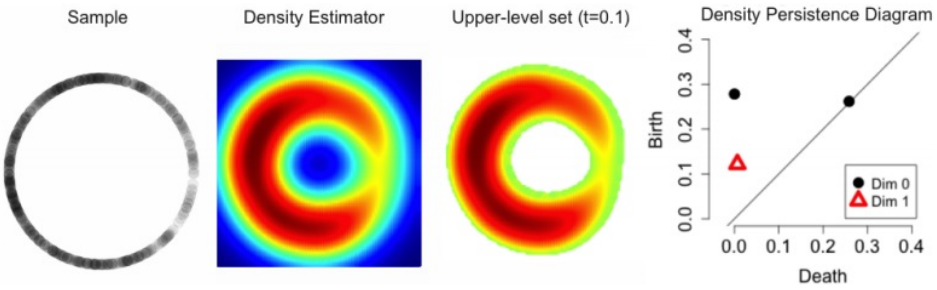

Fig. 18. Sample: 500 puntos de datos de un círculo de radio 1: **Density Estimator** Núcleo Gaussiano con ancho de banda $h = 0{,}3$. **Upper level**: $\widehat{U}_{0,1} = \{x : \widehat{p}_h \geq 0{,}1\}$. **Density Persistence Diagram** resume las características topológicas de los conjuntos de nivel superior del estimador de densidad del núcleo. Los puntos negros representas las dos componentes conexas que aparecieron aproximadamente en $t = 0{,}27$, pero uno de ellos inmediatamente muere por la fusión con otro. El triángulo rojo representa el único agujero 1-dimensional que se forma en $t = 0{,}12$ y muere $t = 0{,}01$.

***Definición* 6** *La persistencia total de grado p de una diagrama P se define como:*

$$Pers_p(P) = \sum_{t \in P} (pers(t))^p. \tag{13}$$

Definición* 7** *Un* ***diagrama de persistencia *es un multiconjunto de puntos en* $\mathbb{R}^2$ *junto con la diagonal* $\Delta = \{(x,y) \in \mathbb{R}^2 | x = y\}$*, donde cada punto en la diagonal tiene multiplicidad infinito.*

Notar que el diagrama se encuentra totalmente contenido en el semiplano por encima de la diagonal Δ dado que la muerte siempre se produce después del nacimiento; en [62] se demuestra como estos diagramas están bien definidos para cualquier espacio métrico y en particular para cualquier espacio métrico compacto. Las características de mayor persistencia son las representadas por los puntos más alejados de la diagonal, mientras que los puntos cercanos pueden ser interpretados como ruidos topológicos.

Un diagrama de persistencia es estable si un pequeño cambio en la función de entrada produce un pequeño cambia en el diagrama. Existen diferentes elecciones de métricas en el espacio de los diagramas de persistencia, análogas a la variedad de métricas en el espacio de funciones. Se trabaja de manera general con una distancia métrica que es análoga a la distancia L^p en el espacio de funciones en un espacio discreto. Una familia natural de métrica se discute en [63]

Definición* 8** ***(Distancia Wasserstein) *La distancia Wasserstein entre dos diagramas de persistencia,* d_1 *y* d_2 *se define:*

$$W_p(d_1,d_2) = \left(inf_\gamma \sum_{x \in d_1} ||t - \gamma(t)||_\infty^p \right)^{\frac{1}{p}}, \tag{14}$$

donde γ *es el conjunto de todas las biyecciones entre* D_1 *y* D_2*; es no vacío debido a la diagonal.*

En el diagrama de persistencia vacío $d_\emptyset$, representa el diagrama que contiene sólo la diagonal. Observar que la $\text{Pers}_p(d) = 2^p (W_p(d, d_\emptyset))^p$ para $t \in d$.

Caso Particular

Para $p = \infty$ se obtiene la distancia bottleneck definida como el ínfimo de los γ para los cuales existe una coincidencia entre los diagramas, de tal forma que dos puntos sólo pueden ser igualados si la distancia

es menor que γ y todos los puntos a una distancia mayor que γ de la diagonal deben ser igualados:

$$B_p(d_1,d_2) = inf_{\gamma} sup_{t\in d_1} ||t-\gamma(t)||_{\infty}. \tag{15}$$

En [56] la métrica W_p es utilizada para definir el siguiente espacio de diagramas de persistencia.

$$D_p = \{d|W_p(d,d_{\emptyset}) < \infty\} = \{d|Pers_p(d) < \infty\} \ p \geq 1. \tag{16}$$

Señalar que la distancia p-Wasserstein es una modificación del concepto clásico de la teoría de probabilidad. Dadas μ y v medidas de probabilidad 35 en un espacio métrico $(\mathbb{X},\rho)$, la distancia p-Wasserstein entre μ y v se define:

$$W_p(\mu,v) = \left(inf_{\gamma\in\Gamma(\mu,v)} \int_{\mathbb{X}x\mathbb{X}} \rho^p(s,t) d_{\gamma}(s,t) \right)^{\frac{1}{p}}, \tag{17}$$

donde $\Gamma(\mu,v)$ es una colección de medidas de probabilidad en $\mathbb{X}x\mathbb{X}$. Requerir finitud del p- momento de la medida de probabilidad μ es similar a requerir la finitud de la persistencia total de grado p de un diagrama d y significa que para algún $t_0 \in \mathbb{X}$ se tiene:

$$\int_{\mathbb{X}} \rho^p(t_0,t) d\mu(t) < \infty. \tag{18}$$

La principal diferencia entre la distancia Wasserstein para los diagramas de persistencia y la distancia para las medidas de probabilidad se debe a la función única de la diagonal en el primer caso.

En [64] son demostrados dos resultados de estabilidad para las funciones Lipschitz en espacios métricos triangulable y compactos. Se formulan dos funciones, la primera en términos de la distancia de Wasserstein entre sus diagramas de persistencia y la segunda en términos de su persistencia total. Se supone que $\mathbb{X}$ es un espacio métrico tal que para cualquier diagrama de persistencia d calculado por un función Liptschiz f con constante Liptschiz $\text{Lip}(f) \leq 1$ y además cumple la condición de *tame* [1], obtenemos $\text{Pers}_p(d) \leq C_{\mathbb{X}}$ que depende sólo de $\mathbb{X}$ implicando que la persistencia total de grado p es acotada.

***Proposición* 1** *(Estabilidad Wasserstein) Si $\mathbb{X}$ es un espacio métrico compacto, triangulable que implica persistencia total de grado k acotada para algún $k \geq 1$ y $f_1, f_2 : \mathbb{X} \to \mathbb{R}$ son funciones Liptschitz tame, entonces $\forall$ dimensiones l y $p \geq k$ tenemos:*

$$W_p(D_l(f_1), D_l(f_2)) \leq C^{\frac{1}{p}} ||f_1 - f_2||_{\infty}^{1-\frac{k}{p}}, \tag{19}$$

donde $C = C_{\mathbb{X}} max\{Lip(f_1), Lip(f_2)^k\}$.

***Teorema* 1** *Si $(\mathbb{X},d)$ es un espacio métrico completo separable y p un número positivo, entonces el espacio métrico (D_p, W_p) es separable y completo.*

5.1. Propiedades del espacio de los diagramas de persistencia

A continuación son expuestas las propiedades particulares que posee el espacio que conforma los diagramas de persistencia para luego poder aplicar inferencia estadística sobre estos espacios mediante objetos estadísticos como: varianzas, esperanzas y probabilidades condicionales.

[1] $f : \mathbb{X} \to \mathbb{R}$ es *tame* si tiene un número finito de valores críticos homológicos y los grupos de homología $H_k(f^{-1}(-\infty,\alpha]$ son de dimensión finita $\forall k \in \mathbb{Z}$ y $\alpha \in \mathbb{R}$

Teorema* 2** (D_p, W_p) ***es completo

Ideas de la demostración
Un espacio métrico se dice completo, si toda sucesión de Cauchy es convergente:
Sea $\{d_n\} \in D_p$ un sucesión de Cauchy:

- Demostrar que $\{d_n\}$ converge en *"persistence-wise"* a un diagrama d^*.
- $d^* \in D_p$
- Demostrar que $\{d_n\}$ converge a d^* en la métrica W_p.

Demostrar cada paso requiere de una serie de suposiciones y lemas que el lector puede revisar en [56].

Teorema* 3** D_p ***es separable

Un espacio topológico es separable si posee un subconjunto denso y numerable, para demostrarlo se trabaja con un subconjunto de diagramas de persistencia con multiplicidad total finita de tal manera que sus puntos tienen coordenadas racionales: $S \subset D_p = \left\{ d \in D_p |\ |d| < \infty \vee \ \ x \in \mathbb{Q}^2 \ \ \forall x \in d \right\}$.

Si $d \in D_p$, entonces $\forall \varepsilon > 0\ \exists\ \alpha > 0; W_p(l_\alpha(d), d_\emptyset)$ y l_α parte α-inferior de d [2]. Si μ_α es la parte α-superior [3] se cumple:

$$W_p(d, \mu_\alpha(d)) \leq W_p(l_\alpha(d), d_\emptyset) < \frac{\varepsilon}{2}. \tag{20}$$

Dado que $\mathbb{Q}^{2|\mu_\alpha(d)|}$ es denso en $\mathbb{R}^{2|\mu_\alpha(d)|}$, existe $d_s \in S$ tal que $W_p(d_s, \mu_\alpha(d)) < \frac{\varepsilon}{2}$. Entonces:

$$W_p(d, d_S) \leq W_p(d, \mu_\alpha(d)) + W_p(d_S, \mu_\alpha(d)) < \varepsilon, \tag{21}$$

esto implica la densidad del conjunto S.

Notar que $S = \bigcup_{m=0}^{\infty} S_m$, donde $S_m = \{d \in S |\ |d| = m\}$. Cada S_m es isomorfo a un subconjunto de $\mathbb{Q}^{2m}$ y este es numerable.
$\rightarrow S$ es numerable.

Compacidad en D_p

En [56] se exige compacidad para los diagramas de persistencia, para esto se requiere de condiciones que debe cumplir D_p. Se analizan casos de diagramas de persistencia que no son compactos en D_p, se definen restricciones para el conjunto $S \subset D_p$ para poder entonces lograr compacidad por la eliminación de estos ejemplos.

***Ejemplo* 1** *$S \subset D_p$, subconjunto de diagramas con un sólo punto fuera de la diagonal de multiplicidad* 1 *y persistencia* $\varepsilon > 0$. *Sea* $\{d_n\} \in S$ *tal que* $b(t) = 2n\varepsilon$, *tenemos(ver fig.19):*

$$W_p(d_n, d_m) = \left((\frac{\varepsilon}{2})^p + (\frac{\varepsilon}{2}) \right)^{\frac{1}{p}} = \frac{\varepsilon}{2} (\frac{1}{2})^{\frac{p}{p-1}}. \tag{22}$$

[2] $t \in l_\alpha \Leftrightarrow t \in d \vee pers(t) < \alpha$
[3] $t \in l_\alpha \Leftrightarrow t \in d \vee pers(t) \geq \alpha$

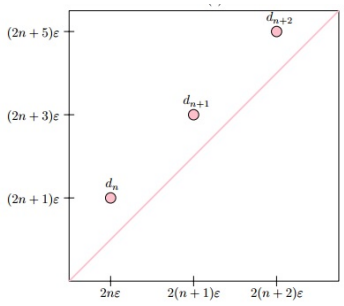

Fig. 19. Gráfico de tres diagramas consecutivos de la sucesión d_n.

Para eliminar este ejemplo se imponen una de las dos condiciones referidas a la acotación para el nacimiento y la muerte de la característica expuesta en [56].

Las condiciones de acotación no son suficientes para asegurar la compacidad relativa como se muestra en el ejemplo siguiente:

***Ejemplo* 2** *$S = \{d|\, W_p(d,d_\emptyset) \leq \varepsilon\} \cap \{d|\, b(t) \geq 0 \vee d(t) \leq C\ \forall t \in d\}$ $(\varepsilon, C \geq 0)$. Sea $\{d_n\} \in S$ diagramas con un sólo punto fuera de la diagonal $t_n = (0, 2^{1-\frac{n}{p}})$ con multiplicidad 2^n. Entonces:*

$$\forall n,m \in \mathbb{N}, m > n;\ \ W_p(d_n,d_m) \geq (2^{m-1}\left(\frac{1}{2}2^{1-\frac{m}{p}}\varepsilon\right)^p)^{\frac{1}{p}} = 2^{\frac{-1}{p}}\varepsilon, \tag{23}$$

ya que al menos tenemos 2^{m-1} puntos de persistencia $2^{1-\frac{m}{p}}\varepsilon$ en la diagonal; por lo tanto una subsucesión de $\{d_n\}$ puede ser Cauchy y S no es relativamente compacto.(ver fig.20).

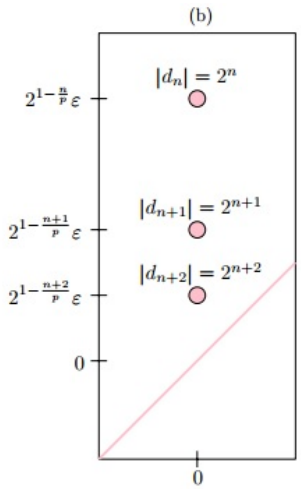

Fig. 20. Cada punto representa un diagrama con un sólo punto fuera de la diagonal cuya multiplicidad crece mientras la persistencia disminuye.

Para lidiar con el problema anterior se introduce la siguiente definición:

***Definición* 9** *$\forall S \subset D_p$ se denomina uniforme si $\forall \varepsilon > 0\ \ \exists \alpha > 0;\ \ W_p(l_\alpha(d), d_\emptyset) \leq \varepsilon\ \ \forall\, d \in S$.*

La exclusión de los casos expuestos anteriormente es suficiente para lograr la acotación general.

***Definición* 10** *$\forall S \subset D_p$ es **totalmente acotado** si y sólo si $\forall \varepsilon > 0$, $\exists$ una ε-red [4] finita de S.*

Notése que :

- Todo conjunto totalmente acotado, es acotado pues es unión finita de conjuntos acotados.
- Si S es totalmente acotado, entonces $[S]$ es también totalmente acotado.
- Todo espacio totalmente acotado es separable, pues para todo n se construye una $\frac{1}{n}$-red finita $\{A_n\}$, cuya unión es un conjunto numerable siempre denso.

Para comprender la definición anterior veamos el siguiente ejemplo:

***Ejemplo* 3** *En l_2 es totalmente acotado de dimensión infinita el **paralelepípedo fundamental o ladrillo de Hilbert** definido por:*

$$\Pi = \left\{ (s_n); |s_n| \leq \frac{1}{2^{n-1}}, \forall n \right\}. \tag{24}$$

Demostración: *Sea $\varepsilon > 0$ y n tal que $\frac{1}{2^{k-1}} < \frac{\varepsilon}{2}$. Relacionamos el vector $s = (s_n)$ con el vector $s^* = (s_1, \cdots, s_n, 0, \cdots, 0)$. Entonces:*

$$d(s,s^*) = \left(\sum_{k=n+1}^{\infty} x_k^2 \right)^{\frac{1}{2}} \leq \left(\sum_{k=n+1}^{\infty} \left(\frac{1}{2^{k-1}} \right)^2 \right)^{\frac{1}{2}} \leq \left(\sum_{k=n+1}^{\infty} \frac{1}{4^k} \right)^{\frac{1}{2}}, \tag{25}$$

es decir;

$$d(s,s^*) \leq \frac{1}{2^{n-1}} < \frac{\varepsilon}{2}. \tag{26}$$

Para cada n fijo, el conjunto Π^* de puntos del tipo de s^* es totalmente acotado, por ser acotado en un espacio de dimensión finita n. Tomando una $\frac{\varepsilon}{2}$-red finita de Π^* ella lo será también de Π, quedando así demostrado que el ladrillo de Hilbert es totalmente acotado.

6. Inferencia estadística con diagramas de persistencia

Se han realizado varios intentos, con diferentes enfoques para estudiar los diagramas de persistencia desde un punto de vista probabilístico; por ejemplo [65,66,67,68]. Con el fin de utilizar los diagramas de persistencia como una verdadera herramienta estadística surgen naturales cuestiones sobre estos resúmenes.

1. ¿Es posible definir medidas de probabilidad sobre los resúmenes?
2. ¿Es posible definir medias y varianzas (Fréchet)?
3. ¿Establecer relaciones entre la distribución de muestreo de los datos y la distribución en el resumen topológico?
4. ¿Calcular media y varianzas (Fréchet) ?
5. ¿Obtener la concentración de la media de Fréchet?

Existen una variedad de razonas para querer caracterizar propiedades estadísticas de los diagramas. Por ejemplo dada una nube de puntos muy grande S, es ventajoso trabajar con submuestras de los datos que producen nubes de puntos más pequeñas $S_1, \cdots, S_n$ y así calcular la media y la varianza del conjunto de diagramas de persistencia obtenido a partir de los n submuestreados conjunto de datos. En términos estadísticos consiste en calcular una estimación bootstrap [69] de diagramas de persistencia de los datos.

[4] $A \subset E$ es una ε-red de S para $\varepsilon > 0$ ssi $\forall s \in S \ \exists a \in A; d(s,a) \leq \varepsilon$

Pero estos procedimientos requieren de una buena definición para la media y varianza de un conjunto de diagramas. Se han realizado algunos progresos en cuanto a la comprensión de la media de las distribuciones de los diagramas de persistencia [56] Mileyko demuestra que el espacio de los diagramas (D_p, W_p) es un espacio Polish [1] y por lo tanto es posible definir la media de Fréchet. En particular se demuestra que la medida de Fréchet de un conjunto finito de diagramas de persistencia siempre existe pero no es necesariamente única. Una posible solución se argumenta en [67] donde se da una alternativa definición probabilística combinando la media con conceptos de teoría de juego; logrando que esta media probabilística sea continua y única.

Un parámetro estadístico es una cantidad que describe atributos de una colección de datos, que resume información acerca de los datos. La teoría estadística define un parámetro estadístico como una función de una muestra donde dicha función es independiente de la distribución de la muestra; es decir puede decirse antes de la realización de los datos. El término estadístico es usado tanto para la función como para el valor de la función en una muestra dada.

Existen tres tipos de parámetros estadísticos: *de centralización, de posición y los de dispersion.* Las medidas de centralización indican en torno a que valor (centro) se distribuyen los datos, incluyen la media, la moda y la mediana. Las medidas de posición dividen un conjunto de datos en grupos con el mismo número de individuos pero es necesario que los datos se encuentren ordenados, incluyen los cuartiles, deciles y percentiles. Por último las medidas de dispersion o variabilidad incluyen la desviación estándar, varianza, desviación media y el rango de los valores (diferencia entre el mayor y el menor de los datos de una distribución estadística) y nos informa sobre cuanto se alejan del centro los valores de la distribución. Las tendencias centrales y sus correspondientes medidas de variabilidad son soluciones para la optimización de las diferentes funciones de costos basadas en la métricas p-Wasserstein.

En [63] se establece la definición natural de la media y la mediana de un conjunto de diagramas considerando las funciones de costo análogas a los de las muestras de los números reales y la definición de estos parámetros estadísticos como las soluciones para la optimización de estas funciones. Es decir se caracterizan las funciones por los mínimos locales y al hacerlo se caracteriza por la media y la mediana. Esto sugiere para direcciones futuras que el trabajo hecho para la media se extienda para la mediana. Pero la discontinuidad y la falta de unicidad de la media y la mediana como muestra Turner en su artículo hace la inferencia estadística mucho más difícil. Una posible solución consiste en considerar un enfoque alternativo probabilístico dada en [65] donde se establece un algoritmo para el cálculo de la media y varianza.

Con este fin se requiere de una distribución de probabilidad $\mathcal{P}_D$ en $(D_p, \mathcal{B}(D_p))$ donde $\mathcal{B}(D_p)$ es la σ-álgebra Borel en D_p.

Dado un espacio de probabilidad $(D_p, \mathcal{B}(D_p), \mathcal{P})$ la cantidad:

$$Var_{\mathcal{P}} = inf_{d \in D_p} \left[F := \int_{D_p} W_p(d_1, d_2)^2 d\mathcal{P}(d_1) < \infty \right], \tag{27}$$

es la **varianza Fréchet** de $\mathcal{P}$ y el conjunto en el cual obtenemos el valor:

$$\mathbb{E}_{\mathcal{P}} = \{ d_2 \in D_p | F(d_2) = Var_{\mathcal{P}} \}, \tag{28}$$

[1] Un espacio se dice "Polish"si es metrizable, con métrica separable y completa.

es la **esperanza Fréchet** también conocida como la media Fréchet. En [56] se demuestra que para naturales restricciones de una distribución de diagramas de persistencia $\mathcal{P}_D$ el diagrama esperado y la varianza sobre estos diagramas están definidos. Surge la interrogante de si existe la esperanza de Fréchet para medidas de probabilidad con soporte compacto:

***Teorema* 4** *(**Mileyko-Mukherjee-Harer**) Sea $\mathcal{P}$ una medida de probabilidad en $(D_p, \mathcal{B}(D_p))$ con segundo momento finito. Si $\mathcal{P}$ tiene un soporte compacto entonces $\mathbb{E}_{\mathcal{P}\neq\emptyset}$.*

6.1. Correspondencia, selección y agrupaciones

En [65] se da un algoritmo para calcular una estimación de la media de Fréchet de un conjunto de diagramas. Dicho algoritmo centra su análisis en la comprensión de un análogo para la distancia Wasserstein con el fin de trabajar con más de dos diagramas.
Un diagrama es representado como una lista de sus puntos fuera de la diagonal $X = [x_1, \cdots, x_k]$. Dado que se trabaja con diagramas con número finito de puntos fuera de la diagonal, se supone implícitamente que esta lista es finita.

***Definición* 11** *Sea $X = [x_1, \cdots, x_k]$ y $Y = [y_1, \cdots, y_k]$ diagramas. Un correspondencia entre X y Y es un biyección $\varphi : X \to Y$. Una correspondencia óptima es aquella que alcanza la distancia Wasserstein.*

Ahora se necesita entender como definir correspondencia cuando existen N-diagramas en lugar de 2. Para ello son definidas selecciones y agrupaciones que restringen las correspondencias cuando $N = 2$.

***Definición* 12** *Dado un conjunto de diagramas $X_1, \cdots, X_n$, una selección es un punto de cada diagrama, donde ese punto puede ser Δ. La selección trivial para un punto particular fuera de la diagonal $x \in X_i$ es la selección m_x la cual elige x para X_i y Δ para cualquier otro diagrama.*
Un agrupamiento es un conjunto de selecciones de modo que cada punto fuera de la diagonal de cada diagrama es parte exactamente de una selección.

Un agrupamiento de N diagramas que tiene k selecciones puede ser almacenado como una matriz G de kxN, donde la entrada $G(i, j) = x$ significa que la selección de orden j tiene el punto $x \in X_i$. Veamos un ejemplo, donde el agrupamiento está dado por la matriz.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccc} D_\bigstar & D_\blacksquare & D_\bullet \end{array} \\ \begin{pmatrix} b & x & f \\ a & \Delta & \Delta \\ \Delta & y & g \\ \Delta & z & \Delta \\ \Delta & \Delta & h \\ c & \Delta & \Delta \end{pmatrix} \end{array}$$

Fig. 21. Agrupamiento Fréchet.

Δ representa la diagonal; tener en cuenta que las agrupaciones se consideran equivalentes hasta un reordenamiento de las selecciones, por la adición o eliminación de cualquier número de $(\Delta, \Delta, \cdots, \Delta)$ filas.

La *media de una selección s* es el punto denotado *means*(s) que minimiza la suma de los cuadrados de las distancias a los elementos de la selección; s consiste de N puntos: $\{p_1, \cdots, p_N\}$ con $p_i = (x_i, y_i)$

fuera de la diagonal y el resto copias de la diagonal Δ. El cálculo brinda el punto:

$$means_X(s) = \frac{1}{2Nk}\left((N+k)\sum_i x_i + (N-k)\sum_i y_i, (N-k)\sum_i x_i + (N+k)\sum_i y_i\right). \tag{29}$$

La *media de una agrupación*, media(G) es un diagrama en D_p con un punto en la media de cada selección. Notar que con la media de la selección se obtiene un punto, mientras que la media de un agrupación aporta un diagrama.

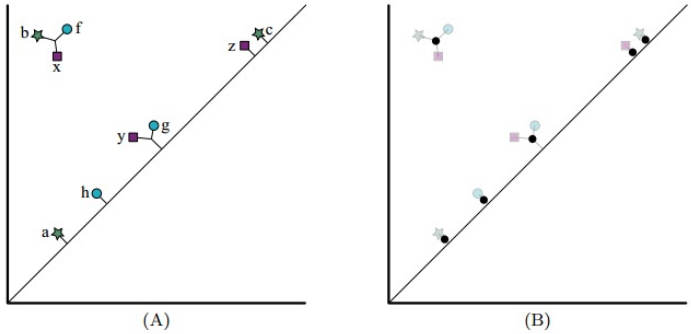

Fig. 22. Ejemplo de una agrupación de tres diagramas de persistencia dados en A. En este ejemplo la agrupación tiene 4 selecciones y la matriz correspondiente 21: Los círculos negros en el diagrama B ofrece el diagrama de media asociado a este grupo en particular.

Turner en [65] muestra que la métrica L^2-Wasserstein en un conjunto de diagramas de persistencia produce un espacio geodésico y que la estructura adicional se puede aprovechar para construir un algoritmo y así poder calcular la media de Fréchet y demostrar La Ley de los Grande Números. En [67] Munch adopta un enfoque diferente e introduce una variante de la media Fréchet como una medida de probabilidad en los diagramas; Si bien esto produce una media única, la solución no es un diagrama de persistencia.

Técnicas para calcular conjuntos de confianzas que permitan separa la señal del ruido topológico se han investigado en[68,70,57]. Alguno autores se enfocan en la métrica Bootleneck(caso especial de la métrica p-Wasserstein con $p = \infty$) señalando que se pueden obtener similares resultados para el caso general bajo supuestos más fuertes sobre el espacio topológico subyacente. En el siguiente capítulo se aborda la problemática mediante la técnica de bootstrap.

7. Bootstrap para diagramas de persistencia

La técnica bootstrap introducida por B.Efron 1979 proporciona estimaciones del error estadístico imponiendo escasas restricciones sobre las variables aleatorias analizadas y estableciéndose como un procedimiento de carácter general, independientemente del estadístico considerado. Consiste en aproximar la precisión de un estimador a partir de una muestra de datos u observaciones, donde la precisión se mide como la inversa de la varianza de un estimador. Esta distribución de estadísticos a través de las muestras definen una distribución muestral empírica, la cual puede ser utilizada para definir estabilidad e hipótesis estadísticas.

Conocer la precisión o el error cuadrático medio (ECM) 42 de estimadores de la varianza suele ser algebraicamente complicado. El método Bootstrap permite obtener una buena aproximación del ECM y otras estimaciones a partir de la muestra, aún sin conocer la distribución de donde provienen los datos.

7.1. Aspectos generales

Considerar las variables aleatorias $X_1, X_2, \cdots, X_n$ igualmente distribuidas con función $\mathcal{F}$ y representar mediante el conjunto $\{x_1, x_2, \cdots, x_n\}$ la muestra correspondiente a extracciones aleatorias sobre las referidas variables. Es válido destacar que los valores correspondientes a las muestras permiten obtener la distribución empírica $\mathcal{F}_e$, que constituye la estimación paramétrica de máxima verosimilitud de la función de distribución $\mathcal{F}$. La media y la varianza de X [71] está dada por:

$$\theta_{\mathcal{F}_e} = E_{\mathcal{F}_e}[X] = \sum_{i=1}^{n} \frac{X_i}{n}, \tag{30}$$

$$\sigma^2_{\mathcal{F}_e} = Var_{\mathcal{F}_e}(X) = \sum_{i=1}^{n} \frac{(X_i - \theta_{\mathcal{F}_e})^2}{n}. \tag{31}$$

También se definen otras medidas y parámetros: mediana, curtosis, coeficientes de asimetría, etc $\cdots$, todos ellos dependientes de la distribución.

Como se conoce la homología persistente permite cuantificar los cambios topológicas de los conjuntos de nivel superior con un multiconjunto de puntos en el plano extendido considerado como el diagrama de persistencia ($\mathcal{P}$). Otro camino para representar las información descrita por $\mathcal{P}$ es la función landscape $\lambda_k : \mathbb{R} \to \overline{\mathbb{R}}$; obtener estos resúmenes puede ser muy difícil directamente. En su lugar se supone que p corresponde a la distribución de probabilidad P. Dada una muestra de tamaño n, creamos un estimador de la función de densidad de probabilidad p_n usando un estimador de densidad por Núcleo. Como n crece, p_n se aproxima a la verdadera probabilidad de densidad. Dado un n suficientemente grande se calcula el diagrama P_n y el landscape λ_n correspondiente a p_n. Pero muchas veces conocer la estimación de un diagrama de persistencia o landscape no es suficiente, saltan a la vista varias interrogantes cómo: que tan cerca puede ser el diagrama estimado del verdadero; una respuesta a esto puede ser construir un conjunto de confidencia para los diagramas y una banda de confianza para los landscapes. Para resolver esta interrogante se utiliza el método Bootstrap:

7.2. Método

En la mayoría de los problemas de estimación, es importante dar un indicador de la precisión de un estimador dado. Un método simple es proporcionar un estimador del sesgo y la varianza del estimador, más preciso es un intervalo de confianza para el estimador. Este capítulo se concentra en el intervalo de confianza bootstrap y de manera general discute el procedimiento bootstrap cómo método para estimar la distribución de un estadístico dado.

Sea $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d, tomando valores en el espacio de medida (X, T, P). Suponer que interesa estimar el parámetro de valor real θ correspondiente a la distribución P de la observación. Estimar θ usando el estadístico $\hat{\theta} = g(X_1, \cdots, X_n)$, estos parámetros pueden ser la media de la población y la media de la muestra. Una explicación teórica más detallada la encontramos [72].

La distribución de la diferencia $\hat{\theta}-\theta$ contiene toda la información necesaria para construir un intervalo de confianza de nivel $(1-\alpha)$ para θ[68]; es decir se tiene un intervalo $[a,b]$ dependiendo de los datos tal que $\mathbb{P}(\theta \in [a,b]) \geq 1-\alpha$. En particular si se conoce la distribución $\mathcal{F}$ de $\hat{\theta}-\theta$ entonces los cuantiles $\mathcal{F}^{-1}(1-\alpha/2)$ y $\mathcal{F}^{-1}(\alpha/2)$ pueden ser calculados. Además los ajustes son obtenidos $a=\hat{\theta}-\mathcal{F}^{-1}(1-\alpha/2)$ y $b=\hat{\theta}-\mathcal{F}^{-1}(\alpha/2)$, resultando el intervalo de confianza $1-\alpha$ para θ:

$$\mathbb{P}(\theta \in [a,b]) = \mathbb{P}\left(F^{-1}(\alpha/2) \leq \hat{\theta}-\theta \leq F^{-1}(1-\alpha/2)\right) = 1-\alpha. \tag{32}$$

Desafortunadamente la distribución de $\hat{\theta}-\theta$ depende de la distribución desconocida P de las observaciones y no puede ser utilizada para evaluar el desempeño de $\hat{\theta}$; por lo tanto el primer paso en el procedimiento Bootstrap debe ser aproximar P. Lo interesante de este paso es la elección del estimador, si se tiene que las observaciones generales son una muestra aleatorias $(X_1,\cdots,X_n)$ de una distribución de probabilidad P, un candidato es la distribución empírica $\mathbb{P}_n = n^{-1}\sum \delta_{X_i}$ de las observaciones, principal para un bootstrap empírico. Obteniendo así una nueva muestra $(X_1^*,\cdots,X_n^*)$; luego estimar la distribución $\mathcal{F}(r)$ con la distribución $\hat{\mathcal{F}}(r) = P_n(\hat{\theta}^*-\hat{\theta} \leq r)$ donde $\hat{\theta}^* = g(X_1^*,\cdots,X_n^*)$.

La distribución $\hat{\mathcal{F}}$ se obtiene por diferentes métodos, en [68] trabajan con el método de aproximación por simulación: para B grande, obtener B valores diferentes de $\hat{\theta}^*$ y aproximar $\hat{\mathcal{F}}(r)$ como: $\widetilde{\mathcal{F}}(r) = \frac{1}{B}\sum_{j=1}^{B} I(\hat{\theta}_j^*-\hat{\theta} \leq r)$. Dado que los cuantiles de $\widetilde{\mathcal{F}}$ aproximan los cuantiles de $\mathcal{F}$, se define el intervalo de confianza estimado como:

$$C_n = \left[\hat{\theta}-\widetilde{\mathcal{F}}_n^{-1}(1-\alpha/2), \hat{\theta}-\widetilde{\mathcal{F}}_n^{-1}(\alpha/2)\right]. \tag{33}$$

El procedimiento anterior se resume en los siguientes pasos: **Procedimiento Bootstrap**

1. Calcular el estimador $\hat{\theta} = g(X_1,\cdots,X_n)$.
2. Obtener $X_1^*,\cdots,X_n^*$ de P_n y calcular $\hat{\theta}^* = g(X_1^*,\cdots,X_n^*)$.
3. Repetir el paso anterior B veces para obtener $\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$.
4. Capturar los cuantiles de $\widetilde{\mathcal{F}}$ y construir el intervalo de confianza C_n

En el procedimiento se pueden cometer dos posibles errores; en el paso de aproximar $\mathcal{F}$ con $\widetilde{\mathcal{F}}$ y luego el paso de calcular $\widetilde{\mathcal{F}}$ por simulación . El error en la simulación se produce en la probabilidad segura del intervalo de confianza resultante, en principio se puede hacer arbitrariamente pequeño, por la elección de las muestras bootstrap lo suficientemente grande.
Formalmente se demuestra que $sup_r|\widetilde{F}(r)-F(r)| \xrightarrow{P} 0$ lo que implica que el intervalo de confianza definido en 33 es asintóticamente consistente a nivel $1-\alpha \leftrightarrow \lim\,(inf)_{n\to\infty}\mathbb{P}(\theta \in C_n) \geq 1-\alpha$.

Cuando una variable aleatoria es una función, en lugar de un valor real, el proceso empírico bootstrap es utilizado para encontrar una banda de confianza para la función $h(t)$. Esto es encontrar un par de funciones $a(t)$ y $b(t)$ tal que la probabilidad que $h(t) \in [a(t),b(t)]\ \forall t$ es al menos $1-\alpha$, la descripción de esta técnica la podemos encontrar [72,73].

7.3. Aplicaciones del Bootstrap

Apoyados en la sección 4.2; se busca encontrar un conjunto de confianza para la distancia de bootleneck i.e, un intervalo $C=[0,c_n]$ tal que $\lim sup_{n\to\infty}\mathbb{P}_h(W_\infty(\hat{\mathcal{P}}_h,\mathcal{P}_h) \in [0,c_n]) \geq 1-\alpha$ con $\alpha \in (0,1)$. Utilizando el Teorema de Estabilidad es suficiente encontrar c_n tal que $\lim sup_{n\to\infty}\mathbb{P}_h(||\hat{p}_h-p_h||_\infty > c_n) \leq \alpha$ [48].

Para encontrar c_n usamos Bootstrap: Sea $\mathcal{F} = \left\{ f_x(u) = \frac{1}{h^D} K\left(\frac{||x-u||}{h}\right) \right\}_{x\in\mathbb{X}}$,basándonos en él método empírico bootstrap se define $p_h(x) = Pf_x$, $\hat{p}_h(x) = P_n f_x$ y $\hat{\theta}$ =sup$_{f_x\in\mathcal{F}}|\mathbb{G}_n f_x| = \sqrt{n}||\hat{p}_h - p_h||_\infty$. El aproximado $1-\alpha$ cuantil q_α se puede obtener a través de la simulación, i.e:

$$q_\alpha = inf\left\{ q : \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{(n)}||\hat{p}_h^j - \hat{p}_h|| \geq q) \leq \alpha \right\}, \tag{34}$$

donde $p_h^j(x)$ denota la distribución de probabilidad correspondiente a la j-muestra bootstrap.

***Teorema* 5** *[48]. Tenemos que:*

$$limsup_{n\to\infty}\mathbb{P}\left(\sqrt{n}||\hat{p}_h - p_h||_\infty > q_\alpha\right) \leq \alpha. \tag{35}$$

Por el teorema de estabilidad se cumple: $lim_{n\to\infty}\mathbb{P}\left(W_\infty(\hat{\mathcal{P}}_h, \mathcal{P}_h) > \frac{q_\alpha}{\sqrt{n}}\right) \leq \alpha.$

El conjunto de confianza C_n es un subconjunto de todos los diagramas de persistencia cuya distancia a $\hat{\mathcal{P}}$ es al menos c_n:

$$C_n = \left\{ \widetilde{\mathcal{P}} : W_\infty(\hat{\mathcal{P}}, \widetilde{\mathcal{P}}) \leq c_n \right\}. \tag{36}$$

Se puede visualizar C_n centrando una caja de longitud $2c_n$ en cada punto p en el diagrama de persistencia. El punto p se considera ruido si la caja correspondiente interseca a la diagonal. Alternativamente se puede visualizar el conjunto de confianza por adicción de una banda $\sqrt{2}c_n$ cerca de la diagonal del diagrama de persistencia; en este caso la interpretación que se obtiene es que los puntos en la banda no son significativamente diferentes del ruido y los que están por encima se pueden interpretar como la representación de una característica topológica significativa. Para más detalle se pueden referir a [48,68].

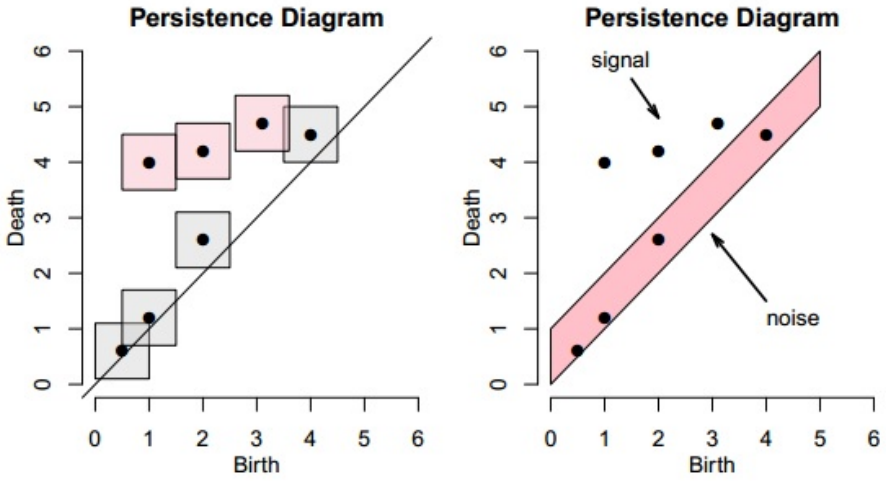

Fig. 23. Primero se obtiene el intervalo de confianza $[0, c_n]$ para $W_\infty(\widehat{P}, P)$, y c_n se puede obtener mediante el método bootstrap.

***Ejemplo* 4** *(Ejemplo 2.2 de [68]) El Toro es intersectado* ($\mathbb{S}^1 x \mathbb{S}^1$) *en* $\mathbb{R}^3$ *y se utiliza el algoritmo de muestreo de rechazo de* ($R = 1{,}5, r = 0{,}8$) *para muestrear* $10{,}000$ *puntos uniformemente del toro. Entonces*

se calcula el diagrama de persistencia $\hat{\mathcal{P}}_h$ usando el núcleo Gaussiano con ancho de banda $h = 0{,}25$ y bootstrap es utilizado para construir un intervalo de confianza $[0;0{,}01]$ con $0{,}95\,\%$ para $W_\infty(\hat{\mathcal{P}}_h, \mathcal{P}_h)$ (ver fig:24). Se puede notar que el conjunto confianza captura correctamente la topología del toro. Es decir sólo los puntos que representan características significativas del toro radican lejos del eje horizontal.

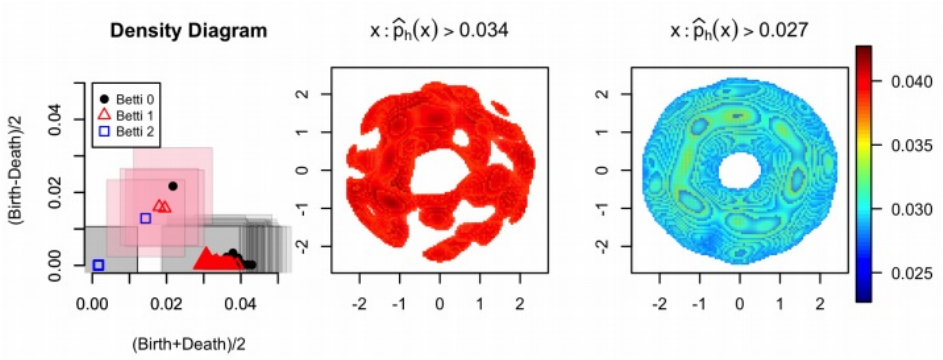

Fig. 24. El primer gráfico representa el diagrama de persistencia de los conjuntos de nivel superior de un estimador de densidad núcleo en el toro $3D$. La cajas de lado $= 2x0{,}001$ cerca de los puntos representan el $95\,\%$ de confianza para $\mathcal{P}_h$, las otras dos figuras representan proyecciones $2D$ de diferentes de conjuntos de nivel superior.

En [74] trabajan sobre un espacio métrico compacto medible (X, ∂_X, μ_X)con una distribución de referencia fijada P. Se define la distancia homológica en X relativa para P como:

$$HD_k^n((X, \partial_X, \mu_X), P) = d_{P_r}(\Phi_h^n(X, \partial_X, \mu_X), P). \tag{37}$$

Consideran MHD_k^n un estadístico robusto relacionado con HD_k^n, para construirlo se comienza con un diagrama de referencia y se calcula la distancia media al diagrama de la submuestra. Para HD_k^n el camino para construir un intervalo de confianza es trabajando con la simulación Monte Carlo. Si se trabaja con MHD_k^n se pueden definir intervalos de confianza, mediante técnicas no-paramétricas estándar para la mediana y media recortada [75]. Para la mediana utilizan estadísticos de orden con el objetivo de determinar los límites de un intervalo que contiene la mediana real con confianza $1 - \alpha$ y un análisis similar se realiza para la media recortada donde el intervalo se obtiene de la desviación estándar de la muestra.

Una aplicación inmediata de los conjuntos de confianzas descritos anteriormente es la formalización de **pruebas de hipótesis** capaces de determinar el ruido topológico de las características importantes. De forma general se tratan los diagramas como pruebas estadísticas no paramétricas, donde dadas dos muestras separadas, interesa el estudio de la prueba que rechaza la hipótesis nula de homogeneidad basado exclusivamente en características homológicas. Algunos resultados de la prueba de hipótesis para homología persistente se presentan en [66,14] estos se basan principalmente en pruebas de permutación pero no proporcionan un riguroso análisis estadístico de la capacidad de estos procedimientos, a continuación se analizan algunos resultados obtenidos dentro del área.

8. Prueba de hipótesis para Análisis Topológico de Datos

Las pruebas de significancia de hipótesis nula [76]; en lo adelante NHST, *por sus siglas en inglés* es una herramienta estadística comúnmente usada que proporciona una medida del nivel de evidencia en contra de

una hipótesis. NHST cuantifica las diferencias entre dos tipos de objetos o procesos subyacentes utilizando los diagramas de persistencia como observaciones. Por ejemplo pueden proporcionar una condición necesaria en cuanto a si todos los diagramas de persistencia particulares se pueden utilizar para la clasificación. Como se estudió en secciones previas el espacio de los diagramas es geométricamente complicado, por lo cual no es posible utilizar cualquiera de los modelos paramétricos para las distribuciones, por lo que no es posible realizar NHST usando un método que requiera de un modelo paramétrico subyacente. El enfoque utilizado por Robinson y Turner en [66] consiste en encontrar una función de pérdida conjunta relevante y luego usar una prueba de aleatorización [77]. La teoría que encierra la prueba de aleatorización muestra que dados dos conjuntos de diagramas extraídos de la misma distribución, el valor de p que se obtiene, es una variable aleatoria con una distribución uniforme sobre un subconjunto uniformemente espaciado de $[0,1]$. Sin embargo no se conoce ninguna teoría sobre si el valor de p es necesariamente pequeño si las distribuciones son diferentes, además debido a que los diagramas de persistencia son resúmenes estadísticos es posible que la distribución de los objetos subyacentes bajo análisis pueda ser diferente, pero las distribuciones correspondientes de los diagramas sean similares.

8.1. Prueba de significancia de hipótesis nula (NHST)

Los pasos de la prueba son los siguientes:

- Seleccionar un parámetro que represente los datos de alguna manera y sobre la que una hipótesis pertinente puede formarse.
- Elegir un estadístico a utilizar para estimar el parámetro.
- Predecir el comportamiento del estadístico bajo la hipótesis nula, tratando de capturar todo el rango de variabilidad que se encuentra implícito en el modelo.
- Comparar el valor observado de la prueba estadística con el comportamiento esperado bajo la hipótesis nula.

Además se requiere nombrar un punto de corte, conocido como el tamaño de la prueba que es por definición la probabilidad de rechazar la hipótesis nula cuando es verdadera. Entonces la probabilidad de observar un resultado tan o más extremo que la prueba estadística observada se calcula basándose en experimentos repetidos y en el supuesto de que la hipótesis nula es verdadera.

El p-valor se define como la probabilidad repetida en fase de experimentación de que un resultado tan o más extremo que se observaría condicionada a la hipótesis nula es verdadera. Los p-valores verdaderos nunca son ceros (con probabilidad uno) y se pueden utilizar para realizar prueba de hipótesis nula seleccionando un umbral α y rechazando la hipótesis nula cuando el valor p es menor a α.

La **potencia de una prueba** de hipótesis nula describe los errores de tipo I y tipo II en términos de la comparación del valor p con el umbral α. Una prueba más potente es mejor para rechazar la hipótesis nula cuando de hecho es falsa. Para H_0 hipótesis nula y H_1 hipótesis alternativa se define la potencia de NHST por:

$$potencia = \mathbb{P}(rechazar\, H_0 | H_1\ verdadera.)$$

No obstante se puede aproximar la evaluación de diferentes pruebas estadísticas mediante el uso de experimentos de simulación, y la evaluación de la proporción de veces que las hipótesis son rechazadas a un nivel nominal basado en diferencias específicas. Se puede comparar estas proporciones directamente entre diferentes pruebas estadísticas para dar una idea de las ventajas relativas de la prueba.

8.2. Pruebas aleatorias

Las Pruebas Aleatorias, también conocidas como Pruebas Realeatorias o Pruebas Permutacionales, fueron el primer tipo de procedimiento de re-muestreo. Este método utiliza muestras del mismo tamaño que la original y no necesita de reasignación de valores. El requisito indispensable en esta prueba es la presencia de algún tipo de aleatorización en el experimento. Suavizan el análisis de la necesidad de nombrar a un modelo formal de las distribuciones subyacentes bajo la hipótesis nula, proporcionando una estimación empírica de la distribución del estadístico de prueba. Sólo se necesita una noción del costo o la función de pérdida y se obtiene entonces así, una pérdida conjunta para una separación de los diagramas en dos conjuntos diferentes. Se puede aleatorizar el etiquetado, de cuales diagramas pertenecen a cuales conjuntos y calcular los costos para diferentes conjuntos de etiquetas.

Se asume que se tiene una colección de n-diagramas de persistencia independientes y un esquema de etiquetado provisional que divide la colección en dos colecciones posiblemente disímiles, X_1contiene n_1 diagramas y X_2 contiene n_2. EL objetivo es evaluar la fuerza de la evidencia de que los procesos que generan la colecciones X_1 y X_2 difieran.

8.3. Prueba estadística

La simulación puede proceder mediante la suma de la distancia por pares de las observaciones que se asignan al azar a un mismo grupo. Cuando las observaciones se encuentran en la recta real, la medida de localización se obtiene minimizando la norma L_2, la localización estimada es la media, y la norma L_2 es una función monótona de la varianza. En [66] se propone que la media o la suma de las varianzas de los dos grupos sería un estadístico de prueba sensible. La expresión usual para la varianza de la muestra está en la forma más cercana para la norma L_2 evaluada en su mínimo, es:

$$\sigma^2_X = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{38}$$

de forma equivalente se puede obtener sin calcular primero la media:

$$\sigma^2_X = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2. \tag{39}$$

Para los diagramas de persistencia etiquetado con L en los conjuntos $\mathcal{X}_1 = \{X_{1,1}, X_{1,2}, \cdots, X_{1,n_1}\}$ y $\mathcal{X}_2 = \{X_{2,1}, X_{2,2}, \cdots, X_{2,n_2}\}$ la prueba estadística análoga es:

$$\sigma^2_{\mathcal{X}_{12}}(L) = \sum_{m=1}^{2}\frac{1}{2n_m(n_m-1)}\sum_{i=1}^{n_m}\sum_{j=1}^{n_m} d_2(X_{m,i} - X_{m,j})^2. \tag{40}$$

Pero la distancia entre las medias no es una prueba estadística adecuada para conjuntos de diagramas de persistencia. Existe un alto costo computacional de calcular la media para cada permutación. Además la media de Fréchet no es necesariamente única lo cual conduce a problemas de definición de esta función de pérdida, cuando no lo es.

Una opción de la función de pérdida para un simple conjunto es la varianza o el costo total(F_2 o F_1). En lugar de esto, Turner [66] utiliza la función de distancia dos a dos, que es mucho más rápida de calcular.

La función de pérdida conjunta correspondiente a la etiqueta L en conjuntos de diagramas en $\{X_{1,1}, X_{1,2}, \cdots, X_{1,n_1}\}$ y $\mathcal{X}_2 = \{X_{2,1}, X_{2,2}, \cdots, X_{2,n_2}\}$ que se utiliza es:

$$costo(L) = \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} d_2(X_{1,i}, X_{1,j})^2 \frac{1}{2m(m-1)} + \sum_{i,j=1}^{n} d_2(X_{2,i}, X_{2,j})^2. \tag{41}$$

Esto es la suma de las distancia media al cuadrado dentro de los conjuntos.

Algoritmo

1. Elegir el número de repeticiones N
2. Calcular $Cost(L_{observada})$ para las etiquetas observadas.
3. Conjunto $Z = 0$
4. Repetir los pasos siguiente N veces:
   - Mezclar aleatoriamente las etiquetas del grupo entre todas las n observaciones para dar el etiquetado L
   - Calcular $Cost(L)$ para la nueva muestra
   - Si $Cost(L) \leq Cost(L_{observada})$ entonces $Z+ = 1$
5. Sea $Z/ = n$
6. Salida Z

La salida Z del algoritmo está fuertemente relacionada a varias probabilidades. Desde que se elige cada uno de los re-etiquetados de manera independiente sabemos que:

$$\mathbb{E}(Z) = \mathbb{P}(Cost(L) \leq Cost(L_{observada})). \tag{42}$$

La Ley de los Grandes Números 44 garantiza que: $Z \to \mathbb{E}(Z)$ con $N \to \infty$ con probabilidad uno. El radio de convergencia es exponencial. Se justifica la salida de un p-valor por el siguiente lema:

***Lema* 2** *Sea* $\{X_{1,1}, X_{1,2}, \cdots, X_{1,n_1}\}$ *y* $\mathcal{X}_2 = \{X_{2,1}, X_{2,2}, \cdots, X_{2,n_2}\}$ *se obtienen diagramas de persistencia i.i.d (ambos conjuntos de la misma distribución) y sea* α *el p- valor calculado por el algoritmo anterior. Entonces* $\forall p \in [0,1]$ *tenemos* $\mathbb{P}(\alpha \leq p) \leq p$. *La prueba del lema la podemos encontrar en [66]*

Ejemplo* 5** ***(Nubes de puntos de diferentes formas muestreados con variación de ruido)
Sea K el círculo unitario y L una parte del círculo con radio $3/5$ *con un círculo de radio* $4/5$. *K y L tienen la misma longitud.*

Fig. 25. Ruido para las muestras.

Muestras K y L con ruido Gaussiano $N(0,\sigma)$ para construir nubes de puntos que contienen 50 puntos i.i.d.

Para cada ejecución de la simulación se crearon 20 nubes de puntos para K y L y se calculan los primeros diagramas de homología persistente para la filtración Rips para K y L. Se simula para encontrar los valores de p para diferentes opciones de ruido.

Se ejecuta la simulación 5 veces cada una para cada incremento 0,01 de σ. Los resultados se muestran en la siguiente figura:

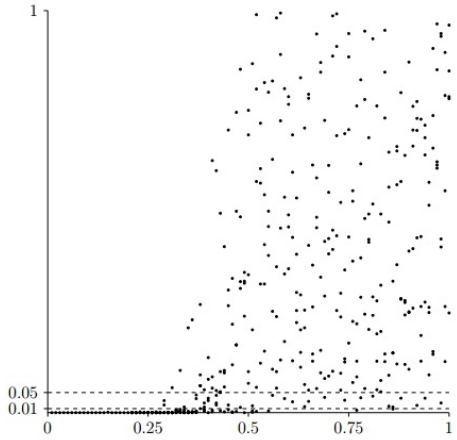

Fig. 26. Valores de p simulado dado el parámetro de ruido, 20 nube de puntos para K y L.

Esta simulación demuestra que para esta K y L cuando el ruido es suficientemente bajo, entonces los p-valores son menores y es válido decir que los diagramas de persistencia provienen de diferentes distribuciones y por lo tanto las formas subyacentes K y L deben ser diferentes. Cuando el ruido se incrementa no podemos rechazar la hipótesis nula. También se puede ver un análisis de la distribución de los valores de p para una simulación muy similar.

En lo expuesto anteriormente se muestra un ejemplo de cómo los métodos no paramétricos pueden ser adaptados para el uso de estadística en los resúmenes topológicos. Existe mucho potencial en la exploración de otros métodos no paramétricos. Alternativamente como dirección futura se pueden considerar diferentes funciones de costo en este marco de pruebas al azar para formar los métodos relacionados con la prueba de hipótesis nula. Otro campo abierto de investigación puede ser realizar la prueba de hipótesis alternativa cuando las observaciones son diagramas de persistencia.

En los capítulos anteriores se fundamenta que no es fácil cuantificar o realizar inferencia estadística en los resúmenes topológicos estándar (código de barra o diagramas de persistencia). A continuación introducimos un nuevo descriptor topológico denominado **persistencia landscape**, representación funcional de los anteriores con el objetivo de poder usar todas las herramientas que nos ofrece el **Análisis Funcional**.

9. Inferencia utilizando persistencia landscape

El nuevo enfoque proporcionado por Peter Bubenik en 2012 [14] define una nuevo descriptor topológico para los datos, asignando los diagramas de persistencia o códigos de barra a ciertas funciones.

Las principales ventajas de este descriptor radican en la representación funcional de los anteriores y la no pérdida de información con respecto a los resúmenes originales. Se define como una secuencia de funciones lineales a trozos y por lo tanto es posible utilizar la estructura de espacio vectorial de su espacio funcional subyacente. Dado que el espacio funcional es un espacio de Banach 29 separable se puede aplicar la teoría de variables aleatorias con valores en dicho espacio y utilizar técnicas y teorías existentes de la estadística no paramétrica. La persistencia landscape es estable [78,64,62] con respecto a la distancia L^p ($1 \leq p \leq \infty$), pequeñas perturbaciones en los datos dan lugar a pequeñas perturbaciones en los pares bajo la elección adecuada de distancia. Esta herramienta proporciona cálculos mucho más rápidos, algunos de estos pueden ser la obtención de la media, cálculo de distancias entre resúmenes topológicos que pueden ser útiles cuando otros métodos son costosos computacionalmente. Variantes de la persistencia landscape tales como siluetas [70] han sido trabajadas.

En el Análisis Topológico de Datos [79], los datos de interés se codifican en un complejo finito filtrado:

$$K_0 \subset K_1 \subset \cdots \subset K_n. \tag{43}$$

Para ser más preciso se aplica homología en algún grado con coeficientes en algún campo dado para la filtración y así obtener una secuencia de espacio vectoriales con dimensión finita y un mapeo lineal:

$$H(K_0) \to H(K_1) \to \cdots \to H(K_n), \tag{44}$$

que se conoce como el módulo de persistencia asociado a la filtración. Cuando el rango de todos los homomorfismos $H(K_t) \to H(K_s), t < s$ son finitos, el módulo se dice que es q-tame [62] y puede ser resumido como un multiconjunto de puntos en el plano real $\{(b_i, d_i)\}$ con $b_i < d_i$ que representan las características homológicas que aparecieron en la filtración en $t = b_i$ y desaparecieron $t = d_i$ obteniendo así un diagrama de persistencia y considerando estos puntos como intervalos $[b_i, d_i)$ se obtienen los códigos de barra explicados anteriormente. Podemos generalizar que $b_i, d_i \in \mathbb{R}$ por estar asociado a una correspondiente secuencia creciente de numeros reales.

Representar los pares nacimiento-muerte (b_i, d_i) del diagrama de persistencia como un multiconjunto finito de puntos: $D = \left\{ (\frac{b_i+d_i}{2}, \frac{d_i-b_i}{2}) \right\}_{i \in I}$, donde I es un conjunto finito. Para definir la persistencia landscape [14], primero construimos la función lineal a pedazos $\Lambda_{(b,d)} : \mathbb{R} \to [0, \infty]$ para cada par nacimiento-muerte.

$$\Lambda_{(b,d)}(x) = \begin{cases} 0 & \text{si } x \notin (b,d) \\ x - b & \text{si } x \in (b, \frac{b+d}{2}] \\ d - x & \text{si } x \in (\frac{b+d}{2}, d) \end{cases} \tag{45}$$

Con el objetivo de definir el *landscape* se consideran de forma equivalente $\lambda(k,x) = \lambda_k : \mathbb{R} \to \overline{\mathbb{R}}$, $p = 1, 2, 3, \cdots$ donde $\lambda_k(x)$ es el k-ésimo mayor valor de $\left\{ \Lambda_{(b_i,d_i)}(x) \right\}_{i=1}^n$, en particular 1-max es la función máxima usual. Se establece $\lambda_k(x) = 0$ si el conjunto $\left\{ \Lambda_{(b_i,d_i)}(x) \right\}_{i=1}^n$ contiene menos de k puntos. Los análisis siguientes son realizados para b y d finitos, luego se explica un extensión para los casos en que b y/o d son infinitos.

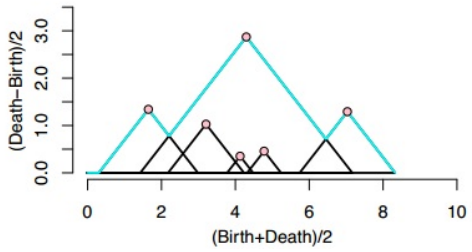

Fig. 27. Los círculos indican los puntos en el diagrama de persistencia D. Cada punto (x,y) corresponde a una función $f_{(b,d)}$, y el landscape $\lambda(k,\cdot)$ es la k-ésima función mayor del gráfico. En particular la curva azul corresponde al landscape $\lambda(1,\cdot)$.

Los *landscapes* cumple las siguientes propiedades:

1. $\lambda_k(t) \geq 0$
2. $\lambda_k(t) \geq \lambda_{k+1}(t)$
3. $\lambda_k(t)$ es 1-Lipstchiz

Notar que las dos primeras propiedades se obtienen directamente de la definición. La 3era propiedad se cumple dado que las funciones $\left\{\Lambda_{(b_i,d_i)}(x)\right\}_{i=1}^{n}$ cumplen la condición de 1-Lipstchiz [14].

9.1. Norma para persistencia landscapes

Sea (S,A,μ) espacio medible 31:

- Si $f : S \to \mathbb{R}$ definida μ-casi donde quiera, para $1 \leq p < \infty$, $\|f\|_p = [\int |f|^p \, d\mu]^{\frac{1}{p}}$
- Para $p = \infty$; $||f||_\infty = sup\ esc\ f = inf\ \left\{a \in \overline{\mathbb{R}}\, |\mu\{s \in S \mid f(s) > a\} = 0\right\}$

Para $1 \leq p \leq \infty$, $\mathcal{L}^p(S) = \{f : S \to \mathbb{R} |\ ||f||_p < \infty\}$. Para obtener un espacio normado se define la relación de equivalencia:

$$\forall f,g \in L^p(\mu),\ f \sim g \ \leftrightarrow\ f = g\ c.s\ (||f-g||_p = 0). \tag{46}$$

Entonces se define $L^p(S) = \mathcal{L}^p(S)/\sim$. Donde para $1 \leq p \leq \infty$, L_p es un espacio vectorial y si $f \in L^p(\mu)$ se cumple que $||f||_p = 0 \ \leftrightarrow\ f = 0\, \mu c.d$

Para los espacios $\mathbb{R}$ y $\mathbb{R}^2$ se usa la medida de Lebesgue. En $\mathbb{N}x\mathbb{R}$ aplicando el Teorema de Fubini 20 y utilizando el producto de la medida contadora en $\mathbb{N}$ y la medida de Lebesgue en $\mathbb{R}$, se obtiene para $1 \leq p < \infty$ y $\lambda : \mathbb{N}x\mathbb{R} \to \overline{\mathbb{R}}$:

$$\|\lambda\|_p^p = \sum_{k=1}^{\infty} \|\lambda_k(t)\|_p^p. \tag{47}$$

Por teorema de Riez-Fisher se cumple que $(L^p(\mu), ||\cdot||_p)$ es un espacio de Banach.

9.2. Punto de vista probabilístico

Para comenzar con el análisis, Bubenik asume que la persistencia landscapes se encuentra en $L_p(S)$ para algún $1 \leq p < \infty$, donde $S = \mathbb{N}x\mathbb{R}$. De manera que $L_p(S)$ es un espacio de Banach separable, además para $p = 2$ cumple que es un espacio de Hilbert, pero no se utiliza esta estructura, se trabaja con espacios

de Banach($p > 2$) con el objetivo de aplicar toda la teoría existente de variables aleatorias en dicho espacio.

Considerar el descriptor topológico como una variable aleatoria con valores en el espacio resumen S. Sea el espacio de probabilidad subyacente $(\Omega, \mathbb{F}, P)$ que consta de un espacio muestral Ω, una σ-álgebra $\mathbb{F}$ de eventos, y una medida de probabilidad P. Uniendo estas construcciones mediante una función obtenemos:

$$X : (\Omega, \mathbb{F}, P) \ \rightarrow \ (S, \mathcal{A}, P'), \tag{48}$$

donde S es el espacio resumen con la suposición de alguna métrica asociada, $\mathcal{A}$ es la correspondiente σ-álgebra de Borel.

Considerar Λ persistencia landscape correspondiente una variable aleatoria Borel con valores en el espacio de Banach separable $L^p(S)$. Entonces para $\omega \in \Omega$, $X(\omega)$ es el dato y $\Lambda(\omega) = \lambda(X(\omega)) =: \lambda$ es el correspondiente resumen estadístico.

Si tenemos una muestra de variables aleatorias $X_1, \cdots, X_n$ *i.i.d* con la misma distribución que X, es útil tener una buena noción de la media μ de X y la media $\overline{X}_n$, y poder demostrar que $\overline{X}_n \rightarrow \mu$ y ser capaz de calcular $\overline{X}_n(w),\ w \in \Omega$.

Sea $\Lambda_1, \cdots, \Lambda_n$ las persistencia landscapes correspondientes a las muestras. Usando la estructura de espacio vectorial de $L^p(S)$, la media landscapes está dada por la media puntual ($\overline{\Lambda}(\omega) = \overline{\lambda}$).

$$\overline{\lambda}_k(t) = \frac{1}{n} \sum_{i=1}^{n} \lambda_k^{(i)}(t). \tag{49}$$

Para poder decir que la media landscape converge a la persistencia landscape esperada, se deben conocer algunos resultados de probabilidad en espacio de Banach explicados en [14], a continuación se exponen los de mayor importancia.

Usando la Ley Fuerte de los Grandes Números 44 y aplicándolo a la persistencia landscape obtenemos:

***Teorema* 6** $\overline{\lambda}^n(X) \rightarrow E(\lambda(X))$ *(c.s) ssi* $E||\lambda(X)|| < \infty$. *Aquí* $||\cdot||$ *corresponde a la norma* L_p.

De manera similar se adapta el Teorema Central del Límite 22 para persistencia landscape:

***Teorema* 7** *Asumiendo* $\lambda(X) \in L^p(S)$ *con* $2 \leq p < \infty$. *Si* $E||\lambda(X)|| < \infty$ *y* $E(||\lambda(X)||^2) < \infty$ *entonces* $\sqrt{n}[\overline{\lambda}^n(X) - E(\lambda(X))]$ *converge débil a una variable aleatoria Gaussiana con matrix de covarianza* $Var[\lambda(X)]$.

Como se menciona en [14] los resultados de los dos teoremas anteriores pueden ser aplicados para obtener una variable aleatoria de valor real que satisface el Teorema Central del Límite.

***Corolario* 1** *Sea* $\lambda(X) \in L^p(S)$ *donde* $2 \leq p < \infty$ *con* $E||\lambda(X)|| < \infty$ *y* $E(||\lambda(X)||^2) < \infty$. *Entonces para cualquier* $f \in L^q(S)$ *con* $\frac{1}{p} + \frac{1}{q} = 1$, *sea*

$$Y = \int_S f\lambda(X) = ||f\lambda(X)||_1. \tag{50}$$

La segunda igualdad muestra que en este caso se trabaja con norma 1. *De esto se deduce que:*

$$\sqrt{n}[\overline{Y}_n - E(Y)] \xrightarrow{d} N(0, Var(Y)), \tag{51}$$

donde d denota la convergencia en distribución 43

9.3. Medida de similitud

A continuación se introduce una explicación de la distancia basada en persistencia landscape y se aplica para demostrar que la *Persistencia Landscape* es una resumen estadístico estable:

Suponer que se tienen 2 muestras y son denotados los landscapes correspondientes: $\lambda_k(t)$ y $\lambda_k^{'}(t)$. Entonces la distancia puede ser medida por la ecuación 52, la cual compara por pares el área debajo de las curvas de nivel k superiores entre λ y $\lambda^{'}$.

$$||\lambda - \lambda^{'}||_p = \left[\sum_k \int_{\mathbb{R}} |\lambda_k(t) - \lambda_k^{'}(t)|^p\right]^{1/p}. \tag{52}$$

Uno de los inconvenientes de la medida de distancia considerada en la anterior ecuación es que sólo los paisajes k superiores se consideran y el resto se descartan.

Para introducir el concepto de estabilidad, recordar que dada una función de valores reales $f : X \to \mathbb{R}$ en un espacio topológico X; $M(f)$ denota el correspondiente módulo de persistencia, donde $M(f)(a) = H(f^{-1}((\infty, a]))$.

***Teorema* 8** *(**Teorema de Estabilidad ∞-Landscape**) Sea $f,g : X \to \mathbb{R}$. Entonces:*

$$\Lambda_\infty(M(f), M(g)) \leq ||f - g||_\infty. \tag{53}$$

El teorema anterior implica la estabilidad de la persistencia landscapes con respecto a la norma del supremo, es válido destacar que las funciones no cumplen ningún supuesto.

A continuación se muestra brevemente como la distancia landscapes aporta límites inferiores para la distancia Bottleneck y Wasserstein [42]:

Sea D un diagrama de persistencia, $Pers_k(D) = \sum_j l_j^k$ persistencia total del diagrama $D = \{x_j\}$. Sea $D^{'}$ un diagrama equivalente al que se le añade una cantidad de puntos en la diagonal según sea necesario, esto es razonable ya que los puntos en la diagonal tienen persistencia cero. Notar que cada diagrama de persistencia tiene un único representante $\widehat{D}$ sin puntos en la diagonal ($|D| = |\widehat{D}|$).

Al permitirnos poder agregar una mayor cantidad de puntos en el diagonal se establecen biyecciones entre los diagramas: $\phi : D \to D'$ que puede ser representado por $\phi : x_j \to x'_j$, donde $j \in J$ con $|J| = |D| + |D'|$, donde $\varepsilon_j = ||x_j - x'_j||_\infty = \max(|b_j - b'_j|, |d_j - d'_j|)$.

La distancia de Bottleneck [78] entre diagramas de persistencia D y D' está dada por:

$$W_\infty(D, D') = inf_{\phi: D \to D'} sup_j \varepsilon_j, \tag{54}$$

donde se toma el ínfimo sobre todas las biyecciones de D a D'. Por lo que para el diagrama vacío ($W_\infty(D, \emptyset) = \frac{1}{2} sup_j l_j$.) Entonces la distancia landscape está acotada por Bottleneck:

***Teorema* 9** *Para los diagramas de persistencia D y D',*

$$\Lambda_\infty(D,D') \leq W_\infty(D,D'). \tag{55}$$

Para $p \geq 1$ la distancia p-Wasserstein [64] está dada por:

$$W_p(D,D') = inf_{\phi:D\to D'}[\sum_j \varepsilon_j^p]^{\frac{1}{p}}. \tag{56}$$

La distancia landscape tiene una mayor relación con la versión ponderada de la distancia Wasserstein:

$$\overline{W}_p(D,D') = inf_{\phi:D\to D'}[\sum_j l_j\varepsilon_j^p]^{\frac{1}{p}}. \tag{57}$$

Supongamos que los diagramas son finitos. El siguiente resultado limita la distancia p-landscape.

***Teorema* 10** *Si $n = |D| + |D'|$ entonces:*

$$\Lambda_p(D,D')^p \leq min_{\phi:D\to D'}\left[\sum_{j=1}^{n} l_j\varepsilon_j + \frac{2}{p+1}\sum_{j=1}^{n}\varepsilon_j^{p+1}\right]. \tag{58}$$

Del teorema se obtiene un límite inferior para la distancia Wasserstein.

***Corolario* 2**

$$W_p(D,D')^p \geq min\left(1, \frac{1}{2}\left[W_\infty(D,\emptyset) + \frac{1}{p+1}\right]^{-1}\Lambda_p(D,D')^p\right). \tag{59}$$

Usando lo expuesto anteriormente se obtiene como resultado final el teorema de estabilidad:

***Teorema* 11** *(**Teorema de estabilidad p-Landscape**) Sea X un espacio métrico compacto triangulable que implica que la persistencia total de grado k esta limitada por algún número real $k \geq 1$, sea f y g funciones Liptchiz tame. Entonces:*

$$\Lambda_p(D(f),D(g))^p \leq C||f-g||_\infty^{p-k}, \tag{60}$$

$\forall p \geq k$, *donde* $C = C_{X,k}||f||_\infty(Lip(f)^k + Lip(g)^k) + C_{X,k+1}\frac{1}{p+1}(Lip(f)^{k+1} + Lip(g)^{k+1})$.

De esta forma el diagrama de persistencia es estable con respecto a la distancia p-landscape si $p > k$ donde X está acotado por $Pers_k(D) = \sum_j l_j^k$. De igual manera la persistencia landscape es estable con respecto a la p-norma si $p > k$.

Se trabaja actualmente en algunos enfoques para inferir estadísticos sobre la persistencia landscape. Por ejemplo la obtención de intervalos de confianza para la media landscape mediante el método Bootstrap [68]. Prueba de hipótesis sobre dos muestras X y Y de v.a con el objetivo de ser capaces de probar hasta que momento la hipótesis nula $\mu_X = \mu_Y$ es verdadera.

9.4. Bootstrap para persistencia landscape

En [70] se demuestra que la persistencia landscape media converge débilmente a un proceso gaussiano y se construye una banda de confianza con 95% para la media landscape usando el método de *multiplier bootstrap*.

***Teorema* 12** *(Banda de confianza para persistencia landscapes [68]) El intervalo $C_n(t)$ indexado por $t \in \mathbb{R}$, se define por $C_n(t) = [\overline{\lambda}_k(t) - \frac{q_\alpha}{\sqrt{n}}, \overline{\lambda}_k(t) + \frac{q_\alpha}{\sqrt{n}}]$ es una banda de confianza para $\mu(t)$:*

$$lim_{n\to\infty}\mathbb{P}(\mu(t) \in C_n(t)\ \forall t) \geq 1-\alpha. \tag{61}$$

Las bandas de confianzas son una herramienta importante para la inferencia estadística, ya que permiten cuantificar y visualizar la incertidumbre acerca de la función de persistencia landscape media μ, y para detectar ruido topológico. Un ejemplo dado por Fabrizio Lecci en [57] lo podemos ver en fig:28.

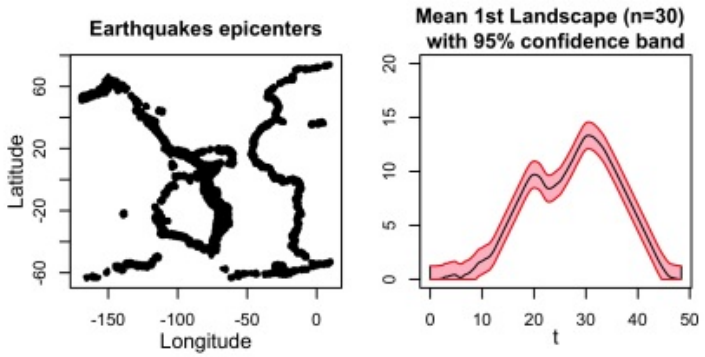

Fig. 28. El primer gráfico muestra 8000 epicentros de los terremotos en latitud/longitud $[-75,75]$x$[100,100]$ de magnitud mayor que 0,5 registrado entre 1970 y 2009. Se muestrearon aleatoriamente $m = 400$ epicentros y se calculo el diagrama de persistencia aproximado de la función distancia (Betti 1). Se repite el procedimiento en 30 momentos, y se calcula la media empírica landscape $\overline{\lambda}_n$. Usando bootstrap multiplicador se obtiene una banda de confianza con 95% para $\mu(t)$.

Se supone que $Y_1, Y_2, \cdots, Y_n$ son muestras (i.i.d.) de una variable aleatoria Y y similarmente $Y'_1, Y'_2, \cdots, Y'_n$ son iid de Y'. Asumimos que se mantienen los supuestos anteriores y Y se define como en 50. Definimos $\mu = E(Y)$ y $\mu' = E(Y')$. Vamos a probar la hipótesis de interés nula: $\mu = \mu'$:

Primero, observar que la media muestral $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n}(Y_i)$ es un estimador insesgado de μ y la varianza de la muestra $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ es un estimador insesgado de $Var(Y)$ similares resultados se aplican para $\overline{Y'}$ y $s_{Y'}^2$.
Suponiendo que las muestras están vinculadas (antes y después) el t-test estadístico para dos muestras se define como:

$$t = \frac{X_D - \mu_0}{s_D/\sqrt{n}} = \frac{\overline{Y} - \overline{Y'}}{\sqrt{\frac{s_Y^2}{n} + \frac{s_{Y'}^2}{n'}}}. \tag{62}$$

Para obtener q_α, se definen un conjunto de variables aleatorias gaussianas con media 0 y varianza 1 $(\xi_1^n = (\xi_1, \cdots, \xi_n))$ y se define el proceso de *multiplier bootstrap*.

$$\widetilde{\mathbb{G}}_n(f_t) = \widetilde{\mathbb{G}}_n(\lambda_1^n, \xi_1^n)(f_t) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i(f_t(\lambda_i) - \overline{\lambda}_n(t)),\ t \in [0,T]. \tag{63}$$

Sea $q(\alpha)$ como el único valor tal que:

$$\mathbb{P}\left(sup_t|\widetilde{\mathbb{G}}_n(f_t)| > q_\alpha|\ \lambda_1,\cdots,\lambda_n\right) = \alpha. \tag{64}$$

Donde $q(\alpha)$ puede ser aproximado por simulación Monte Carlo. Sea $\widetilde{\theta} = sup_{t\in[0,T]}|\widetilde{\mathbb{G}}_n(f_t)|$ una muestra bootstrap; repetimos el proceso B-veces, se obtienen B valores de rendimiento $\widetilde{\theta}_1,\cdots,\widetilde{\theta}_n$:

$$\widetilde{q}(\alpha) = inf\left\{z : \frac{1}{B}\sum_{j=1}^{B} I(\widetilde{\theta}_j > z) \leq \alpha\right\}. \tag{65}$$

Se toma B tan grande como se quiera para que el error de Monte Carlo sea arbitrariamente pequeño. Por lo tanto cuando se utiliza el método bootstrap uno ignora el error en el aproximación. Una explicación más detallada de este método la encontramos en [70,68]. Actualmente se investigan métodos para estimar la diferencia entre la media landscape de submuestras y la media del conjunto de datos originales.

10. Función rango

La función rango es otro enfoque funcional para la homología persistente que consiste en una función con valores enteros, de dos variables reales y se puede considerar como una función de distribución acumulativa del diagrama de persistencia. Dado que la función rango es justamente una función, es posible aplicar técnicas estadísticas para analizar su distribución. Dicha función está relacionada con la función tamaño [80] y también se ha definido para persistencia multidimensional en el caso de las filtraciones que se construyen a partir de dos o más parámetros. Este enfoque funcional al igual que la persistencia landscape proporciona un marco más simple para inferir estadísticos, por ejemplo medias y varianzas de los descriptores topológicos. En este capítulo se exponen los conocimientos necesarios para poder definir el nuevo enfoque apoyándonos en [81].

Se retorna a la definición de grupos de homología persistente y cuantificarlos por su rango mediante los llamados números Betti:

$$\beta_k(a,b) := rang\ H_k(a,b), \quad a < b. \tag{66}$$

Como se conoce el grupo de homología k-ésimo de K se define como:

$$H_k(K) := Z_k(K)/B_k(K). \tag{67}$$

Considerando la definición de grupos de homología persistente de una filtración. Una filtración $K = \{K_r | r \in \mathbb{R}\}$ es un complejo simplicial numerable indexado sobre los números reales tales que K_a es un complejo simplicial y $K_a \subseteq K_b$ para $a \leq b$. Se desea describir como la topología de la filtración cambia cuando el parámetro aumenta. Para $a \leq b$ tenemos la inclusión de complejos simpliciales: $i : K_a \to K_b$ que induce los mapas de inclusión:

$$i_B : B_k(K_a) \to B_k(K_b) \quad i_Z : Z_k(K_a) \to Z_k(K_b). \tag{68}$$

Estas inclusiones inducen homomorfismos que generalmente no son inclusiones en los grupos de homología:

$$i_k^{a\to b} : H_k(K_a) \to H_k(K_b). \tag{69}$$

La imagen de $i_k^{a\to b}$ consta de las clases de equivalencia de los ciclos que estaban presentes en K_a, donde la equivalencia homológica es medida con respecto a los límites en K_b. Por lo tanto se define el grupo de homología persistente para $a \leq b$ como:

$$H_k(a,b) = i(Z_k(K_a))/(B_k(K_b) \cap i(Z_k(K_a))). \tag{70}$$

Observar que $H_k(a,a) = H_k(a)$ y que $\beta_k(a,a) = \beta_k(K_a)$ los cual motiva a el uso de la misma simbología para la función rango como una generalización de los números Betti.

Sea $\mathbb{R}^{2+} := \{(x,y) \in (-\infty \cup \mathbb{R})x(\mathbb{R} \cup \infty) : x < y\}$. Se define la función rango k-dimensional correspondiente a la filtración de K como:

$$\beta_k(K) : \mathbb{R}^{2+} \to \mathbb{Z} \tag{71}$$

$$(a,b) \to dim\, H_k(a,b).$$

Existen dos argumentos para el nombre *rango*; $B_k(K)(a,b)$ es el rango de $H_k(a,b)$ visto cómo un módulo sobre $\mathbb{Z}_2$. Alternativamente dado que $H_k(a,b)$ es isomorfo a $i_k^{a\to b}(H_k(K_a))$ sabemos que $\beta_k(K)(a,b)$ es el rango de la función $i_k^{a\to b} : H_k(K_a) \to H_k(K_b)$.

Se denomina a una filtración *tame* si $\dim H_k(a,b)$ es finita para toda $a < b$ y se restringe el análisis a las filtraciones *tame*, que se satisface con cualquier aplicación que implique datos finitos.

La función rango contiene la misma información que los diagramas de persistencia y los códigos de barra usados frecuentemente. En nuestro caso decimos que una clase de homología $\alpha \in H_k(K_{b(\alpha)})$ nace en b_α si está en el co-núcleo de $i_k^{s\to b(\alpha)}$ para cualquier $s < b(\alpha)$; es decir si α no es la imagen de cualquier ciclo que se produce a principios de la filtración. La clase de homología de α muere en $d(\alpha)$ si para $b(\alpha) < t < d(\alpha)$ tenemos $\alpha \notin ker(i_k^{b(\alpha)\to t})$ pero $\alpha \in ker(i_k^{b(\alpha)\to d(\alpha)})$. De manera informal podemos pensar en el proceso de morir como el ciclo de convertirse en un límite o la fusión de dos ciclos. La diferencia $d(\alpha) - b(\alpha)$ se denomina persistencia del ciclo α. Algunos ciclos pueden nunca morir y se les denomina como *clases esenciales*.

La funciones rangos pueden ser tratadas como una función acumulativa del diagrama de persistencia. Esto es debido a que como la cantidad $\beta_k(a,b)$ es igual al número de puntos del diagrama de persistencia en la región $(-\infty,a]\text{x}[b,\infty)$. Debido a la correspondencia es fácil obtener la función rango de los diagramas de persistencia y viceversa.

Si β es una función rango entonces para $a \leq c \leq b \leq d$ tenemos:

$$\beta(c,b) - \beta(a,b) - \beta(c,d) + \beta(a,d) \geq 0. \tag{72}$$

Como se demostró en el capítulo anterior, el conjunto de funciones landscape forman un espacio de Banach, en este nuevo enfoque Turner [81] demuestra que las funciones rango forman un subconjunto de un espacio de Hilbert30 de funciones. Los espacio de Hilbert constituyen una generalización del concepto de espacio euclídeo, permite que nociones de técnicas algebraicas y geométricas aplicables a espacios de dimensión dos y tres se extiendan a espacios de dimensión arbitraria, incluyendo espacios de dimensión finita.

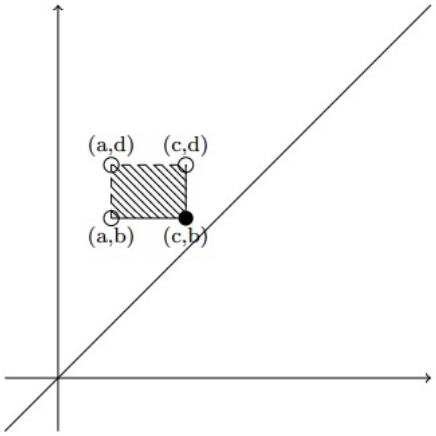

Fig. 29. $\beta(c,b) - \beta(a,b) - \beta(c,d) + \beta(a,d)$ para la función rango β es el número de puntos en el diagrama de persistencia correspondiente que vemos en la región sombreada.

10.1. Métrica asociada

La función rango surge en el espacio de funciones de valores reales del espacio métrico Riemanniano $(\mathbb{R}^{2+}, g)$. Considerar una métrica $d(\cdot,\cdot)$ en el espacio de funciones definido por:

$$d(f,h)^2 = \int_{\mathbb{R}^{2+}} (f-h)^2 d\mu, \tag{73}$$

donde μ es la medida en $\mathbb{R}^{2+}$ correspondiente a la métrica g. Existen diferente elecciones para la métrica y por lo tanto también para su correspondiente medida. Pero surge el problema de que dependiendo de la elección de μ, la distancia por pares entre las funciones rango puede ser infinita. Por ejemplo, seleccionando la medida de Lebesgue la distancia entre funciones rango sólo puede ser finita cuando sus respectivos conjuntos de *clases esenciales* tienen igual tiempo de nacimiento. De lo contrario existe una región infinita en el plano donde las funciones de rango difieren y la medida de Lebesgue de la región es infinita. Para atacar el problema se consideran diferentes medidas en $\mathbb{R}^{2+}$ obtenida por la multiplicación de la medida de Lebesgue por una función de peso:

$$\mu((x,y)) = \phi(y-x)\lambda((x,y)). \tag{74}$$

Tomando ϕ en función de $(y-x)$ implica que la medida μ es una función de la persistencia o la vida útil de cada clase de homología. Dada la función de ponderación la función distancia está dada por:

$$d_\phi(f,h)^2 = \int_{x<y} (f-h)^2 \phi(y-x) dx dy. \tag{75}$$

El siguiente lema establece condiciones suficientes para garantizar que las funciones rango están separadas a una distancia finita. Luego dado un conjunto de funciones rango, por pares con distancia finita, podemos considerar el *espacio afín* en el que se encuentran para obtener la media y realizar análisis de componentes principales.

***Lema* 3** *Sea* $\phi : [0,\infty] \to [0,\infty]$ *tal que* $\int_0^\infty \phi(t)dt < \infty$. *Sea* f_K *y* f_L *las funciones rango construidas a partir de las filtraciones* K_t *y* L_t. *Si* K_∞ *y* L_∞ *son complejos simpliciales finitos con* $H_*(K_\infty) = H_*(L_\infty)$ *y* $H_*(K_{-\infty}) = H_*(L_{-\infty})$ *entonces* $d_\phi(f_K, f_L) < \infty$.

Fijando una función $h : \mathbb{R}^{2+} \to \mathbb{R}$ y considerando el espacio de funciones, denotado por $A(h)$, que se encuentran a una distancia finita de h. Este es un *espacio afín* que se puede transformar en un espacio de vectores, $V(h)$, centrado sobre h: $V(h) = \{f - h : f \in A(h)\}$. Además se puede definir un producto interno en $V(h)$ utilizando:

$$\langle v_1, v_2 \rangle = \int_{\mathbb{R}^{2+}} v_1 v_2 \, d\mu. \tag{76}$$

Sea $\beta^1, \beta^2, \cdots, \beta^n$ funciones de rango de dimensión k. Se define la función rango media:

$$\overline{\beta}(a,b) = \frac{1}{n} \sum_{i=1}^{n} \beta^i(a,b) \quad \forall (a,b) \in \mathbb{R}^{2+}. \tag{77}$$

La función rango media es poco probable que sea una función de homología persistente, ya que probablemente contiene algunos valores enteros, pero sin embargo se puede demostrar que conserva la propiedades de monotonía de funciones de homología persistente.

En [81] ofrecen los primeros pasos para realizar **Análisis de Componentes Principales** mediante el uso de la matriz de productos internos de un conjunto de funciones centradas. A continuación se analiza el caso cuando lo datos son de dimensión finita. Se trabaja con una adaptación de PCA para datos funcionales, y se representa cada función por sus valores, donde cada una se convierte en una fila de la matriz X y cada columna es una coordenada en el que se evalúan las funciones. Las componentes principales son los vectores propios de $X^T X$ ordenados por el mayor valor propio correspondiente. Las entradas en la matriz producto son funciones cuyo producto interno se define con respecto a la función de ponderación 75, dicha representación interpreta las matrices como operadores lineales de la forma siguiente:

Dado un conjunto de n funciones rango $\{f_1, f_2, \cdots, f_n\}$ se define los operadores lineales.

$$X : L^2(\mathbb{R}^{2+}, g) \to \mathbb{R}^n$$

$$g \to (< g, f_1 >, < g, f_2 >, \cdots, < g, f_n >)$$

y

$$X^T : \mathbb{R}^n \to L^2(\mathbb{R}^{2+}, g)$$

$$(a_1, a_2, \cdots, a_n) \to \sum_{i=1}^{n} a_i f_i$$

Para encontrar las funciones de ponderación de las componentes principales $\varsigma_1, \varsigma_2, \cdots, \varsigma_n$. Se calcula primeramente los valores y vectores propios de la matriz de productos internos $(< f_i, f_j >)_{i,j=1}^n$ y entonces el conjunto ς_k es la función de norma unitaria que es un múltiplo escalar de $X^T w_k$, (w_k vectores propios)

La funciones rango como descriptor topológico significa un paso de avance en el análisis de componentes principales con el objetivo de poder distinguir entre procesos puntuales espaciales generados por diferentes modelos. Pero siguen existiendo cuestiones teóricas sobre su análisis por explorar. Como la robustez de la distancia por pares entre las funciones de rango que están bajo diferentes elecciones de la función de ponderación 75 entre otras.

11. Categorización de la homología persistente

En los últimos años la topología algebraica ha experimentado un proceso de abstracción de sus fundamentos teóricos, este proceso ofrece la posibilidad de aclarar algunas ideas y pruebas, permitiendo generalizar y aplicar todos sus conceptos. El desarrollo y uso de la teoría de categorías es una parte crítica dentro de este proceso. Pensado cómo un marco de referencia en el cual muchas de las definiciones admiten generalizaciones, y donde se puede considerar más modos de persistencia. ¿Por qué aplicar teoría de categorías?:

- Permite un tratamiento uniforme para varias versiones de persistencia.
- Ofrece pruebas simples, comunes a algunos resultados básicos de persistencia.
- Elimina supuestos.
- Aplica persistencia para enfoques más generales.
- Permite que el funtor homología sea sustituido por otros funtores.
- Proporciona un marco para nuevas aplicaciones.

Bubenick y Scott [82] han trabajado sobre categorización de la homología persistente. Este enfoque si bien se ha centrado sobre los diagramas de tipo finito $(\mathbb{R},\leq)$ puede ser aplicable a las categorías más generales de diagramas de espacios vectoriales. Estudian la categoría de los funtores $(\mathbb{R},\leq) \to Vect_k$ y se demuestra que la categoría de los módulos de persistencia es abeliana. Se aprovecha esto para demostrar el teorema de estabilidad: para funciones arbitrarias no necesariamente continuas ($\mathbb{X} \to \mathbb{R}$) de un espacio topológico, y para cualquier funtor H de un espacio topológico a una categoría de diagramas reales indexados en una categoría abeliana D. La distancia de intercalado entre los diagramas generada por la aplicación de H para las filtraciones de los conjuntos subnivel de las funciones es acotado superiormente por la distancia L_∞ de las funciones. Además Bubenick y Scott demuestran que muchas de las categorías que emergen de forma natural en la homología persistente son abelianas.

A continuación se ofrecen algunas definiciones básicas de la teoría de categorías, útiles para poder entender los razonamientos que se muestran en el capítulo:

***Definición* 13** *Una categoría* $\boldsymbol{C}$ *consta de los siguiente datos:*

1. *Un conjunto Ob(*$\boldsymbol{C}$*) de objetos,* $(A,B,\cdots)$
2. *Para cada par de objetos de* $\boldsymbol{C}$*, conjunto de morfismos Mor(*$\boldsymbol{C}$*),* $(f,g,\cdots)$
3. *Para cada terna de objetos* (A,B,C) *de* $\boldsymbol{C}$*, una ley de composición:*

$$\circ : \boldsymbol{C}(A,B)\, x\, \boldsymbol{C}(B,C) \to \boldsymbol{C}(A,C)$$

tal que:

- *Axioma* 1*:* $f : A \to B$*;* $g : B \to C$*;* $H : C \to D$

$$h \circ (g \circ f) = (h \circ g) \circ f$$

- *Axioma* 2*: para cada objeto A existe un morfismo*

$$1_A : A \to A \in \boldsymbol{C}(A,A)$$

tal que para cualquier $f : A \to B \wedge g : C \to A$

$$1_A \circ g = g \ \wedge \ f \circ 1_A = f$$

Ejemplos de categorías:

- Una categoría $\mathbf{Vec}^K$ de espacios vectoriales sobre un campo K. Los objetos son espacios vectoriales sobre K y los morfismos son transformaciones lineales. Una generalización de esto es cuando R es un anillo conmutativo con identidad y consideramos las categoría $\mathbf{Mod}_R$ de módulos unitarios sobre R y morfismos de módulos.
- Todo espacio topológico (X,T) es una categoría. La clase de elementos son los abiertos de la topología $T(X)$, si $U,V \in T(X)$ y $U \subseteq V$ entonces $Hom_X(U,V) = \left\{ i_U^V \right\}$ donde i_U^V es la función de inclusión de U a V. En caso contrario, $Hom_X(U,V) = \emptyset$.
- Una categoría Gr^f dos grafos dirigidos y homomorfismos de grafos dirigidos, con la ley de composición usual.
- Sea $(X,\leq)$ un conjunto pre-ordenado, considerar una categoría $\mathbf{C}^{(X,\leq)}$ que tiene como objetos los elementos de X, siendo:

$$\mathbf{C}^{(X,\leq)}(x,y) = \begin{cases} \{x \to y\} & si\, x \leq y \\ \\ 0 & otro\ \ caso \end{cases}$$

Se consideran dos ejemplos clásicos en el que se aplica la homología persistente y se muestran como encajan en el marco categórico. Además se puede notar como los diagramas indexados por $[n]$, $(\mathbb{Z}_+,\leq)$ y $(\mathbb{Z},\leq)$ son casos especiales de los diagramas indexados por $(\mathbb{R},\leq)$. Finalmente la homología persistente es definida:

Sea K un complejo simplicial finito, con una filtración:

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K. \tag{78}$$

Entonces, se obtiene un diagrama de espacios topológicos indexado por $[n]$, i.e. $K \in \mathbf{Top}^{[n]}$, con $K(i) = K_i$ y $K(i \leq j)$ dado por la inclusión.

Sea H_k el funtor de homología simplicial de grado k con coeficientes en un campo $\mathbb{F}$. Entonces H_kK es un diagrama $[n]$-indexado de espacios vectoriales de dimensión finita. Esto es, $H_kK(i) = H_k(K_i,\mathbb{F})$y $H_kK(i \leq j)$ es el mapa inducido en la homología con la inclusión $K_i \to K_j$. Así que $H_kK \in \mathbf{Vec}^{[n]}$ (categoría de un espacio vectorial de dimensión finita). Es posible resumir la homología en todos los grados para obtener $HF \in \mathbf{Vec}^{[n]}$, dada por $HF(i) = \oplus_k H_k(K_i,\mathbb{F})$.

11.1. Conjuntos de subnivel

Sea X un espacio topológico, y sea $f : X \to \mathbb{R}$ función de valores reales no necesariamente continua en X. Sea $a \in \mathbb{R}$, considerar los conjuntos subnivel:

$$f^{-1}((-\infty,a]) = \{x \in X | f(x) \leq a\}. \tag{79}$$

Se trabaja como un espacio topológico considerando la topología del subespacio. Notemos que si $a \leq b$ entonces $f^{-1}(-\infty, a] \subseteq f^{-1}(-\infty, b]$ y la inclusión es una aplicación continua. Los datos se pueden agrupar en un diagrama de espacios topológicos indexado en $(\mathbb{R}, \leq)$, $F \in \mathbf{Top}^{(\mathbb{R},\leq)}$. Para $a \in \mathbb{R}$, se define $F(a) = f^{-1}(\infty, a]$, y para $a \leq b$ se define $F(a \leq b)$ como la inclusión $f^{-1}(-\infty, a] \to f^{-1}(-\infty, b]$. Es fácil comprobar que esto define una funtor $F : (\mathbb{R}, \leq) \to \mathbf{Top}$.

Sea H_k el k-ésimo funtor de homología singular con coeficiente en algún campo $\mathbb{F}$. Entonces $H_kF(a) = H_k(f^{-1}(-\infty, a], \mathbb{F})$, y para $a \leq b$, $H_kF(a \leq b)$ induce el mapa en la homología por la inclusión $f^{-1}(-\infty, a] \to f^{-1}(-\infty, b]$, si f posee la propiedad de que para todo $a \in \mathbb{R}$, $H_k(f^{-1}(-\infty, a], \mathbb{F})$ es un espacio vectorial de dimensión finita, entonces $H_kF \in \mathbf{Vec}^{(\mathbb{R},\leq)}$.

Si f tiene las propiedades que $\forall a \in \mathbb{R}, H_*(f^{-1}(-\infty, a], \mathbb{F})$ es de dimensión finita, entonces $HF \in \mathbf{Vec}^{(\mathbb{R},\leq)}$ esta dado por $HF(a) = \oplus_k H_k(f^{-1}(-\infty, a], \mathbb{F})$.

11.2. Diagramas por $[n], (\mathbb{Z}_+, \leq)$ y $(\mathbb{Z}, \leq)$

En lo expuesto anteriormente se ha considerado la categoría indexada sobre $(\mathbb{R}, \leq)$. Sin embargo se incluyen los casos para $[n], (\mathbb{Z}_+, \leq)$ y $(\mathbb{Z}, \leq)$ mediante la siguiente observación. Considerar $F \in \mathbf{Top}^{[n]}$; entonces es posible extender F a un diagrama indexado $(\mathbb{R}, \leq)$ como sigue. El funtor inclusión $\mathbf{i} : [n] \to (\mathbb{R}, \leq)$ dada por $\mathbf{i}(j) = j$ tiene un funtor retracción $\mathbf{r} : (\mathbb{R}, \leq) \to [n]$ dada por:

$$\mathbf{r}(\text{a})= \begin{cases} 0 \quad si \quad a \leq 0 \\ \lfloor a \rfloor \; si \; 0 < a < n. \\ n \quad si \quad a \geq n \end{cases} \quad (80)$$

Por lo tanto la función compuesta $F\mathbf{r}$ es un elemento de $\mathbf{Top}^{(\mathbb{R},\leq)}$ y $F\mathbf{ri} = F$. Se define de manera análoga los funtores retracción para $(\mathbb{Z}_+, \leq)$ y $(\mathbb{Z}, \leq)$.

***Definición* 14** *Homología Persistente*
Dado un diagrama $F \in \mathbf{Top}^{(\mathbb{R},\leq)}$, se define el k-ésimo grupo de homología p-persistente de $F(a)$ como la imagen del mapa $H_kF(a \leq a+p)$.

***Definición* 15** *Módulo de persistencia*
Diagramas en $\mathbf{Vec}^{[n]}$, $\mathbf{Vec}^{(\mathbb{Z}_+,\leq)}$, $\mathbf{Vec}^{(\mathbb{R},\leq)}$ se conocen como los módulos de persistencia.

11.3. *Interleavings* de diagramas

Considerando la categoría $(\mathbb{R}, \leq)$, cuyos objetos son números reales y el conjunto de morfismos de a a b consiste de un único morfismo si $a \leq b$ y en otro caso es vacío. Para $b \geq 0$, se define $T_b : (\mathbb{R}, \leq) \to (\mathbb{R}, \leq)$ a el funtor dado por $T_b(a) = a + b$ y se define $\eta_b(a) : a \leq a + b$. Notar que $T_bT_c = T_{b+c}$ y que $\eta_b\eta_c = \eta_{b+c}$.

Sea D una categoría; $\varepsilon \geq 0$ y sea $F, G \in D^{(\mathbb{R},\leq)}$:

***Definición* 16** *Un ε-intercalado de F y G consiste de transformaciones naturales $\varphi : F \to GT_\varepsilon$ y $\psi : G \to FT_\varepsilon$, i.e. tal que:*

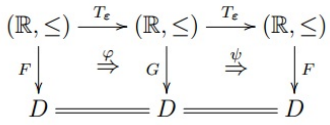

Fig. 30. *Interleaving*.

$$(\psi T_\varepsilon)\varphi = F\eta_{2\varepsilon} \ \wedge \ (\varphi T_\varepsilon)\psi = G\eta_{2\varepsilon}$$

Si (F,G,φ,ψ) es una ε-intercalados, entonces F y G están ε-intercalados. La existencia de las transformaciones naturales ψ y φ implican la existencia de diagramas conmutativos para todo $a \leq b$ [82].

***Definición* 17** *Decimos que $d(F,G) \leq \varepsilon$ si F y G están ε-intercalados. Explícitamente:*

$$d(F,G) = inf\{\varepsilon \geq 0 \mid F \wedge G \ \ están \ \ \varepsilon - intercalados\}, \tag{81}$$

donde el conjunto $d(F,G) = \infty$ si F y G no están ε-intercalados para cualquier $\varepsilon \geq 0$.

***Teorema* 13** *La función d definida anteriormente es un pseudo-métrica extendida en cualquier subconjunto de la clase de diagramas indexado-$(\mathbb{R},\leq)$ en D.*

***Corolario* 3** *Si los diagramas cuya distancia intercalación es 0 son identificados, entonces d es una métrica extendida en este conjunto de las clases de equivalencia.*

Desde el punto de vista categórico, los cálculos de homología persistente se llevan a cabo en los diagramas en la categoría **Vec** de los espacios vectoriales de dimensión finita sobre un campo fijo F. En [82] encontramos un estudio acerca de los diagramas $(\mathbb{R},\leq)$-indexado en **Vec** y se definen algunas configuraciones usuales en la persistencia topológica: diagramas de persistencia, códigos de barra y la distancia bottleneck. El principal resultado que se muestra, es una isometría del conjunto de códigos de barra con la distancia de bottleneck y el conjunto de objetos de $\mathbf{Vec}^{(\mathbb{R},\leq)}$ con la distancia intercalación.

***Definición* 18** *Dado un intervalo $I \subseteq \mathbb{R}$, se define el diagrama $\chi_i \in \mathbf{Vec}^{(\mathbb{R},\leq)}$ por:*

$$\chi_I(a) = \begin{cases} \mathbb{F} & si\ a \in I \\ 0 & otro\ \ caso \end{cases}, \qquad \chi_I(a \leq b) = \begin{cases} Id_{\mathbb{F}} & si\ a,b \in I \\ 0 & otro\ \ caso \end{cases}. \tag{82}$$

El diagrama $F \in \mathbf{Vec}^{(\mathbb{R},\leq)}$ tiene tipo finito si $F \cong \oplus_{k=1}^{N}\chi_{I_k}$. Destacar que $\chi_{\mathbb{R}}$ y $\chi_{\emptyset}$ son los funtores constantes $\mathbb{F}$ y 0 respectivamente.

11.4. Diagramas de persistencia y códigos de barra

A continuación se definen los resúmenes topológicos para diagramas de tipo finito en $\mathbf{Vec}^{(\mathbb{R},\leq)}$, donde dichos diagramas son una categorización de los códigos de barra finito.

***Definición* 19** *Asumiendo que $F \in \mathbf{Vec}^{(\mathbb{R},\leq)}$ es de tipo finito. Un código de barra es un multiconjunto de intervalos. El código de barra de F es el multiconjunto $\{I_k\}_{k=1}^{n}$, donde $F \cong \oplus_{k=1}^{n}\chi_{I_k}$. El diagrama de persistencia de F es el multiconjunto $\{(a_k,b_k)\}_{k=1}^{n}$ donde $a_k \leq b_k$ son los puntos finales de I_k.*

Notar que un código de barra es un multiconjunto finito de intervalos, no un multiconjunto de intervalos finitos.

***Corolario* 4** *Categorización de los códigos de barra*
Existe una biyección entre las clases de isomorfismos de diagramas de tipo finito en $\boldsymbol{Vec}^{(\mathbb{R},\leq)}$ *y códigos de barra finitos.*

11.5. Distancia Bottleneck

A continuación se define la distancia bottleneck entre dos códigos de barra en términos de la distancia intercalación.

***Definición* 20** *Dados los multiconjuntos A y B, se define el multiconjunto* A_B *siendo la unión disjunta de A y el multiconjunto que contiene el intervalo vacío con cardinalidad* $|B|$*. Una biyección estable en dos multiconjuntos A y B es una biyección,* $f : A_B \to B_A$*.Se puede escribir como:* $f : A \rightleftharpoons B$

***Definición* 21** *Sea A y* A' *dos códigos de barra. Se define la distancia bottleneck entre B y* B' *por:*

$$d_B(A,A') = inf_{f:A\rightleftharpoons A'} sup_{I\in domf} d(\chi_I, \chi_{f(I)}). \tag{83}$$

En el lado derecho de 83 se tiene la distancia intercalado. Analizando algunas proposiciones que demuestra Bubenick en [82] se deriva que esta nueva definición de la distancia bottleneck es equivalente a la que se muestra en [78]. Con el análisis anterior realizado estamos en condiciones de definir la categorización para el espacio métrico de diagramas de persistencia.

***Teorema* 14** *Sea B el conjunto de códigos de barra finito,* d_B *la distancia bottleneck y d la distancia intercalación. El mapeo* χ *definido por* $\chi(\{I_k\}_{k=1}^n) = \oplus_{k=1}^n \chi_{I_k}$ *aporta una inmersión isométrica de espacios métricos:*

$$\chi : (B, d_B) \to (\boldsymbol{Vec}^{(\mathbb{R},\leq)}, d). \tag{84}$$

En este capítulo fueron mostrados algunos pasos para estudiar la persistencia considerando diagramas indexados por $(\mathbb{R},\leq)$. Si embargo existen versiones de la persistencia en el que el objeto de estudio se puede analizar como diagramas con una categoría de indexación más general. Por ejemplo ser capaces de considerar los diagramas indexados por $(\mathbb{R}^n,\leq)$ para persistencia multidimensional y la categoría $\cdot \to \cdot \leftarrow \cdot \to \cdots \leftarrow \cdot$ para persistencia zig-zag de la cual se analizan algunos conceptos para su comprensión en capítulos posteriores.

12. Métodos de submuestreo para homología persistente

La complejidad con respecto al tiempo y el espacio de los algoritmos para homología persistente es uno de los principales obstáculos en aplicar técnicas de TDA sobre problemas de alta dimensión. Para solucionar el problema de los costos computacionales Chazal en [83] propone la siguiente estrategia: dada una nube de puntos grande, escoger varias submuestras, calcular el landscape para cada submuestra y luego combinar la información. En efecto, contrariamente a los diagramas de persistencia, la persistencia landscape puede ser promediado de una manera directa.

Una medida de probabilidad μ es definida en un espacio métrico $(\mathbb{M},\rho)$ y el soporte de μ es un conjunto compacto $\mathbb{X}_\mu$. En las explicaciones que damos a continuación se asume que el diámetro de $\mathbb{M}$ es finito y es acotado por $\frac{T}{2}$ donde T es la misma constante que definimos en el capítulo anterior.

12.1. Enfoque: Muestras múltiples

Sea m un número entero positivo, y $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}_\mu$ muestra de m puntos de la medida μ. La correspondiente persistencia landscape es λ_X y se denota por Ψ_μ^m la medida inducida por $\mu^{\otimes m}$ en el espacio de las persistencias landscapes. La esperanza puntual de la persistencia landscape bajo dicha medida se define: $\mathbb{E}_{\Phi_\mu^m}[\lambda_X(t)]$, $t \in [0,T]$. El promedio landscape tiene una contraparte natural empírica que puede ser utilizado como su estimador insesgado.

Sea $(S_1^m, \cdots, S_n^m)$ n muestras independientes de tamaño m para μ y se define el promedio landscape como:

$$\overline{\lambda_n^m}(t) = \frac{1}{n}\sum_{i=1}^{n} \lambda_{S_i^m}(t), \ \ \forall\, t \in [0,T] \tag{85}$$

y se propone el uso de $\overline{\lambda_n^m}$ para estimar $\lambda_{\mathbb{X}_\mu}$.

Además del promedio, también considerar el uso de la muestra más cercana a $\mathbb{X}_\mu$ en distancia Hausdorff. El *método de muestreo* más cercano consiste en seleccionar una muestra de m puntos de $\mathbb{X}$ lo más cerca posible a $\mathbb{X}_\mu$ y luego utilizar la muestra para construir un landscape que aproxima $\lambda_{\mathbb{X}_\mu}$; donde La muestra más cercana es:

$$\widehat{C}_n^m = arg_{S \in \{S_1^m, \cdots, S_n^m\}} min\, d_H(S, \mathbb{X}_\mu) \tag{86}$$

y la correspondiente función landscape es $\widehat{\lambda}_n^m = \lambda_{\widehat{C}_n^m}$. El método requiere que el soporte de μ ser una cantidad conocida.

***Notas* 1** *En [83] se demuestra que lo descrito anteriormente es válido para el caso en que μ es una medida discreta con soporte $\mathbb{X}_N = \{x_1, \cdots, x_N\} \subset \mathbb{R}^D$. Este marco puede ser muy común en la práctica cuando una medida continua desconocida se aproxima por una medida uniforme discreta.*

Chazal [83] demuestra que el landscape promedio es una cantidad interesante dado que contiene información topológica acerca de la medida en cuestión μ a partir del cual los datos son generados. En particular se comparan landscapes promedios correspondientes a dos medidas que están cerca una de la otra en la métrica Wasserstein. Llegando a la conclusión que el comportamiento promedio de los landscapes de conjuntos de m puntos muestreados de acuerdo a una medida es estable con respecto a la distancia Wasserstein.

***Teorema* 15** *Sea $X \sim \mu^{\otimes m}$ y $Y \sim \nu^{\otimes m}$ donde μ y ν son dos medidas de probabilidad en $\mathbb{M}$ con soporte compacto. Sea $p \geq 1$ tenemos:*

$$\left\| \mathbb{E}_{\Phi_\mu^m}[\lambda_X] - \mathbb{E}_{\Phi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2m^{\frac{1}{p}} W_{\rho,p}(\mu,\nu). \tag{87}$$

***Notas* 2** ▪ *Para medidas que no están definidas en el mismo espacio métrico, la inecuación del teorema anterior se puede extender a la métrica Gromov-Wasserstein:*

$$\left\| \mathbb{E}_{\Phi_\mu^m}[\lambda_X] - \mathbb{E}_{\Phi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2m^{\frac{1}{p}} GW_{\rho,p}(\mu,\nu). \tag{88}$$

Los resultados del teorema anterior son útiles por dos razones. En primer lugar nos dice que para un m fijo, la esperanza (comportamiento topológico) de un conjunto de m puntos desarrolla alguna información estable acerca de la medida subyacente a partir de la cual se generaron los datos. En segundo lugar proporciona un límite inferior para la distancia Wasserstein entre dos medidas, basado en la señales topológicas de muestras de m puntos.

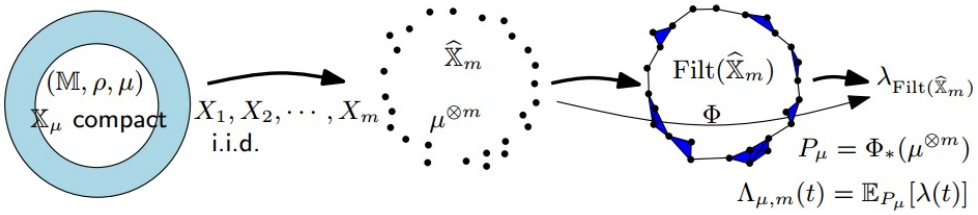

Fig. 31. Muestras múltiples.

12.2. Experimento

Chazal y colaboradores [83] aplican el método al problema de las actividades humanas distintivas, se utiliza para clasificar 19 actividades realizadas por 8 personas que llevan unidades de sensores en el pecho, los brazos y las piernas. Para cada actividad existen $7,500$ mediciones consecutivas que se tratan como una nube de puntos 3D en el espacio Euclidiano. Para $n = 80$, se tienen submuestras de tamaño $m = 200$ puntos de una nube de puntos para cada actividad, se construye el *landscape* de la submuestras más cercanas, el promedio *landscape* (dim 1) y la matriz de disimilitud basada en la distancia l_∞ de los landscapes promedio ver fig: 32 .

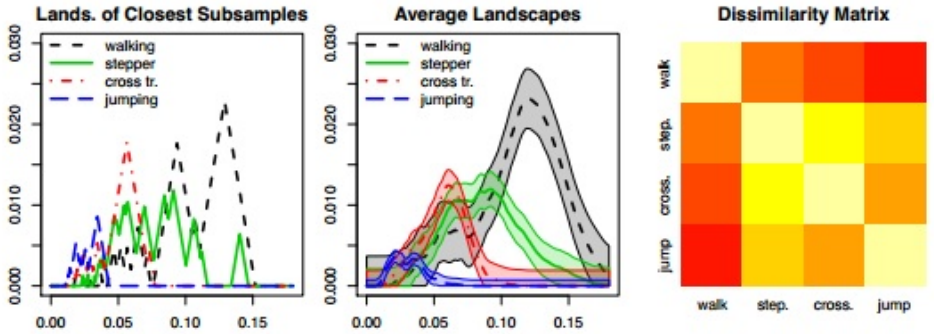

Fig. 32. Construcción de landscapes de las muestras más cercanas, los landscapes promedio con una banda de confianza de 95 % y la matriz de disimilitud de los pares de distancia l_∞ entre los landscapes promedio.

Como conclusión Chazal presenta un método para aproximar la homología persistente de un conjunto utilizando submuestras y proporciona resultados de estabilidad para los nuevos descriptores topológicos y límites sobre el riesgo de los estimadores que se proponen.

A continuación se describe una nueva metodología para estudiar la persistencia de las características topológicas a través de una familia de espacios denominada *persistencia zigzag* . Se fundamenta sobre la base de los resultados clásicos, generalizando la teoría de gran éxito conocida como homología persistente y se dirige a situaciones no cubiertas por la teoría.

13. Persistencia Zigzag

En secciones previas se han analizado resultados relacionados con la homología persistente y métodos eficientes para obtener resúmenes topológicos. La homología persistente ofrece garantías para inferir la

homología de un espacio topológico desconocido, pero en ocasiones la teoría de la persistencia es limitada. Un espacio topológico de un nube de puntos utilizando sólo cálculos de distancias conduce a construcciones de complejos Cech que son extremadamente grandes. Se plantea que en ocasiones los grandes complejos considerados pueden no ser suficientes para obtener un diagrama de persistencia válido. Por otro lado se considera rígida en el sentido en que los mapas simpliciales de izquierda a derecha en una filtración restringe el área de aplicación a la aproximación multiescala de los espacios topológicos presentados. Esto motivo a la introducción de la teoría de *homología persistente zigzag* que generaliza la teoría de persistencia y ofrece una nueva metodología para estudiar persistencia de características topológicas a través de una familia de espacios o conjuntos de datos de nube de puntos; para mayores detalles se pueden revisar los textos: [84,85,86].

La estructura en persistencia zigzag se activa cada vez que se construye un diagrama de zigzag en espacios topológicos o espacios vectoriales: una secuencia de espacios $X_1, \cdots, X_n$ donde cada par adyacente es conectado por un mapa $X_i \to X_{i+1}$ ó $X_i \leftarrow X_{i+1}$. La novedad de este enfoque es que la dirección de cada mapa de enlace es arbitraria, en contraste con la teoría habitual de persistencia donde todos los mapas apuntan en la misma dirección. Si se trabaja con una secuencia de complejos simpliciales, los mapas de inclusión como se conoce inducen mapas lineales entre los espacios homológicos asociado:

$$H_d(X_1,k) \leftrightarrow H_d(X_2,k) \leftrightarrow \cdots \leftrightarrow H_d(X_n,k). \tag{89}$$

En otras palabras todas las flechas corresponden a adición o supresión de símplices. El algoritmo es esencialmente un extensión del algoritmo de homología persistente presentado anteriormente. Una diferencia notable con el algoritmo estándar de persistencia es que no introduce o extrae nuevos símplices en el final del zigzag sino más bien en el medio.

En [84] se desarrolla una teoría matemática de la persistencia para diagramas zigzag, se describen escenarios en topología aplicada donde es natural considerar diagramas zigzag y se desarrollan algoritmos para calcular persistencia zigzag. Además se introduce el Principio del Diamante, herramienta de cálculo análogo a el teorema de Mayer-Vietoris en la topología algebraica clásica.

Definición* 22 *Persistencia Zigzag
Sea $\mathbb{V}$ *un módulo zigzag de tipo arbitrario:* $(V_1 \xleftrightarrow{p_1} V_2 \xleftrightarrow{p_2} \cdots \xleftrightarrow{p_{n-1}} V_n)$

$$Pers(\mathbb{V}) = \left\{ [b_j, d_j] \subseteq \{1, \cdots, n\} \,|\, j = 1, \cdots, N \right\}. \tag{90}$$

Un importante resultado de la teoría de persistencia zigzag, es que un módulo zigzag tiene una descomposición en intervalos:

$$\mathbb{V} \cong \mathbb{I}(b_1,d_1) \otimes \mathbb{I}(b_2,d_2) \otimes \cdots \otimes \mathbb{I}(b_m,d_m). \tag{91}$$

Gráficamente la $Pers(\mathbb{V})$ puede representarse en los resúmenes topológicos que estudiamos anteriormente. Ejemplo fig.33

Con los módulos de persistencia, hay varias maneras equivalentes para reconocer la existencia de una característica. Se muestra en la siguiente proposición.

***Proposición* 2** *Sea* $\mathbb{V}$ *un módulo de persistencia de longitud n, y* $1 \leq p \leq q \leq n$. *Son equivalentes:*

1. *El mapa* $V_p \to V_q$ *es no vacío*

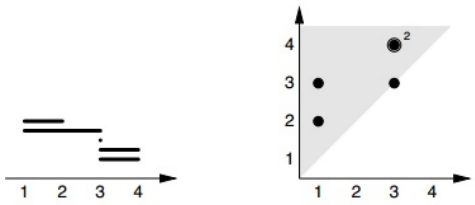

Fig. 33. Código de barras y diagramas de persistencia: representación de la persistencia $\{[1,2],[1,3],[3,3],[3,4]\}$ de un módulo zigzag de longitud 4.

2. $\exists\, x_i \neq 0 \in V_i$ *para* $p \leq i \leq q$ *tal que* $x_{i+1} = f_i(x_i)$ *para* $p \leq i < q$
3. *Existe un submódulo de* $\mathbb{V}[p,q]$ *isomorfo a* $\mathbb{I}[p,q]$.
4. *Existe un sumando de* $\mathbb{V}[p,q]$ *isomorfo a* $\mathbb{I}[p,q]$ *(característica sobre* $[p,q]$*)*

Una estrategia para entender y construir descomposiciones de un τ-módulo $\mathbb{V}$ por un proceso iterativo, moviéndose de izquierda a derecha y retener la información necesaria en cada etapa. La mayor parte de esta filtración se codifica como una filtración en el extremo derecho de V_n.

La filtración derecha $R(\mathbb{V})$ es calculada de manera incremental y resulta en una filtración en V_n, con un tiempo de nacimiento asociado a cada espacio de cociente. Una filtración en V_i se denota:

$$\mathcal{R}_i = (R_i^0, R_i^1, \cdots, R_i^i), \tag{92}$$

donde $R_i^0 \leq R_i^1 \leq \cdots \leq R_i^i$ y $R_i^i = V_i$. El cociente $R_i^1/R_i^0, R_i^2/R_i^1, \cdots, R_i^i/R_i^{i-1}$ cada una están asociadas con un tiempo de nacimiento b_i^j $(j = 0, \cdots, i)$, las cuales se guardan en el vector:

$$\mathbf{b}_i = (b_i^1, b_i^2, \cdots, b_i^1). \tag{93}$$

Se puede escribir cómo el cociente:

$$\mathcal{R}_i' = (R_i^1/R_i^0, R_i^2/R_i^1, \cdots, R_i^i/R_i^{i-1}). \tag{94}$$

El cálculo de la filtración derecha depende de la dirección del mapa. Si tenemos $\mathcal{R}_i$ y $\mathbf{b}_i$, entonces:

- Si $V_i \xrightarrow{f_i} V_{i+1}$, entonces

$$\mathcal{R}_{i+1} = (f_i(R_i^0), f_i(R_i^1), \cdots, f_i(R_i^i), V_{i+1}) \tag{95}$$

$$\mathbf{b}_{i+1} = (b_i^1, b_i^2, \cdots, b_i^i, i+1).$$

- Si $V_i \xleftarrow{g_i} V_{i+1}$, entonces

$$\mathcal{R}_{i+1} = (0, g_i^{-1}(R_i^0), g_i^{-1}(R_i^0), \cdots, g_i^{-1}(R_i^i)) \tag{96}$$

$$\mathbf{b}_{i+1} = (i+1, b_i^1, b_i^2, \cdots, b_i^i).$$

Dado que se asume que los complejos simpliciales consecutivos difieren al menos de un símplice, el cambio en dimensión entre V_i y V_{i+1} es como máximo 1. Similarmente la dimensión del espacio cociente puede ser 0 ó 1. La dimensión de V_i es el rango del grupo de homología para K_i (número de Betti, $\beta(K_i)$):

$$dim(V_i) = rank(H(K(i))) = \beta(K_i) \leq i. \tag{97}$$

Por ejemplo la dimensión de los espacios cociente puede ser una secuencia de 0 y 1.

$$dim(R_i^1/R_i^0, R_i^2/R_i^1, \cdots, R_i^i/R_i^{i-1}) = (0,0,1,1,0,\cdots,1,0). \tag{98}$$

Se debe tener en cuenta que la elección de una clase de homología de cada uno de los espacios cociente no nulos resulta en una base para V_i.

***Ejemplo* 6** *Estas son las filtraciones derecha asociadas para* 4 *casos con longitud* 3*:*

$$R(V_1 \xrightarrow{f_1} V_2 \xrightarrow{f_2} V_3) = (0, f_2f_1(V_1), f_2(V_2), V_3)$$
$$R(V_1 \xrightarrow{f_1} V_2 \xleftarrow{g_2} V_3) = (0, g_2^{-1}(0), g_2^{-1}f_1(V_1), V_3)$$
$$R(V_1 \xleftarrow{g_1} V_2 \xrightarrow{f_2} V_3) = (0, f_2g_1^{-1}(0), f_2(V_2), V_3)$$
$$R(V_1 \xleftarrow{g_1} V_2 \xleftarrow{g_2} V_3) = (0, g_2^{-1}(0), g_2^{-1}g_1^{-1}(0), V_3)$$

Ver fig34 para una representación esquemática:

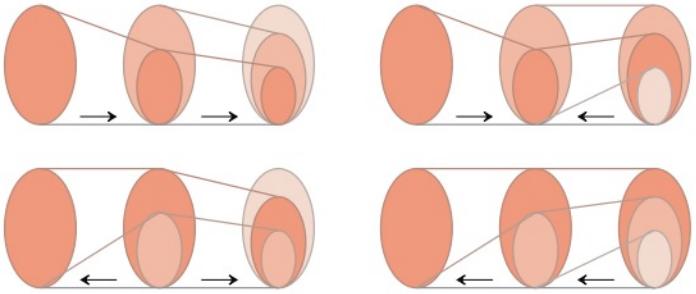

Fig. 34. Representación para los 4 casos de longitud 3: ff, fg, gf, gg.

El principio del diamante [87]
Considere el siguiente diagrama de espacio vectoriales y aplicaciones lineales entre ellos: Sea V^+ y V^-

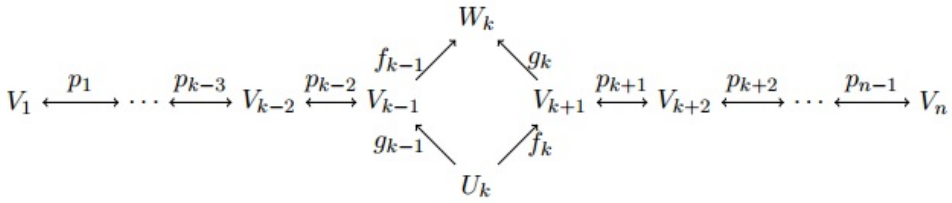

Fig. 35. Principio del Diamante.

los módulos zigzag superior e inferior:

$$V^+ = V_1 \xleftrightarrow{p_1} \cdots \xleftrightarrow{p_{k-2}} V_{k-1} \xrightarrow{f_{k-1}} W_k \xleftarrow{g_k} V_{k+1} \xleftrightarrow{p_{k+1}} \cdots \xleftrightarrow{p_{n-1}} V_n. \tag{99}$$

$$V^- = V_1 \xleftrightarrow{p_1} \cdots \xleftrightarrow{p_{k-2}} V_{k-1} \xrightarrow{g_{k-1}} U_k \xleftarrow{f_k} V_{k+1} \xleftrightarrow{p_{k+1}} \cdots \xleftrightarrow{p_{n-1}} V_n. \tag{100}$$

Donde el Principio del Diamante aporta la siguiente relación entre la persistencia zigzag de V^+ y V^-

***Teorema* 16** *Dado V^+ y V^- como anteriormente, suponemos que el diamante medio es exacto. Existe una biyección parcial entre Pers(V^+) y Pers(V^-) con los intervalos matcheados de acuerdo a la siguiente regla:*

- *Intervalos de tipo $[k,k]$ no se matchean,*
- *Intervalos del tipo $[b,k]$ se matchean con los de tipo $[b,k-1]$ y viceversa para $b \leq k-1$,*
- *Intervalos del tipo $[k,d]$ se matchean con los de tipo $[k+1,d]$ y viceversa para $d \geq k+1$,*
- *Intervalos del tipo $[b,d]$ se matchean con los de tipo $[b,d]$ en todos los otros casos.*

Fig. 36. Visualización del principio del diamante.

Una demostración del Principio del Diamante realizado por Carlsson y Silva la encontramos en [84].

Aplicaciones

En [85] Carlsson explora el uso de la homología zigzag para estudiar la información topológica en conjuntos de datos de nube de puntos. Podemos ver

Bootstrapping topológico

Dado un conjunto de datos X interesa comprender la homología de todo el conjunto de datos de las muestras y como las muestras se relacionan unas con otras. Trabajan con un enfoque estadístico mediante el método de bootstrap, donde obtienen información sobre un conjunto de datos mediante la realización de muestreos:

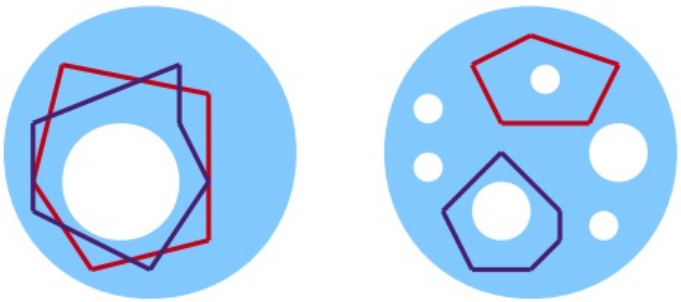

Fig. 37. Bootsrapping Topológico.

Donde interesa saber si las clases de homología de las muestras están midiendo las mismas características homológicas (como a la izquierda) o diferentes características (como a la derecha). Para evaluar la compatibilidad de dos muestras X_i y X_j consideran la unión $X_i \cup X_j$. Donde trabajando con los complejos Vietoris-Rips se obtiene el complejo simplicial filtrado. Los complejos tienen la propiedad que:

$$V_\varepsilon(X_i) \subset V_\varepsilon(X_i \cup X_j) \supset V_\varepsilon(X_j). \tag{101}$$

Al aplicar el funtor $H_p(-)$ se obtiene una descomposición en intervalos del diagrama zigzag que brinda información relevante acerca de la compatibilidad de la características homológicas del complejo.

Umbralización

Suponer que se tiene un conjunto de datos X y una función de filtración parametrizada $f(-,\theta) : X \to \mathbb{R}$. Un ejemplo puede ser un estimador de densidad el cual está parametrizado por algún tipo de parámetro de anchura o de varianza. El problema de interés es el estudio homológico de como se comporta la función de filtración en el conjunto de datos para diferentes valores de los parámetros.

$$X_f[\theta, T] = \{x \in X\}; \tag{102}$$

x es uno de los puntos $T\,\%$ superior clasificados por la función de filtración.

Se necesita conocer como las muestras se relacionan entre sí a medida que cambiamos el parámetro, por lo que se consideran las secuencias de muestras para una secuencia de parámetros $\{\theta_i\}$:

$$\cdots \leftarrow X_f[\theta_i, T] \to X_f[\theta_i, T] \cup X_f[\theta_{i+1}, T] \leftarrow X_f[\theta_{i+1}, T] \to \cdots \tag{103}$$

Se puede aplicar una inclusión preservando la construcción del complejo filtrado y calculando la homología zigzag. Por ejemplo es posible obtener un código de barras para:

$$\cdots \leftarrow H_p(V(X_f[\theta_i, T])) \to H_p(V(X_f[\theta_i, T] \cup X_f[\theta_{i+1}, T])) \leftarrow H_p(V(X_f[\theta_{i+1}, T])) \to \cdots \tag{104}$$

La existencia de intervalos de longitud positiva sugiere la preservación de características homológicas a través de varios valores del parámetro θ. Del mismo modo se considera la secuencia de intersecciones análoga a las anteriores.

$$\cdots \to H_p(V(X_f[\theta_i, T])) \leftarrow H_p(V(X_f[\theta_i, T] \cup X_f[\theta_{i+1}, T])) \to H_p(V(X_f[\theta_{i+1}, T])) \leftarrow \cdots \tag{105}$$

El cálculo de la homología de las secuencias anteriores revela información importante sobre cómo el parámetro afecta a las propiedades homológicas del conjunto de datos.

Comparación de complejos Witness

Tauzz y Carlson realizan una comparación de selecciones *landmark* para los complejos witness. Estos complejos witness permiten estimar las propiedades topológicas de una nube de puntos con sólo trabajar con un subconjunto de los puntos actuales.

Un subconjunto $L \subset X$ es designado como un conjunto *landmark*, y los puntos son los vértices en la construcción witness que denotamos por $W(X, L)$ (los puntos en el complemento influyen en la construcción del complejo pero no aparecen en él). Surgen cuestiones de cómo la homología persistente de la aproximación complejos witness se relaciona con la homología persistente de todo el conjunto de datos. Si bien no es posible responder a esta cuestión completamente sin calcular la homología para toda la nube de puntos, se discute un método para comparar diferentes submuestras. Se construye un diagrama zigzag de espacios topológicos:

$$\cdots \to W(X, L_i) \leftarrow W(X; L_i, L_{i+1}) \to W(X, L_{i+1}) \leftarrow W(X; L_{i+1}, L_{i+2}) \to \cdots \tag{106}$$

En [88] Adams y Carlsson abordan un problema de evasión para redes de sensores mobiles en la que los sensores no conocen su ubicación y en su lugar sólo miden la conectividad local de los datos. Utilizan

la persistencia zigzag para producir un criterio de poder discriminatorio equivalente que también permita el cálculo streaming, que es una característica importante para las redes de sensores en movimiento durante un largo período de tiempo.

A efectos del análisis de datos, uno puede hacer crecer los puntos que son lo suficientemente denso y luego observar cómo los cambios en la densidad afectan a los módulos de persistencia resultantes, pero no se puede proporcionar a los científicos algo tan simple como un diagrama de persistencia. Idear descriptores eficaces para homología persistente multidimensional es uno de los desafío centrales para TDA.

14. Vineyards

La mayoría de los trabajos del análisis topológico de datos se han centrado en el estudio de nubes de puntos estáticas. Este capítulo introduce una extensión de la teoría de homología persistente para sistemas variables en el tiempo. En particular dada una nube de puntos se puede construir su resumen topológico, por ejemplo los diagramas de persistencia. Dado que el diagrama varía continuamente a medida que la nube de puntos varía de forma continua, se estudia el espacio de los diagramas de persistencia variables en el tiempo, llamados vineyards introducidos por [89].

En [90] Munch demuestra que con una buena elección de métrica, los *vineyards* son estables para pequeñas perturbaciones en sus nubes de puntos asociadas. Se define una nueva media para un conjunto de diagramas de persistencia basado [56]. El aporte principal de la tesis de Munch es un aplicación de la homología persistente a las predicciones de comportamiento. Se crean vectores de comportamiento para el seguimiento de agentes en imágenes de satélite y el uso de la homología 0-dimensional para agrupar agentes por comportamientos. Además se construye una estructura de datos flexible para almacenar y consultar los datos con el fin de permitir el desarrollo de nuevos e interesante vectores de comportamiento para estudiar.

La idea básica es analizar como los grupos de homología de un espacio topológico cambian mientras el espacio cambia, y poder inferir alguna información acerca del espacio original.

14.1. Estabilidad de bottleneck para vineyards

Una nube de puntos dinámica $\mathbb{X}(t) = \{x_1(t), \cdots, x_N(t)\}$ es una nube de puntos que se mueve de forma continua durante una cantidad de tiempo finito. Por simplicidad se asume que el tiempo varía entre 0 y 1. Por lo tanto una nube de puntos dinámica es un mapa:

$$\begin{aligned} [0,1] \to \quad & (\mathbb{R}^d)^N. \\ t \to \{x_1(t), \cdots, x_N(t)\}&. \end{aligned} \tag{107}$$

Es natural considerar los diagramas de persistencia variables que surgen de estas nubes de puntos dinámicas. Entonces dada una nube de puntos dinámica $\mathbb{X}$ existe un diagrama de persistencia $D(\mathbb{X}(t))$ para cada tiempo t. Esta familia de diagramas se denomina vineyards:

$$V(\mathbb{X}) = \{D(\mathbb{X}(t)) |\, t \in [0,1]\}. \tag{108}$$

$$[0,1] \rightarrow \quad D_\infty.$$
$$t \rightarrow D(\mathbb{X}(t));$$

donde D_∞ representa el espacio de los diagramas de persistencia.

***Corolario* 5** *Si una nube de punto dinámica $\mathbb{X}(t)$ es continua con respecto a la distancia de Hausdorff, el correspondiente vineyard $V(\mathbb{X})$ es continuo con respecto a la distancia bottleneck.*

Con el fin de demostrar la estabilidad para los vineyards se necesitan métricas para el espacio de nubes de punto dinámica y para el espacio de los vineyards. Dado que existe una noción de distancia entre los diagramas para cada tiempo t así como entre las nubes de puntos estáticas para cada tiempo, dichas métricas pueden ser integradas sobre el tiempo y así obtener nuevas métricas homólogas a las anteriores pero con la condición de variables en el tiempo.

***Definición* 23** *Sean los vineyards $V(\mathbb{X})$ y $V(\mathbb{Y})$ que se obtienen de las nubes de puntos dinámicas $\mathbb{X}$ y $\mathbb{Y}$, la métrica bottleneck integrada esta dada por:*

$$I[W_\infty](V(\mathbb{X}),V(\mathbb{Y})) := \int_0^1 W_\infty(D(\mathbb{X}(t)),D(\mathbb{Y}(t)))dt, \tag{109}$$

y la métrica de Hausdorff integrada:

$$I[H](\mathbb{X},\mathbb{Y}) := \int_0^1 H(\mathbb{X}(t),\mathbb{Y}(t))dt, \tag{110}$$

pero para demostrar que en efecto las funciones definidas son las métricas, primero se debe demostrar que la distancia de Hausdorff y por lo tanto la distancia Bottleneck son continuas cuando las nubes de puntos dinámicas son continuas, además debemos demostrar los axiomas de la definición de métrica. Estas definiciones y pruebas extensas podemos encontrarlas en [90].

***Teorema* 17** *Teorema de estabilidad para vineyards Dado dos nubes de puntos dinámicas finitas $\mathbb{X}$ y $\mathbb{Y}$,*

$$I[W_\infty](V(\mathbb{X}),V(\mathbb{Y})) \leq I[H](\mathbb{X},\mathbb{Y}). \tag{111}$$

Dado que las traslaciones y rotaciones de las nubes de puntos no cambian los diagramas de persistencia, se obtiene el mismo teorema de estabilidad si sustituimos la métrica de las nubes de puntos dinámicas por el mínimo sobre todas las rotaciones y traslaciones de las nubes de puntos.

14.2. Probabilidad en vineyards

Con un algoritmo para calcular la media de un conjunto de diagramas descrito en el capítulo, se quiere calcular la media de un conjunto de diagramas de persistencia variables en el tiempo.

***Definición* 24** *El espacio de vineyards abstracto se define:*

$$V = \{v : [0,1] \rightarrow D_p |\ v\ \ continua\}, \tag{112}$$

el espacio de los mapas del intervalo unidad a D_p, donde v es continuo con respecto a W_p

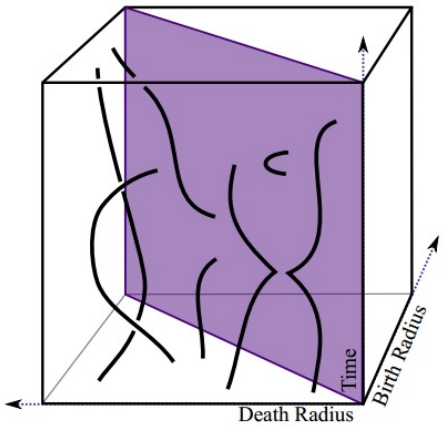

Fig. 38. Ejemplo de un vineyard. Para cada tiempo, dado en el eje z, existe un diagrama de persistencia. Dado que los *vineyards* surgen de nubes de puntos continuas son continuas, cada punto en el diagrama traza un camino llamado *vine*. Estos *vines* pueden tener puntos finales en los tiempos de origen o destino, o en el plano que se proyecta a la diagonal.

14.3. Media de Fréchet para vineyards

Considerar el ejemplo de la Figura 39 ; donde tenemos dos diagramas de persistencia superpuestos, (puntos 1, 2 de un diagrama, y puntos a y b del otro). Dado que los 4 puntos forman exactamente un cuadrado, los pares para la distancia Wasserstein pueden ser $\{(a,1),(b,2)\}$ ó $\{(a,2),(b,1)\}$. Entonces se tienen dos diagramas que dan un mínimo de la función Fréchet: el diagrama con u y v, ó el diagrama con x y y.

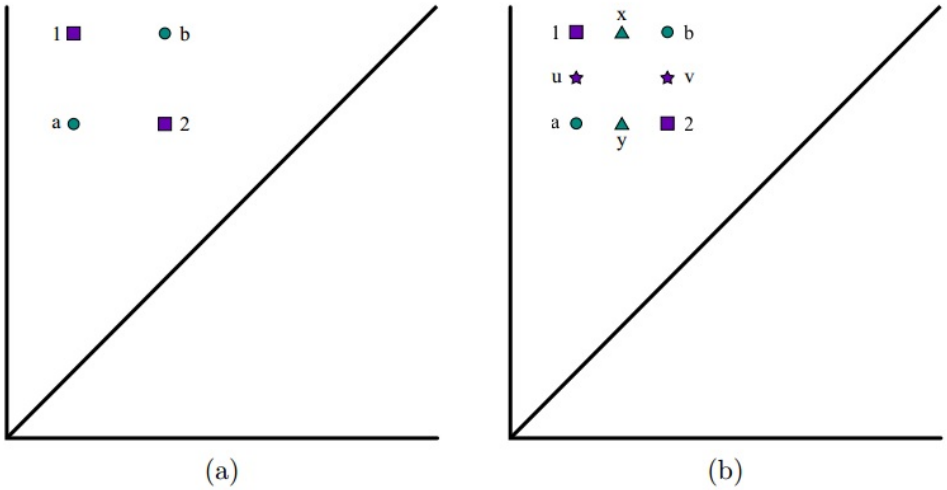

Fig. 39. Dado que la correspondencia Wasserstein no es única, la media Fréchet tampoco. Existen dos medias posibles dadas en la fig:b.

Si dos *vineyards* pasan a través de esta configuración, la media de los *vineyards* construida por efectuar la media en cada tiempo no será continua. Considerar por ejemplo dos *vineyards* de dos puntos cada uno iniciado en Figura:40(a) y se mueve a lo largo de las líneas de puntos en la configuración de la Figura: 40(b). En la curva de la línea de puntos, los puntos están en las esquinas de un cuadrado, así como en el ejemplo de la Figura:39, existen dos posibles elecciones para la media

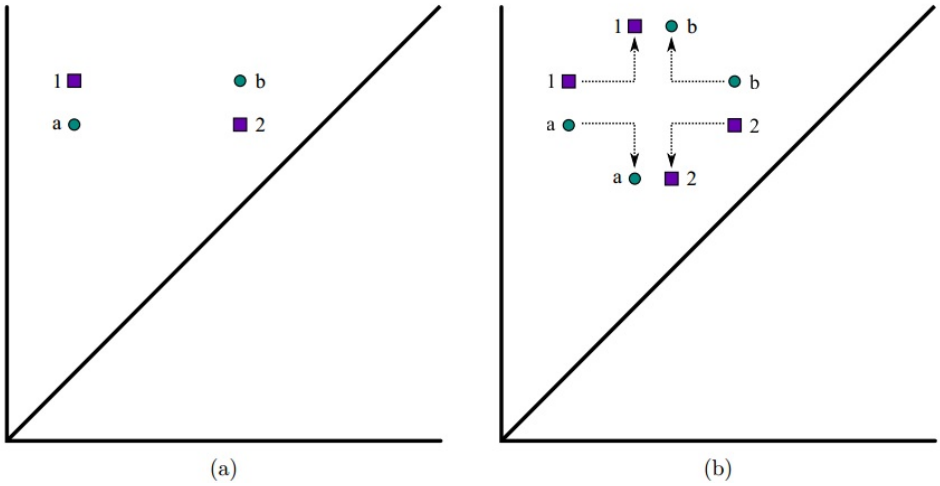

Fig. 40. Dos *vineyards* cuya media puntual no es continua. La media es continua hasta que los puntos llegan a la línea, donde forman un cuadrado y la media salta de forma discontinua.

En [67] se demuestra que la media es en efecto una distribución en el espacio de los diagramas que es una característica de la distribución de los diagramas de la que surgió. Esta nueva definición proporciona una herramienta estadística útil para el análisis de datos topológicos. Varias preguntas siguen siendo una interrogante que pueden ser analizadas en trabajos futuros. Aprovechar esta nueva definición en el campo de las estadísticas tradicionales; de manera particular probar la ley de los grandes números, el teorema central del límite, entre otras. Las investigaciones hasta el momento centran su atención en la media y varianza como medidas cuantitativas, derivadas de un conjunto de datos; trabajar con otras medidas como la mediana y las medidas de posición pueden ofrecer otras informaciones no descubiertas.

15. Conclusiones

La topología computacional juega un papel fundamental para agrupar las investigaciones sobre topología algebraica, geometría computacional, análisis de datos y otras áreas científicas relacionadas. El trabajo introduce las principales técnicas que han experimentado un crecimiento de manera especial en el área de análisis de datos. Estas técnicas como se observa a lo largo del trabajo se han desarrollado desde un enfoque estadístico, donde su utilidad y la capacidad que tienen para ser aplicadas en problemas prácticos hace inminente un camino hacia nuevas oportunidades de investigación. Se reportan aplicaciones exitosas en el área de reconocimiento de patrones y en otras que se encuentran referenciadas en la sección 2.

Debido a lo novedoso de la teoría existen todavía cuestiones abiertas detectadas a lo largo de la investigación sobre el campo. En primer lugar el método de selección del modelo impide hacer frente a los complejos simpliciales heterogéneos, y no existe una forma explícita para este caso ya que la ecuación

que la define es mucho más complicada que en el caso homogéneo.

Por otra parte el tema de la homología persistente ha sido utilizada para estudiar la homología de los conjuntos de nivel de una función dada. Por ejemplo como se expone en 4.2 fundamentado sobre [48] donde se analiza que se podría estimar la homología de $\mathbb{X}$ a partir de estimadores de densidad. Pero el análisis de riesgo completo del estimador de densidad del Núcleo cuando es utilizado para homología persistente del soporte de la distribución no se ha propuesto hasta el momento. Además se puede pensar en usar otros tipos de núcleos, como por ejemplo los núcleos de Laplace y Triángulo son opciones naturales. Para ambos, los resultados coinciden con los del Núcleo de Gauss; la distancia bajo el Núcleo de Laplace es también una métrica, pero no se conoce que es para el Núcleo del Triángulo. Sin embargo el segundo sería más interesante dado que tiene soporte acotado, y puede ser más fácil computacionalmente.

Los razonamientos estadísticos parecen ser ineludibles en TDA, ya que las aplicaciones involucran algún tipo de muestreo. De entrada TDA es un problema de inferencia bajo incertidumbre: si se pretende descubrir la estructura topológica de un objeto, ¡es porque ésta no se conoce!. Existe gran dificultad para formalizar modelos, parámetros, etc. (conceptos convencionales de inferencia que no necesariamente se aplican de inmediato). Uno de los problemas de origen es la complejidad de los espacios involucrados y para ellos aplicamos la teoría de probabilidad.

Un problema en la inferencia topológica está asociado a los parámetros, por ejemplo elegir el ancho de banda es una pregunta compleja. Es conocido que estas elecciones dependen de la geometría del soporte, pero por supuesto en la práctica estas cantidades son desconocidas. Existen algunas ideas relacionadas sobre este tema como el seguimiento de la evolución de la persistencia de las características homológicas y la variabilidad del parámetro de ajuste. Es posible aplicar las ideas expuestas anteriormente sobre los métodos de sub-muestreo 12 con el objetivo de seleccionar m de una manera eficiente. Realizar una estimación de la distancia entre μ (la media landscape para submuestras) y λ (landscape para el conjunto de datos original) y analizar que información es posible extraer de los resultados no se ha investigado. Dentro del área de TDA se encuentran trabajos que abordan el tema de prueba de hipótesis, de gran importancia debido a que diferentes tipos de problemas de toma de decisiones, pruebas o experimentos pueden formularse como prueba de hipótesis; pero no se han encontrado artículos que aborden la realización de pruebas de hipótesis alternativa cuando las observaciones son diagramas de persistencia y también de forma general desarrollar pruebas de hipótesis para la comparación de nubes de puntos.

En el reporte son presentadas ideas relacionadas con la media y la varianza de un conjunto de diagramas, algunas direcciones futuras de trabajo pueden estar centradas en extender los resultados a otros estadísticos como la mediana y esperanzas condicionales. Los intervalos de confianza expuestos brindan protección contra errores de tipo I i.e falsas detecciones. Es importante investigar el poder de los métodos para detectar características topológicas reales y del mismo modo poder cuantificar los límites min-max para homología persistente. Sería interesante construir intervalos de confianza para otros parámetros topológicos como para el grado total de persistencia.

Debido a diferentes necesidades que surgen de los problemas prácticos se han comenzado a desarrollar metodologías para generalizar la persistencia. En un primer caso: la metodología zigzag que revela información importante acerca de los conjuntos de datos no lineales. Pero saltan interrogantes: ¿en que medida puede la persistencia zigzag ser optimizada?; ¿puede superarse la brecha en el rendimiento entre

la persistencia estándar y la persistencia zigzag?. El campo de los *vineyards* no ha sido explotado por los investigadores, surgen interrogantes de si es posible aprovechar está nueva definición en el campo de las estadísticas tradicionales; en particular ¿se podrá probar la ley de los grandes números y el teorema central del límite.?

Actualmente se centra la atención sobre el tema de la persistencia multidimensional donde proponer un enfoque estadístico se hace bastante complejo. La no existencia de una invariante completa para este caso y el alto costo de la distancias entre módulos de persistencia multidimensional entre otros hace el estudio más complejo.

Trabajos futuros

- Definir los complejos simpliciales heterogéneos y analizar su aplicación en problemas prácticos.
- Combinar herramientas de TDA con *machine learning* para problemas práticos en la visión por computadora.
- Establecer un marco mediante la teoría de categorías donde se puedan considerar diferentes modos de persistencia.
- Extender las herramientas explicadas en el reporte para los métodos de submuestreos.
- Desarrollar pruebas de hipótesis para comparar nubes de puntos, y trabajar con pruebas de hipótesis alternativa.
- Analizar diferentes representaciones funcionales de los descriptores topológicos y establecer ventajas y desventajas para su posterior uso en problemas prácticos.
- Centrar la atención en la persistencia multidimensional que presenta problemas sin solución y parcialmente resueltos, por ejemplo: la existencia de una invariante completa para este caso.

Referencias bibliográficas

1. Edelsbrunner, H., Harer, J.: Computational topology: an introduction. American Mathematical Soc. (2010)
2. Singh, G., Mémoli, F., Carlsson, G.E.: Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: SPBG. (2007) 91–100
3. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. Discrete and Computational Geometry **28**(4) (2002) 511–533
4. Hiraoka, Y., Nakamura, T., Hirata, A., Escolar, E.G., Matsue, K., Nishiura, Y.: Hierarchical structures of amorphous solids characterized by persistent homology. Proceedings of the National Academy of Sciences **113**(26) (2016) 7035–7040
5. Chan, J.M., Carlsson, G., Rabadan, R.: Topology of viral evolution. Proceedings of the National Academy of Sciences **110**(46) (2013) 18566–18571
6. Zeppelzauer, M., Zieliński, B., Juda, M., Seidl, M.: Topological descriptors for 3d surface analysis. In: International Workshop on Computational Topology in Image Context, Springer (2016) 77–87
7. Dłotko, P., Wanner, T.: Topological microstructure analysis using persistence landscapes. Physica D: Nonlinear Phenomena (2016)
8. Kusano, G., Fukumizu, K., Hiraoka, Y.: Persistence weighted gaussian kernel for topological data analysis. arXiv preprint arXiv:1601.01741 (2016)
9. Venkataraman, V., Ramamurthy, K.N., Turaga, P.: Persistent homology of attractors for action recognition. arXiv preprint arXiv:1603.05310 (2016)
10. Nicolau, M., Levine, A.J., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences **108**(17) (2011) 7265–7270
11. Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P., Vaccarino, F.: Homological scaffolds of brain functional networks. Journal of The Royal Society Interface **11**(101) (2014) 20140873
12. Chung, M.K., Bubenik, P., Kim, P.T.: Persistence diagrams of cortical surface data. In: Information Processing in Medical Imaging, Springer (2009) 386–397

13. Kovacev-Nikolic, V., Bubenik, P., Nikolić, D., Heo, G.: Using persistent homology and dynamical distances to analyze protein binding. Statistical applications in genetics and molecular biology **15**(1) (2016) 19–38
14. Bubenik, P.: Statistical topological data analysis using persistence landscapes. (2014)
15. Bendich, P., Chin, S., Clarke, J., DeSena, J., Harer, J., Munch, E., Newman, A., Porter, D., Rouse, D., Strawn, N., et al.: Topological and statistical behavior classifiers for tracking applications. arXiv preprint arXiv:1406.0214 (2014)
16. Sato, M.: Can tda be a new risk measure? an application to finance of persistent homology. An Application to Finance of Persistent Homology (January 3, 2016) (2016)
17. Taylor, J.E., Worsley, K.J.: Detecting sparse signals in random fields, with an application to brain mapping. Journal of the American Statistical Association **102**(479) (2007) 913–928
18. Adler, R.J., Taylor, J.E.: Random fields and geometry. Springer Science & Business Media (2009)
19. Turner, K., Mukherjee, S., Boyer, D.M.: Persistent homology transform for modeling shapes and surfaces. Information and Inference (2014) iau011
20. Bendich, P., Edelsbrunner, H., Kerber, M.: Computing robustness and persistence for images. IEEE transactions on visualization and computer graphics **16**(6) (2010) 1251–1260
21. Carstens, C., Horadam, K.: Persistent homology of collaboration networks. Mathematical problems in engineering **2013** (2013)
22. Xia, K., Zhao, Z., Wei, G.W.: Multiresolution topological simplification. Journal of Computational Biology **22**(9) (2015) 887–891
23. Offroy, M., Duponchel, L.: Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. Analytica chimica acta **910** (2016) 1–11
24. Babichev, A., Dabaghian, Y.: Persistent memories in transient networks. arXiv preprint arXiv:1602.00681 (2016)
25. Basso, E., Arai, M., Dabaghian, Y.: Gamma synchronization of the hippocampal spatial map—topological model. arXiv preprint arXiv:1603.06248 (2016)
26. Curto, C.: What can topology tell us about the neural code? arXiv preprint arXiv:1605.01905 (2016)
27. Dotko, P., Hess, K., Levi, R., Nolte, M., Reimann, M., Scolamiero, M., Turner, K., Muller, E., Markram, H.: Topological analysis of the connectome of digital reconstructions of neural microcircuits. arXiv preprint arXiv:1601.01580 (2016)
28. Yoo, J., Kim, E.Y., Ahn, Y.M., Ye, J.C.: Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. Journal of neuroscience methods **267** (2016) 1–13
29. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete & Computational Geometry **33**(2) (2005) 249–274
30. Edelsbrunner, H., Mücke, E.P.: Three-dimensional alpha shapes. ACM Transactions on Graphics (TOG) **13**(1) (1994) 43–72
31. Giesen, J., John, M.: The flow complex: a data structure for geometric modeling. In: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics (2003) 285–294
32. Mark de Berg, M.: van krefeld, m. overmars, and o. schwarzkopf. Computational geometry: Algorithms and applications. Springer **2** (2000)
33. De Silva, V., Carlsson, G.: Topological estimation using witness complexes. Proc. Sympos. Point-Based Graphics (2004) 157–166
34. Boissonat, Chazal, Y.: Computational Topology Inference. (2015)
35. Bobrowski, O., Mukherjee, S.: The topology of probability distributions on manifolds. Probability Theory and Related Fields **161**(3-4) (2014) 651–686
36. Owada, T., Adler, R.J.: Limit theorems for point processes under geometric constraints (and topological crackle). arXiv preprint arXiv:1503.08416 (2015)
37. Vick, J.W.: Homology theory: an introduction to algebraic topology. Volume 145. Springer Science & Business Media (2012)
38. Hatcher, A.: Algebraic topology cambridge university press. Cambridge, UK (2002)
39. Borsuk, K.: On the imbedding of systems of compacta in simplicial complexes. Fundamenta Mathematicae **1**(35) (1948) 217–234
40. Penrose, M.: Random geometric graphs. Volume 5. Oxford University Press Oxford (2003)
41. Vietoris, L.: Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen. Mathematische Annalen **97**(1) (1927) 454–472
42. Chazal, F., De Silva, V., Oudot, S.: Persistence stability for geometric complexes. Geometriae Dedicata **173**(1) (2014) 193–214
43. Boissonat, Chazal, Y.: Computational Topology Inference. Volume 5. Buscar (2015)
44. De Silva, V., Ghrist, R.: Coverage in sensor networks via persistent homology. Algebraic & Geometric Topology **7**(1) (2007) 339–358
45. Ghrist, R.: Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society **45**(1) (2008) 61–75
46. Botnan, M.B.: Three approaches in computational geometry and topology: Persistent homology, discrete differential geometry and discrete morse theory. (2011)

47. Lazar, E.A., Mason, J.K., MacPherson, R.D., Srolovitz, D.J.: Statistical topology of three-dimensional poisson-voronoi cells and cell boundary networks. Physical Review E **88**(6) (2013) 063309
48. Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., Singh, A., et al.: Confidence sets for persistence diagrams. The Annals of Statistics **42**(6) (2014) 2301–2339
49. Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., Singh, A.: Statistical inference for persistent homology: Confidence sets for persistence diagrams. arxiv preprint. arXiv **1303** (2013)
50. Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., Ringach, D.L.: Topological analysis of population activity in visual cortex. Journal of vision **8**(8) (2008) 11–11
51. Kasson, P.M., Zomorodian, A., Park, S., Singhal, N., Guibas, L.J., Pande, V.S.: Persistent voids: a new structural metric for membrane fusion. Bioinformatics **23**(14) (2007) 1753–1759
52. Chazal, F., Cohen-Steiner, D., Guibas, L.J., Mémoli, F., Oudot, S.Y.: Gromov-hausdorff stable signatures for shapes using persistence. In: Computer Graphics Forum. Volume 28., Wiley Online Library (2009) 1393–1403
53. Chazal, F., Guibas, L.J., Oudot, S.Y., Skraba, P.: Persistence-based clustering in riemannian manifolds. Journal of the ACM (JACM) **60**(6) (2013) 41
54. De Silva, V., Ghrist, R.: Homological sensor networks. Notices of the American mathematical society **54**(1) (2007)
55. Bubenik, P., Kim, P.T., et al.: A statistical approach to persistent homology. Homology, Homotopy and Applications **9**(2) (2007) 337–362
56. Mileyko, Y., Mukherjee, S., Harer, J.: Probability measures on the space of persistence diagrams. Inverse Problems **27**(12) (2011) 124007
57. Lecci, F., Cisewski, J., Chazal, F., Rinaldo, A., Tibshirani, R., Wasserman, L.: Statistical inference for topological data analysis. (2014)
58. Chen, Y.C., Wang, D., Rinaldo, A., Wasserman, L.: Statistical analysis of persistence intensity functions. arXiv preprint arXiv:1510.02502 (2015)
59. Bobrowski, O., Adler, R.J.: Distance functions, critical points, and the topology of random cech complexes. Homology, Homotopy and Applications **16**(2) (2014) 311–344
60. Bendich, P., Galkovskyi, T., Harer, J.: Improving homology estimates with random walks. Inverse Problems **27**(12) (2011) 124002
61. Sheather, S.J., et al.: Density estimation. Statistical Science **19**(4) (2004) 588–597
62. Chazal, F., De Silva, V., Glisse, M., Oudot, S.: The structure and stability of persistence modules. arXiv preprint arXiv:1207.3674 (2012)
63. Turner, K.: Means and medians of sets of persistence diagrams. arXiv preprint arXiv:1307.8300 (2013)
64. Cohen-Steiner, D., Edelsbrunner, H., Harer, J., Mileyko, Y.: Lipschitz functions have l p-stable persistence. Foundations of computational mathematics **10**(2) (2010) 127–139
65. Turner, K., Mileyko, Y., Mukherjee, S., Harer, J.: Fréchet means for distributions of persistence diagrams. Discrete & Computational Geometry **52**(1) (2014) 44–70
66. Robinson, A., Turner, K.: Hypothesis testing for topological data analysis. arXiv preprint arXiv:1310.7467v2 (2016)
67. Munch, E., Turner, K., Bendich, P., Mukherjee, S., Mattingly, J., Harer, J., et al.: Probabilistic fréchet means for time varying persistence diagrams. Electronic Journal of Statistics **9**(1) (2015) 1173–1204
68. Chazal, F., Fasy, B.T., Lecci, F., Rinaldo, A., Singh, A., Wasserman, L.: On the bootstrap for persistence diagrams and landscapes. arXiv preprint arXiv:1311.0376 (2013)
69. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap, monographs on statistics and applied probability, vol. 57. New York and London: Chapman and Hall/CRC (1993)
70. Chazal, F., Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L.: Stochastic convergence of persistence landscapes and silhouettes. In: Proceedings of the thirtieth annual symposium on Computational geometry, ACM (2014) 474
71. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC press (1994)
72. Van der Vaart, A.W.: Asymptotic statistics. Volume 3. Cambridge university press (2000)
73. Kosorok, M.: Introduction to empirical processes and semiparametric inference. 2008
74. Blumberg, A.J., Gal, I., Mandell, M.A., Pancia, M.: Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. arXiv preprint arXiv:1206.4581 (2012)
75. Tsao, M., Zhou, J.: A nonparametric confidence interval for the trimmed mean. Journal of Nonparametric Statistics **14**(6) (2002) 665–673
76. Ares, V.M.: La prueba de significación de la «hipótesis cero» en las investigaciones por encuesta. Metodología de encuestas **1**(1) (1999) 47–68
77. Edgington, E., Onghena, P.: Randomization tests. CRC Press (2007)
78. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. Discrete & Computational Geometry **37**(1) (2007) 103–120
79. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society **46**(2) (2009) 255–308

80. Verri, A., Uras, C., Frosini, P., Ferri, M.: On the use of size functions for shape analysis. Biological cybernetics **70**(2) (1993) 99–107
81. Robins, V., Turner, K.: Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. arXiv preprint arXiv:1507.01454 (2015)
82. Bubenik, P., Scott, J.A.: Categorification of persistent homology. Discrete & Computational Geometry **51**(3) (2014) 600–627
83. Chazal, F., Fasy, B.T., Lecci, F., Michel, B., Rinaldo, A., Wasserman, L.: Subsampling methods for persistent homology. arXiv preprint arXiv:1406.1901 (2014)
84. Carlsson, G., De Silva, V.: Zigzag persistence. Foundations of computational mathematics **10**(4) (2010) 367–405
85. Tausz, A., Carlsson, G.: Applications of zigzag persistence to topological data analysis. arXiv preprint arXiv:1108.3545 (2011)
86. Maria, C.: Algorithms and data structures in computational topology. PhD thesis, Université Nice Sophia Antipolis (2014)
87. Kališnik, S.: Persistent Homology and Duality. PhD thesis, UNIVERSITY OF LJUBLJANA (2013)
88. Adams, H., Carlsson, G.: Evasion paths in mobile sensor networks. The International Journal of Robotics Research **34**(1) (2015) 90–104
89. Cohen-Steiner, D., Edelsbrunner, H., Morozov, D.: Vines and vineyards by updating persistence in linear time. In: Proceedings of the twenty-second annual symposium on Computational geometry, ACM (2006) 119–126
90. Munch, E.: Applications of persistent homology to time varying systems. PhD thesis, Duke University (2013)

Anexo 1: Conceptos topológicos

***Definición* 25** *(**Topología, Espacio Topológico**) Sea X un conjunto y sea $P(X)$ la colección de los subconjuntos de X. Se dice que $T \subset P(X)$ es una topología sobre X si cumple:*

- $\emptyset \in T$ *y* $X \in T$*;*
- $A_1 \in T$, $A_2 \in T \rightarrow A_1 \cup A_2 \in T$*;*
- $\forall_{\alpha \in I}$, $A_\alpha \in T \rightarrow \cup_{\alpha \in I} A_\alpha \in T$, *I conjunto.*

***Notas* 3** *La hipótesis $\emptyset \in T$ puede omitirse basándose en que, $\cup_{i \in \emptyset} A_i = \emptyset$ y $\cap_{i \in \emptyset} A_i = X$.*

Los elementos de T se llaman **abiertos**. Si X es un conjunto y T una topología sobre X, al par (X,T) se denomina **espacio topológico**. Si no hay riesgo de confusión, lo denotamos simplemente por X y sus elementos se denominan **puntos**.

***Definición* 26** *(**Espacio métrico**) Sea E un conjunto. Un aplicación $d : ExE \rightarrow R_+$ es una distancia o métrica en E si para todos $x,y,z \in E$ se cumple:*

1. $d(x,y) \leq 0$ *y* $d(x,y) = 0 \Leftrightarrow x = y$*;*
2. $d(x,y) = d(y,x)$
3. $d(x,y) \leq d(x,z) + d(z,y)$ *(desigualdad triangular)*

*El par (E,d) se llama **espacio métrico**.*

***Definición* 27** *Sean (E,d_E) y (F,d_F) espacios métricos y $f : E \rightarrow F$.*

- *f es **continua** en $x_0 \in E$ si y sólo si* $\forall \varepsilon > 0 \ \exists \ \delta > 0$, *tal que si* $d_E(x,x_0) < \delta$ *entonces* $d_F(f(x),f(x_0)) < \varepsilon$.
- *f es **continua** en E si lo es en todo $x \in E$*
- *f es un **homeomorfismo** si es invertible y tanto ella como su inversa son continuas.*
- *f es una **isometría** si para todos $x,y \in E$ es $d_E(x,y) = d_F(f(x),f(y))$*

***Definición* 28** *Sea (E,d) un espacio métrico*

- *Se dice que la sucesión $\{x_n\} \subset E$ es **fundamental o de Cauchy** si y sólo si para todo $\varepsilon > 0$ existe N, tal que $d(x_n,x_m) < \varepsilon \ \forall n,m \geq N$.*
- *El espacio métrico (E,d) es **completo** si en él toda sucesión de Cauchy es convergente.*

***Ejemplo* 7** *$(C[a,b],d_\infty)$ es completo, donde $C[a,b]$ el espacio de las funciones continuas en $[0,1]$*

***Definición* 29** *(**Espacio de Banach**) El espacio normado $(E,||\cdot||)$, donde $||\cdot|| : E \rightarrow \mathbb{R}$ es una norma en E, es de Banach si E es completo.*

La importancia de los espacios de Banach radica en su completitud, que es uno de los conceptos más frecuentemente explotados en Análisis Funcional. La razón de esto radica básicamente en el **Teorema de Baire** .

***Teorema* 18** *(**Teorema de Baire**) Sea E un espacio métrico completo y E_n una sucesión de abiertos densos en E, entonces $\cap E_n$ es densa en E.*

***Definición* 30** *(**Espacio de Hilbert**) Un espacio euclidiano $(E,\langle\cdot,\cdot\rangle)$ se llama **espacio de Hilbert** si es completo, de dimensión infinita y separable (E contiene un subconjunto numerable siempre denso).*

***Definición* 31** *(**Espacio Medible**) Sea X un conjunto:*

- *Un conjunto $T \subset P(X)$ se llama tribu o σ-álgebra si se cumple:*
  1. *$X \in T$;*
  2. *$A \in T \rightarrow A^c \in T$*
  3. *$(A_n) \subset T \rightarrow \bigcup_n A_n \in T$*
- *Los elementos de T se llaman **conjuntos medibles**.*
- *El par (X,T) se llama **espacio medible**.*

***Definición* 32** *Si (X,τ) es un espacio topológico, la σ-álgebra generada por la topología $\sigma(\tau)$ se denomina **σ-álgebra boreliana**.*

***Teorema* 19** *Sea $f : X \rightarrow Y$ y T una σ-álgebra sobre Y, entonces se cumple:*

- *$f^{-1}(T)$ es una σ-álgebra en X.*
- *Si $T = \sigma(S)$ entonces $f^{-1}(\sigma(S)) = \sigma(f^{-1}(S))$*

***Definición* 33** *(**Funciones Medibles**) Dada $f : X \rightarrow \mathbb{R}$ se denota:*

$$\{f \prec a\} = \{x \in X; f(x) \prec a\}, \tag{113}$$

donde $\prec$ puede ser $<, \leq, >, \geq, =, \neq$.

*Si (X,T) e (Y,U) son espacios medibles, se dice que $f : X \rightarrow Y$ es **medible** si $f^{-1}(U) \subset T$ es decir, si la preimagen de todo conjunto medible en Y es medible en X.*

***Definición* 34** *(**Funciones simples**) Sea (X,T) un espacio medible. Una función $s : X \rightarrow \mathbb{R}$(ó $\mathbb{C}$) se dice **simple** si es medible y solamente toma un número finito de valores diferentes.*

***Representación natural**: $s = \sum_{k=0}^{n} a_k 1_{A_k}$; $A_k = \{s = a_k\}$*

***Definición* 35** *Una **medida** sobre un espacio medible (X,T) es una función de conjuntos $\mu : T \rightarrow \overline{\mathbb{R}}_+$ que cumple:*

- *$\mu(\emptyset) = 0$*
- *μ es σ-aditiva, es decir:*

$$(A_n \emptyset) \subset T \rightarrow \ \mu(\cup_n A_n) = \sum_n \mu(A_n).$$

La terna (X,T,μ) se denomina espacio de medida.

Si $\mu(X) = 1$, se dice que μ es una probabilidad.

***Definición* 36** *En un espacio de medida (X,T,μ), la medida μ se dice completa si $\mu(A) = 0$ y $B \subset A$ implica que $B \in T$ (y por tanto $\mu(B) = 0$)*

Integral de funciones medibles

***Definición* 37** *Una función $f \in M(X,\mathbb{C})$ se dice integrable si se cumple:*

$$||f||_1 = \int |f| d\mu < \infty. \tag{114}$$

Se define además el conjunto:

$$L_1(\mu) = \{f \in M(X,\mathbb{C}) \ \ integrables\}. \tag{115}$$

***Definición* 38** *Se dice que una propiedad P se cumple **casi donde quiera** respecto a μ (μ c.d.) si existe $A \in T$ tal que $\mu(A) = 0$*

Los espacios L_p

***Definición* 39** *1. Si $f \in M(X,\mathbb{C})$ y $0 < p < \infty$ se define:*

$$||f||_p = (\int |f|^p d\mu)^{\frac{1}{p}}, \tag{116}$$

y se denota

$$L_p(\mu) = \{f \in M(X,\mathbb{C}); ||f||_p < \infty\}. \tag{117}$$

2. Si $f \in M(X,\overline{\mathbb{R}}_+)$ se define el ***supremo esencial*** *de f como*

$$supescf = inf\left\{a \in \overline{\mathbb{R}}_+; \mu\{f > a\} = 0\right\}. \tag{118}$$

3. Si $f \in M(X,\mathbb{C})$ se define

$$||f||_\infty = supesc|f|, \tag{119}$$

y se denota

$$L_\infty(\mu) = \{f \in M(X,(C)); ||f||_\infty < \infty\}. \tag{120}$$

***Teorema* 20** *Teorema de Fubini*
Sean (X,T,μ) y (Y,U,λ) espacios de medida σ-finitos y sea $f(x,y)$ una función $T \otimes U$-medible con funciones parciales f_x y f^y.

- *Si $0 \leq f \leq \infty$, entonces*

$$x \mapsto \int_Y f_x \, d\lambda \;\; es \;\; T - medible, \tag{121}$$

$$x \mapsto \int_X f^y \, d\mu \;\; es \;\; U - medible, \tag{122}$$

y se cumple

$$\int_X \int_Y f(x,y) d\lambda(y) d\mu(x) = \int_{XxY} f(x,y) d(\mu \otimes \lambda)(x,y). \tag{123}$$

- *Si $f \in L_1(\mu \otimes \lambda)$, entonces $f_x \in L_1(\lambda)$ μ c.d. y $f^y \in L_1(\mu)$ λ c.d., lo que implica*

$$\int_Y f_x \, d\lambda \in L_1(\mu) \;\; y \;\; \int_X f^y d\mu \in L_1(\lambda), \tag{124}$$

y se cumple el paso anterior.

Anexo 2: Conceptos estadísticos

En TDA se dispone de una muestra con un gran volumen de datos, es difícil realizar una interpretación de los mismos a partir de un proceso inferencial. Por lo tanto se necesita reducir de alguna manera el volumen de los datos para ello usualmente se acostumbra a calcular algunos estadísticos como por ejemplo: la media, la mediana, el valor más grande o el más pequeño, la varianza muestral por citar algunos. Los estadígrafos o estadísticos juegan un papel importante en la teoría de la inferencia estadística, son funciones medibles de los datos que contribuyen en algún sentido al análisis que se pretende realizar. La definición formal es la siguiente:

***Definición 40** (**Estadígrafos**) Sea $(\Omega, \mathcal{P}_\theta, P_\theta)$ un modelo estadístico. Diremos que T es un estadígrafo si es una función medible de las observaciones que no depende del parámetro desconocido θ y se define como*

$$T : \Omega \subset R^n \to \Lambda \subset R^k \qquad k \in N, \tag{125}$$

$$\mathbf{x} \quad \rightsquigarrow T(\mathbf{x}). \tag{126}$$

Debido a que un estadístico define una forma de reducción o resumen de los datos, el investigador solamente utiliza el valor observado de un estadístico $T(\mathbf{x})$ en lugar de la muestra $\mathbf{x} = (x_1, \cdots, x_n)$, considerará como iguales dos muestras $\mathbf{x}$ y $\mathbf{y}$ si satisfacen que $T(\mathbf{x}) = T(\mathbf{y})$, aunque los valores muestrales sean diferentes. Visto desde otra perspectiva, la reducción de los datos en términos de un estadístico puede verse a partir del establecimiento de una relación de equivalencia:

$$\mathbf{x}\, R\, \mathbf{y} \quad \Longleftrightarrow \quad T(\mathbf{x}) = T(\mathbf{y}). \tag{127}$$

Esta relación de equivalencia particiona al espacio muestral en clases de equivalencia definidas como

$$D_t = \{\mathbf{x} \in \Omega : T(\mathbf{x}) = t\}. \tag{128}$$

Otra noción de la inferencia estadística asociada a la idea de la reducción de los datos es el concepto de completitud, que en cierto modo es un requerimiento aún más fuerte.

***Definición 41** (**Completitud**) Sea $(\Omega, \mathcal{P}_\theta, P_\theta)$ un modelo estadístico y T un estadígrafo, diremos que es completo (acotadamente completo) para la familia de distribución de T indizada por θ si y solo si para toda función h (acotada) que satisfaga que $\mathbb{E}_\theta(h(S)) = 0$ se cumple entonces que $h \equiv 0$, excepto por un conjunto de probabilidad cero, i.e.:*

$$\mathbb{E}_\theta(h(S)) = 0, \quad \forall \theta \in \Theta \Rightarrow\ P_\theta(h(S) = 0) = 1. \tag{129}$$

El concepto de completitud se encuentra asociado a la familia de distribuciones condicionales del estadígrafo dado los diferentes valores del parámetro, es una condición de clase y se usa para buscar unicidad entre los estadígrafos.

La variabilidad de un estimador es evaluada por lo que se denomina Error Cuadrático Medio (ECM):

***Definición 42** (**Error Cuadrático Medio (ECM)**) Sea $(\Omega, \mathcal{P}_\theta, P_\theta)$ un modelo estadístico y $\mathbf{X} = (X_1, \cdots, X_n)$ una muestra, sea T un estimador de $\tau(\theta)$, entonces se define el error cuadrático medio del estimador como*

$$ECM_{\tau(\theta)}(T) = \mathbb{E}_\theta(T(\mathbf{X}) - \tau(\theta))^2. \tag{130}$$

Convergencia en probabilidad, casi segura y en distribución

Recordemos algunos resultados básicos de convergencia necesarios para los conceptos que se verán más adelante

Definición 43 *(**Modos de convergencia**) Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\{X_n\}$ una sucesión de variables aleatorias, entonces*

1. *(*Casi segura*) Diremos que X_n converge casi seguramente a X, o con probabilidad uno si:*

$$P\left(\{\omega \in \Omega : X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)\}\right) = 1, \tag{131}$$

$$\{\omega \in \Omega : X_n(\omega) \to X(\omega)\} = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n>N} \left\{ |X_n(\omega) - X(\omega)| < \frac{1}{k} \right\} \in \mathcal{F}, \tag{132}$$

y se denota como

$$X_n \xrightarrow[n\to\infty]{c.s.} X. \tag{133}$$

2. *(*Puntual casi seguramente*) Sea $X_{n,\theta} = X_n(\omega,\theta)$ y $X_\theta = X(\omega,\theta)$ con $\theta \in \Theta$ se dice que $X_{n,\theta}$ converge puntualmente casi seguramente a X_θ si*

$$P\left(\lim_{n\to\infty} |X_{n,\theta} - X_\theta| = 0\right) = 1, \qquad \forall \theta \in \Theta. \tag{134}$$

3. *(*Uniforme casi seguramente*) Sea $X_{n,\theta} = X_n(\omega,\theta)$ y $X_\theta = X(\omega,\theta)$ con $\theta \in \Theta$ se dice que $X_{n,\theta}$ converge uniformemente casi seguramente a X_θ si*

$$P\left(\lim_{n\to\infty} \sup_{\theta\in\Theta} |X_{n,\theta} - X_\theta| = 0\right) = 1. \tag{135}$$

4. *(*En probabilidad*) Diremos que X_n converge en probabilidad a X o estocásticamente si:*

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0 \qquad \forall \varepsilon > 0. \tag{136}$$

y se denota por

$$X_n \xrightarrow[n\to\infty]{P} X. \tag{137}$$

5. *(*Puntual en probabilidad*) Sea $X_{n,\theta} = X_n(\omega,\theta)$ y $X_\theta = X(\omega,\theta)$ con $\theta \in \Theta$ se dice que $X_{n,\theta}$ converge puntualmente en probabilidad a X_θ si*

$$\lim_{n\to\infty} P(|X_{n,\theta} - X_\theta| > \varepsilon) = 0 \qquad \forall \varepsilon > 0,\ \forall \theta \in \Theta. \tag{138}$$

6. *(*Uniforme en probabilidad*) Sea $X_{n,\theta} = X_n(\omega,\theta)$ y $X_\theta = X(\omega,\theta)$ con $\theta \in \Theta$ se dice que $X_{n,\theta}$ converge uniformemente en probabilidad a X_θ si*

$$\lim_{n\to\infty} P\left(\sup_{\theta\in\Theta} |X_{n,\theta} - X_\theta| > \varepsilon\right) = 0 \qquad \forall \varepsilon > 0. \tag{139}$$

7. *(*En media*) Diremos que X_n converge en media de orden p a X si:*

$$\mathbb{E}(|X_n - X|^p) \to 0, \tag{140}$$

y se denota por

$$X_n \to [n \to \infty] L_p X. \tag{141}$$

Cuando $p = 2$ la convergencia se dice que es en media cuadrática.

*8. (*En distribución*) Diremos que X_n con función de distribución F_n converge en distribución o en ley a X cuya función de distribución se denota por F_X si:*

$$P(X_n \leq x) = F_n(x) \to F_X(x), \tag{142}$$

en todo punto de continuidad de F_X y se denota por

$$X_n \xrightarrow[n\to\infty]{\mathcal{D}} X.$$

***Definición 44** (**Ley de los Grandes Números**) Se dice que la sucesión $\{X_n\}$ de variables aleatorias satisface la Ley de los Grandes Números con respecto a las funciones $\{g_n\}$, $g_n = g_n(X_1, \cdots, X_n)$, si existe una sucesión de constantes $\{b_n\}$ tales que*

1. Ley Fuerte de los Grandes Números

$$g_n(X_1, \cdots, X_n) - b_n \xrightarrow[n\to\infty]{c.s.} 0. \tag{143}$$

2. Ley Débil de los Grandes Números

$$g_n(X_1, \cdots, X_n) - b_n \xrightarrow[n\to\infty]{P} 0. \tag{144}$$

Usualmente se toma

$$g_n(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=k}^{n} X_k, \qquad b_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i). \tag{145}$$

En cuanto al Teorema Central del Límite los resultados clásicos son los siguientes

***Teorema 21** (**Moivre-Laplace**) Sea $\{X_n\}$ una sucesión de variables aleatorias Bernoulli i.i.d tales que $\mathbb{E}(X_k) = p$ y $V(X_k) = p(1-p)$ para todo $k = 1, 2, \cdots, n$. Sea $S_n = \sum_{k=1}^{n} X_k$, entonces*

$$\frac{\sum_{k=1}^{n} X_k - np}{\sqrt{np(1-p)}} \xrightarrow[n\to\infty]{\mathcal{D}} Z \sim N(0,1). \tag{146}$$

***Teorema 22** (**Linderbeg-Lévy**) Sea $\{X_n\}$ una sucesión de variables aleatorias tales que $\mathbb{E}(X_k) = \mu_k < +\infty$ para todo $k = 1, 2, \cdots, n$ y $V(S_n) < +\infty$, donde $S_n = \sum_{k=1}^{n} X_k$, entonces*

$$\frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{\sqrt{V(S_n)}} \xrightarrow[n\to\infty]{\mathcal{D}} Z \sim N(0,1). \tag{147}$$

Si las variables son i.i.d. tales que $\mathbb{E}(X_k) = \mu < +\infty$ y $V(X_k) = \sigma^2 < +\infty$ para todo $k = 1, 2, \cdots, n$ entonces

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{\mathcal{D}} Z \sim N(0,1). \tag{148}$$

Existen diversas situaciones donde a un investigador le resulta de mayor utilidad un conjunto de posibles valores de un parámetro que una sola cantidad producida por una estimación del mismo, dando lugar a la estimación por intervalos.

Definición 45 (Intervalo de Confianza) *Sea* $\mathbf{X} = (X_1, \cdots, X_n)$ *una muestra aleatoria cuya distribución depende de un parámetro* θ *desconocido. Diremos que una estimación por intervalo para una función* $g(\theta)$ *del parámetro escalar* θ*, es cualquier par de funciones* $L(\mathbf{X})$ *y* $U(\mathbf{X})$ *que satisfacen que* $L(\mathbf{x}) \leq U(\mathbf{x})$ *para todo punto muestral* $\mathbf{x} \in \Omega$ *del espacio muestral, entonces* $IC_{1-\alpha} = [L(\mathbf{x}); U(\mathbf{x})]$ *es un intervalo de confianza para* $g(\theta)$ *con confianza* $1-\alpha$ *si la probabilidad de cubrimiento o la probabilidad de que el intervalo aleatorio cubra al verdadero valor del estimando satisface que*

$$P\big(L(\mathbf{X}) \leq g(\theta) \leq U(\mathbf{X})\big) = 1-\alpha. \tag{149}$$

Cuando la muestra es observada y el estimador por intervalo es evaluado se tiene entonces la estimación por intervalo. Usualmente $1-\alpha$ *se conoce como nivel de confianza del intervalo.*

Definición 46 (Límites de Confianza)

Sea $\mathbf{X} = (X_1, \cdots, X_n)$ *una muestra aleatoria cuya distribución depende de un parámetro* θ *desconocido. Diremos que* T^{α} *es un límite de confianza superior para* $g(\theta)$ *si*

$$P\big(T^{\alpha}(\mathbf{X}) \geq g(\theta)\big) = 1-\alpha. \tag{150}$$

De manera similar se define el límite de confianza inferior T_{α}*:*

$$P\big(T_{\alpha}(\mathbf{X}) \leq g(\theta)\big) = 1-\alpha. \tag{151}$$

RT_088, enero 2017
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba