Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Maestría en Estadística Matemática

TÍTULO DE LA TESIS: **Test de hipótesis sobre homología persistente utilizando la distancia de Fermat**

Tesis presentada para optar al Título de Magister de la Universidad de Buenos Aires en Estadística Matemática

MAESTRANDO: Diego Javier Battocchio

Director: Pablo Groisman
Lugar de Trabajo: Buenos Aires - UBA FCEyN - Instituto del Cálculo
Fecha de presentación del ejemplar: 2024/10/28
Fecha de Defensa: 2025/08/29

FIRMA DEL DIRECTOR FIRMA DEL MAESTRANDO

Resumen

Esta tesis se centra en el análisis de datos topológicos (TDA) y, específicamente, en el desarrollo de tests de hipótesis para detectar características topológicas de una variedad a partir de muestras i.i.d. de puntos. Se introducen conceptos clave de TDA y se describen técnicas actuales para realizar tests sobre diagramas de persistencia, los cuales resumen la homología de un conjunto de datos, facilitando la identificación de ciclos de primer grado (curvas no deformables a un punto). Tradicionalmente, estas pruebas se basan en la norma euclídea, pero aquí se explora el uso de la distancia de Fermat, que ha demostrado ser beneficiosa en conjuntos con topología subyacente. Mediante simulaciones de regiones de confianza sobre datos sintéticos y reales, se comparan ambas distancias, demostrando que la distancia de Fermat supera a la euclídea en la detección de características topológicas, con resultados estadísticamente significativos. Además, se muestra que la distancia de Fermat es computacionalmente escalable frente a la dimensionalidad de los datos. Como aporte adicional, se desarrolló una biblioteca en R para facilitar investigaciones futuras sobre el uso de la distancia de Fermat en este contexto.

Palabras clave: TDA, topología, homología persistente, distancia de Fermat, prueba de hipótesis, diagramas de persistencia, regiones de confianza.

FIRMA DEL DIRECTOR FIRMA DEL MAESTRANDO

Abstract

Hypothesis testing for persistent homology using the Fermat distance

This dissertation focuses on topological data analysis (TDA) and, specifically, on the development of hypothesis tests to detect topological features of a manifold based on i.i.d. samples. Key TDA concepts are introduced, and state-of-the-art techniques for hypothesis testing on persistence diagrams are described. These diagrams summarize the homology of a dataset, helping to identify first-degree cycles (curves that cannot be deformed to a point). Traditionally, such tests rely on the Euclidean norm, but this work explores the use of the Fermat distance, which has shown benefits in datasets with underlying topology. Through simulations of confidence sets on synthetic and real datasets, both distances are compared, demonstrating that Fermat distance outperforms Euclidean distance in detecting topological features, with statistically significant results. Moreover, Fermat distance is shown to be computationally scalable with respect to data dimensionality. As an additional contribution, an R library was developed to facilitate future research on Fermat distance in this context.

Keywords: TDA, topology, persistent homology, Fermat distance, hypothesis testing, persistence diagramas, confidence sets.

DIRECTOR'S SIGNATURE STUDENT'S SIGNATURE

Tabla de contenidos

Capítulo 1

Introducción

En la presente Tesis se aborda, de forma general, el análisis topológico de datos o TDA por sus siglas en inglés (*topological data analysis*) (Chazal y Michel 2021) y en forma particular los tests de hipótesis para detectar características topológicas de una variedad basados en muestras independientes e idénticamente distribuidas (i.i.d) de puntos sobre la misma. Se realiza una introducción al TDA y se describen las técnicas del estado del arte en el área de test de hipótesis sobre diagramas de persistencia. Siendo estos diagramas una medida de resumen de la homología de un conjunto de datos, se imponen como una herramienta fundamental para evaluar estadísticamente la existencia de ciclos de primer grado, es decir, curvas que no pueden ser deformadas en un punto (Wikipedia 2022). Durante el proceso de construcción de los diagramas de persistencia se hace uso de una medida de distancia sobre el espacio que se busca describir, que resulta ser crucial a la hora de calcular los diagramas, obteniéndose resultados muy distintos dependiendo de la distancia utilizada. Tradicionalmente, esta medida de distancia se basa en la norma euclídea, pero en este trabajo se explora el uso de la *distancia de Fermat* que ha demostrado ser beneficiosa en conjuntos de datos con una topología subyacente (Sapienza, Groisman, y Jonckheere 2018). Particularmente, se realizan simulaciones de las regiones de confianza mediante las técnicas del estado del arte sobre varios conjuntos de datos sintéticos utilizados frecuentemente en la literatura de TDA (Fasy et al. 2014), así también como en conjuntos de datos reales que puedan tener utilidad práctica (Chazal et al. 2017) utilizando tanto la norma euclídea como la distancia de Fermat para realizar estos cálculos. Para el caso de los conjuntos sintéticos, en los cuales se permite obtener diferentes muestras de la misma distribución, se busca también obtener una estimación computacional de la potencia de estas pruebas de hipótesis a la hora de detectar agujeros en los espacios topológicos de los cuales los conjuntos de datos fueron obtenidos.

Los resultados obtenidos muestran que efectivamente la distancia de Fermat logra capturar de forma estadísticamente significativa y con mayor potencia la homología de algunas distribuciones sintéticas comunes en la literatura, donde las técnicas análogas que utilizan la distancia euclídea fallan. Se obtienen también resultados superiores a los de las otras técnicas para conjuntos de datos reales. Por otro lado, se muestra cómo las técnicas basadas en la distancia de Fermat logran introducir estas ventajas en los resultados de forma escalable computacionalmente en términos de la dimensionalidad del espacio de origen de los datos, siendo esto uno de los principales problemas de otras técnicas que han sido propuestas para mejorar los resultados que se obtienen al utilizar la distancia euclídea.

Como aporte adicional, se desarrolló una biblioteca en el lenguaje de programación R que podrá ser utilizada para futuros trabajos sobre la distancia de Fermat en este lenguaje.

A modo de compromiso con la reproducibilidad en la ciencia, la totalidad de esta Tesis se encuentra en GitHub.

Capítulo 2

Conceptos Utilizados y Marco Teórico

2.1 Distancia de Fermat

En una gran cantidad de aplicaciones de aprendizaje automático resulta de crucial importancia la elección de una medida de distancia apropiada para representar el espacio, como por ejemplo, durante la construcción de diagramas de persistencia desarrollados durante esta Tesis. Un enfoque a la hora de elegir esta medida es dejar que la misma se infiera a partir de los datos, en vez de ser elegida arbitrariamente. Este enfoque resulta de particular interés cuando los datos proviene de una variedad desconocida de dimensión menor, siendo esto una situación típica en distintos tipos de aplicaciones (Bengio, Courville, y Vincent 2013). Para atacar este problema usaremos la distancia de Fermat, que se define a continuación:

Sea una muestra de N puntos

$$\mathcal{S}_N = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{M},$$

donde $\mathcal{M} \subseteq \mathbb{R}^D$ es una variedad de dimensión $d \geq 1$ y tal que $d \ll D$. Sea además $l(\cdot, \cdot)$ una función de distancia definida en $\mathcal{M} \times \mathcal{M}$ (siendo una opción típica la distancia euclídea, pero pudiendo ser considerada cualquier otra distancia). Se define entonces, para $\lambda \geq 1$, la distancia de Fermat para el par de puntos $\mathbf{p}, \mathbf{q} \in \mathcal{M}$ como:

$$d_F(\mathbf{p}, \mathbf{q}) = \mathcal{D}_{\mathcal{S}_N}(\mathbf{p}, \mathbf{q}) = \min_{K} \min_{\mathcal{S}_N^K} \sum_{i=1}^{K-1} l\left(\mathbf{x}^i, \mathbf{x}^{i+1}\right)^{\lambda}, \tag{2.1}$$

donde $\mathcal{S}_N^K$, $K \geq 2$, representa todas las secuencias posibles de puntos compuestas por K elementos $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K$ tomados de $\mathcal{S}_N$ con

$$\mathbf{x}^1 = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{S}_N} l(\mathbf{p}, \mathbf{x}), \tag{2.2}$$

$$\mathbf{x}^K = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{S}_N} l(\mathbf{q}, \mathbf{x}). \tag{2.3}$$

La minimización se realiza sobre todos los valores posibles de K (Sapienza, Groisman, y Jonckheere 2018). Esta distancia captura por sí sola características topológicas del espacio (Fernández et al. 2024), por lo que se espera que ayude a la posterior inferencia sobre el diagrama de persistencia construido.

Una de las principales ventajas de la distancia de Fermat es que si $\mathcal{S}_N$ es una muestra i.i.d, distribuida a partir de $f : \mathcal{M} \longrightarrow \mathbb{R}$ una función de densidad desconocida con soporte en $\mathcal{M}$, entonces se demuestra en (Fernández et al. 2024; Sapienza, Groisman, y Jonckheere 2018) que

$$\lim_{N\to\infty} N^{\frac{\lambda-1}{D}} \mathcal{D}_{\mathcal{S}_N}(\mathbf{p},\mathbf{q}) = c_{d,D} \inf_{\Gamma\subset\mathcal{M}} \int_\Gamma \frac{1}{f^{\frac{\lambda-1}{D}}},$$

es decir, la distancia de Fermat es un estimador consistente de la distancia geodésica sobre $\mathcal{M}$ en la que los caminos se ponderan por el valor de la función de densidad. Esto significa que la distancia de Fermat recupera tanto características de la topología del espacio $\mathcal{M}$ como de la función de distribución a partir de la cual se muestrea $\mathcal{S}_N$.

Figura 2.1: Ilustración de la distancia de Fermat (con l la distancia euclídea y $\lambda = 2$) entre los puntos destacadas en rojo y verde. Se observa que la misma se obtiene al recorrer la zona con alta densidad de datos del conjunto, sumando las bolas de distancia entre sucesivos saltos. Contrario a lo que se obtendría al utilizar la distancia euclídea, donde la misma resultaría dada por una línea recta que une ambas muestras de forma directa, pasando por el espacio vacío de la parábola de datos.

2.1.1 Aproximación por *k*-NN

Teniendo en cuenta que la carga computacional del cálculo de la distancia de Fermat es elevada, siendo esta $\mathcal{O}(N^3)$ para todas las distancias entre pares de un conjunto de N puntos, una heurística sencilla para reducir este tiempo, sin perdida de garantías en cuanto a los resultados asintóticos, es tomar solo los k vecinos más cercanos de $\mathbf{x}^i$ como posibles $\mathbf{x}^{i+1}$ durante la construcción de las secuencias $\mathcal{S}_N^K$ sobre las cuales se realiza la minimización. Formalmente, si se define $\mathcal{N}_k(\mathbf{x})$ como el conjunto de k vecinos más cercanos (k-NN) de $\mathbf{x}$, entonces la distancia de Fermat se define análogamente en (2.1) pero siendo $\mathcal{S}_N^K$, $K \geq 2$ todas las secuencias posibles de puntos compuestos de K elementos tomados de $\mathcal{S}_N$ tal que $\mathbf{x}^{i+1} \in \mathcal{N}_k(\mathbf{x}^i)\ \forall i$.

Con esta modificación, el cómputo de la distancia de Fermat para todos los pares de un conjunto de N puntos tiene una complejidad temporal asintótica de $\mathcal{O}(N^2 \log N)$, sumado a esto, es posible almacenar la matriz de distancia de forma mucho más eficiente, ya que la misma pasa de ser densa,

con N^2 elementos distinto de cero, a ser rala contando con $N * k$ elementos no nulos, pudiéndose esto almacenar de forma mucho más eficiente.

2.1.2 Aproximación por *Landmarks*

Otro mecanismo de aproximación para la distancia de Fermat se obtiene a partir de la utilización de *landmarks* en el cálculo del camino más corto entre nodos dentro de un grafo. Teniendo en cuenta que el problema de encontrar la distancia de Fermat entre dos puntos puede verse como el problema análogo de encontrar el camino más corto entre dos nodos en el grafo en el que cada punto del conjunto representa un nodo y sus aristas están dadas por las distancias entre los respectivos puntos elevadas al factor λ, es decir, un grafo completo. A continuación se describe la técnica de *Landmarks* para el cálculo eficiente del camino más corto entre nodos de un grafo:

Sea un grafo $G(V, E)$ con n vértices y m aristas. Dados dos vértices $s, t \in V$, se define $d_G(s, t)$ como la longitud del camino más corto entre dos vértices $s, t \in V$. Sea además un conjunto ordenado de d vértices $D = \{u_1, u_2, ..., u_d\}$ del grafo G, que llamamos *landmarks*. La idea principal es representar cada otro vértice en el grafo como un vector de distancias de camino más corto al conjunto de *landmarks*. En particular, cada vértice $v \in V$ se representa como un vector $\phi(v)$ en $\mathbb{R}^d$:

$$\phi(v) = [d_G(v, u_1), d_G(v, u_2), ..., d_G(v, u_d)]^T = [\phi_1(v), \phi_2(v), ..., \phi_d(v)]^T.$$

Teniendo en cuenta que la distancia de camino más corto en grafos es una métrica y, por lo tanto, satisface la desigualdad triangular. Es decir, dados tres nodos s, u y t, se cumplen las siguientes desigualdades:

$$|d_G(s, u) - d_G(u, t)| \leq d_G(s, t) \leq d_G(s, u) + d_G(u, t). \tag{2.4}$$

Resulta una observación importante que si u pertenece al camino más corto de s a t, entonces la cota superior se cumple con igualdad, es decir $d_G(s, t) = d_G(s, u) + d_G(u, t)$

Haciendo uso entonces de las representaciones $\phi(v)$ y $\phi(t)$ y mediante las desigualdades expresadas en (2.4) podemos acotar $d_G(s, t)$ como

$$\max_i |\phi_i(s) - \phi_i(t)| \leq d_G(s, t) \leq \min_j \{\phi_j(s) + \phi_j(t)\}.$$

Si se definen $L = \max_i |\phi_i(s) - \phi_i(t)|$ y $\min_j \{\phi_j(s) + \phi_j(t)\} = U$ podemos entonces estimar $d_G(s, t)$ como

- La estimación superior $\tilde{d}_{G,U}(s, t) = U$.
- La estimación del punto medio $\tilde{d}_{G,M}(s, t) = \frac{L+U}{2}$.
- La estimación de la media geométrica $\tilde{d}_{G,G}(s, t) = \sqrt{L \cdot U}$.

Nótese que en todos los casos, la estimación es muy rápida, ya que solo se necesitan $O(d)$. La cota superior $\tilde{d}_{G,U}(s, t) = U$ resulta la mejor estimación según la experimentación realizada en (Potamias et al. 2009). Adicionalmente, en (Potamias et al. 2009) se muestra que la elección aleatoria de *landmarks* cuenta con buenas garantías frente a la elección óptima, siendo esta última un problema computacionalmente demasiado costoso.

2.1.3 Implementación

La distancia de Fermat, tanto en su versión tradicional como en la aproximación por k vecinos más cercanos y por *landmarks*, se encuentra implementada en GitHub para *Python* y *R*, habiendo sido la librería para este último lenguaje construida especialmente para el desarrollo de esta Tesis.

2.2 Análisis Topológico de Datos (TDA)

El Análisis Topológico de Datos o TDA es una área recientemente surgida de la ciencia de datos cuya principal motivación es la idea de que la topología y geometría proveen un acercamiento poderoso para inferir información robusta, cualitativa y a veces cuantitativa sobre la estructura de los datos (Chazal y Michel 2021). El objetivo de esta sección es brindar una introducción a las herramientas y conceptos del TDA que utilizaremos para el desarrollo de esta Tesis.

2.2.1 Espacios métricos

Las cualidades topológicas y geométricas de un espacio están usualmente asociadas a espacios continuos, por lo que los datos representados como un conjunto finito de observaciones no revelan directamente ninguna información topológica por sí mismos. Una forma natural de destacar la estructura topológica de los datos es "conectar" los puntos que están "cerca" con el objetivo de exhibir una estructura global subyacente de los datos. La "cercanía" entre puntos se mide mediante una función de distancia, por lo que resulta conveniente considerar a los conjuntos de datos como muestras de un espacio métrico.

Espacio Métrico. Una espacio métrico (M, ρ) es un conjunto M con una función $\rho : M \times M \longrightarrow \mathbb{R}^{\geq 0}$, llamada distancia, tal que para cualquier terna de puntos $\mathbf{x}, \mathbf{y}, \mathbf{z} \in M$ se cumple:

I. No negatividad: $\rho(\mathbf{x}, \mathbf{y}) \geq 0$ y $\rho(\mathbf{x}, \mathbf{y}) = 0$ solo sí $\mathbf{x} = \mathbf{y}$.
II. Simetría: $\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}, \mathbf{x})$.
III. Desigualdad triangular: $\rho(\mathbf{x}, \mathbf{z}) \leq \rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z})$.

Puede demostrarse (Sapienza, Groisman, y Jonckheere 2018) que la distancia de Fermat desarrollada en la Sección 2.1 cumple las condiciones para ser una distancia según esta definición, por lo que pueden estudiarse las cualidades topológicas de un espacio métrico en base a esta misma.

Distancia de Hausdorff

Dados dos subconjuntos compactos $A, B \subseteq M$ la distancia de Hausdorff $d_H(A, B)$ entre A y B se define como el valor δ más pequeño tal que para cualquier $\mathbf{a} \in A$ existe un $\mathbf{b} \in B$ tal que $\rho(\mathbf{a}, \mathbf{b}) \leq \delta$ (ver Figura 2.2)

En otras palabras, si para todo subconjunto compacto $C \subseteq M$ denotamos $d(\cdot, C) : M \longrightarrow \mathbb{R}^+$ como la función de distancia a C, definida como

$$d(\mathbf{x}, C) := \inf_{\mathbf{c} \in C} \rho(\mathbf{x}, \mathbf{c}), \quad \forall \mathbf{x} \in M.$$

Puede probarse que la distancia de Hausdorff entre A y B está definida mediante cualquiera de las siguientes igualdades

$$\begin{aligned} d_H(A, B) &= \max\{\sup_{\mathbf{b} \in B} d(\mathbf{b}, A), \sup_{\mathbf{a} \in A} d(\mathbf{a}, B)\} \\ d_H(A, B) &= \sup_{\mathbf{x} \in M} |d(\mathbf{x}, A) - d(\mathbf{x}, B)| = ||d(\cdot, A) - d(\cdot, B)||_\infty. \end{aligned} \tag{2.5}$$

La distancia de Hausdorff provee una forma conveniente de cuantificar la proximidad entre diferentes conjuntos de datos provenientes del mismo espacio métrico.

2.2.2 Homología

La homología es un concepto fundamental en diversas ramas de las matemáticas, particularmente en topología y geometría algebraica. Esta área de estudio se centra en comprender y clasificar las formas y estructuras espaciales de manera abstracta, permitiendo así la comparación y el análisis de diferentes espacios geométricos. La homología asocia entonces a cada espacio una serie de grupos, llamados grupos de homología, que reflejan características importantes del mismo, como pueden ser componentes conexas y agujeros de diversas dimensiones. Esta asignación se hace a partir de la definición de los conceptos de

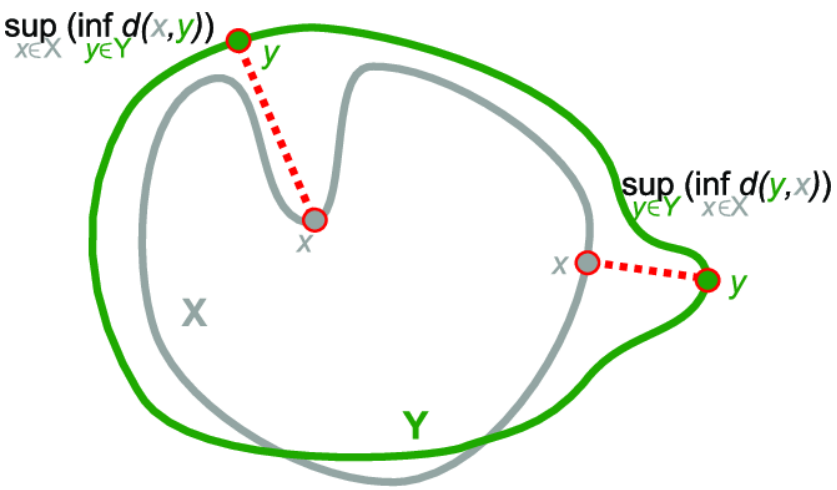


Figura 2.2: Se ilustra la distancia de Hausdorff para el caso de dos curvas X e Y, representando los subconjuntos $A, B \subseteq M$ siguiendo la notación de (2.5). Si para todo punto en la curva gris se busca su punto más cercano en la curva verde, para posteriormente conservar la máxima de esas distancias, se obtienen los puntos x e y en la parte superior de la imagen. Por otro lado, si este proceso se realiza tomando ahora en cuenta todos los puntos de la curva verde, buscando su par más cercano en la curva gris, para posteriormente solo conservar la mayor de esas distancias, se obtiene la distancia marcada en línea punteada roja a la izquierda. Habiendo obtenido estas dos distancias, la mayor de ellas será la distancia de Hausdorff. Crédito de la imagen: (Pellerin 2014).

"cadenas", "ciclos" y "bordes": una cadena es una combinación formal de elementos geométricos, como pueden ser puntos, curvas o superficies. Un borde se define como una cadena que actúa de frontera de una cadena de dimensión superior, se define entonces el operador borde ∂ que mapea cadenas de dimensión k (o k-cadenas) a $(k-1)$-cadenas. A partir de la noción de borde, se define a los ciclos como cadenas c sin frontera, es decir $\partial c = 0$. Estos representan intuitivamente "agujeros" en diversas dimensiones. En particular, a lo largo de este trabajo, haremos foco mayoritariamente en la homología de grado uno, denotada como Π_1, que puede ser interpretada, de la forma más intuitiva posible, como la cantidad de ciclos unidimensionales diferentes, donde por "diferentes" nos referimos a que no pueden ser deformados dentro de la variedad hasta ser iguales entre sí. Si la variedad $\mathcal{M}$ presenta entonces un único agujero de estas características, esto se denota como:

$$\Pi_1(\mathcal{M}) = \mathbb{Z},$$

siendo uno la dimensión del espacio de homología en grado uno de la variedad, lo que significa que hay un solo ciclo independiente que no puede deformarse dentro de la estructura, en notación matemática:

$$\dim(\Pi_1(\mathcal{M})) = \dim(\mathbb{Z}) = \beta_1(\mathcal{M}) = 1.$$

A la cantidad β_1, o en general, β_k para homologías de grado superior, se la conoce como el k-ésimo número de *Betti*.

Otro grado de homología que mencionaremos brevemente será el de las componentes conexas, denotado como Π_0. Intuitivamente, una componente conexa representa todos los puntos que pueden ser unidos mediante una curva sin salirse de la variedad.

2.2.3 Homología persistente

A partir de la definición general de Homología desarrollada en la Sección 2.2.2, el problema fundamental que se busca abordar con el Análisis Topológico de Datos (TDA) es inferir la homología de una variedad $\mathcal{M}$ a partir de una muestra $\mathcal{S}_N$, obtenida mediante una función de densidad f con soporte en $\mathcal{M}$. Como

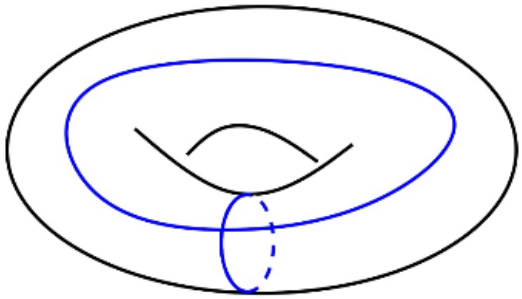


Figura 2.3: Se ilustra la homología de primer grado de un toro, marcando en azul las dos curvas del mismo que no pueden ser continuamente deformadas la una en la otra. Cualquier otra curva cerrada que elijamos sobre el toro, podrá ser deformada en alguna de estas dos. Crédito de la imagen: (Chazal y Michel 2021).

se mencionó brevemente en la Sección 2.2.1, esto se realiza mediante la elección de una medida de distancia. Concretamente, si se define una *bola* $B(\mathbf{x}_i, \epsilon)$ como la región del espacio que se encuentra a distancia menor a ϵ del punto $\mathbf{x}_i$, es decir:

$$B(\mathbf{x}_i, \epsilon) = \{\mathbf{x} \mid d(\mathbf{x}_i, \mathbf{x}) < \epsilon\},$$

con $d(\cdot, \cdot)$ una función de distancia. Se define entonces el espacio generado por la unión de estas bolas de radio ϵ centradas en cada una de las muestras de $\mathcal{S}_N$:

$$\mathcal{S}_N^\epsilon = \cup_{\mathbf{x} \in \mathcal{S}_N} B(\mathbf{x}, \epsilon).$$

Se demuestra en (Fasy et al. 2014) que cuando $\mathcal{M}$ es una variedad compacta y suave, bajo condiciones poco restrictivas sobre la distancia de Hausdorff (definida en la Sección 2.2.1) entre la muestra y la variedad de la que proviene $d_H(\mathcal{S}_N, \mathcal{M})$, y una elección apropiada del valor ϵ, entonces el espacio $\mathcal{S}_N^\epsilon$ es topológicamente equivalente a $\mathcal{M}$ (Niyogi, Smale, y Weinberger 2008). Esto significa que podemos inferir la topología de $\mathcal{M}$ a partir de $\mathcal{S}_N^\epsilon$. Se ilustra este concepto en la Figura 2.4.

En la práctica, este enfoque resulta en dos problemas fundamentales:

1. A pesar de ser posible, resulta complejo inferir la topología a partir de $\mathcal{S}_N^\epsilon$. Es por esto que se introducen los complejos simpliciales, los cuales desarrollaremos en la Sección 2.2.3, para eficientizar el cómputo.
2. La elección del ϵ apropiado no resulta obvia. Como se observa en la Figura 2.4 la homología de $\mathcal{S}_N^\epsilon$ varía con la elección de ϵ. La respuesta a este problema viene dada por las técnicas de homología persistente que, esencialmente, consisten en evaluar la homología de $\mathcal{S}_N^\epsilon$ no solo para un valor de ϵ, si no para un rango de valores lo suficientemente grande. La homología persistente busca resumir cómo las cualidades topológicas de $\mathcal{S}_N^\epsilon$ cambian a medida que varía el valor de ϵ. Por ejemplo los agujeros pueden aparecer o cerrarse a medida que el radio de las bolas aumenta.

Este procedimiento da como resultado un gráfico conocido como "diagrama de persistencia", descrito a continuación.

Diagrama de persistencia

El procedimiento de variar ϵ y evaluar las cualidades topológicas de $\mathcal{S}_N^\epsilon$ para cada valor da como resultado un tiempo de nacimiento b_i en el que la cualidad topológica i aparece en $\mathcal{S}_N^\epsilon$ y un valor de

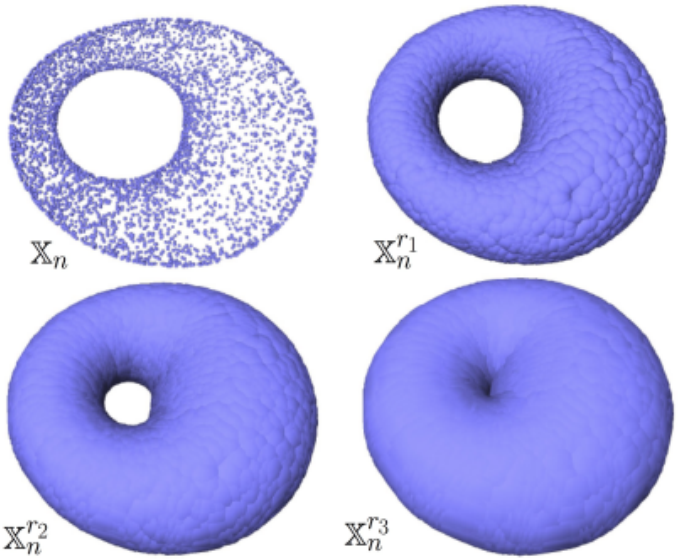


Figura 2.4: Se ilustra en la imagen el conjunto de datos $\mathcal{S}_N$ (esquina superior izquierda, denotado como $\mathbb{X}_n$) y los espacios $\mathcal{S}_N^\epsilon$ (denotados en la imagen como $\mathbb{X}_n^{r_i}$) para tres diferentes valores de ϵ. Se observa cómo el conjunto de datos parece provenir de una variedad que presenta un agujero en su interior. Esto deja de ser apreciable para el valor de ϵ más grande en la esquina inferior derecha. Crédito de la imagen: (Chazal y Michel 2021).

muerte d_i, en el que esa misma cualidad desaparece. Un caso de esto sería, por ejemplo, el valor de ϵ en el que un agujero aparece y el valor en el que todo el espacio de este agujero es tapado por las bolas $B(\mathbf{x}_i, \epsilon)$. Tomando como referencia la Figura 2.4, se observa que para el primer valor de ϵ ya es apreciable el agujero, por lo que ya se tendría el valor de nacimiento $b_i \leq \epsilon_1$, mientras que para el último valor del radio de las bolas en la esquina inferior derecha el agujero ya no es apreciable, por lo que a esta altura se obtendría el valor de muerte $d_i \leq \epsilon_3$ de la cualidad.

Al combinar estos valores de nacimiento y muerte obtenemos un par ordenado (b_i, d_i) para cada cualidad topológica. Si son dispuestos conjuntamente en un plano dan lugar al diagrama de persistencia. Este procedimiento constructivo se ilustra en la Figura 2.5, donde se representa para distintos valores de ϵ el nacimiento de una cualidad topológica como una barra, correspondiendo el color rojo a una componente conexa y el azul a un agujero de primer orden. En un inicio, cada punto es su propia componente conexa, pero a medida que el radio de la bola comienza a aumentar, solo se preserva una única componente conexa correspondiente a la unión de todas las bolas. Por su parte, al momento en el que las bolas logran unirse por primera vez, se observa el nacimiento de los dos agujeros, representados como dos barras azules. El agujero inferior es más pequeño y sus puntos son más dispersos, por lo que esa cualidad topológica nace después y muere antes. En la esquina inferior izquierda de la imagen se observa cómo ya todas las cualidades encontraron su ϵ de muerte, por lo que pueden ser finalmente representadas en el diagrama de persistencia. La altura de cada barra se preserva por sobre la diagonal, indicando el eje x el momento de nacimiento y el y el de muerte de la cualidad.

El diagrama de persistencia representa una medida de resumen sobre las características topológicas de los datos utilizados. Puntos cercanos a la diagonal representan cualidades topológicas con corto tiempo de vida y serán consideradas como "ruido topológico". Esto se ilustra en la Figura 2.6, donde se observa la secuencia constructiva del diagrama de persistencia para una circunferencia ruidosa. En el proceso "nacen" cualidades topológicas espurias, que no corresponden a lo que debería ser el diagrama de persistencia de una circunferencia ideal (una única componente de primer grado). Si bien estas son efectivamente componentes que nacen y mueren en diferentes ϵ, su tiempo de vida es muy inferior al del agujero real, por lo que se posicionarán más cerca de la diagonal en el diagrama de persistencia resultante. Las aplicaciones están usualmente interesadas en cualidades que pueden ser distinguidas del

ruido, es decir, cualidades que persisten para un amplio rango de valores ϵ (Fasy et al. 2014; Maletić, Zhao, y Rajković 2016; Perea 2018; Emrani, Gentimis, y Krim 2014). El rol de los test de hipótesis el de determinar, bajo cierto nivel de confianza, cuáles características topológicas son distintas a ruido.

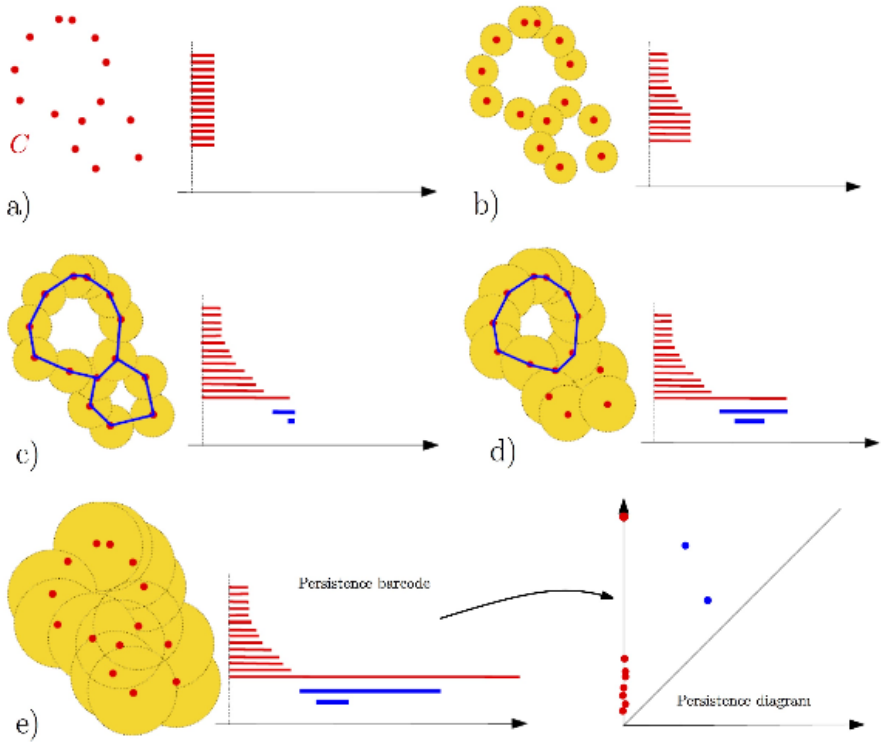


Figura 2.5: Ilustración del proceso constructivo del diagrama de persistencia para un conjunto de datos provenientes de una figura similar a la de un número ocho. Distintos valores de ϵ se corresponden con la representación del conjunto $\mathcal{S}_N^\epsilon$ construido a partir de los datos y a su lado un diagrama de barras donde cada barra representa el tiempo de vida de cada una de las cualidad topológicas que nacieron hasta el momento. Crédito de la imagen: (Chazal y Michel 2021).

Distancia entre diagramas de persistencia

Sean $\mathcal{X}$ y $\mathcal{Y}$ dos diagramas de persistencia, resulta natural preguntarse cómo estos pueden ser comparados entre sí. Más aún, si los conjuntos de datos que dieron lugar a los diagramas de persistencia $\mathcal{X}$ y $\mathcal{Y}$ son muy distintos, ¿Serán también estos diagramas resultantes muy disimiles? Para responder estas preguntas se debe definir la noción de distancia entre diagramas de persistencia conocida como distancia *bottleneck* (Edelsbrunner y Harer 2010).

Para definir la distancia *bottleneck* entre $\mathcal{X}$ e $\mathcal{Y}$ consideraremos una biyección $\eta : \mathcal{X} \to \mathcal{Y}$, es decir, una función que le asigna a cada punto $x \in \mathcal{X}$ un punto $y \in \mathcal{Y}$ y viceversa. Guardaremos el supremo de las distancias entre puntos correspondientes a cada diagrama, midiendose esta distancia como

$$||x - y||_\infty = \max\{|x_1 - y_1|, |x_2 - y_2|\}.$$

Se define la distancia *bottleneck* $W_\infty(\mathcal{X}, \mathcal{Y})$ en (2.6)

$$W_\infty(\mathcal{X}, \mathcal{Y}) = \inf_{\eta:\mathcal{X}\to\mathcal{Y}} \sup_{x\in\mathcal{X}} ||x - \eta(x)||_\infty. \tag{2.6}$$

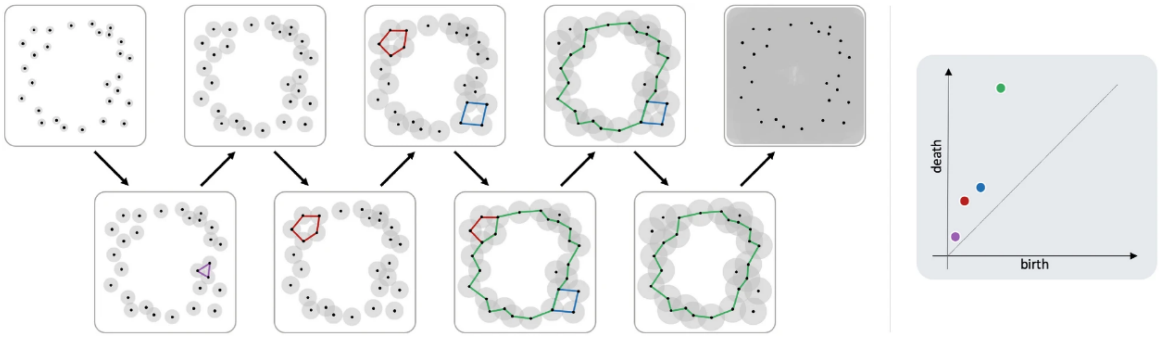


Figura 2.6: Construcción del diagrama de persistencia para datos provenientes de una circunferencia ruidosa. Se incluyen solamente las cualidades topológicas de primer grado, ignorando las componentes conexas. Cada ciclo se ilustra desde el ϵ de su nacimiento hasta su muerte con un color diferente, esta duración resultante es finalmente representada bajo el mismo color en el diagrama de persistencia final. Crédito de la imagen: (Bobrowski y Skraba 2023).

Cabe mencionar la posibilidad de que $\mathcal{X}$ e $\mathcal{Y}$ no cuenten con la misma cantidad de puntos, en cuyo caso, se considera que ambos diagramas de persistencia también poseen infinitos puntos en su diagonal, estos puntos se utilizarán en la medida que sean necesarios para formar la biyección $\eta : \mathcal{X} \to \mathcal{Y}$.

La distancia *bottleneck* se calcula como la mayor distancia entre dos pares de puntos correspondientes a la biyección que minimiza esta máxima distancia.

En la Figura 2.7 se ilustra el cálculo de esta distancia, en la que se dibujó un cuadrado de lado dos veces la distancia *bottleneck* centrado en cada uno de los puntos de $\mathcal{X}$, de forma que este cuadrado contenga a su par correspondiente en $\mathcal{Y}$

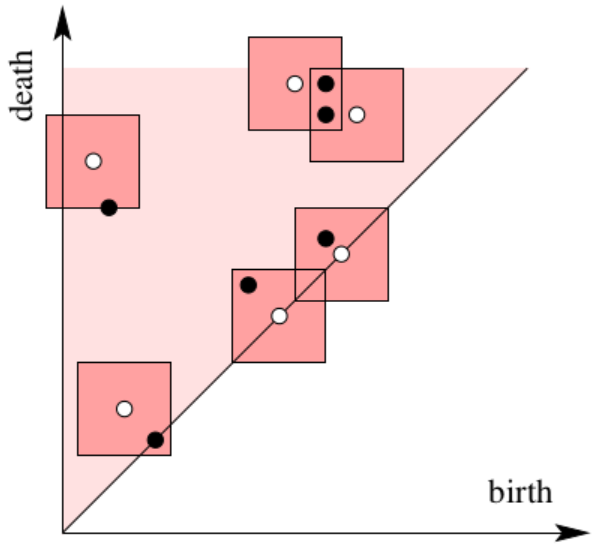


Figura 2.7: Superposición de dos diagramas de persistencia con puntos agregados en la diagonal para completar la biyección. La distancia *bottleneck* es equivalente a la mitad del lado de los cuadrados que ilustran la biyección. Crédito de la imagen: (Edelsbrunner y Harer 2010).

Resulta evidente que $W_\infty(\mathcal{X}, \mathcal{Y}) = 0$ solo sí $\mathcal{X} = \mathcal{Y}$, más aún, $W_\infty(\mathcal{X}, \mathcal{Y}) = W_\infty(\mathcal{Y}, \mathcal{X})$ y $W_\infty(\mathcal{X}, \mathcal{Z}) \leq W_\infty(\mathcal{X}, \mathcal{Y}) + W_\infty(\mathcal{Y}, \mathcal{Z})$; por lo que la distancia *bottleneck* cumple todos los axiomas de una métrica y merece ser llamada una distancia (Edelsbrunner y Harer 2010).

Una desventaja de la distancia *bottleneck* es su falta de sensibilidad a detalles de la biyección por fuera de la distancia entre puntos a máxima distancia de la misma. Para sobrellevar esta desventaja, se introduce la distancia de *Wasserstein* de grado q. Esta distancia reemplaza el cálculo del supremo en la distancia *bottleneck* mediante la suma de las q-ésima potencias de las distancias $||x - \eta(x)||_\infty$, nuevamente, minimizando para todas las biyecciónes posibles y finalmente tomando la raiz q-ésima, como se ilustra en (2.7)

$$W_q(\mathcal{Y}, \mathcal{X}) = \left(\inf_{\eta: \mathcal{X} \to \mathcal{Y}} \sum_{x \in \mathcal{X}} ||x - \eta(x)||_\infty^q \right)^{1/q}. \tag{2.7}$$

Como la similitud entre la notación de las ecuaciones (2.6) y (2.7) lo sugiere, la distancia *bottleneck* es el límite de la distancia *Wasserstein* para q tendiendo a infinito. Análogamente a la distancia *bottleneck*, es fácil verificar que W_q satisface los requerimientos de una métrica y también merece ser llamada distancia (Edelsbrunner y Harer 2010)

Estas distancias definidas son de suma importancia ya que se logra probar para cada una de ellas, mediante los teoremas de estabilidad, que las mismas están acotadas por la distancia real que existe entre los espacios topológicos subyacentes que dan lugar a los conjuntos de datos sobre los cuáles se calculan los diagramas de persistencia (Edelsbrunner y Harer 2010)

Complejos simpliciales

Como se mencionó previamente, calcular la homología $H(\mathcal{M})$ directamente a partir de $\mathcal{S}_N^\epsilon$ resulta una tarea difícil. Esto se realiza mediante la construcción del complejo simplicial a partir de $\mathcal{S}_N$. Un complejo simplicial es un conjunto de símplices, siendo estos generalizaciones de un triángulo en dimensiones arbitrarias, definidos al conectar puntos a menos distancia que ϵ. En particular, un complejo simplicial muy utilizado en topología computacional es el complejo de Čech. El complejo, denotado como $\check{C}ech(\mathcal{S}_N, \epsilon)$, representa el conjunto de símplices σ con vértices $\mathbf{v}_1, \cdots, \mathbf{v}_k \in \mathcal{S}_N$ tales que

$$\bigcap_{i=1}^{k} B(\mathbf{v}_i, \epsilon) \neq \emptyset.$$

Es decir, un símplice σ se incluye en el complejo de Čech solo sí la intersección entre todas las bolas de radio ϵ centradas en los vértices de σ es no nula. Evaluar la intersección de bolas resulta muy costoso computacionalmente (Dantchev y Ivrissimtzis 2012), en especial a medida que la dimensión del espacio crece, ya que para realizar esto no basta con evaluar las distancias de a pares entre los vértices y se requiere realizar operaciones mas complejas.

Este complejo resulta muy importante ya que el Teorema del Nervio garantiza que $\mathcal{S}_N^\epsilon$ y $\check{C}ech(\mathcal{S}_N, \epsilon)$ son homotópicamente equivalentes (Fasy et al. 2014; Bauer et al. 2023), es decir, comparten la misma homología.

Por cuestiones de eficiencia computacional, es común aproximar el complejo de Čech con el complejo de Vietoris-Rips (Fasy et al. 2014), denotado como $V(\mathcal{S}_N, \epsilon)$, que consiste en los símplices con vértices en $\mathcal{S}_N$ de diámetro máximo 2ϵ. En otras palabras, el símplice σ es incluido en el complejo si cada par de vértices en σ está separado como máximo a distancia 2ϵ. Este complejo resulta de cálculo más eficiente que el de Čech ya que solo las distancias de a pares entre los puntos son necesarias, y el mismo cumple la siguiente igualdad (Fasy et al. 2014):

$$\check{C}ech(\mathcal{S}_N, \epsilon) \subset V(\mathcal{S}_N, \epsilon) \subset \check{C}ech(\mathcal{S}_N, \sqrt{2}\epsilon).$$

Por lo que si se evalúa para un rango de valores de ϵ, las conclusiones a las que se llegará en cuanto a la homología de $\mathcal{S}_N$ serán equivalentes. En particular, el diagrama de persistencia será análogo. En la Figura 2.8 se ilustra la diferencia entre estos dos complejos simpliciales.

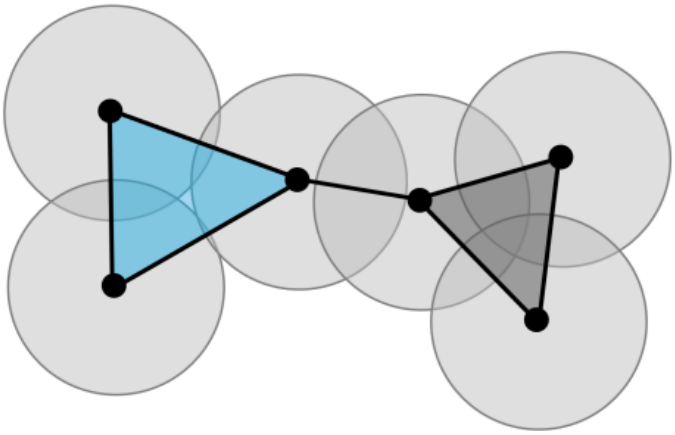


Figura 2.8: Conjunto de datos de 6 puntos sobre el cual se construyen los complejos de Čech ($\check{C}ech(\mathcal{S}_N, \epsilon)$) y Vietoris-Rips ($V(\mathcal{S}_N, \epsilon)$). El símplice gris corresponde a ambos complejos, ya que la intersección entre las bolas de los tres puntos que lo conforman es no nula, por su parte, el símplice celeste solo pertenece a $V(\mathcal{S}_N, \epsilon)$, ya que las bolas que lo conforman tienen intersecciones no nulas de a pares (la distancia entre ellos es menor a 2ϵ), pero no en conjunto. Crédito de la imagen: (Bobrowski y Krioukov 2022).

Diagramas de persistencia de conjuntos de nivel

Hasta el momento se discutió cómo estimar la homología de $\mathcal{M}$ a partir de la unión de bolas $B(\mathbf{x}_i, \epsilon)$. Como se demuestra que esta última conserva la homología del espacio original para un ϵ desconocido, él mismo se varía construyendo el diagrama de persistencia. Este proceso fue ilustrado en la Figura 2.5. Una observación que puede hacerse de este procedimiento es que la unión de bolas $B(\mathbf{x}_i, \epsilon)$ puede definirse, alternativamente, como los conjuntos de nivel inferiores de una función $f(\mathbf{x})$, siendo esta el mínimo de la función de distancia entre el punto $\mathbf{x}$ y los datos de la muestra, es decir:

$$L_\epsilon = \cup_{\mathbf{x}\in\mathcal{S}_N} B(\mathbf{x}, \epsilon) = \{\mathbf{x} \mid \min_{\mathbf{x}_i\in\mathcal{S}_N} d(\mathbf{x}_i, \mathbf{x}) < \epsilon\}.$$

Esta observación, que se ilustra en la Figura 2.9, abre las puertas a que el diagrama de persistencia pueda expresarse en términos más generales reemplazando la función $\min_{\mathbf{x}_i\in\mathcal{S}_N} d(\mathbf{x}_i, \mathbf{x})$ por alguna otra función arbitraria $g(\mathbf{x})$, cuyos conjuntos de nivel serán utilizados para construir el diagrama de persistencia y así estimar la homología de $\mathcal{M}$. En la Figura 2.10 se observa un ejemplo de este procedimiento para una función escalar arbitraria. Para algunas funciones g, como por ejemplo la correspondiente a la estimación de densidad de probabilidad por ventanas (Fasy et al. 2014), la homología reproducida conserva las cualidades topológicas del espacio original pero con ventajas adicionales como pueden ser mayor robustez ante datos atípicos o ruido. Es por esto que la inferencia de las cualidades topológicas del espacio original $\mathcal{M}$ puede realizarse directamente sobre un diagrama de persistencia construido con los conjuntos de nivel de alguna otra función $g(x)$. En términos prácticos y computacionales, este diagrama de persistencia se construye a partir de evaluar g sobre una grilla de puntos dentro del espacio a analizar, y una vez obtenidos estos valores realizar una interpolación para construir el diagrama (Fasy et al. 2014; Dlotko y Wanner 2018; The GUDHI Project 2015).

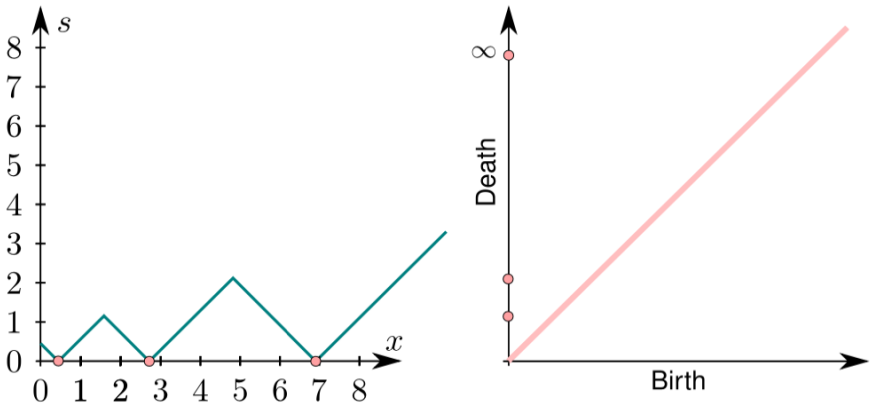


Figura 2.9: Diagrama de persistencia para una muestra de tres puntos a partir de la función de mínima distancia a algún elemento del conjunto (denotada como $s = min_{x_i \in \mathcal{S}_N} d(x_i, x)$ en la figura). Se observa que comenzando con $s < 0$, se tienen tres componentes centradas en los puntos (es por eso que en el diagrama de persistencia todas las componentes comienza en cero), esto cambia cuando se alcanza el conjunto de nivel $s \leq 1$, donde las bolas de las dos primeras muestras se unirían, bajo este enfoque se muestra que la función deja por debajo el primer triángulo, uniendo los soportes para los cuales esas dos muestras estaban a menos de uno de distancia. Finalmente, lo mismo sucede al alcanzar $s \leq 2$, donde todos los conjuntos de nivel se unen, formándose así la última componente conexa que se mantiene hasta ∞. Crédito de la imagen: (Fasy et al. 2014).

2.3 Intervalos de confianza

En estadística frecuentista, un intervalo de confianza es un rango estimado para un parámetro desconocido. El intervalo de confianza se calcula a un nivel designado, siendo el nivel de confianza de 95% el más utilizado. El nivel de confianza representa la proporción teórica de intervalos de confianza que contiene al verdadero valor del parámetro. Por ejemplo, de todos los intervalos computados al nivel 95% para diferentes conjuntos de datos del mismo tamaño obtenidos a partir de la misma distribución, el 95% de ellos deberían contener al verdadero valor (Illowsky y Dean 2017)

Formalmente, sea X una muestra aleatoria proveniente de una distribución de probabilidad p de parámetro unidimensional θ, siendo θ la cantidad a estimar. Un intervalo de confianza de nivel $1 - \alpha$ para el parámetro θ es el intervalo $(\theta_L(X), \theta_U(X))$ determinado por las variables aleatorias $\theta_L(X)$ y $\theta_U(X)$ tales que la siguiente igualdad se cumple para todo θ:

$$\mathbb{P}\{\theta_L(X) \leq \theta \leq \theta_u(X)\} = 1 - \alpha. \tag{2.8}$$

Para el caso de parámetros de dimensionalidad mayor, es decir $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$, se habla de "regiones de confianza", siendo la intuición de estas equivalente a la del caso unidimensional, pero admitiendo regiones $A(X)$ de cualquier forma, es decir $\mathbb{P}\{\boldsymbol{\theta} \in A(X)\} = 1 - \alpha$.

2.3.1 Relación entre intervalos de confianza y pruebas de hipótesis

Resulta importante para el desarrollo de esta Tesis analizar la relación que existe entre test de hipótesis e intervalos de confianza, ya que utilizaremos los intervalos de confianza construidos sobre diagramas de persistencia para evaluar, mediante pruebas de hipótesis, la existencia de cualidades topológicas. Recordando la definición de intervalo de confianza dada por (2.8), imaginemos que queremos verificar que nuestro parámetro es distinto a algún valor θ_0, como por ejemplo $\theta_0 = 0$, resulta intuitivo pensar que podemos calcular un intervalo de confianza de nivel $1 - \alpha$ para este parámetro y verificar si el mismo contiene o no a nuestro valor de interés θ_0. En caso de no contenerlo, y recordando que, por definición, el intervalo de confianza calculado con el estadístico de la muestra tiene una probabilidad de

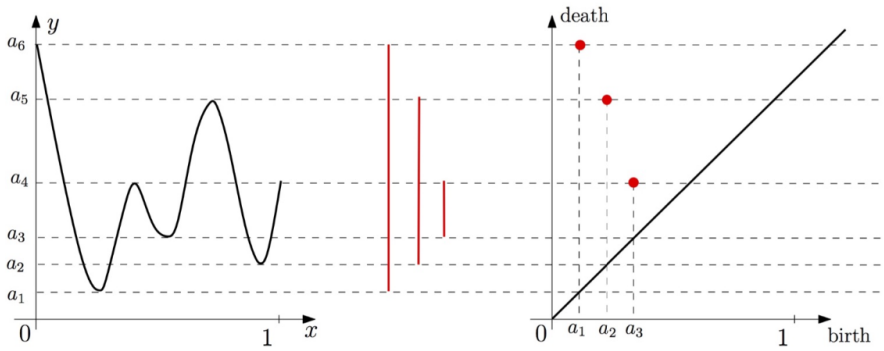


Figura 2.10: Diagrama de persistencia de los conjuntos de nivel de una función arbitraria $y(x)$. Se observa que, comenzando en $y = 0$, el conjunto de nivel $y < 0$ es nulo, esto cambia cuando se llega al valor $y < a_1$, en donde el primer intervalo comienza a aparecer. En los valores $y < a_2$ e $y < a_3$ aparecen nuevos intervalos en los mínimos locales de la función. Cuando se alcanza $y < a_4$ dos de estos intervalos se unen, y al alcanzarse $y < a_5$ este último se junta con el iniciado en $y < a_1$ que termina en $y < a_6$. De esta forma, para las cualidades topológicas resultantes, se obtienen los tiempos de nacimiento $b_i = a_1, a_2, a_3$ y sus respectivos tiempos de muerte $d_i = a_6, a_5, a_4$, siendo estos los puntos que se grafican en el diagrama de persistencia. Crédito de la imagen: (Chazal y Michel 2021).

$(1 - \alpha)$ de contener a nuestro parámetro real, podemos descartar la hipótesis $H_0 : \theta = \theta_0$ con un nivel de confianza $1 - \alpha$.

2.3.2 Cálculo de intervalos de confianza

En algunas situaciones controladas los intervalos de confianza pueden calcularse de forma analítica con fórmula cerrada, como es el típico caso del intervalo de confianza para la media de una distribución gaussiana. En otros casos, estos intervalos resultan de muy difícil o incluso imposible cálculo, al no tener expresión cerrada. Es por esto que surgen estrategias computacionales de simulación para calcularlos, de las cuales destacan aquellas que utilizan remuestreo para obtener nuevos conjuntos de datos y estimar así los intervalos de confianza. Dos de las estrategias más conocidas de remuestreo son el *Bootstrap* y el sub-muestreo, que serán utilizadas para estimar nuestros intervalos de confianza de cualidades topológicas. A continuación se describen ambas técnicas.

Bootstrap

El *Bootstrap* es una técnica que busca hacer inferencia sobre una población a partir de una muestra de la misma. La principal idea es que es posible obtener información de la población mediante remuestreo de la muestra original. Dado que la población es desconocida, el verdadero error del estadístico de la muestra contra el valor de la población es desconocido, pero en los remuestreos *bootstrap* la "población" es de hecho la muestra, que si es conocida, por lo que la calidad de la inferencia sobre la muestra a partir del remuestreo es medible

Más formalmente, *bootstrap* trata al problema de inferencia sobre la verdadera distribución de probabilidad P dada la muestra original como análogo al problema de inferencia sobre la distribución empírica $\hat{P}$ a partir de la muestras *bootstrap*, es decir, remuestreo con reposición de igual tamaño que la muestra original (Efron y Tibshirani 1994; Wikipedia 2024). La calidad de la inferencia respecto a $\hat{P}$ puede ser evaluada ya que esta distribución es conocida, y si $\hat{P}$ es una aproximación razonable de P, entonces la calidad de inferencia sobre la misma distribución P puede ser inferida. A fines prácticos, dada una muestra i.i.d $\mathcal{S}_N = \{x_1, \ldots, x_N\}$ se obtienen M muestras *bootstrap* $\mathcal{S}_N^j$, $j = 0, \ldots, M$ a partir del remuestreo con reposición de tamaño N sobre $\mathcal{S}_N$. A partir de estas muestras *bootstrap*, la distribución del estadístico $\hat{\theta} = T(\mathcal{S}_N)$ puede ser inferida analizando la distribución empírica de $\hat{\theta}^j = T(\mathcal{S}_N^j)$

Sub-muestreo

Otra técnica proveniente del remuestreo, que es utilizada como alternativa al *bootstrap*, es el sub-muestro. Esta técnica se basa en utilizar subconjuntos de la muestra original para evaluar el estimador en cuestión y obtener una noción de variabilidad del mismo (Babu 1992). Formalmente, dada una muestra $\mathcal{S}_N = \{x_1, \ldots, x_N\}$, el sub-muestreo consiste en obtener muestras aleatorias, sin reposición, de tamaño b de $\mathcal{S}_N$, con $b < N$. Es decir, obtener aleatoriamente secuencias de los $\binom{N}{b}$ conjuntos posibles que se forman al tomar b elementos de $\mathcal{S}_N$, a la que llamaremos $\mathcal{S}^j_{N,b},\ j = 0, \ldots, M$ y calcular el estadístico a partir de estas sub-muestras $\hat{\theta}^j_b = T(\mathcal{S}^j_{N,b})$. Al repetir este procedimiento una cantidad M lo suficientemente grande podemos aproximar la distribución del estadístico, y en particular, un intervalo de confianza para el mismo al tomar los percentiles de la distribución empírica de $\hat{\theta}^j_b$, análogamente a lo realizado para el *bootstrap*.

Este método resulta consistente en casos en los que el *bootstrap* no lo es (Babu 1992), pero introduce la complejidad de la elección de el tamaño de sub-muestras b.

2.4 Regiones de confianza para diagramas de persistencia

Resulta natural preguntarse cómo se aplican las regiones de confianza, desarrollados en la Sección 2.3, a los diagramas de persistencia introducidos en la Sección 2.2.3. En este caso, no es un parámetro de una distribución sobre el cual se busca obtener una estimación de la región de confianza, sino de las cualidades topológicas de la variedad $\mathcal{M}$ del que provienen nuestros datos. Para lograr esto se busca, dado un nivel $1 - \alpha$, encontrar el estadístico θ_n tal que

$$\lim_{n\to\infty} \inf \mathbb{P}(0 \leq W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \leq \theta_n) \geq 1 - \alpha, \tag{2.9}$$

con $\mathcal{P}$ el diagrama de persistencia de la variedad $\mathcal{M}$, definido a partir de los conjuntos de nivel $\{\mathbf{x} \mid d_{\mathcal{M}}(\mathbf{x}) < \epsilon\}$ de la función de distancia a la variedad, expresada como:

$$d_{\mathcal{M}}(\mathbf{x}) = \inf_{\mathbf{y}\in\mathcal{M}} ||\mathbf{x} - \mathbf{y}||,$$

siendo $\hat{\mathcal{P}}$ el diagrama de persistencia construido con el conjunto de datos $\mathcal{S}_N$ que busca estimar $\mathcal{P}$.

La región de confianza resultante puede visualizarse centrando una caja de lado $2\theta_n$ en cada punto $p = (b_i, d_i)$ del diagrama de persistencia. El punto en cuestión es considerado indistinguible de ruido si la caja correspondiente, definida como $\{q \in \mathbb{R}^2 : ||q - p||_\infty \leq \theta_n\}$, interseca con la diagonal. Alternativamente, la región de confianza puede visualizarse añadiendo una banda de ancho $\sqrt{2}\theta_n$ alrededor de la diagonal del diagrama de persistencia. La interpretación de esta banda sería la siguiente: puntos dentro de la banda no son significativamente diferentes de ruido, puntos por fuera de la banda representan una característica topológica significativa. Dicho en otras palabras, si la región de confianza de un punto toca la diagonal, no podemos descartar que el tiempo de vida de esa cualidad topológica sea nulo, por lo que consideramos que es en realidad ruido. En la Figura 2.11 se ilustran las regiones de confianza de ambas formas.

Es posible entonces construir pruebas de hipótesis para diagramas de persistencia de la siguiente forma: asumamos que estamos interesados en probar, con un nivel de significancia $1-\alpha$, que una dada cualidad topológica está presente en nuestra variedad $\mathcal{M}$, de la cual provienen los datos. Para lograr esto, basta con evaluar si la componente $p_i = (b_i, d_i)$ del diagrama de persistencia $\hat{\mathcal{P}}$ es significativa. Esto puede expresarse como la siguiente hipótesis nula:

$$H^i_0 : l_i = d_i - b_i = 0,$$

que se buscará rechazar en pos de aceptar la hipótesis alternativa, dada por

$$H_1^i : l_i > 0.$$

Este *test* se realiza en simultáneo para todas las cualidades topológicas del diagrama $\hat{\mathcal{P}}$ con un nivel $1-\alpha$. Continuando el razonamiento ilustrado en la Figura 2.11, si una cualidad topológica p_i se encuentra a una distancia menor a $\sqrt{2}\theta_n$ de la diagonal, entonces la hipótesis nula H_0^i para esa cualidad topológica no puede ser rechazada. Otra forma de entender este razonamiento es la siguiente: todos los posibles diagramas de persistencia $\tilde{\mathcal{P}}$ que están dentro de nuestra región de confianza de nivel $1-\alpha$ se denotan como

$$\mathcal{C}_n = \{\tilde{\mathcal{P}} : W_\infty(\tilde{\mathcal{P}}, \hat{\mathcal{P}}) < \theta_n\}, \tag{2.10}$$

por lo que podemos mover cada uno de los puntos p_i del diagrama de persistencia dentro de la caja de lado $2\theta_n$ centrada en p_i sin salirnos de esa región de confianza. Si al realizar este procedimiento no podemos llevar al punto p_i a la diagonal, esto significa que la cualidad topológica que representa p_i no puede ser interpretada como ruido bajo un *test* de nivel $1-\alpha$, ya que no existe un $\tilde{\mathcal{P}}$ que no posea esa cualidad. Entonces, si bien la región de confianza se construye para todo el diagrama de persistencia, es posible extraer información de cada una de las cualidades topológicas p_i individualmente de esta forma.

Resulta importante destacar que la región de confianza resultante de (2.9) es asintótica, por lo que en términos prácticos los resultados obtenidos mantendrán este nivel estadístico sólo para muestras n suficientemente grandes. El valor de n necesario dependerá de la homología subyacente y de la función de distancia utilizada. Se espera que al utilizar la distancia de Fermat se obtengan resultados más consistentes con la verdadera homología de la variedad $\mathcal{M}$ para menores n.

Una observación relevante que se realiza en (Fasy et al. 2014) sobre este procedimiento es que esta forma dicotómica de clasificar las componentes topológicas en "señal" o "ruido" no es la única posible, ya que en realidad la región de confianza construida da lugar, para un tamaño de muestra n, a un conjunto de variedades con diagramas de persistencia, expresados en (2.10). Esta definición de $\mathcal{C}_n$ da lugar a formas mucho más elaboradas de cuantificar la incertidumbre del diagrama de persistencia $\hat{\mathcal{P}}$ construido.

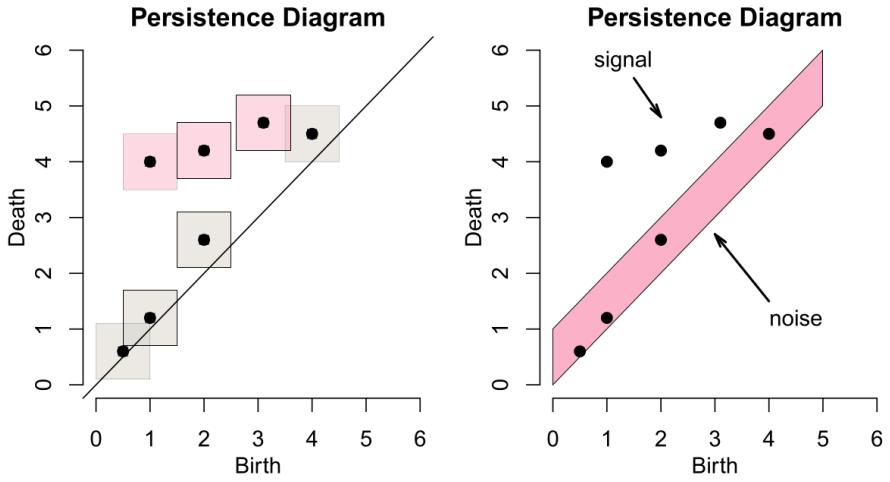


Figura 2.11: A la izquierda se observan los puntos del diagrama de persistencia en donde una caja de lado $2\theta_n$ se centra en cada uno de ellos, representando la región de confianza para la cualidad topológica representada por ese punto. En color gris, se observan aquellas cualidades cuya región de confianza se interseca con la diagonal, ilustrando que estos puntos no son distinguibles de ruido, por otro lado, en color rojo se ilustran las regiones que no tienen contacto con la diagonal, siendo estas cualidades topológicas significativas. A la derecha se ilustra otra forma de interpretar estas regiones, posando una banda de tamaño $\sqrt{2}\theta_n$ sobre la diagonal que cubre toda la región del diagrama de persistencia en la que un punto presente en ella tendría su región de confianza intersecando la diagonal. Se observa que todos los puntos con caja gris en la figura de la izquierda aparecen dentro de la banda roja en la figura derecha. Créditos de la imagen: (Fasy et al. 2014).

Capítulo 3

Métodos y Desarrollo

En esta sección abordaremos los distintos métodos computacionales y estadísticos que utilizaremos para obtener los resultados de este trabajo. Asimismo, se presentarán los conjuntos de datos sobre los cuales estos métodos serán utilizados.

3.1 Conjuntos de datos sintéticos

Es común en la literatura de topología computacional enfocarse en conjuntos de datos sintéticos simples, sobre los cuales se corren los métodos de interés. Al ser estos utilizados por diferentes trabajos los resultados pueden ser fácilmente contrastados (Fasy et al. 2014; Chazal et al. 2017). A continuación se muestran los conjuntos de datos sintéticos que se utilizarán durante la Capítulo 4. En cada caso, se generarán dos nuevos conjuntos de datos por medio de agregar datos atípicos y ruido, respectivamente, para lograr una mayor variabilidad en los conjuntos que puedan dificultar la tarea a las técnicas desarrolladas.

3.1.1 Circunferencia uniforme

El más sencillo de los conjuntos de datos a utilizar es la circunferencia uniforme, en donde la distancia al centro de los puntos es fija mientras que el ángulo proviene de una distribución uniforme en el intervalo $(-\pi, \pi)$, es decir

$$\angle \mathbf{x} \sim \mathcal{U}(-\pi, \pi), \qquad |\mathbf{x}| = r.$$

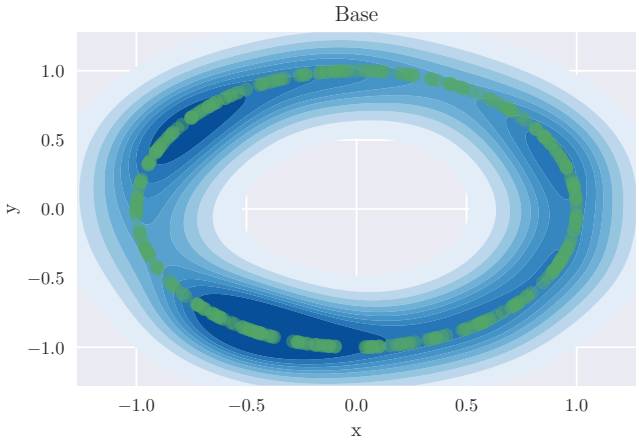
En la Figura 3.1 se observa una muestra de 300 elementos provenientes de esta distribución, junto con sus versiones con ruido agregado y datos atípicos, respectivamente. Se incluyen además los conjuntos de nivel de la función de densidad estimada sobre la muestra.

3.1.2 Circunferencia gaussiana

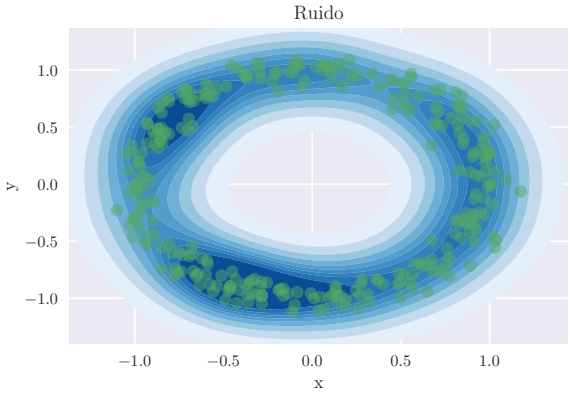
Una variación utilizada de la circunferencia uniforme, que busca construir regiones con menor densidad de datos aunque manteniendo el soporte del cual estas mismas se obtienen, es la circunferencia gaussiana. La definición de este conjunto de datos es análoga a la del caso uniforme, pero en este caso el ángulo $\angle \mathbf{x}$ se obtiene de una distribución normal truncada entre $-\pi$ y π, es decir

$$\angle \mathbf{x} \sim \text{TruncNormal}(-\pi, \pi, \mu, \sigma),$$

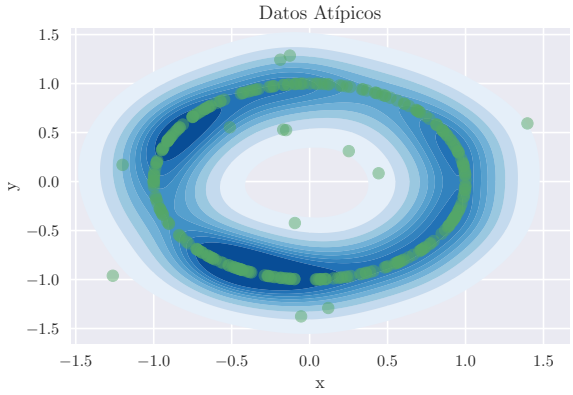
donde la media μ será igual a 0, mientras que la desviación estándar, σ, se elige de tal forma que el 95% de las muestras de una gaussiana tradicional estén comprendidas dentro del intervalo $(-\pi, \pi)$. En



(a) Muestreo uniforme



(b) Ruido agregado



(c) Datos atípicos

Figura 3.1: Conjunto de datos sintéticos correspondientes a la circunferencia uniforme.

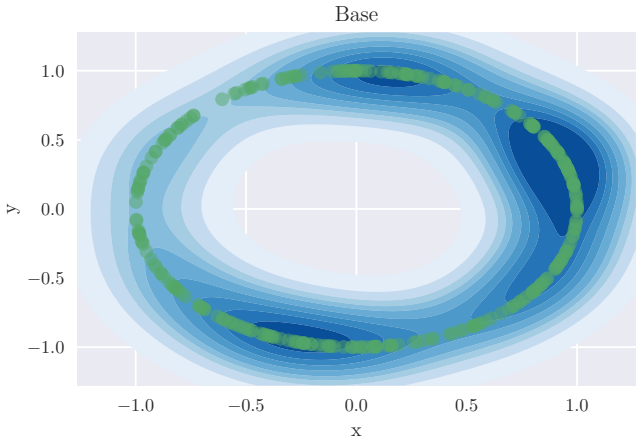
la Figura 3.2 se observa el conjunto de datos provenientes de esta distribución junto con sus análogos con ruido y datos atípicos.

3.1.3 Anteojos

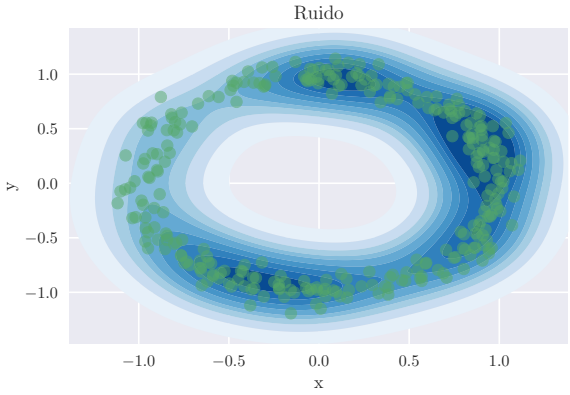
El conjunto de datos que llamaremos Anteojos (*eyeglasses* en inglés) se ilustra en la Figura 3.3, junto con sus análogos con ruido gaussiano y datos atípicos. Esta topología resulta similar a un óvalo de Cassini y debe su nombre a su parecido con unos anteojos. La misma presenta dos círculos conectados mediante una apertura en ellos. Se utiliza extensivamente durante el cálculo de regiones de confianza para datos sintéticos (Fasy et al. 2014), ya que la topología en realidad posee un único agujero de grado uno, pero resulta complejo para los algoritmos detectar esta topología, usualmente resultando en dos agujeros significativos, dependiendo de la cantidad de datos y el grado de apertura.

3.1.4 Círculo con densidad dependiente del radio

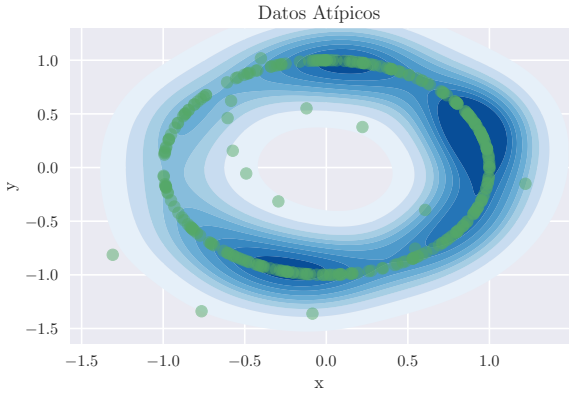
Hasta el momento todos los conjuntos de datos presentados poseen un agujero, pero resulta de interés analizar cómo se comportan los diagramas de persistencia y las respectivas regiones de confianza calculados sobre los mismos para un conjunto de datos que no posee agujeros de grado 1. El objetivo será entonces utilizar los datos en el círculo con densidad dependiente del radio para validar la capacidad de las regiones de confianza obtenidas, mediante cada uno de los métodos, para rechazar la hipótesis de existencia de agujeros, verificando así la tasa de errores tipo I o nivel estadístico.



(a) Muestreo uniforme



(b) Ruido agregado



(c) Datos atípicos

Figura 3.2: Conjunto de datos sintéticos correspondientes a la circunferencia gaussiana.

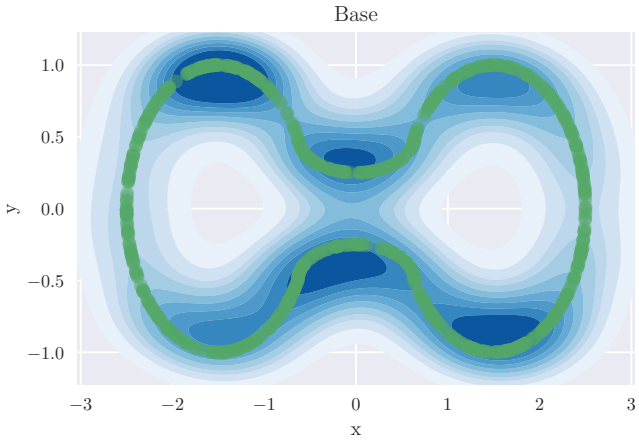
El conjunto de datos que sumaremos a nuestras simulaciones para realizar esta tarea será un círculo en donde la densidad de los puntos que lo rellenan es dependiente del radio. Es decir, sea un punto en dos dimensiones parametrizado por su distancia al origen, r, y su ángulo θ con respecto al eje x, se obtienen muestras bajo la siguiente distribución

$$\theta \sim \mathcal{U}(0, 2\pi), \quad r \sim r_{max} \cdot \mathcal{U}(0,1)^{\frac{1}{r_{power}}}.$$

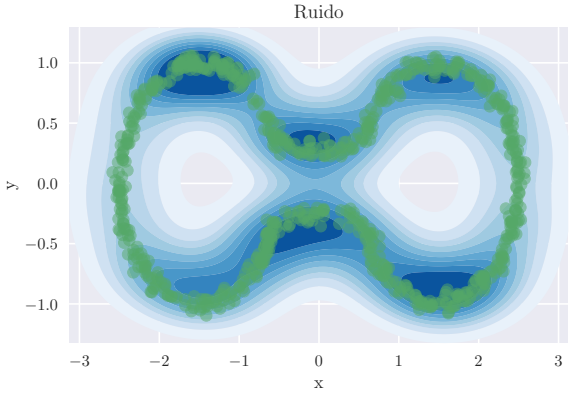
Resulta importante observar que el área de una corona circular crece cuadráticamente con el radio, por lo que si queremos obtener un muestreo uniforme en la superficie del círculo debemos establecer $r_{power} = 2$. De esta forma la distribución de r será tal que se obtenga más concentración en los radios superiores, de forma exacta para compensar el crecimiento cuadrático del área a medida que se aumenta el radio. Siguiendo este razonamiento para valores mayores, es decir $r_{power} > 2$, se obtiene una concentración más fuerte de muestras en los radios superiores, mientras que para $r_{power} < 2$ se obtienen más muestras cercanas al centro. Esto se debe a que al tomar una potencia cada vez menor (a medida que r_{power} crece en el denominador) de un valor entre cero y uno (los obtenidos mediante la distribución $\mathcal{U}(0,1)$) el resultado es cada vez más cercano a uno. Finalmente, r_{max} solo tiene el objetivo de escalar los límites de la distribución de r, obteniendo así muestras para radios superiores a uno.

Será de especial interés analizar los casos de $r_{power} > 2$ ya que se espera que los métodos trabajados logren descartar la existencia de agujeros a pesar de la baja en la densidad de muestreo cerca del origen.

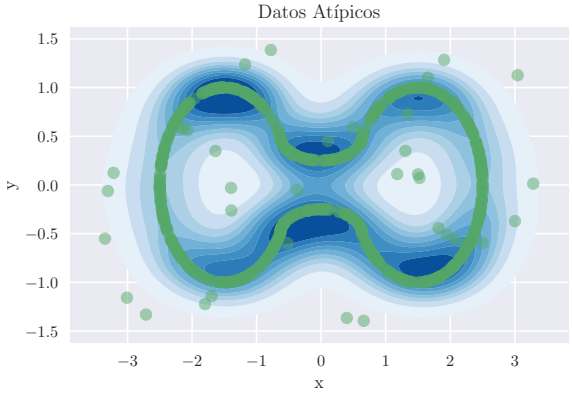
En la Figura 3.4 se ilustra el conjunto de datos para los valores $r_{max} = 1.5$ y $r_{power} = 2.6$



(a) Muestreo uniforme



(b) Ruido agregado



(c) Datos atípicos

Figura 3.3: Conjunto de datos sintéticos correspondientes a los anteojos.

3.2 Conjuntos de datos reales

Adicionalmente al uso de los conjuntos de datos sintéticos descritos en la Sección 3.1 se buscará también validar los resultados obtenidos en cuanto a la detección de agujeros de primer grado en un conjunto de datos reales utilizado en (Chazal et al. 2017), cuya estructura se describe en (Pettersen et al. 2014). El *dataset* consiste en mediciones correspondientes a la posición de jugadores de fútbol dentro de la cancha a lo largo de un partido, en el que se obtiene un punto cada un intervalo de tiempo determinado. A este conjunto de datos, siguiendo lo realizado en (Chazal et al. 2017), se le agrega un borde artificial demarcando los límites de la cancha con el objetivo de que los espacios en donde el jugador no tuvo incidencia sean agujeros con bordes en los límites de la cancha y las posibles ubicaciones del jugador en el conjunto.

En la Figura 3.5 se observan los datos para cuatro jugadores correspondientes a diferentes posiciones de juego en los que se evidencian espacios donde el jugador no tuvo incidencia, o al menos de forma recurrente.

3.3 Regiones de confianza

Haciendo uso de los conjuntos de datos descritos en la Sección 3.1, calcularemos las regiones de confianza para los diagramas de persistencia resultantes empleando tres métodos. Dos de ellos propuestos por la literatura, y analizados extensivamente en (Fasy et al. 2014), y un método adicional, que consiste en reemplazar la distancia euclídea utilizada en uno de ellos por la distancia de Fermat. El objetivo es

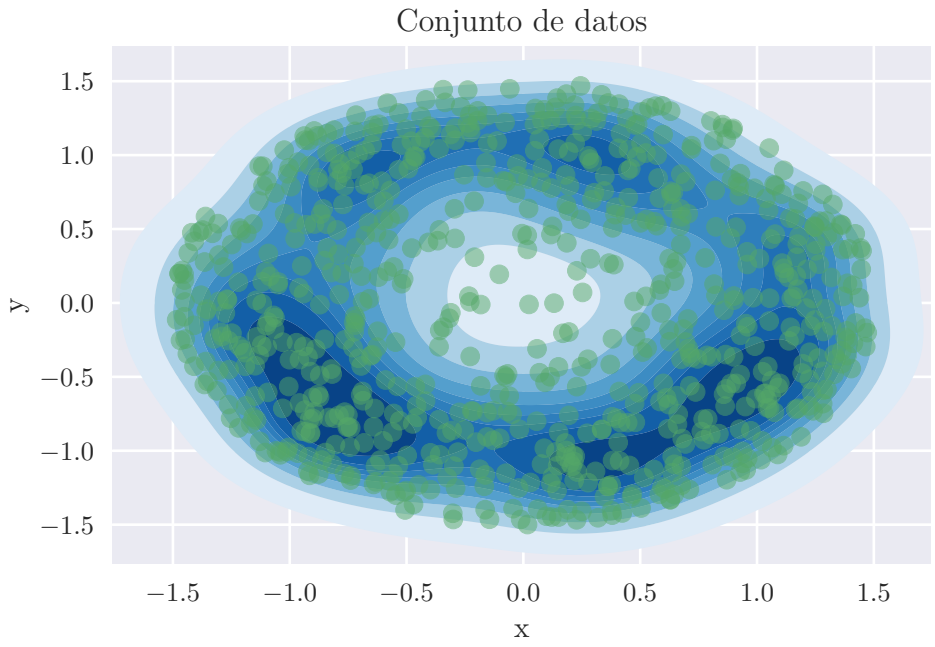


Figura 3.4: Conjunto de datos sintéticos correspondientes al círculo con densidad dependiente del radio.

observar las diferencias obtenidas en las regiones de confianza y las cualidades topológicas detectadas. A continuación se describen los métodos a utilizar para calcular estas regiones de confianza.

3.3.1 Sub-muestreo con distancia euclídea

El primer método de cálculo de regiones de confianza para diagramas de persistencia será mediante el sub-muestreo con distancia euclídea, desarrollado en (Fasy et al. 2014). Este método consiste en usar la técnica de submuestreo, introducida de forma teórica en la Sección 2.3.2, para obtener sub-muestras $\mathcal{S}_{N,b}^j$ del conjunto original mediante las cuales calcular directamente la distancia de Hausdorff basada en una distancia euclídea. Es decir, sea $\mathcal{S}_{N,b}^j$ una sub-muestra sin reposición de tamaño b sobre $\mathcal{S}_N$ y sea $\theta_j = d_H(\mathcal{S}_N, \mathcal{S}_{N,b}^j)$ la distancia de Hausdorff entre $\mathcal{S}_N$ y $\mathcal{S}_{N,b}^j$, definimos

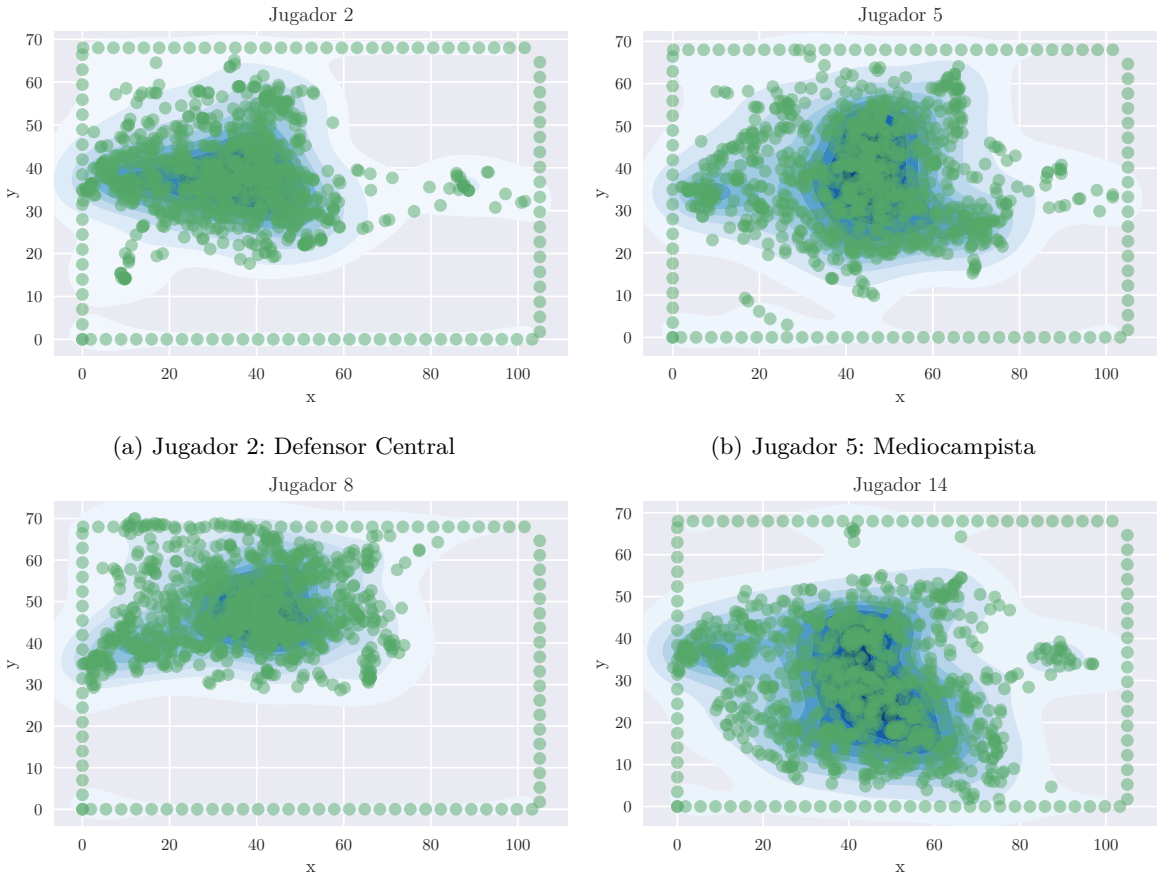
$$L_b(t) = \frac{1}{M} \sum_{j=1}^{M} I(\theta_j > t),$$

con $I(.)$ la función indicadora. Si definimos $c_b = 2L_b^{-1}(\alpha)$ entonces puede demostrarse el siguiente resultado (Fasy et al. 2014)

$$\mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > c_b) \leq \mathbb{P}(d_H(\mathcal{S}_N, \mathcal{M}) > c_b) \leq \alpha + \mathcal{O}\left(\frac{b}{N}\right)^{1/4},$$

es decir, la caja de lado $2c_b$ centrada en los puntos del diagrama de persistencia $\hat{\mathcal{P}}$ es una región de confianza de nivel asintótico $1 - \alpha$ para el diagrama de persistencia $\mathcal{P}$.

Haremos uso del siguiente algoritmo (descrito en *pseudo*-código) para obtener nuestra primer región de confianza para los diagramas de persistencia calculados:



(a) Jugador 2: Defensor Central

(b) Jugador 5: Mediocampista

(c) Jugador 8: Lateral Izquierdo

(d) Jugador 14: Mediocampista

Figura 3.5: Conjunto de datos reales correspondientes a la posición medida de distintos jugadores a lo largo de un partido de fútbol. Los diferentes jugadores ocupan diferentes roles en el equipo para mostrar diferentes patrones de posiciones.

Input: $\mathcal{S}_N,\ b,\ \alpha,\ M$
$j \leftarrow 1$;
$\theta \leftarrow array(M)$;
while $j \leq M$ **do**
  $\mathcal{S}^j_{N,b}$: Sub-muestro sin reposición de $\mathcal{S}_N$;
  $\theta[j] \leftarrow d_H(\mathcal{S}_N, \mathcal{S}^j_{N,b})$;
end
return 2 * quantile(θ, $1-\alpha$)

3.3.2 Sub-muestreo con distancia de Fermat

El método descrito en la Sección 3.3.1 hace uso de la distancia de Hausdorff para conseguir una cota superior de $\mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > c_b)$, pero resulta importante destacar que esta distancia se define a partir de una medida de distancia subyacente, que se evidencia en la definición de la distancia de Hausdorff para nubes de puntos

$$d_H(\mathcal{S}_N, \mathcal{S}^k_{N,b}) = \max\Big(\max_{\mathbf{x_i}\in\mathcal{S}_N} \min_{\mathbf{x_j}\in\mathcal{S}^k_{N,b}} d(\mathbf{x_i}, \mathbf{x_j}),\ \max_{\mathbf{x_i}\in\mathcal{S}^k_{N,b}} \min_{\mathbf{x_j}\in\mathcal{S}_N} d(\mathbf{x_i}, \mathbf{x_j})\Big). \tag{3.1}$$

Esta expresión puede reducirse teniendo en cuenta que, al ser $\mathcal{S}^k_{N,b}$ un subconjunto de $\mathcal{S}_N$, para todo

$\mathbf{x_i} \in \mathcal{S}^k_{N,b}$ se tiene que $\min_{\mathbf{x_j} \in \mathcal{S}_N} d(\mathbf{x_i}, \mathbf{x_j}) = 0$, ya que el mínimo se obtiene en exactamente la misma muestra, por lo que (3.1) se reduce a

$$d_H(\mathcal{S}_N, \mathcal{S}^k_{N,b}) = \max_{\mathbf{x_i} \in \mathcal{S}_N} \min_{\mathbf{x_j} \in \mathcal{S}^k_{N,b}} d(\mathbf{x_i}, \mathbf{x_j}),$$

donde la maximización se alcanzará en alguno de aquellos puntos que no hayan sido sub-muestreados. A partir de esta descripción de la distancia de Hausdorff resulta evidente que si se reemplaza la función de distancia euclídea d, utilizada en la Sección 3.3.1, por la distancia de Fermat, todas las garantías teóricas obtenidas en (Fasy et al. 2014) seguirán siendo válidas ya que las mismas no dependen de la función elegida para este cómputo. De esta forma, el algoritmo desarrollado para el método de la Sección 3.3.1 no debe ser modificado más allá del cambio en la medida de distancia.

En la práctica, tanto para la distancia de Fermat como para la distancia euclídea de la sección precedente, la distancia de Hausdorff definida en (3.1) se modifica con el objetivo de hacerla más robusta a conjuntos reales y posibles muestras atípicas. Siguiendo el mecanismo propuesto en (Dang-Nguyen y Do-Hong 2019), se reemplazan los "máximos" del cómputo por una versión podada de los mismos, es decir, un percentil apropiado que en la práctica se busca como el máximo percentil que logra que los resultados no cambien abruptamente al variar el parámetro. Es decir, si al modificar un 1% este percentil el resultado obtenido cambia drásticamente, significa que el método está siendo efectivo en eliminar distancias atípicas y aún tiene margen de mejora, por lo que debemos seguir buscando el valor apropiado. A partir de esta observación, la distancia de Hausdorff que utilizaremos para nubes de puntos resulta:

$$d_H(\mathcal{S}_N, \mathcal{S}^k_{N,b}) = \text{Percentil}^{\gamma}_{\mathbf{x_i} \in \mathcal{S}_N} \min_{\mathbf{x_j} \in \mathcal{S}^k_{N,b}} d(\mathbf{x_i}, \mathbf{x_j}).$$

Los percentiles utilizados se buscarán en el intervalo $0.91 \leq \gamma \leq 0.99$.

3.3.3 Bootstrap con estimación por densidad

Este enfoque es distinto a los tratados anteriormente ya que no se basa en una distancia propiamente dicha. El mismo consta de construir un estimador de densidad suave a partir de los datos para posteriormente calcular el diagrama de persistencia a partir de una filtración de las curvas de nivel superior de la función del estimador de densidad (Fasy et al. 2014).

Sea $\mathcal{S}_N$ una muestra i.i.d obtenida a partir de la medida de probabilidad F con soporte en una variedad $\mathcal{M}$, se define

$$f_h(x) = \int_{\mathcal{M}} \frac{1}{h^D} K\left(\frac{||x-u||_2}{h}\right) dF(u),$$

f_h representa entonces la densidad de la medida de probabilidad F_h, que es la convolución $F_h = F \star \mathbb{K}_h$ con $\mathbb{K}_h(A) = h^{-D}\mathbb{K}(h^{-1}A)$ y $\mathbb{K}(A) = \int_A K(t)dt$. Esto básicamente significa que la medida de probabilidad F_h es una versión suavizada de F. Nuestro objetivo será el de computar el diagrama de persistencia de los conjuntos de nivel superiores de f_h, el cual denotaremos $\mathcal{P}_h$.

Estos conjuntos de nivel son relevantes ya que se demuestra que conservan la información topológica de $\mathcal{M}$ y que son más estables al costo de omitir detalles más sutiles de la topología del espacio original (Fasy et al. 2014). Esto resulta de especial importancia en los casos en los que la muestra $\mathcal{S}_N$ se obtiene, no directamente de la densidad f sobre $\mathcal{M}$, sino con un agregado de ruido o, posiblemente, de datos atípicos. Este método podría ser capaz de ignorar los detalles ruidosos y concentrarse mejor en las características topológicas relevantes del espacio subyacente.

Dado que f_h es desconocido, utilizaremos su estimador usual:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{N} \frac{1}{h^D} K\left(\frac{||x-x_i||_2}{h}\right),$$

que será evaluado en una grilla de puntos, a partir del cual construiremos el diagrama de persistencia.

Para obtener la región de confianza, repetiremos el procedimiento realizado en las secciones precedentes para utilizar *bootstrap*. El algoritmo se muestra a continuación

Input: $\mathcal{S}_N,\ \alpha,\ M$
$j \leftarrow 1$;
$\theta \leftarrow array(M)$;
$\hat{f}_h(x)$: Estimador de densidad de f;
$\hat{\mathcal{P}}$: Diagrama de persistencia de una grilla de $\hat{f}_h(x)$;
while $j \leq M$ **do**
  $\mathcal{S}_N^j$: Sub-muestro de tamaño N con reposición de $\mathcal{S}_N$;
  $\hat{f}_h(x)^j$: Estimador de densidad de f basado en la submuestra;
  $\hat{\mathcal{P}}^j$: Diagrama de persistencia de una grilla de $\hat{f}_h(x)^j$;
  $\theta[j] \leftarrow W_\infty(\hat{\mathcal{P}}, \hat{\mathcal{P}}^j)$;
end
return 2 * quantile(θ, $1-\alpha$)

Se demuestra que este método resulta en *tests* más potentes frente a las realizadas haciendo uso de la distancia euclídea (Fasy et al. 2014). Será de interés verificar su rendimiento frente a los *test* de hipótesis que se obtengan a partir de la distancia de Fermat.

Capítulo 4

Resultados

A continuación se disponen los resultados obtenidos de las regiones de confianza para los conjuntos de datos introducidos en la Sección 3.1. Se observa en primera instancia un ejemplo de resultado de diagrama de persistencia junto con su región de confianza para cada método desarrollado en la Sección 3.3 y posteriormente se procede a repetir el experimento una cantidad M de veces para analizar la potencia. Estimada como la tasa de aciertos a la hora de rechazar la hipótesis nula, H_0, dado que la misma es falsa, para distintas muestras de una misma variedad subyacente. Dicho en otras palabras, decidir que hay cualidades topológicas de primer grado (agujeros) en el conjunto de datos, dado que efectivamente las hay. Este análisis se realiza para cada uno de los métodos utilizados sobre los distintos espacios topológicos sintéticos presentados, disponiendo los resultados en forma de tablas.

4.1 Conjuntos de Datos Sintéticos

4.1.1 Diagramas de persistencia y regiones de confianza

A continuación se presentan, individualmente para cada uno de los conjuntos de datos, los diagramas de persistencia y regiones de confianza computadas sobre los mismos para cada uno de los métodos.

Circunferencia uniforme

En las Figuras 4.1, 4.2 y 4.3 se observan, respectivamente, los resultados obtenidos para el *dataset* original, con ruido agregado y con datos atípicos. En cada figura se observa el conjunto de datos muestreado sobre su densidad empírica, junto con los tres diagramas de persistencia obtenidos, correspondientes a cada uno de los métodos desarrollados, sobre los cuales se indica la región de confianza resultante. Para este conjunto de datos, todos los métodos logran detectar correctamente la existencia de un agujero, no viéndose este resultado afectado por la presencia de ruido ni de datos atípicos. Vale la pena destacar que el método que emplea la distancia de Fermat muestra también una componente de grado 0 para cada uno de los datos atípicos, esto resulta interesante ya que podríamos detectar estos *outliers* mirando el diagrama de persistencia. Esto tiene mucho sentido ya que los outliers se encuentran lejos de la masa de datos, por lo que podrían ser parte de otra componente conexa de la topología. Este método resulta entonces el único que logra detectar este comportamiento de forma explícita.

Circunferencia Gaussiana

Análogamente a los resultados de obtenidos para el muestreo uniforme sobre la circunferencia (Sección 4.1.1), en las Figuras 4.4, 4.5 y 4.6 se observan los resultados para el caso del muestreo gaussiano sobre la circunferencia. Los resultados son también similares a los obtenidos en el caso uniforme, ya que en todas las ocasiones se logra detectar de forma apropiada los agujeros de la topología subyacente. Esto demuestra que ninguno de los métodos se ve notoriamente afectado por la diferencia de densidad en el muestreo de las diferentes zonas. KDE podría ser el método más afectado

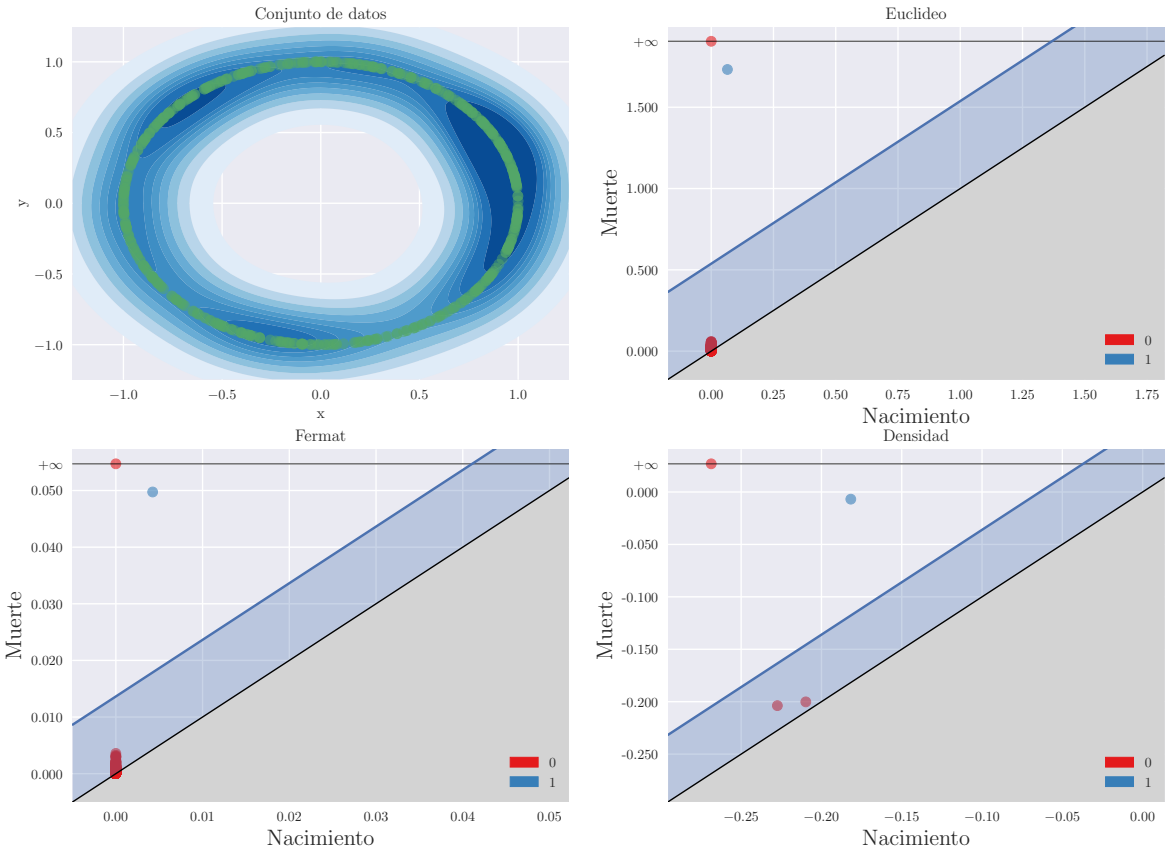


Figura 4.1: Conjunto de datos y regiones de confianza obtenidas para la circunferencia uniforme.

dada las hipótesis con las cuales este método se desarrolla. Para los hiperparámetros y cantidad de muestras probados el mismo no muestra dificultades para descartar correctamente la hipótesis nula en este caso. Resulta destacable que, nuevamente, Fermat es el único método que evidencia los datos atípicos como conjuntos conexos por encima de la región de confianza.

Anteojos

El conjunto de datos de anteojos se analiza en las Figuras 4.7, 4.8 y 4.9 para las variantes base, con ruido agregado y con datos atípicos, respectivamente. Se observa que los resultados son altamente favorables para Fermat en todos los casos, siendo el único capaz de detectar un único agujero de primer grado en todos los casos. Puntualmente, para el caso base (Figura 4.7) se ve que tanto el método euclídeo como KDE detectan muy claramente dos agujeros significativos, esto resulta así para distintos valores probados para el hiperparámetro h en el caso de KDE, que se varió con el objetivo de obtener mejores resultados para este método. Fermat mantiene su correcta interpretación detectando un único agujero significativo con mucha claridad. Para el conjunto de datos con ruido agregado (Figura 4.8) se observa una historia análoga, en la que ninguno de los métodos evaluados cambia el resultado obtenido. El caso de datos atípicos añadidos resulta interesante, ya que se observa en la Figura 4.9 que el método euclídeo logra detectar un único agujero significativo de primer grado. Esto se debe, lamentablemente, no a una mejora atribuible a el método en cuestión, sino a la distribución de datos atípicos, que logra rellenar el agujero derecho del anteojos, de forma tal que el método euclídeo termine solamente detectando como un agujero significativo el izquierdo, que si bien es el resultado esperando nominalmente, se obtiene por las razones equivocadas. Para KDE la historia es análoga a la de los casos anteriores, ya que continúa detectando dos agujeros significativos. Fermat logra nuevamente, no solo detectar un único agujero significativo de primer orden, sino que también logra atribuirle una componente conexa significativa a cada uno de los datos atípicos, como sucedió en cada uno de los conjuntos de datos anteriores cuando

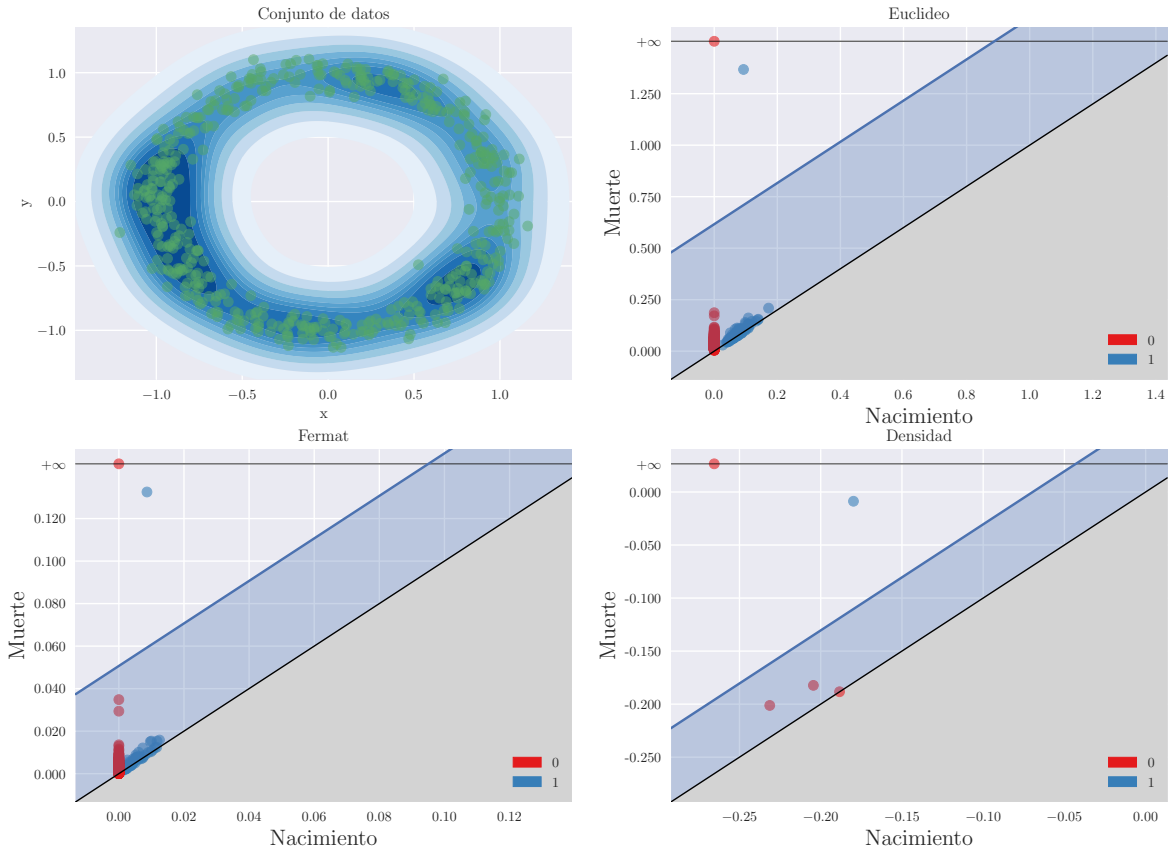


Figura 4.2: Conjunto de datos y regiones de confianza obtenidas para la circunferencia uniforme con ruido agregado.

se le adicionan esta contaminación. Aunque la detección de *outliers* como componentes conexas no es nuestro foco en este trabajo, es un resultado interesante que merece ser analizado en más detalle.

Círculo con densidad dependiente del radio

En la Figura 4.10 se observan los resultados obtenidos, según cada uno de los métodos, para el círculo relleno con densidad dependiente del radio. Recordemos que, como fue introducido en la Sección 3.1.4, en el mismo se espera que los distintos métodos logren aceptar la hipótesis nula, es decir, que no se detecte un agujero de primer grado en la topología subyacente. Se observa en la Figura 4.10 que el método de densidad (KDE) es el único que no logra rechazar la hipótesis nula, mostrando un agujero de grado uno significativo, es decir, por encima de la región de confianza. Este resultado resulta muy relevante ya que nos da a entender que las cualidades de robustez a la hora de calcular los diagramas de persistencia en los casos anteriores, especialmente en presencia de ruido o datos atípicos, pueden estar resultando positivamente a expensas de una distorsión en la topología subyacente que hace que se pierda la significancia a la hora de evaluar topologías sin estos agujeros, perdiéndose la noción de lo que es el nivel de un *test* estadístico.

4.1.2 Potencia

Si ahora se repite el procedimiento realizado en la Sección 4.1.1 un cantidad $M = 50$ de veces con el objetivo de analizar la frecuencia en la que la región estimada no contiene efectivamente a la cantidad de agujeros reales, siendo en nuestro caso un agujero real para todos los conjuntos de datos analizados, entonces se obtiene la potencia estimada de la prueba de hipótesis sobre una variedad, es decir, la probabilidad de rechazar la hipótesis nula H_0 dado que esta es falsa. Como se detalló en la Sección 2.4,

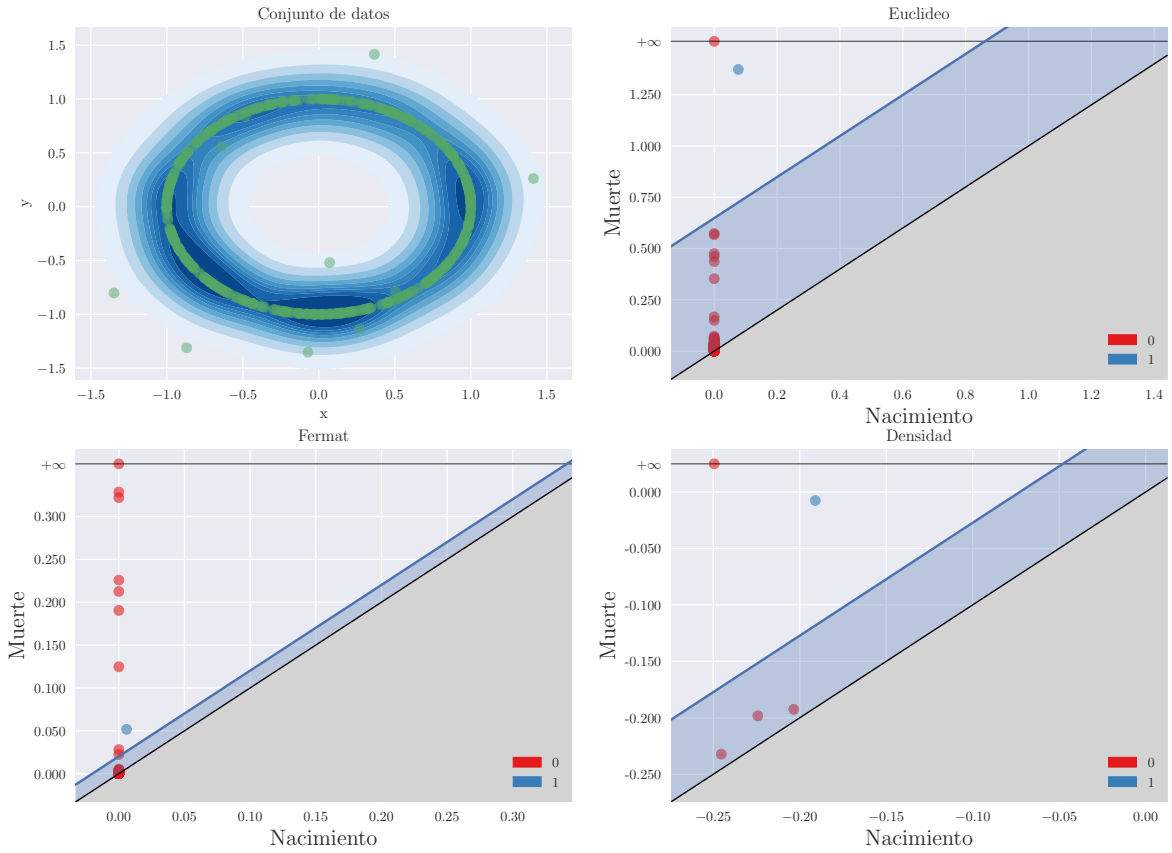


Figura 4.3: Conjunto de datos y regiones de confianza obtenidas para la circunferencia uniforme con muestras atípicas agregadas.

se define una hipótesis nula H_0^i para cada una de las cualidades topológicas que resultan del diagrama de persistencia, manifestándose estas como un punto p_i con tiempo de vida $l_i = d_i - b_i$ en dicho gráfico. Estas hipótesis se expresan entonces como:

$$H_0^i : l_i = d_i - b_i = 0,$$

donde las mismas se evalúan en simultáneo con nivel $1-\alpha$. Como se mencionó en la Sección 2.4 y en (Fasy et al. 2014), esta interpretación de la región de confianza construida sobre el diagrama de persistencia que consiste en caracterizar de forma dicotómica, como "ruido" o "señal", a las cualidad topológicas individuales no es la única interpretación posible. La región de confianza construida proporciona un conjunto $\mathcal{C}_n$ posible de variedades $\mathcal{M}$ de las cuales nuestro conjunto de datos finito de N elementos podría haberse muestreado. Esto significa que el nivel de $1-\alpha$ fijado para la prueba no solo es asintótico, sino que es para la variedad de la cual se obtiene nuestro conjunto de datos, no para la muestra de datos específica. De esta forma, resulta interesante ver cómo distintas muestras de una misma variedad varían en su pertenencia al conjunto $\mathcal{C}_n$ y cómo las distintas técnicas utilizadas para construir el diagrama logran modificar este conjunto de variedades posibles a nivel $1-\alpha$ para el número de muestras utilizado.

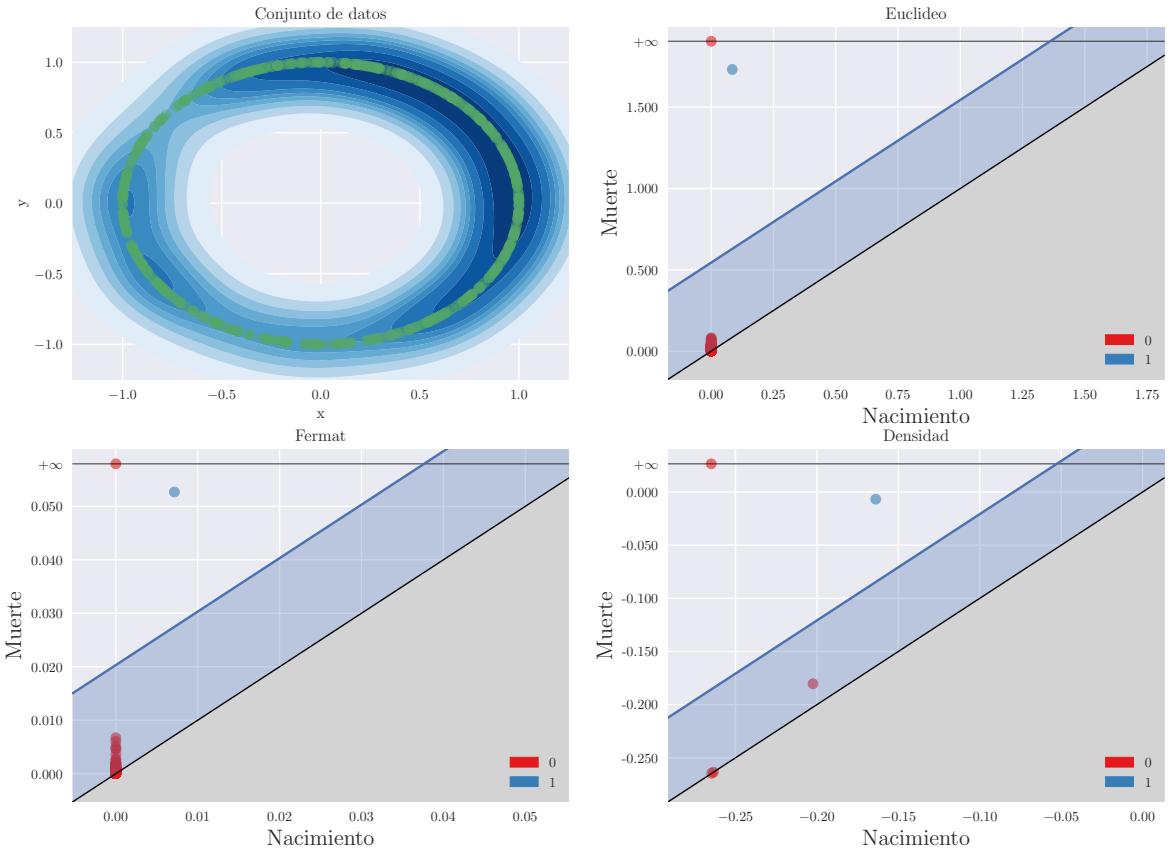


Figura 4.4: Conjunto de datos y regiones de confianza obtenidas para la circunferencia gaussiana

Tabla 4.1: Resultados de potencia estimada para los tres métodos sobre los conjuntos de datos originales. Cada fila de la tabla muestra, según el método que se indica a la izquierda, el porcentaje de veces $(\frac{\#}{M} * 100)$ que cada cantidad de agujeros fue detectada de forma significativa en las corridas realizadas.

Método	Agujeros	Circunferencia Uniforme Porcentaje de Detecciones	Circunferencia Gaussiana Porcentaje de Detecciones	Anteojos Porcentaje de Detecciones
Euclídeo	1	100	100	0
Euclídeo	2	0	0	100
Fermat	1	100	100	100
Fermat	2	0	0	0
KDE	1	100	100	0
KDE	2	0	0	100

En la Tabla 4.1 se observan los resultados para los conjuntos de datos base, sin presencia de ruido o datos atípicos agregados. Se observa que el único método que logra detectar, el 100% de las ocasiones, que la topología subyacente consiste en un único agujero es Fermat. Si bien tanto el método Euclídeo como KDE logran acertar para los casos de circunferencias de muestreo uniforme y gaussiana, respectivamente, fracasan para el conjunto de datos con forma de anteojos, en el cual ambos métodos logran encontrar siempre dos agujeros significativos. Los resultados obtenidos comienzan a demostrar que las intuiciones construidas sobre los diagramas de persistencia individuales de la Sección 4.1.1 generalizan para diferentes muestras de las mismas topologías. Esto no resulta asombroso, ya que las topologías sintéticas estudiadas son bastante estables, es decir, por más que cambien los puntos exactos muestreados es fácil reconstruir la misma topología.

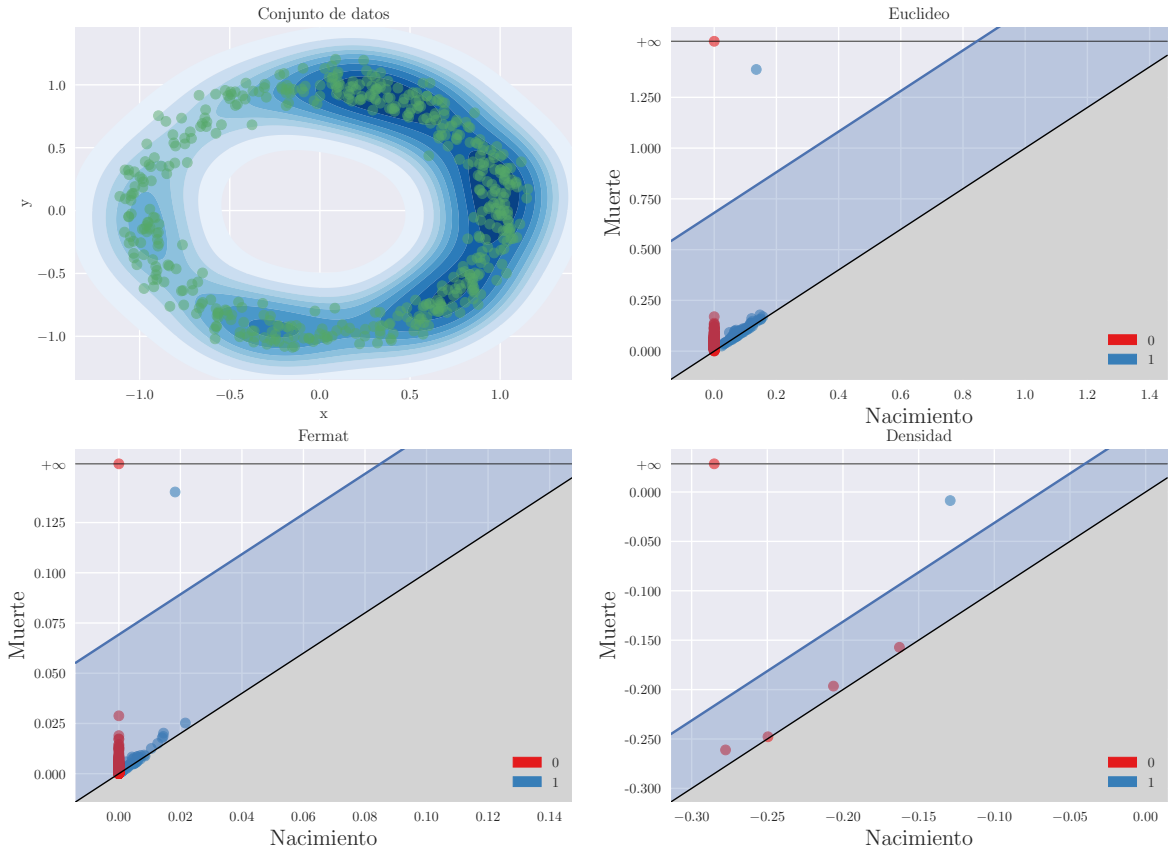


Figura 4.5: Conjunto de datos y regiones de confianza obtenidas para la circunferencia gaussiana con ruido agregado.

Tabla 4.2: Resultados obtenidos para los conjuntos de datos con ruido agregado, análogamente a lo obtenido en la Tabla 4.1 para los conjuntos base

Método	Agujeros	Circunferencia Uniforme Porcentaje de Detecciones	Circunferencia Gaussiana Porcentaje de Detecciones	Anteojos Porcentaje de Detecciones
Euclídeo	1	100	100	0
Euclídeo	2	0	0	100
Fermat	1	100	100	98
Fermat	2	0	0	2
KDE	1	100	100	0
KDE	2	0	0	100

Se muestran en la Tabla 4.2 los resultados para los conjuntos de datos con ruido agregado. La conclusiones resultan altamente similares a las obtenidas en la Tabla 4.1 para los conjuntos base, con la única salvedad de que, si bien sigue siendo el único método en correctamente detectar a los anteojos como una topología de un único agujero significativo con alta potencia, el método de Fermat detecta en limitadas ocasiones (2%) dos agujeros significativos en los Anteojos. Esto se debe, posiblemente, a que la presencia de ruido en el canal que une a las dos circunferencias principales de los anteojos puede llegar “engañar” al método reduciendo el ancho de este canal. Para los casos de circunferencias ninguno de los tres métodos se ve afectado por la presencia del ruido en la varianza estudiada ($\sigma = 0.075$).

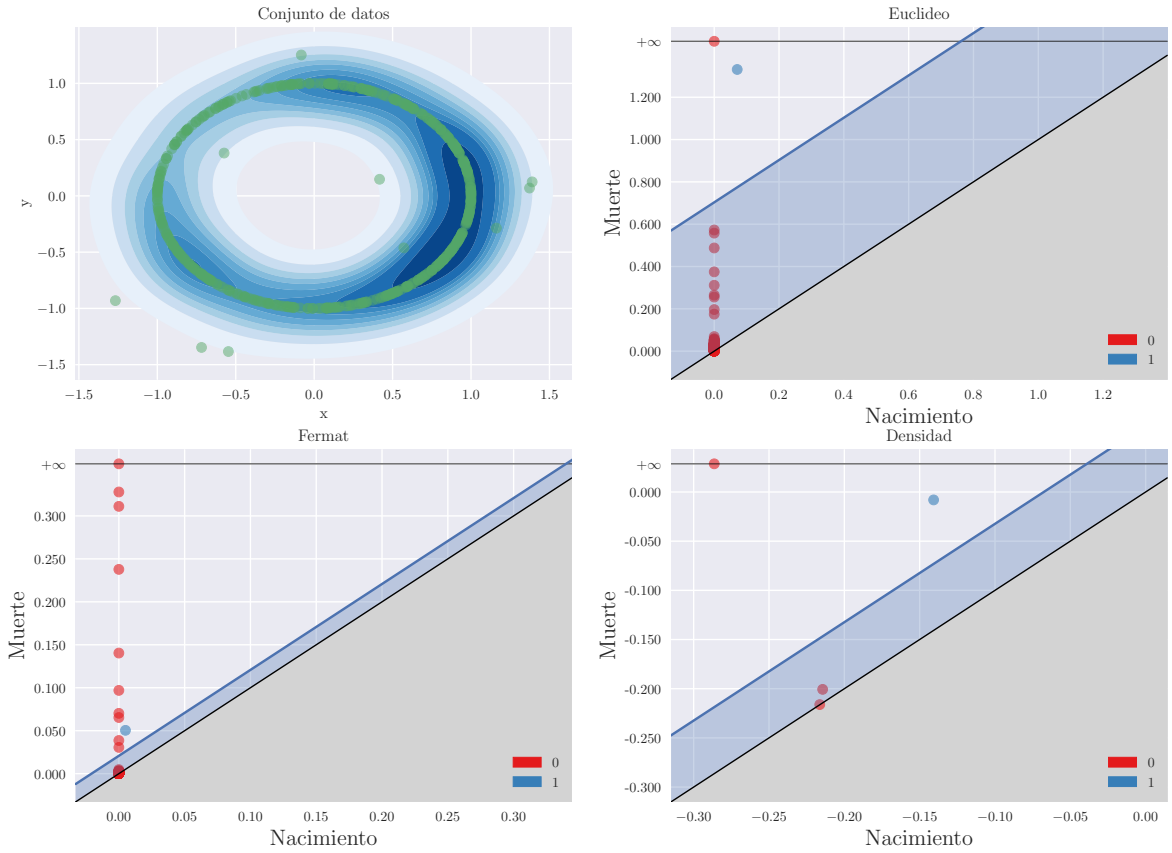


Figura 4.6: Conjunto de datos y regiones de confianza obtenidas para la circunferencia gaussiana con muestras atípicas agregadas.

Tabla 4.3: Resultados de potencia estimada obtenidos para los conjuntos de datos con datos atípicos agregados

Método	Agujeros	Circunferencia Uniforme Porcentaje de Detecciones	Circunferencia Gaussiana Porcentaje de Detecciones	Anteojos Porcentaje de Detecciones
Euclídeo	0	0	0	8
Euclídeo	1	100	100	80
Euclídeo	2	0	0	12
Euclídeo	3	0	0	0
Fermat	0	0	0	0
Fermat	1	98	100	84
Fermat	2	2	0	14
Fermat	3	0	0	2
KDE	0	0	0	0
KDE	1	100	100	0
KDE	2	0	0	100
KDE	3	0	0	0

En la Tabla 4.3 se observan los resultados para el caso de datos atípicos agregados. Para este ejercicio, se observa un poco más de variabilidad en los resultados obtenidos corrida a corrida. Por ejemplo, para el caso del método Euclídeo, se observa que además de seguir detectando apropiadamente un único agujero para las circunferencias de muestreo uniforme y gaussiano, ahora también logra detectar correctamente, en el 80% de los casos, que los Anteojos presentan un único agujero significativo. Esto

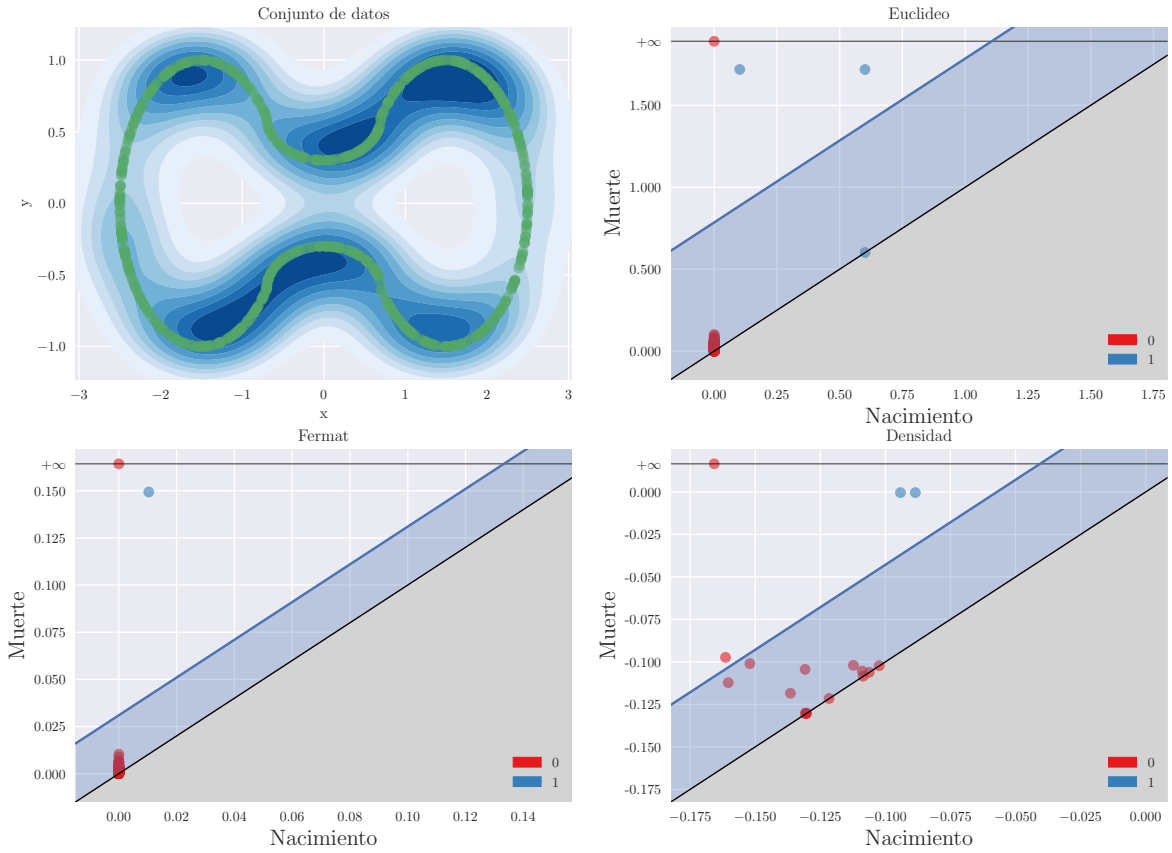


Figura 4.7: Conjunto de datos y regiones de confianza obtenidas para los anteojos

se debe seguramente a que los datos atípicos que se encuentran en el espacio de relleno de los dos agujeros principales de la topología logran reducir el área efectiva del agujero, llevándola una proporción similar a la del túnel que los une, por lo que la interpretación de la reconstrucción mediante homología persistente del método Euclídeo es que hay un único agujero de aproximadamente el tamaño del túnel. La aparición de estos datos atípicos también afecta considerablemente al método de Fermat, en el cual aparecen incluso equivocaciones, aunque en un número muy reducido de ocasiones (2%), para la circunferencia de muestreo uniforme. De nuevo, el conjunto de datos más afectados resulta ser el de Anteojos, ya que una cantidad finita de datos atípicos bien ubicados (como puede ser en el medio del túnel), resulta en detecciones espurias de más agujeros significativos.

Tabla 4.4: Resultados de potencia estimada obtenidos para el conjunto de datos del círculo relleno con densidad dependiente del radio, en el que el resultado esperado sería la aceptación de la hipótesis nula, es decir, detectar cero agujeros

Método	Agujeros	Porcentaje de Detecciones
Euclídeo	0	100
Euclídeo	1	0
Fermat	0	100
Fermat	1	0
KDE	0	50
KDE	1	50

Resulta importante destacar que los resultados obtenidos para el método de ventanas de densidad (KDE) no varían en ningún caso, como se observa en las tablas 4.1, 4.2 y 4.3. Un primer instinto podría

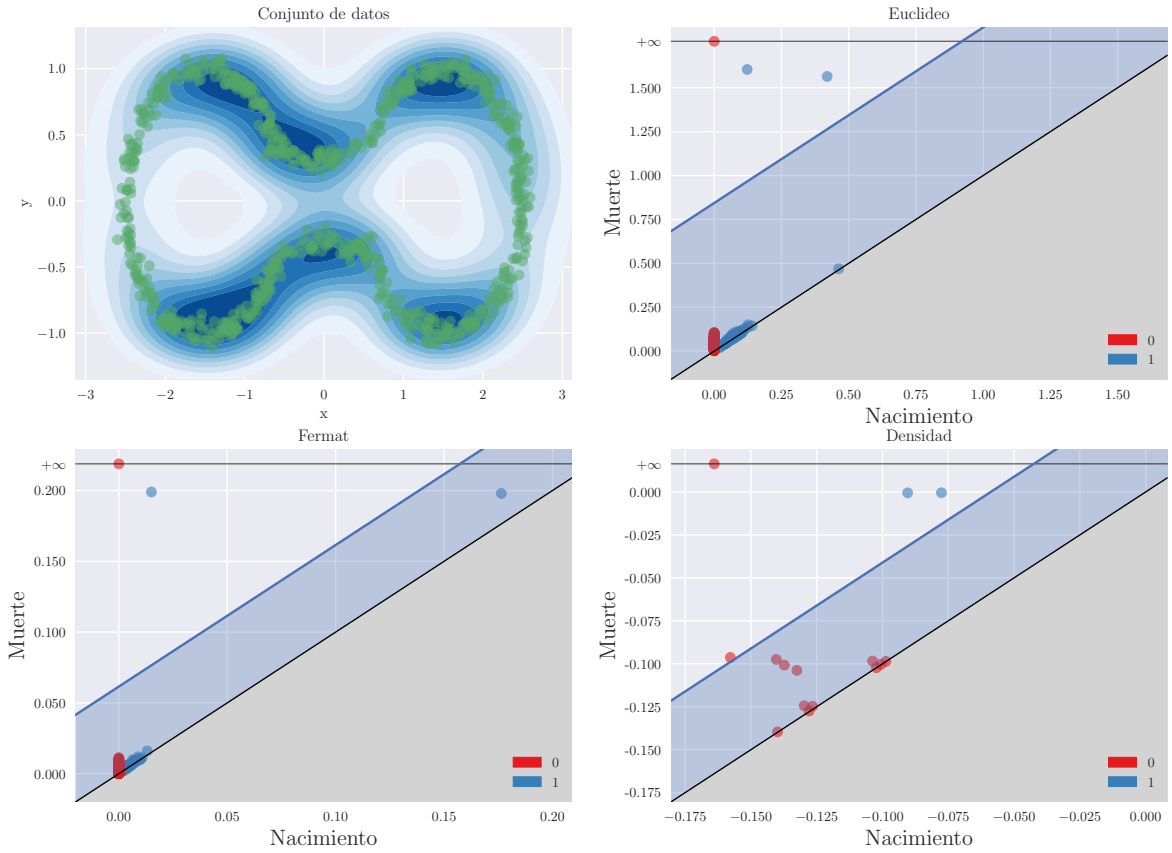


Figura 4.8: Conjunto de datos y regiones de confianza obtenidas para los anteojos con ruido agregado

indicarnos que esto se debe a una propiedad de robustez del método, como se analiza en el trabajo que introduce este algoritmo (Fasy et al. 2014), tal que el ruido y los datos atípicos no afectan las buenas propiedades del mismo. Si bien esto tiene mucho sentido, ya que al usar ventanas de densidad la topología sobre la cual se realiza el test es en realidad una versión suavizada de la original, ocultándose las cualidades ruidosas del conjunto y centrándose en las características principales de la topología subyacente, es también posible que en ese proceso se escondan o se pierdan cualidades que eran de interés para reconstruir la topología subyacente verdadera de la variedad $\mathcal{M}$ bajo estudio. Esto puede evidenciarse en la Tabla 4.4, en la que ahora las corridas se realizan sobre diferentes conjuntos de datos obtenidos de una distribución que muestrea sobre un círculo en el que la densidad de probabilidad es dependiente del radio, como se introdujo en la Sección 3.1.4. Se observa que tanto el método Euclídeo como el de Fermat detectan correctamente que la topología subyacente no presenta ningún agujero, mientras que el método de ventanas de densidad (KDE) descarta la existencia de agujeros solo el 50% de las veces. La explicación más plausible para este comportamiente es que, debido a la distorsión generada por KDE en la topología original y a la baja densidad cercana al origen de la topología, para el número de muestras analizado la región de confianza $1 - \alpha$ no logra incluir a la variedad que genera este conjunto de datos en el conjunto $\mathcal{C}_n$ de posibles variedades.

4.2 Conjuntos de Datos Sintéticos en Dimensiones Superiores

Una observación importante es que en ningún momento del trabajo se discutió analizar conjuntos de datos, sintéticos o no, cuya dimensión de origen sea superior a dos, es decir $D > 2$. Esto se debe a que una parte fundamental de la presente Tesis es contrastar el comportamiento del método de Fermat con el de KDE, y este último presenta un grave problema a la hora de ser evaluado en conjuntos de datos provenientes de dimensiones superiores. Esto se debe a que para obtener el diagrama de

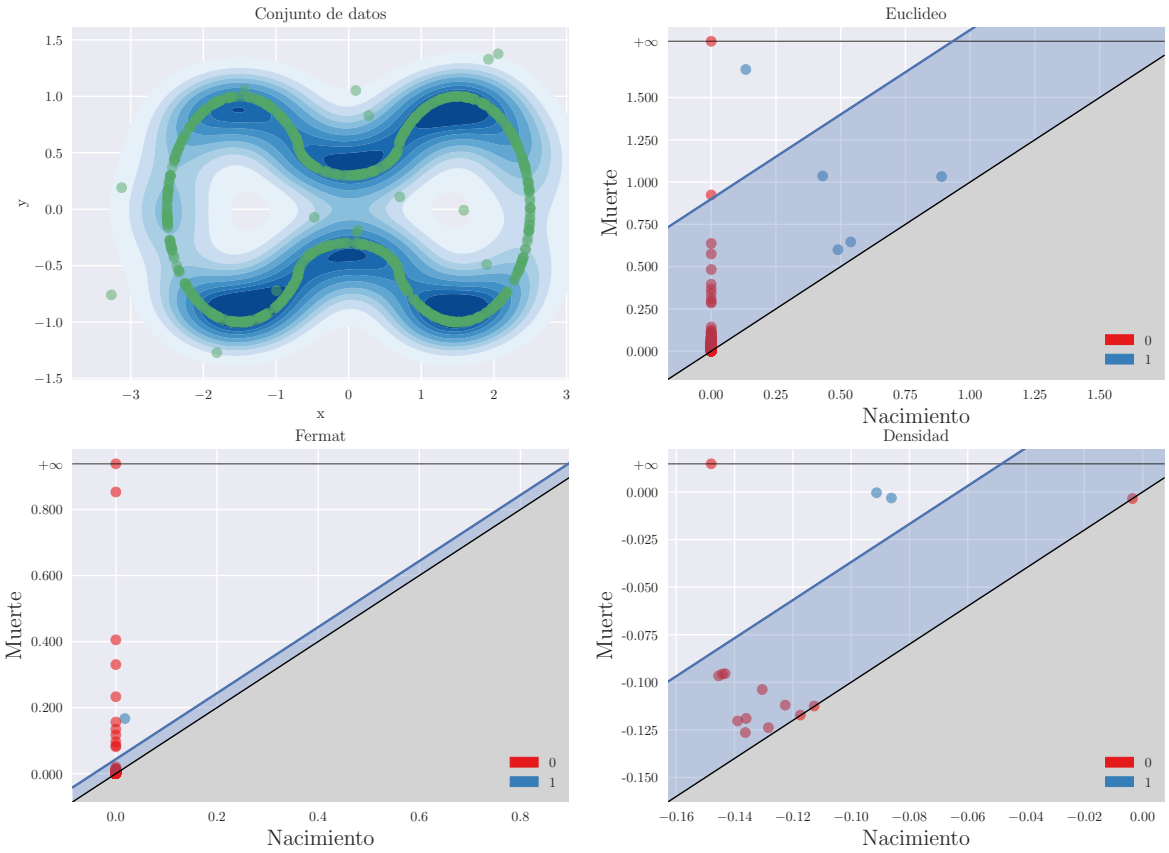


Figura 4.9: Conjunto de datos y regiones de confianza obtenidas para los anteojos con datos atípicos agregados

persistencia mediante KDE, o cualquier otro conjunto de nivel de una función, como se introdujo en la Sección 2.2.3, se requiere trazar una grilla de M_h puntos por dimensión, sobre la cual evaluar la función cuyos conjuntos de nivel son de interés. Teniendo en cuenta que la cantidad de estos puntos condiciona la sensibilidad que se tiene sobre el espacio, resulta importante que M_h sea lo más alto posible, por ejemplo para los casos discutidos en las secciones precedentes, se utilizó $M_h = 100$. Teniendo en cuenta que la cantidad de puntos a evaluar y que serán posteriormente utilizados en la construcción del diagrama de persistencia crecen de forma exponencial con la dimensión del espacio ($k = M_h^D$), y que la complejidad algorítmica de la construcción del diagrama de persistencia escala también de forma exponencial con la cantidad de dimensiones ($\mathcal{O}(k^D) = \mathcal{O}(M_h^{D^2})$) (Dlotko y Wanner 2018), Para los recursos computacionales trabajados una dimensión $D = 3$ ya se torna computacionalmente muy costosa. Resulta importante destacar que esto no sucede para los métodos que se basan en distancias, ya que para estos la complejidad computacional en la construcción del diagrama de persistencia solo depende de la cantidad de puntos, que no depende entonces de la dimensionalidad del espacio.

A modo de ejemplo, podemos ilustrar esta ventaja de los métodos basados en distancia para un conjunto de datos de mayor dimensionalidad construido de forma sencilla: Tomamos el conjunto de datos de la circunferencia uniforme, introducido en la Sección 3.1, y agregamos dimensiones con valores constantes a cada muestra. Para un ejemplo de dos puntos, este procedimiento se vería de la siguiente forma: si tenemos dos puntos muestreados de la circunferencia $x_1 = [0, 1]$ y $x_2 = [1, 0]$, obtenemos muestras en mayores dimensiones como $\hat{x}_1 = [0, 1, \sigma_D^1, \ldots, \sigma_D^L]$ y $\hat{x}_2 = [1, 0, \sigma_D^1, \ldots, \sigma_D^L]$. Lo que se obtiene es la misma circunferencia pero embebida en un espacio de dimensión mayor, en donde la muestra se encuentra ahora viviendo en un plano distinto. En la Figura 4.11 se ilustra este procedimiento para $L = 1$, dando lugar a una circunferencia suspendida en un plano del tercer eje de $\mathbb{R}^3$.

A partir de este nuevo conjunto de datos, podemos realizar el cálculo del diagrama de persistencia y

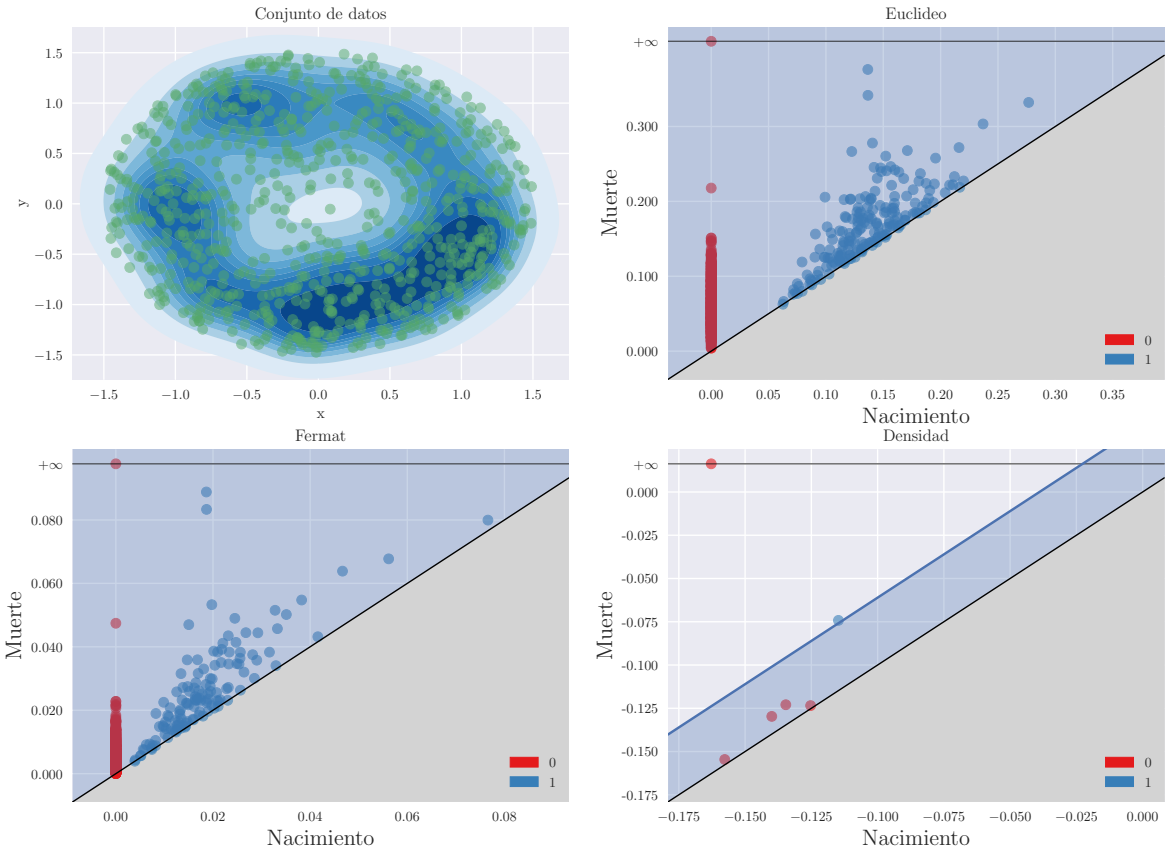


Figura 4.10: Conjunto de datos y regiones de confianza obtenidas para el círculo relleno

su respectiva región de confianza variando la dimensión D y observando cómo este procedimiento se vuelve más computacionalmente costoso. En la Tabla 4.5 se ilustra este fenómeno, donde se muestra el tiempo demorado, medido en segundos, para cada método en obtener la región de confianza al tomar las dimensiones $D = 2, 3, 4$. Se observa claramente que los métodos Euclídeo y Fermat no varían en el tiempo demorado, pero KDE crece exponencialmente en el mismo. Para realizar estas pruebas se utilizó $M_h = 30$, siendo este valor menor al que se utilizó en el resto de los experimentos para conjuntos de datos en dimensión dos ($M_h = 100$), esto se debe a que de otra forma el tiempo requerido para el cómputo con $D = 4$ hubiese sido innecesariamente excesivo.

Tabla 4.5: Costo computacional, representado en segundos que se demora en la construcción del diagrama de pesistencia y su correspondiente región de confianza para los distintos métodos estudiados y tres dimensionalidades diferentes. Se observa cómo los métodos basados en distancia (Euclídeo y Fermat) no varían en el tiempo demandando, mientras que KDE escala exponencialmente.

Dimensiones	Euclídeo	Fermat	KDE
2	8.99	8.07	2.28
3	8.43	8.00	31.65
4	9.06	8.37	1501.22

4.3 Conjuntos de Datos Reales

El trabajar con conjuntos de datos sintéticos nos permite evaluar nuestros resultados a lo largo de diferentes muestras de la misma distribución, como se realizó en la Sección 4.1.2, pero resulta también

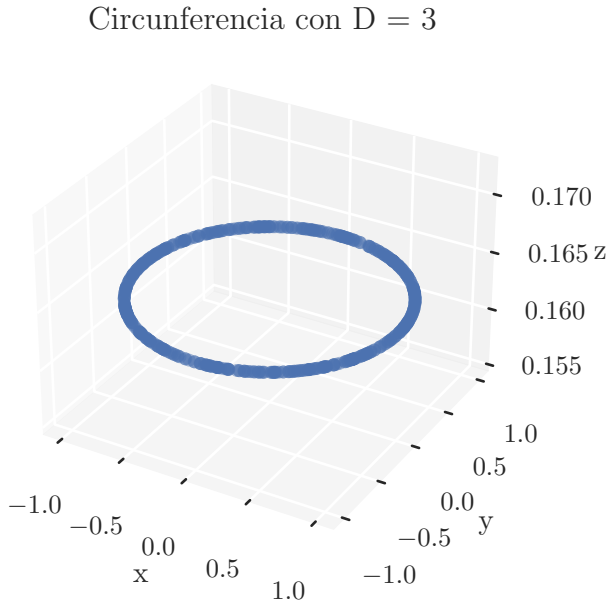


Figura 4.11: Conjunto de datos correspondientes a la circunferencia uniforme con un agregado de una variable adicional que toma un valor constante para todas las muestras, con el objetivo de suspender el conjunto de datos en un plano diferente de un espacio de dimensionalidad superior

muy importante analizar cómo estos métodos se comportan con conjuntos de datos reales, de los cuales interpretaciones relevantes del dominio de aplicación puedan extraerse a partir de las cualidades significativas del diagrama de persistencia para cada uno de los métodos. En línea con el trabajo realizado en (Chazal et al. 2017) utilizaremos el conjunto de datos reales allí utilizado, que como se explicó brevemente en la Sección 3.2, consiste en mediciones correspondientes a la posición de jugadores de fútbol dentro de la cancha a lo largo de un partido, en el que se obtiene un punto cada un intervalo de tiempo determinado. Con el objetivo de obtener agujeros en las zonas donde los jugadores no participan activamente, se agregan artificialmente puntos en los bordes del conjunto de datos, correspondientes a los límites de la cancha (Pettersen et al. 2014). Los diferentes jugadores ocupan diferentes espacios en la cancha, en función de la posición que ocupan en el juego, por lo que analizaremos por separado los diferentes jugadores elegidos

4.3.1 Jugador 2 (Defensor Central)

El jugador con identificador número dos corresponde según una inspección visual a un defensor central, ya que se observa en el gráfico de dispersión en la Figura 4.12 que el mismo ocupa mayoritariamente posiciones en el centro del sector correspondiente al lado del arco de su equipo, aunque realiza algunas maniobras ofensivas por el centro al arco rival, posiblemente para cabecear ofensivamente en los tiros de esquina. En esta muestra, el método euclídeo detecta un único agujero significativo, seguido por un agujero que por poco no logra ser significativo, mientras que Fermat detecta dos bien marcados. Una interpretación es que en ambos casos esos dos puntos del diagrama de persistencia para ambos métodos corresponde a las partes inferior y superior de la zona ofensiva del campo de juego, aunque solo Fermat determina que ambas son significativas. Por su parte, KDE no logra detectar ninguna componente significativa, a pesar de que se varió ampliamente los hiperparámetros con los que se obtiene este diagrama no se logró que KDE obtuviera ningún resultado relevante.

4.3.2 Jugador 5 (Mediocampista)

Para el caso del jugador número 5, puede observarse mediante inspección visual de la Figura 4.13 que el comportamiento del jugador corresponde al de un mediocampista. Se observa que el lateral inferior

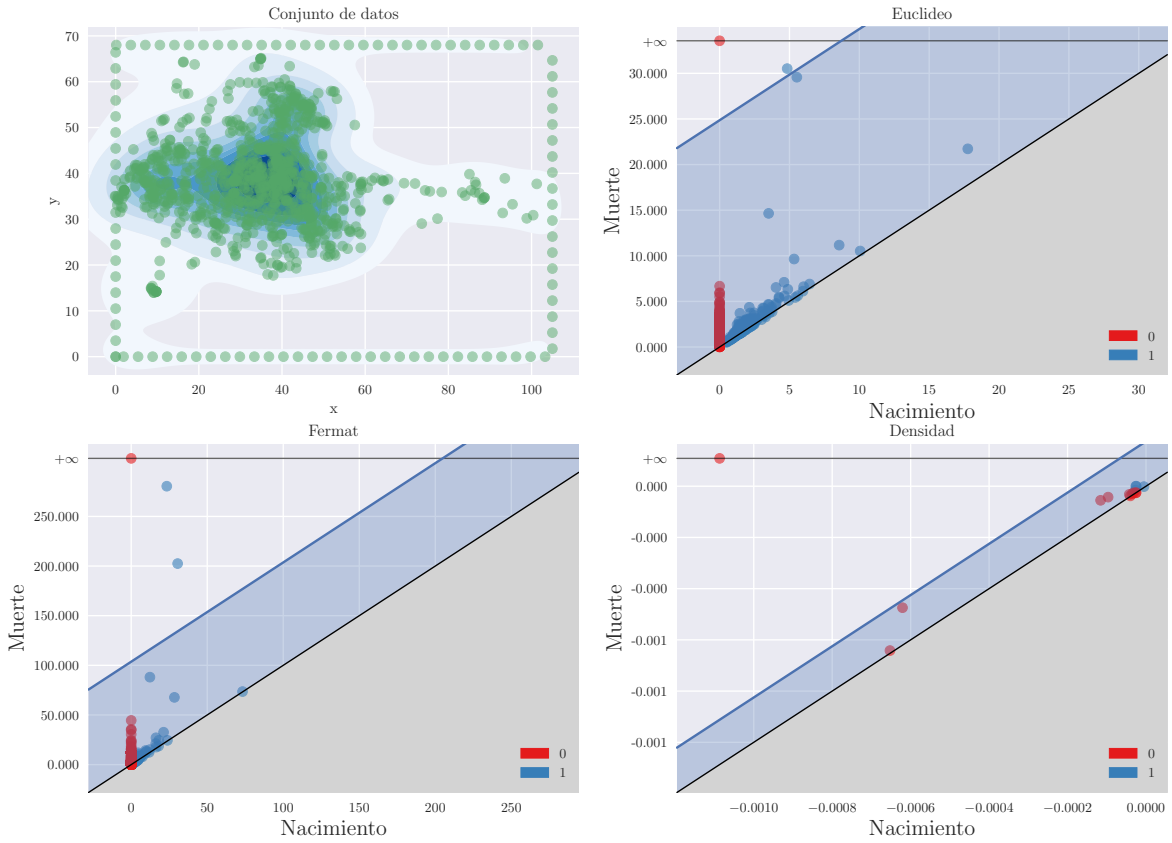


Figura 4.12: Conjunto de datos y regiones de confianza obtenidas para el jugador identificado como el número 2, que corresponde al comportamiento de un defensor central

del campo carece de puntos y que también existen zonas de baja densidad en las esquinas superiores derecha e izquierda, aunque esta última presenta una zona desocupada más reducida. Los distintos métodos obtienen distintos resultados: Como ya se mencionó para el caso del jugador número 2, KDE no logra hacer ningún tipo de detección relevante, sin importar cómo se ajusten los hiperparámetros. El método euclídeo detecta dos agujeros significativos, que no resulta obvio a qué agujeros visuales se corresponden, mientras que Fermat detecta cuatro de estos, que muy posiblemente se correspondan con las cuatro esquinas del campo de juego.

4.3.3 Jugador 8 (Lateral Izquierdo)

El caso del jugador número 8 resulta el más sencillo de todos, ya que en la inspección visual del diagrama de dispersión en la Figura 4.14 se evidencia de forma clara que el jugador solo ocupa, aunque con buena densidad, la esquina superior izquierda, adentrándose hasta no mucho más que la mitad del campo rival. Esto significa que el comportamiento del jugador se corresponde con el de un lateral izquierdo de corte defensivo, y se espera que los distintos algoritmos detecten con facilidad la existencia de un único agujero significativo. Los métodos de Fermat y Euclídeo no presentan problemas en detectar esta topología subyacente, descartando de la región significativa todo salvo un agujero de primer orden. Análogamente a lo obtenido en los jugadores anteriores, KDE no logra resultados satisfactorios.

4.3.4 Jugador 14 (Mediocampista)

En la Figura 4.15 se observa el gráfico de dispersión y los resultados obtenidos según cada uno de los métodos para el jugador número 14. Se observa que el comportamiento de juego del mismo se corresponde con el de un mediocampista, y que exhibe un comportamiento muy similar al del jugador

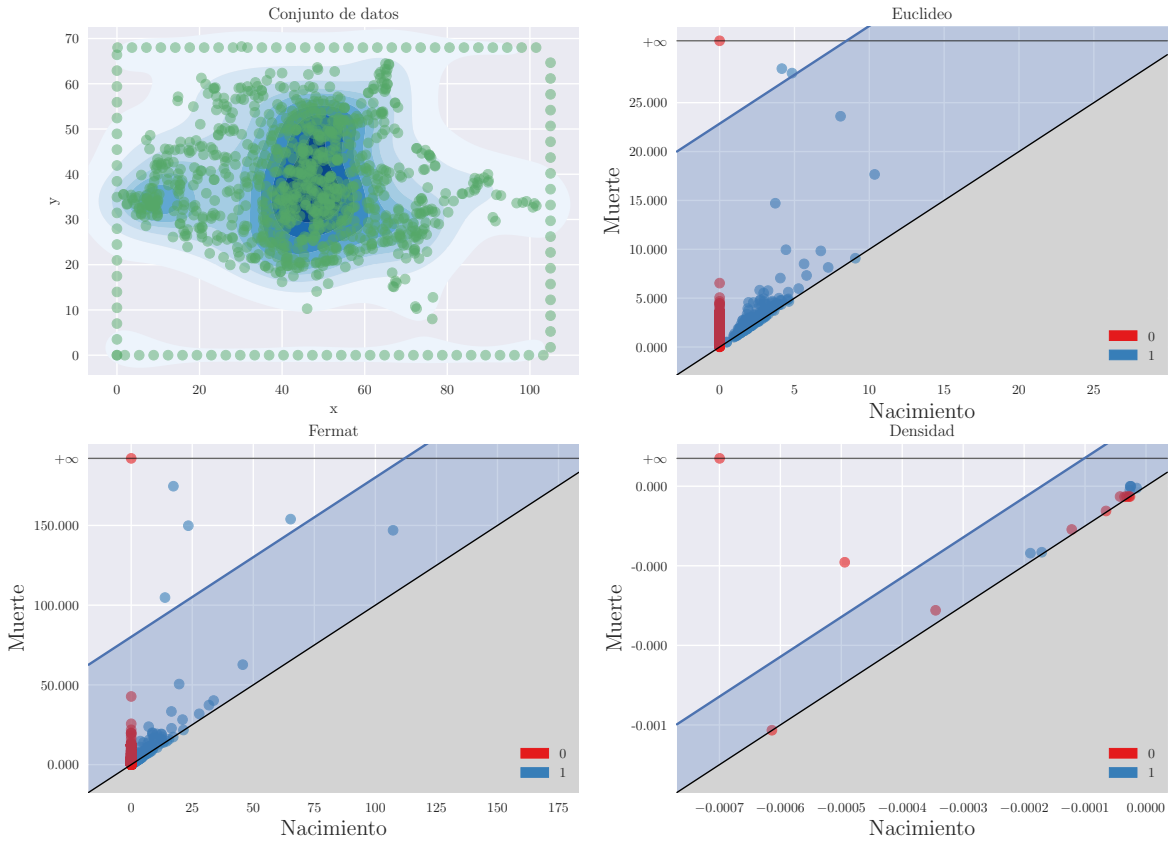


Figura 4.13: Conjunto de datos y regiones de confianza obtenidas para el jugador identificado como el número 5, que corresponde al comportamiento de un mediocampista

número 5. Esta similitud visual en el comportamiento se manifiesta también en los diagramas de persistencia obtenidos, ya que por ejemplo el método de Fermat detecta, al igual que para el jugador 5 y con una composición muy similar, cuatro agujeros significativos que seguramente estén asociados a las cuatro esquinas del campo de juego. El caso del método Euclídeo es similar, ya que el diagrama de persistencia muestra similitud con el obtenido para el jugador 5 en el que cinco puntos se alejan de la diagonal considerablemente, sin embargo, en este caso solo un agujero se detecta como significativo, a diferencia del caso del jugador 5 en el que se detectaron dos

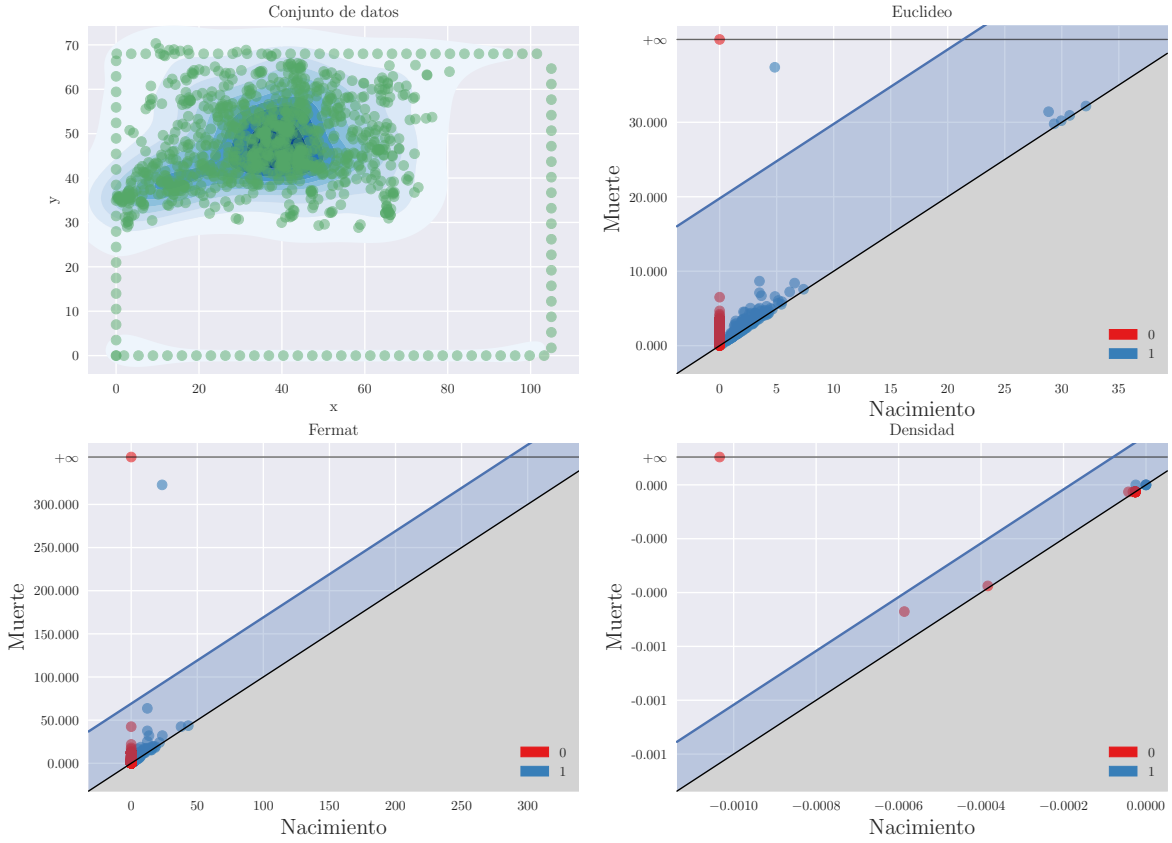


Figura 4.14: Conjunto de datos y regiones de confianza obtenidas para el jugador identificado como el número 8, que corresponde al comportamiento de un lateral izquierdo

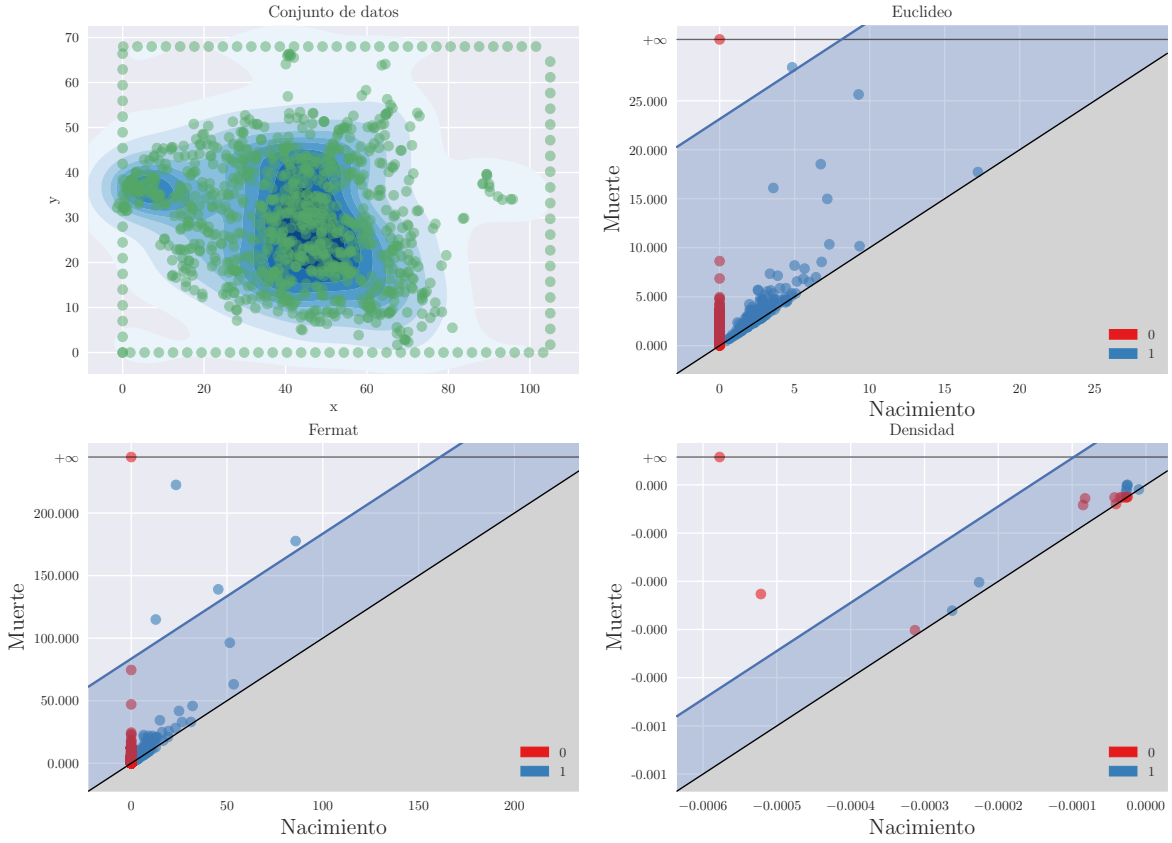


Figura 4.15: Conjunto de datos y regiones de confianza obtenidas para el jugador identificado como el número 14, que corresponde al comportamiento de un mediocampista

Capítulo 5

Conclusiones y Próximos pasos

A lo largo del presente trabajo se buscó realizar una introducción teórica de los métodos estadísticos existentes aplicados al análisis topológico de datos (Fasy et al. 2014), en particular dentro del área de homología persistente, con el objetivo de inferir la topología subyacente, descrita a partir de los agujeros de primer orden que la misma presente, de un conjunto de datos. En la Capítulo 2 se introdujo cómo la función de distancia elegida cumple un rol fundamental en esta tarea, por lo que en la Sección 2.1 se presentó la Distancia de Fermat, presentándose ésta como una alternativa a la distancia euclídea, con la característica de que logra codificar en las distancias de los puntos entre sí de un conjunto de datos de forma más apropiada la topología subyacente de los mismos (Fernández et al. 2024; Sapienza, Groisman, y Jonckheere 2018). En la Capítulo 3 se detalló cómo esta distancia se utilizó para construir diagramas de persistencia, con sus respectivas regiones de confianza, de los distintos conjuntos de datos introducidos, tanto sintéticos como reales. Los resultados obtenidos a lo largo de la Capítulo 4 muestran cómo se comparan los resultados del método de Fermat contra los obtenidos según los métodos discutidos en la bibliografía (Fasy et al. 2014), los cuales llamamos método Euclídeo y KDE. En particular, en la Sección 4.1, donde se analizaron los conjuntos de datos sintéticos, se observan las primeras diferencias y similitudes entre los distintos métodos.

Resulta importante destacar que para los resultados expuestos a lo largo de todo el trabajo, distintos valores de hiperparámetros fueron probados, conservando el conjunto de valores que mejores resultados obtenía en las pruebas realizadas para cada uno de los métodos. Una línea de trabajo interesante que se desprende de los resultados obtenidos es el impacto de los distintos valores de hiperparámetros en los mismos. Esta experimentación requiere de un muy alto poder computacional, dada la complejidad de los algoritmos utilizados para construir los diagramas de persistencia y teniendo en cuenta que la construcción de regiones de confianza sobre los mismos requiere realizar esta operación decenas de veces. Particularmente, para el caso del hiperparámetro λ utilizado como potencia en el cálculo de la distancia de Fermat, el mismo demostró ser bastante robusto ya que su valor más intuitivo $\lambda = 2$ logra buenos resultados, encontrándose como resultado heurístico que un rango de valores entre 1.8 y 2.7 para λ funcionan también de buena manera. Teniendo en cuenta la interpretación de este hiperparámetro y su impacto directo en el cálculo de la distancia, este rango de valores funcionales resulta muy amplio e intuitivo.

Para todos los conjuntos de datos sintéticos que presentan un agujero se observan resultados consistentes, en los que Fermat logra determinar correctamente la presencia del mismo para la totalidad de los *datasets*, mientras que KDE y Euclídeo fallan para el conjunto de Anteojos, detectando dos agujeros, pero lo logran para las circunferencias. Problemas discutidos en la bibliografía (Fasy et al. 2014; Chazal et al. 2017) como son la no uniformidad en la densidad de muestreo sobre la topología (que impacta principalmente en el la circunferencia con muestreo Gaussiano) no resultan ser problemáticos para los métodos estudiados.

Los conjuntos de datos fueron perturbados con ruido gaussiano y datos atípicos, para el caso de ruido, los resultados se mantienen prácticamente análogos al conjunto original, como se observa en la Tabla 4.2.

Resultaría interesante como continuación de este trabajo analizar cómo estos resultados varían en función de la varianza del ruido agregado.

Distinto es el caso de datos atípicos agregados, ya que para los modelos basados en distancia, que reconstruyen la topología a partir de complejos simpliciales, una muestra bien ubicada es suficiente para cerrar un ciclo de los mismos, distorsionando la topología recuperada. Así se evidencia para las simulaciones realizadas en la Tabla 4.3, en donde puede verse cómo KDE es el único no afectado, ya que tanto Fermat como Euclídeo logran detectar en entre un 16% (Fermat) y 20% (Euclídeo) una topología distinta la verdadera, con mayor o menor cantidad de agujeros. Merece una mención especial la capacidad del método de Fermat de detectar de forma significativa para el conjunto de datos con muestras atípicas que la topología presenta más componentes conexas, teniendo en cuenta que para todos nuestros conjuntos de datos se espera una única componente de grado cero. Más aún, en todos los casos analizados, la cantidad de componentes conexas coincide con la cantidad de datos atípicos, lo que motiva un trabajo futuro de estos métodos como mecanismo de detección de muestras atípicas.

Si bien el método KDE es el único no afectado por la presencia de datos atípicos para los conjuntos de datos con agujeros, esto toma una especial perspectiva cuando se analiza el conjunto de datos del círculo con densidad dependiente del radio. En (Fasy et al. 2014) se establece que la distorsión del espacio original que realiza KDE mantiene la topología principal subyacente a la vez que hace al método más robusto al ruido, suavizando ruido y datos atípicos, pero con este conjunto de datos se observa que la región de confianza obtenido para un nivel de 95% no parece ser consistente con detectar agujeros en solo el 5% de las corridas, más aún, el método es incapaz de aceptar la hipótesis nula en la mitad de las corridas, sabiendo que esta es verdadera. Esto, junto con la dificultad a la hora de extender este método a dimensiones superiores por cuestiones computacionales (Tabla 4.5) representan dos desventajas muy claras que el método propuesto en (Fasy et al. 2014) tienen frente al uso de Fermat como métrica.

A la hora de analizar conjuntos de datos reales (Sección 4.3), es fácil verificar uno de los problemas que la bibliografía menciona sobre KDE (Chazal et al. 2017). Al no tener nuestro conjunto de datos agujeros sin la presencia de un borde, debe adicionarse uno de forma artificial, pero la densidad de muestras que se le asignen a este borde determinará en buena medida si KDE considera o no a estas muestras como parte de la topología subyacente a analizar, ya que el suavizado propio de este método podría simplemente obviar. Como contraparte de esta necesidad está la creciente carga computacional de agregar más puntos al conjunto de datos original. Para los valores utilizados, los resultados de KDE muestran nuevamente una debilidad de este método, consistente a lo obtenido en (Chazal et al. 2017). Por su parte, el método de Fermat logra detectar una cantidad de agujeros significativos muy similar a lo que arroja una inspección visual del problema, por lo que se considera que el método es efectivo para este caso de uso, más aún, la similitud en los resultados obtenidos para los diagramas de persistencia entre jugadores diferentes que comparten posición resulta muy interesante. El método Euclídeo por su parte logra un diagrama de persistencia saludable pero no logra una detección de agujeros significativos que coincidan apropiadamente con la inspección visual de los conjuntos de datos. Resulta una interesante línea de investigación analizar el comportamiento de estos métodos para otro tipo de conjuntos de datos reales.

A partir de todos los resultados obtenidos se concluye entonces que utilizar los métodos basados en distancias con Fermat como medida de las mismas logra solucionar muchos de los problemas que presenta el utilizar distancia euclídea o el método de ventanas de densidad para analizar una variedad de conjuntos de datos.

FIRMA DEL DIRECTOR FIRMA DEL MAESTRANDO

Referencias

Babu, Gutti Jogesh. 1992. «Subsample and half-sample methods». *Annals of the Institute of Statistical Mathematics* 44 (4): 703-20. https://doi.org/10.1007/BF00053399.

Bauer, Ulrich, Michael Kerber, Fabian Roll, y Alexander Rolle. 2023. «A unified view on the functorial nerve theorem and its variations». *Expositiones Mathematicae* 41 (4): 125503. https://doi.org/https://doi.org/10.1016/j.exmath.2023.04.005.

Bengio, Yoshua, Aaron Courville, y Pascal Vincent. 2013. «Representation Learning: A Review and New Perspectives». *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (8): 1798-828. https://doi.org/10.1109/TPAMI.2013.50.

Blumberg, Andrew J., Itamar Gal, Michael A. Mandell, y Matthew Pancia. 2014. «Robust Statistics, Hypothesis Testing, and Confidence Intervals for Persistent Homology on Metric Measure Spaces». *Found. Comput. Math.* 14 (4): 745-89. https://doi.org/10.1007/s10208-014-9201-4.

Bobrowski, Omer, y Dmitri Krioukov. 2022. «Random Simplicial Complexes: Models and Phenomena». En *Higher-Order Systems*, 59-96. Springer International Publishing. https://doi.org/10.1007/978-3-030-91374-8_2.

Bobrowski, Omer, y Primoz Skraba. 2020. «Homological Percolation: The Formation of Giant k-Cycles». *International Mathematics Research Notices* 2022 (8): 6186-6213. https://doi.org/10.1093/imrn/rnaa305.

———. 2023. «A universal null-distribution for topological data analysis». *Scientific Reports* 13 (julio). https://doi.org/10.1038/s41598-023-37842-2.

Bubenik, Peter. 2012. «Statistical topological data analysis using persistence landscapes». https://doi.org/10.48550/ARXIV.1207.6437.

Bubenik, Peter, y Peter T. Kim. 2007. «A statistical approach to persistent homology». *Homology, Homotopy and Applications* 9 (2): 337-62. https://doi.org/hha/1201127341.

Chazal, Frédéric, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, y Larry Wasserman. 2014. «Stochastic Convergence of Persistence Landscapes and Silhouettes». En *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*, 474-83. SOCG'14. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2582112.2582128.

Chazal, Frédéric, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, Alessandro Rinaldo, y Larry Wasserman. 2017. «Robust topological inference: distance to a measure and kernel distance». *J. Mach. Learn. Res.* 18 (1): 5845-84.

Chazal, Frédéric, y Bertrand Michel. 2021. «An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists». *Front Artif Intell* 4 (septiembre): 667963.

Dang-Nguyen, Chau, y Tuan Do-Hong. 2019. «Robust Line Hausdorff Distance for Face Recognition». En *2019 International Symposium on Electrical and Electronics Engineering (ISEE)*, 103-7. https://doi.org/10.1109/ISEE2.2019.8921218.

Dantchev, Stefan, y Ioannis Ivrissimtzis. 2012. «Efficient construction of the Čech complex». *Computers and Graphics* 36: 708-13. https://doi.org/10.1016/j.cag.2012.02.016.

Dlotko, Pawel, y Thomas Wanner. 2018. «Rigorous cubical approximation and persistent homology of continuous functions». *Computers & Mathematics with Applications*, marzo. https://hal.science/hal-01706695.

Edelsbrunner, Herbert, y John Harer. 2010. *Computational Topology - an Introduction*. American Mathematical Society.

Efron, B., y R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs

on Statistics & Applied Probability. Taylor & Francis. https://books.google.com.br/books?id=gLlpIUxRntoC.

Emrani, Saba, Thanos Gentimis, y Hamid Krim. 2014. «Persistent Homology of Delay Embeddings and its Application to Wheeze Detection». *IEEE Signal Processing Letters* 21 (4): 459-63. https://doi.org/10.1109/LSP.2014.2305700.

Fasy, Brittany Terese, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, y Aarti Singh. 2014. «Confidence sets for persistence diagrams». *The Annals of Statistics* 42 (6): 2301-39. https://doi.org/10.1214/14-AOS1252.

Fernández, Ximena, Eugenio Borghini, Gabriel Mindlin, y Pablo Groisman. 2024. «Intrinsic persistent homology via density-based metric learning». *J. Mach. Learn. Res.* 24 (1).

Groisman, Pablo, Matthieu Jonckheere, y Facundo Sapienza. 2022. «Nonhomogeneous Euclidean first-passage percolation and distance learning». *Bernoulli* 28 (1): 255-76. https://doi.org/10.3150/21-BEJ1341.

Illowsky, B., y S. Dean. 2017. *Introductory Statistics*. Samurai Media Limited. https://books.google.com.ar/books?id=25-RtAEACAAJ.

Iniesta, Raquel, Carr Ewan, Carrière Mathieu, Yerolemou Naya, Michel Bertrand, y Chazal Frédéric. 2022. «Topological Data Analysis and its usefulness for precision medicine studies». *SORT-Statistics and Operations Research Transactions* 46 (1): 115-36. https://doi.org/10.2436/20.8080.02.120.

Maletić, Slobodan, Yi Zhao, y Milan Rajković. 2016. «Persistent topological features of dynamical systems». *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26 (5): 053105. https://doi.org/10.1063/1.4949472.

Niyogi, Partha, Stephen Smale, y Shmuel Weinberger. 2008. «Finding the Homology of Submanifolds with High Confidence from Random Samples». *Discrete & Computational Geometry* 39 (1): 419-41. https://doi.org/10.1007/s00454-008-9053-2.

Pellerin, Jeanne. 2014. «Accounting for the geometrical complexity of geological structural models in Voronoi-based meshing methods». Tesis doctoral. https://doi.org/10.13140/RG.2.1.2719.2169.

Perea, Jose A. 2018. «Topological Time Series Analysis». *ArXiv* abs/1812.05143. https://api.semanticscholar.org/CorpusID:56148259.

Pérez, Julián Burella, Sydney Hauke, Umberto Lupo, Matteo Caorsi, y Alberto Dassatti. 2021. «giotto-ph: A Python Library for High-Performance Computation of Persistent Homology of Vietoris-Rips Filtrations». arXiv. https://doi.org/10.48550/ARXIV.2107.05412.

Pettersen, Svein, Dag Johansen, Håvard Dagenborg, Vegard Berg-Johansen, Vamsidhar Gaddam, Asgeir Mortensen, Ragnar Langseth, Carsten Griwodz, Håkon Stensland, y Pål Halvorsen. 2014. «Soccer Video and Player Position Dataset». En *Proceedings of the 5th ACM Multimedia Systems Conference, MMSys 2014*. https://doi.org/10.1145/2557642.2563677.

Potamias, Michalis, Francesco Bonchi, Carlos Castillo, y Aristides Gionis. 2009. «Fast shortest path distance estimation in large networks». En *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 867-76. CIKM '09. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1645953.1646063.

Sapienza, Facundo, Pablo Groisman, y Matthieu Jonckheere. 2018. «Weighted Geodesic Distance Following Fermat's Principle». En *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:34393656.

Silva, Vin de, y Robert Ghrist. 2007. «Coverage in sensor networks via persistent homology». *Algebraic & Geometric Topology* 7 (1): 339-58. https://doi.org/10.2140/agt.2007.7.339.

The GUDHI Project. 2015. *GUDHI User and Reference Manual*. GUDHI Editorial Board. http://gudhi.gforge.inria.fr/doc/latest/.

Wikipedia. 2022. «Homology (mathematics) — Wikipedia, The Free Encyclopedia». http://en.wikipedia.org/w/index.php?title=Homology%20(mathematics)&oldid=1098660019.

———. 2024. «Bootstrapping (statistics) — Wikipedia, The Free Encyclopedia». http://en.wikipedia.org/w/index.php?title=Bootstrapping%20(statistics)&oldid=1244613103.

Wong, Chi-Chong, y Chi-Man Vong. 2021. «Persistent Homology based Graph Convolution Network for Fine-grained 3D Shape Segmentation». En *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7078-87. https://doi.org/10.1109/ICCV48922.2021.00701.