CIMAT

CENTRO DE INVESTIGACIONES EN MATEMÁTICAS
MAESTRÍA EN PROBABILIDAD Y ESTADÍSTICA

UN ROL PARA LA CURVA DE BETTI EN PROBLEMAS DE CIENCIA DE DATOS

Por:
Licenciado Rafael José González De Gouveia

Director:
Dr. Miguel Nakamura Savoy
Centro de Investigaciones en Matemáticas.

Guanajuato, 3 de diciembre de 2017

Resumen

Un rol para la curva de Betti
en problemas de Ciencia de Datos

Rafael José González De Gouveia

Centro de Investigaciones en Matemáticas. 2017

Se responde a preguntas sobre la utilidad y eficacia del Análisis Topológico de Datos (ATD) en la solución de problemas de Ciencia de Datos. Se utilizan las Curvas de Betti (CB). Estas curvas resumen la forma de una nube de datos, es decir, cualidades como clusters (componentes conexas), estructura filamentosa o agujeros. Una esfera y un toro tienen formas diferentes porque sus agujeros son diferentes. Para estudiar la aplicación de las CB se escogieron casos en los cuales es posible compararlas con herramientas usuales en Ciencia de Datos.

En el primer caso se aborda la bondad del ajuste en procesos puntuales bajo la hipótesis de aleatoriedad espacial completa. El estadístico alternativo de prueba utiliza diferencias entre curvas de Betti para cuantificar discrepancias entre procesos puntuales. Esta prueba se contrasta con la prueba de cuadrantes y la prueba de desviaciones absolutas máximas (o MAD por sus siglas en inglés), ampliamente conocidas en la literatura de procesos puntuales. Como segundo problema, se clasifican nubes de puntos muestreadas sobre objetos 3D. Se aborda utilizando vectores de Betti. Este concepto es original de este trabajo y resulta útil para hacer clasificación con topología. Este método es comparado con una aproximación en ciencia de datos, utilizando directamente los datos en un algoritmo de máquina de soporte vectorial. Estos experimentos cuantifican el aporte de utilizar métodos topológicos contra las técnicas conocidas en Ciencia de datos.

Agradecimientos

Agradezco la culminación de esta tesis a:

- Conacyt por el apoyo recibido en la realización de esta maestría.
- Centro de Investigaciones en Matemáticas. Especialmente le agradezco a mi asesor Miguel Nakamura y al profesor Víctor Pérez Abreu por su mentoría a lo largo de esta etapa.
- Instituto de Matemáticas de la Universidad Nacional Autónoma de México por el apoyo recibido durante la estancia de investigación realizada en París en febrero de 2017, Proyecto 265667 FORDECYT "Programa para un avance global e integrado de la matemática mexicana".
- Rolando Biscay por sugerir el uso de las curvas de Betti en el Capítulo 3.
- Rogelio Ramos por ser parte de los sinodales en la defensa esta tesis.
- Yair Hernández y Gilberto Flores por facilitarme el algoritmo para simular puntos de la botella de Klein en el Capítulo 3.
- Avner Bar-Hen del Conservatoire National des Arts et Métiers por su apoyo en la comprensión de los procesos puntuales.
- mis compañeros de estudio Zitlalli y Oscar.
- otros compañeros de CIMAT como Lily, Adrián y Larry.
- mi familia: Sagrario, M. Teresa, M. Victoria, Dina y Rafael.

Rafael González de Gouveia

Índice general

1

Introducción

El análisis topológico de datos (ATD) es una disciplina que tiene como objetivo estudiar estructuras complejas de información. Los objetos que se analizan comprenden números, vectores, matrices, archivos de texto o imágenes (Carlsson, 2009). El ATD estudia la forma de los datos a través de diversos conceptos de topología. La forma de los datos se refiere a las características topológicas, como la cantidad de componentes conexas, ciclos o agujeros. La cantidad de estas características se denominan números de Betti. Los números de Betti se pueden calcular mediante la homología persistente como en Edelsbrunner and Harer (2010), una técnica de topología algebraica. Esta tesis estudia en detalle las curvas de Betti, que son una estadística funcional que se deriva de la homología persistente. En este trabajo se abordan problemas de ciencia de datos con curvas de Betti.

El ATD ha tenido una notable evolución en los últimos años. Las primeras apariciones pueden verse desde Edelsbrunner et al. (1995), quien trabaja en aplicaciones de topología en el análisis de proteínas. El ATD fue impulsado en gran medida por Carlsson (2009). En este artículo se exponen ventajas de la disciplina y aplicaciones

en reconocimiento de patrones. Trabajos más actuales incluyen a (Chowdhury and Mémoli, 2016), quienes estudian redes. Giusti et al. (2015) estudian el cerebro a través de la topología cliqué. Perea and Carlsson (2014) estudian texturas de imágenes basados en la topología de la botella de Klein.

Diversas escuelas han logrado avances significativos en el desarrollo del ATD. Por ejemplo, el proyecto DATASHAPE en INRIA, dirigido por F. Chazal en Francia, o el proyecto TopStat de Carnegie Mellon University, dirigido por L. Wasserman. Otro aporte significativo es el paquete TDA en R dirigido por B. Fassy de la Universidad de Montana en EEUU, (Fasy et al., 2014). Esta evolución y crecimiento del ATD ha dado lugar a un fuerte interés en CIMAT por entender estos temas que están en la frontera del conocimiento. Por lo tanto, resulta relevante estudiar con mayor profundidad el ATD.

En este trabajo entenderemos la Ciencia de Datos (CD) como una disciplina que tiene el objetivo de analizar datos complejos (complejos por su volumen o variedad) utilizando técnicas convencionales de análisis multivariado. Los problemas abordados por la CD abarcan bondad del ajuste, clasificación, regresión y análisis de componentes principales, entre otros. La evolución y el interés en la CD se ha visto incrementado con la llegada de las computadoras; los problemas estadísticos han aumentado en tamaño y complejidad a lo largo del tiempo. Entre ellos podemos destacar el análisis de conglomerados para estudiar la conducta humana, ver Tryon (1939) y Cattell (1943), y la clasificación de números escritos a mano, el trabajo con base de datos el MNIST, ver LeCun et al. (1998). Un compendio amplio de antecedentes puede consultarse en Hastie et al. (2009). Es de nuestro interés enriquecer la CD mediante técnicas basadas en ATD.

Tanto el ATD como la CD buscan extraer nuevos conocimientos de datos complejos. Esta relación da lugar a preguntas naturales entre ambas disciplinas. Una primera interrogante es cómo resolver problemas de CD empleando herramientas de ATD. Esta cuestión plantea identificar escenarios donde se puedan utilizar técnicas de ambas disciplinas y luego describir cómo se resolvería el problema con ATD. Una segunda interrogante consiste en comparar, en los casos identificados, si la técnica

de ATD tiene mejor desempeño. Esta comparación demanda diseñar un experimento en donde ambas técnicas puedan examinarse y evaluar numéricamente qué método es superior en cada caso. El objetivo de este trabajo es responder a estas preguntas.

El aporte de esta tesis es dar un rol novedoso al ATD a través de las curvas de Betti. Estas curvas son estadísticas funcionales que resumen el comportamiento de los números de Betti de una nube de datos. El uso de las curvas de Betti no es muy generalizado debido a que se pierde parte de la información topológica. En este trabajo se verá que la información que se pierde resulta irrelevante para los problemas abordados. Las curvas de Betti proporcionan rapidez en los cálculos y simplicidad en la interpretación.

Como nota de interés, otras preguntas planteadas en torno al ATD han sido abordadas por tesis de estudiantes de la licenciatura en matemáticas de la Universidad de Guanajuato y de la maestría en probabilidad y estadística de CIMAT. Una interrogante referente a si el ATD puede dar nueva información ha sido parcialmente respondida por Ibarra-Rodríguez (2015), quien trabaja en filogenética, y por González-Cucurachi (2016), quien aborda el estudio de nichos ecológicos. Otra pregunta referente a la velocidad de estos métodos han sido respondida por Peréz-Angulo (2016), quien ha considerado diversos arreglos de datos y algoritmos para medir su tiempo de computo.

1.1 Objetivo

De las interrogantes que surgen del contexto mencionado, son de interés particular las siguientes dos preguntas:

1. ¿Cómo se puede comparar el uso de las curvas de Betti con respecto a herramientas estándar de CD en escenarios específicos?

2. ¿En qué medida puede ser mejor utilizar las curvas de Betti en lugar de otras herramientas?

La primera pregunta implica identificar y describir estos escenarios. Éstos se buscan de tal forma que permitan un análisis comparativo entre ATD y CD. Luego de

una búsqueda amplia se consideraron dos situaciones: la primera, estudiando el problema de bondad del ajuste de procesos puntuales, y la segunda, analizando el problema de clasificación de objetos 3D. Estos escenarios están sujetos a implementación mediante un experimento de simulación y cuantificación a través de alguna medida de rendimiento.

En el primer caso, es de interés probar la hipótesis de aleatoriedad espacial completa en procesos puntuales (Cressie, 2015). Se utilizan las curvas de Betti para construir una estadística de prueba. Luego se utilizan técnicas de Monte Carlo (Hope, 1968) para construir una prueba de hipótesis. El enfoque estándar de procesos puntuales se estudió en CNAM-París con ayuda de el Dr. Avner Bar-Hen durante una estancia de investigación. Este enfoque incluye probar la hipótesis de aleatoriedad espacial completa con técnicas como la prueba de Quadrat (Cressie, 2015) y la prueba MAD (Baddeley et al., 2014). En este escenario se cuestiona si es posible tener una prueba más potente utilizando técnicas de ATD.

En el segundo caso, es de interés clasificar objetos 3D. Estos objetos 3D son nubes de datos muestreadas de manera uniforme sobre esferas, cubos y botellas de Klein. Se utilizan las curvas de Betti para construir un algoritmo de clasificación. Para este algoritmo se introduce el concepto de vectores de Betti, explicado en el Capítulo 3. El enfoque "usual" en Ciencia de Datos consiste en tomar los datos, sin transformarlos, para realizar la clasificación. En ambos casos se clasifica utilizando la máquina de soporte vectorial (Hastie et al., 2009). Como parte del objetivo, se quiere investigar qué tan buen clasificador resulta considerar las características topológicas. Además, es de interés saber si ocurre un sobreajuste, esto es, si el algoritmo es bueno tanto para los datos de entrenamiento como para otros datos desconocidos.

En ambos casos se diseñan y ejecutan experimentos de simulación. En el problema de bondad del ajuste, hacemos variar que tan agrupados o separados están los puntos considerando distintos procesos puntuales. Para el caso de clasificación, se comparan por parejas las figuras: esfera, cubo y botella de klein. Los escenarios se escogieron de tal forma que fuesen sencillos para ilustrar el uso de las curvas de Betti en problemas de Ciencia de Datos pero a la vez complejos para no dar lugar a un

ejemplo trivial.

Con estos experimentos se pretende medir el aporte de cada técnica en los escenarios identificados. En el primer caso estudiamos la bondad del ajuste a través de la potencia de la prueba. En el segundo caso utilizamos el error de clasificación estimado por validación cruzada. Estas medidas de rendimiento permiten estudiar la contribución de técnicas de ATD en CD, respondiendo a la segunda pregunta.

En resumen, en esta tesis se plantean dos grandes objetivos:

1. Identificar y describir escenarios comparativos en problemas estadísticos específicos para evaluar técnicas de ATD y CD.

2. Diseñar y ejecutar experimentos de simulación para cuantificar el aporte de técnicas de ATD en CD.

Con estos objetivos se busca entender el aporte que tienen las curvas de Betti con respecto a técnicas de CD. A continuación se describen y definen algunas nociones básicas de topología algebraica y análisis topológico de datos. Éstas pueden ser omitidas en una primera lectura y retornar a ellas si es necesario.

1.2 Antecedentes técnicos

En esta sección se describen de manera elemental conceptos utilizados en el desarrollo de esta tesis. Se presentan algunas ideas de topología algebraica que forman la base del ATD, como la homología persistente. Para finalizar, se introducen las curvas de Betti, concepto fundamental en este trabajo.

Las curvas de Betti se presentan en la Sección 1.2.3. Éstas resumen la topología de una nube de datos. Aunque este concepto ha sido tratado por diversos autores como Van de Weygaert et al. (2011) o Giusti et al. (2015), éstos se enfocan en la aplicación a problemas en astronomía y neurología. Estos autores no consideran la componente estocástica del fenómeno, que ocurre dado que la nube de datos es una muestra aleatoria. En esta tesis, las curvas de Betti tienen un enfoque más estadístico. Se introduce la noción de curva de Betti media y curva de Betti promedio. Como se verá,

esta noción resulta útil para resolver los problemas de bondad del ajuste y de clasificación, mencionados en los objetivos.

Se supone un nivel intemedio de estadística y topología. Se sugiere Wasserman (2013) como lectura inicial para la parte estadística. Hastie et al. (2009) es una buena referencia para entender el problema de clasificación. Se recomienda ampliamente leer las notas del curso de ATD impartido en CIMAT en enero de 2016 (Biscay et al., 2016), disponibles en la página atd.cimat.mx. Estas notas ofrecen una introducción completa y accesible de la disciplina de ATD, además de proporcionar una importante lista de referencias bibliográficas.

1.2.1 Nociones de topología algebraica

La topología algebraica es una rama de las matemáticas que tiene como objetivo encontrar invariantes algebraicos que clasifiquen espacios topológicos. Utiliza herramientas del álgebra abstracta. Los conceptos estudiados en este resumen incluyen: simplejos, independencia geométrica, complejo simplicial, grupos de cadenas, operadores de frontera, grupo de ciclos, grupo frontera, grupo de homología y números de Betti. Se sigue una línea de conceptos similar a Munkres (1984):

Los simplejos son los bloques fundamentales con los que se construye la topología simplicial de un conjunto de datos. Estos bloques abarcan espacios como puntos, segmentos de linea, triángulos o tetraedros. Un complejo es construido a partir de los simplejos "pegándolos" de manera especial. Decimos que un conjunto $\{a_0, \ldots, a_p\}$ es geométricamente independiente, o que tiene independencia geométrica, si los puntos $a_1 - a_0, \ldots, a_p - a_0$ son linealmente independientes. Por ejemplo, dos puntos distintos en $\mathbb{R}^n$ forman un conjunto geométricamente independiente, tres puntos no colineales, cuatro puntos no coplanares y así sucesivamente.

Sea $\{a_0, \ldots, a_p\}$ un conjunto geométricamente independiente en $\mathbb{R}^n$. Definimos un p-simplejo σ generado por $a_0, \ldots, a_p$ como el conjunto de todos los puntos en $\mathbb{R}^n$,

tales que

$$x = \sum_{i=0}^{p} t_i a_i, \qquad \text{donde} \qquad \sum_{i=0}^{p} t_i = 1, \qquad t_i \geq 0.$$

Denotaremos un p-simplejo como $[a_0, a_1, \ldots, a_p]$. En la Figura (1.1) se muestran los p-simplejos para $p = 0, 1, 2, 3$ en $\mathbb{R}^3$.

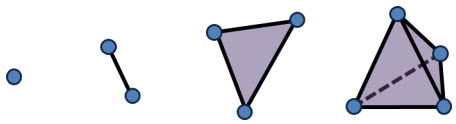

Figura 1.1: De izquierda a derecha, p-simplejos de orden $p = 0, 1, 2, 3$ en $\mathbb{R}^3$. Imagen tomada de Wright (2016).

Un complejo simplicial K en $\mathbb{R}^n$ es una colección de simplejos en $\mathbb{R}^n$ tales que: toda cara de K está en K y la intersección de cualesquiera dos simplejos de K es cara de cada uno de ellos. Un ejemplo de complejo simplicial puede verse en Figura (1.2).

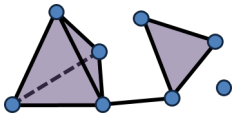

Figura 1.2: Ejemplo de un complejo simplicial. Imagen tomada de Wright (2016).

El grupo de p-cadenas es un espacio lineal formado por una combinación lineal de simplejos de orden p de un complejo simplicial K con coeficiente en {0,1}, esto es, el grupo de p-cadenas denotado por $C_p(K)$ es,

$$C_p(K) = \left\{ c = \sum_{i=1}^{k} r_i \sigma_i : \qquad r_i = 0, 1 \right\}.$$

La frontera de un simplejo es una función lineal que va de $C_p(K)$ a $C_{p-1}(K)$, tal que $\sigma \mapsto \partial(\sigma)$. Donde σ es un simplejo y

$$\partial(\sigma) = \sum_{i=0}^{p} [v_0, \ldots, \hat{v}_i, \ldots, v_p],$$

donde $[v_0, \ldots, \hat{v}_i, \ldots, v_p]$ es el $(p-1)$-simplejo sin el vértice v_i. El grupo de ciclos se define como

$$Z_p(K) = \ker(\partial : C_p \to C_{p-1}) = \{c \in C_p(K) : \partial(c) = 0\}.$$

El grupo frontera se define como

$$B_p(K) = \text{Im}(\partial : C_{p+1} \to C_p) = \{x \in C_p(K) : \exists c' \in C_{m+1}(K), \partial(c') = c\}.$$

La siguiente relación se satisface para cada p:

$$B_p(K) \subset Z_p(K) \subset C_p(K).$$

El grupo de homología es el grupo cociente entre el grupo de ciclos y el grupo frontera. Este grupo esta formado por clases de equivalencia que contienen la información de las características topológicas de K,

$$H_p(K) = \frac{Z_p(K)}{B_p(K)}.$$

Los números de Betti representan las dimensiones de los grupos de homología. Estos son utilizados en la topología algebraica para distinguir entre espacios topológicos basándose en la conectividad y agujeros n-dimensionales. De manera intuitiva miden el número de componentes conexas, ciclos y hoyos de un complejo simplicial K.

Tipos de complejos simpliciales

En el ATD se utilizan diferentes tipos de complejos simpliciales que sirven para estudiar la forma de los datos. El complejo de Vietoris Rips es el más conocido en la literatura dado que es sencillo de calcular (Zomorodian, 2005). Sin embargo, el tiempo en computarlo es considerable para muestras grandes: para la fecha en que se escribió esta tesis, nubes de datos en $\mathbb{R}^2$ con 100 puntos requieren un tiempo cercano a 80 segundos, con el paquete TDA en R y una computadora de escritorio usual. Por otro lado, los complejos alfa tienen la ventaja de que ser mucho más rápidos en su cálculo que los de Vietoris Rips para dimensiones pequeñas (Edelsbrunner and Harer, 2010), $\mathbb{R}^2$ y $\mathbb{R}^3$. El mismo cálculo que se hizo para los complejos de Vietoris Rips se realizó en 0.04 segundos para los complejos alfa, bajo condiciones idénticas.

Por esta razón de velocidad, los complejos alfa son utilizados en el desarrollo de este trabajo.

El complejo de Vietoris Rips para un radio r de un conjunto finito de puntos S se define como

$$\text{VR}(S, r) = \{\sigma \subset S : \text{ diam}(\sigma) \leq 2r\},$$

donde σ es un simplejo y $\text{diam}(\sigma) = \sup_{x_i, x_j \in \sigma} \{d(x_i, x_j)\}$. Un ejemplo puede apreciarse en la Figura (1.3).

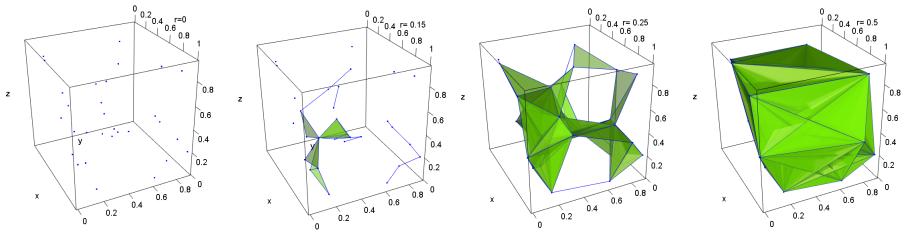

Figura 1.3: Complejo de Vietoris Rips en tres dimensiones para diferentes radios. A medida que el radio aumenta, se agregan más simplejos al complejo simplicial.

Los complejos alfa surgen de las celdas de Voronoi. Estos son estudiados en detalle en Edelsbrunner and Harer (2010). Sea S un conjunto finito de puntos en $\mathbb{R}^d$. Una celda de Voronoi de un punto $u \in S$, denotada por V_u, es el conjunto de todos los puntos que están más cerca de u que de cualquier otro puntos de S. En notación matemática esto es,

$$V_u = \left\{x \in \mathbb{R}^d | \, \|x - u\| \leq \|x - v\| \, , \forall v \in S\right\}.$$

Cada punto en S tiene asociada una celda de Voronoi; estas celdas son conjuntos convexos que representan una partición de todo el espacio, como se muestra en la Figura (1.4, izquierda). Se llama diagrama de Voronoi al conjunto de todas las celdas de Voronoi.

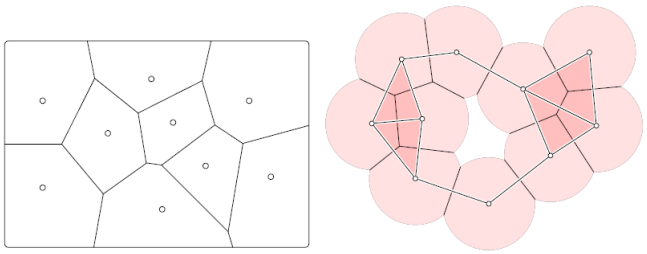

Figura 1.4: Celdas de Voronoi (izquierda). Complejo alfa (derecha). Imagen tomada de Edelsbrunner et al. (1995).

Sea S un conjunto finito de puntos en $\mathbb{R}^d$ y r un número real no negativo. Denotemos por $B_u(r)$ al disco cerrado con centro en u y radio r para cada $u \in S$. Definimos $R_u(r)$ como la intersección de $B_u(r)$ y V_u, donde V_u es la celda de Voronoi del punto $u \in S$. El complejo alfa se define como

$$\text{Alfa}(r) = \left\{ \sigma \subset S | \bigcap_{u \in \sigma} R_u(r) \neq \emptyset \right\}.$$

Por ejemplo, un 1-simplejo que conecta a los puntos u y v está en el complejo alfa de radio r de un conjunto de puntos S, si la intersección de los discos $B_u(r)$ y $B_v(r)$ y las celdas V_u y V_v es no vacía. Esto se ilustra en la Figura (1.4, derecha). En casos donde se trabaje en $\mathbb{R}^2$ y $\mathbb{R}^3$, los complejos alfa resultan mucho más rápidos de calcular. Por esta razón se utilizan en el desarrollo de esta tesis

1.2.2 Nociones de homología persistente

La homología persistente es un método de topología algebraica que estima números de Betti. Para obtener una idea del método observe la Figura (1.5): es de interés ser capaz de identificar que los datos de la figura provienen de un objeto circular. En términos de números de Betti se quiere verificar que $\beta_0 = 1$ y $\beta_1 = 1$, esto es, que tiene una componente conexa y un ciclo.

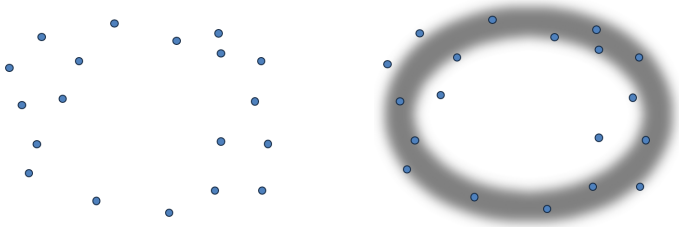

Figura 1.5: Nube de datos con forma ovalada (izquierda). Espacio topológico en el que viven los datos (derecha). Imagen tomada de Wright (2016).

Para calcular los números de Betti se conectan puntos cercanos a través de complejos simpliciales. El procedimiento se ilustra en la Figura (1.6): (1) se escoge una distancia $r \geq 0$; (2) se construye el complejo simplicial de radio r asociado a la nube de puntos; (3) se calcula la dimensión de los grupos de homología, es decir, los números de Betti. El cálculo del grupo de homología recurre a técnicas avanzadas de topología algebraica. Su explicación va más allá de los objetivos de este trabajo, pero puede decirse que está basado en el fondo en técnicas de álgebra lineal numérica. Para su estudio se recomienda leer Edelsbrunner and Harer (2010).

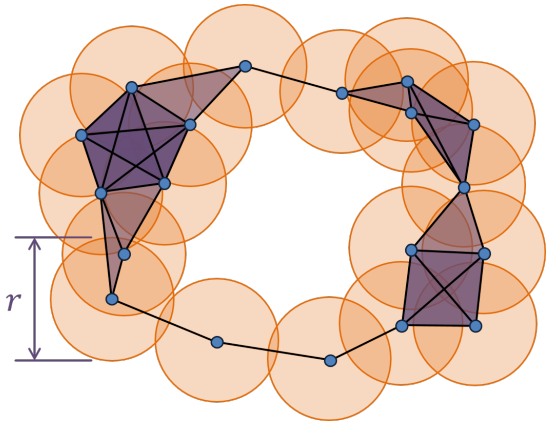

Figura 1.6: Cálculo de números de Betti para un radio determinado: (1) Escoger distancia r, (2) construir complejo simplicial y (3) calcular números de Betti. Imagen tomada de Wright (2016).

La homología persistente captura los cambios que ocurren en los números de Betti. En la Figura (1.7) se observa un agujero que nace para la distancia r_1 y que muere

para la distancia r_2. Se llama persistencia a la diferencia entre estas distancias, esto es, $r_2 - r_1$. Los nacimientos y muertes de características topológicas se representan a través del código de barras y el diagrama de persistencia.

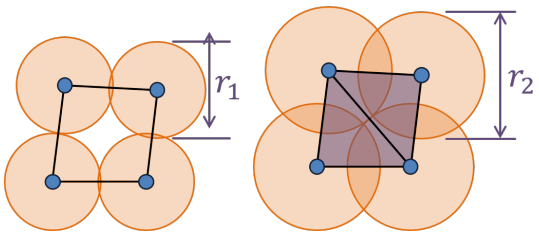

Figura 1.7: Complejo simplicial para distancias diferentes. Imagen tomada de Wright (2016).

El código de barras es la representación de los puntos de nacimientos y muertes como barras horizontales. La Figura (1.8) muestra los complejos simpliciales asociados a diferentes radios y su código de barra. Esta visualización es capaz de agrupar el comportamiento de los números de Betti para diferentes radios. Es usual llamar a las barras cortas como ruido topológico y a las barras largas como características reales del espacio de donde fueron muestreados los datos.

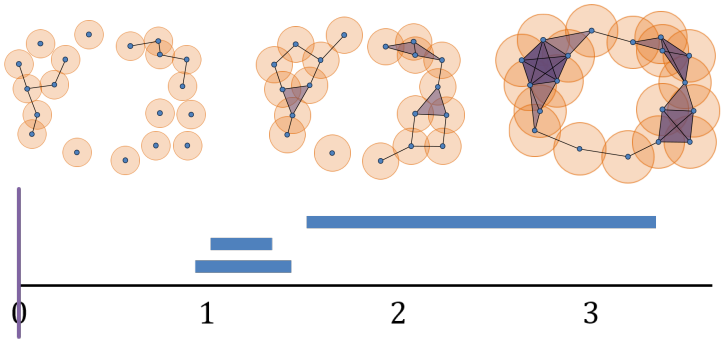

Figura 1.8: Complejos simpliciales para diferentes radios y su código de barras. Imagen tomada de Wright (2016).

El diagrama de persistencia es una descripción alternativa al código de barras. En lugar de representar los nacimientos y muertes como una barra, dibujamos el punto en un plano cartesiano, como se muestra en la Figura (1.9). El eje x corresponde a

los nacimientos y el eje y corresponde a las muertes. Los puntos que estén cercanos a la diagonal representan ruido mientras que los que estén alejados de la diagonal representan características reales del espacio ambiente.

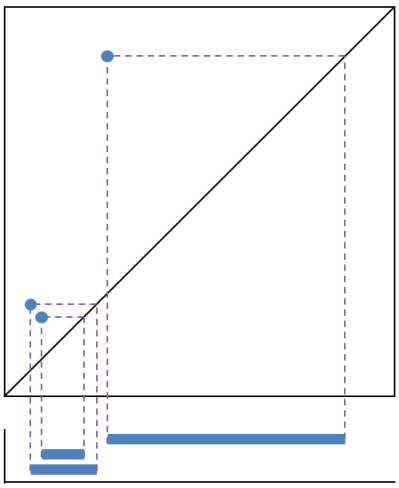

Figura 1.9: Diagrama de persistencia (arriba) y código de barras asociado (abajo). En esta figura se observa la relación entre ambas visualizaciones. Imagen tomada de Wright (2016).

1.2.3 Curvas de Betti

Las curvas de Betti son funciones que se construyen a partir del diagrama de persistencia. Se trata de funciones aleatorias debido a que dependen de una nube de puntos que sucede al azar. Éstas pueden verse como procesos estocásticos que cuentan el número de Betti, como función del radio r de un complejo simplicial. Se denotan por β_i, donde i es la dimensión del grupo de homología. Así, las curvas β_0 y β_1 corresponden al número de componentes conexas y al número de ciclos, respectivamente. Una definición formal, bajo la presunción de que se ha observado una nube de puntos y que se ha calculado el diagrama de persistencia, es la siguiente.

Definición: La curva de Betti de orden (o dimensión) i es el proceso estocástico con valores enteros indexado por $r > 0$, denotado por β_i, donde $\beta_i(r)$ es el número de Betti de orden i asociado al complejo simplicial de radio r.

Es posible obtener directamente la curva de Betti a partir del diagrama de persistencia. El valor de la curva de Betti de dimensión i para un radio r se calcula contando los puntos que corresponden a la dimensión i en el diagrama tales que los nacimientos (el eje x) son menores que r y las muertes (el eje y) son mayores que r. Los puntos en esta región corresponden características topológicas de dimensión i que han nacido pero no han muerto para un cierto radio r, es decir, que persisten para el radio r. En la Figura (1.10) se muestra el conteo que da lugar a las curvas de Betti de dimensión 1 para $r = d_1$ y $r = d_2$. Computacionalmente, ésta es una forma rápida y eficiente de calcular la curva de Betti, siempre que cuente como insumo con el diagrama de persistencia. Mientras que el diagrama de persistencia se calcula utilizando el paquete TDA del lenguaje de programación R (Fasy et al., 2014), el algoritmo para generar las curvas de Betti fue desarrollado específicamente para los fines de esta tesis, y éste también ha sido escrito en R.

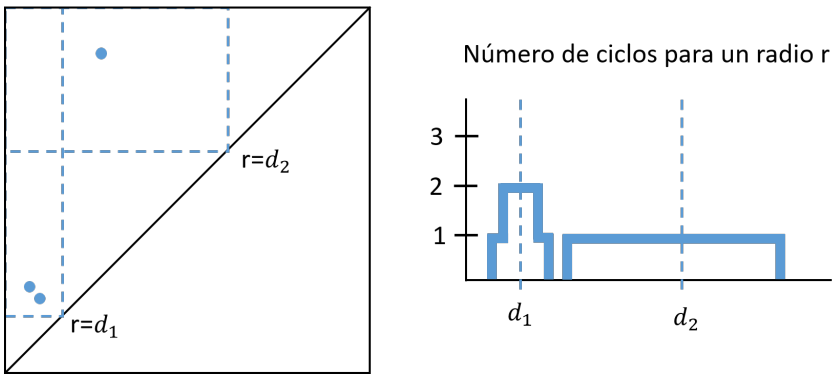

Figura 1.10: Izquierda, diagrama de persistencia. Derecha, Curva de Betti de grado 1, utilizando el algoritmo de conteo basado en regiones rectangulares dibujadas sobre el diagrama de persistencia.

El uso de las curvas de Betti tiene ventajas y desventajas. Mientras que los diagramas de persistencia recogen toda la historia de las características topologícas, las curvas de Betti pierden el seguimiento del nacimiento y muerte de cada componente. En efecto, cuando una componente muere, no es posible saber a qué nacimiento corresponde. Esto puede ocasionar que distintas nubes de datos den lugar a la misma curva de Betti. Por otro lado, la representación en términos de funciones univariadas

de las curvas de Betti hace más sencilla y conveniente el resumen de la topología de las nubes de datos. Otra ventaja es que su cálculo es mucho más rápido que otros resúmenes estadísticos como los panoramas de persistencia (Bubenik, 2015) o las funciones de rango de persistencia (Robins and Turner, 2016). Estas ventajas y desventajas resultan en una utilidad positiva para efectos de este trabajo.

La curva de Betti ha sido referida previamente en la literatura. Van de Weygaert et al. (2011) las utiliza para estudiar la topología del universo a gran escala. Bobrowski and Kahle (2014) presentan algunas propiedades de convergencia. Giusti et al. (2015) realiza una aplicación en neurología, quien da el nombre de curvas de Betti.

Curva de Betti media y curva de Betti promedio

Dado que las curvas de Betti son procesos estocásticos, podemos hablar de su valor esperado. Así, la curva de Betti media de orden i, denotada por β_i^*, es el valor esperado de la curva de Betti, esto es,

$$\beta_i^*(r) = \mathbb{E}(\beta_i(r)) = \sum_{k=0}^{\infty} k\mathbb{P}(\beta_i(r) = k), \qquad \forall r \in [0, \infty).$$

En general, esta curva media de Betti no conduce a una expresión analítica conocida. La razón es que depende explícitamente de la distribución de β_i, de quien no se conoce expresión cerrada, y depende implícitamente de la distribución de la nube de puntos, la cuál no necesariamente ha sido especificada. Por lo tanto, si la curva de Betti media ha de ser utilizada como un descriptor topológico de un tipo de nube de datos, sería útil contar con una manera de estimar la curva de Betti media.

La curva de Betti promedio, denotada por $\overline{\beta}_i$, es una estimación de la curva de Betti media, basada en una muestra de m nubes de datos. Dadas m diferentes curvas de Betti de orden i, $\beta_i^{(1)}, \ldots, \beta_i^{(m)}$, definimos la curva de Betti promedio de orden i, en el valor $r \in (0, \infty)$ como

$$\overline{\beta}_i(r) = \frac{1}{m} \sum_{k=1}^{m} \beta_i^{(k)}(r). \tag{1.1}$$

Esta curva promedio toma valores en los números reales no negativos y también es un proceso estocástico indexado por r. Se interpreta como el número esperado de características topológicas de orden i como función del radio r.

Teóricamente se obtiene una relación entre la curva de Betti media y la curva de Betti promedio en el radio r, dada por la ley de los grandes números. En efecto, esta afirma que para cada $r \in (0, \infty)$, $\overline{\beta}_i(r) \overset{\text{c.s.}}{\to} \beta_i^*(r)$. Por lo tanto, la curva promedio es una buena aproximación a la curva media siempre que m sea grande.

A diferencia de las curvas de Betti, las curvas promedios tienen formas suaves. En la Figura (1.11) se observan las curvas de Betti promedio estimadas con 1000 nubes de datos. Para $\beta_0(t)$, se aprecia cómo en promedio empezamos con tantas componentes como datos hayan, y la curva decae conforme aumenta el radio. Para $\beta_1(t)$, empieza con cero agujeros, luego aumenta a un máximo y al final vuelve a cero. Las curvas de Betti promedio ayudan a describir la topología de las nubes de datos estudiadas en conjunto.

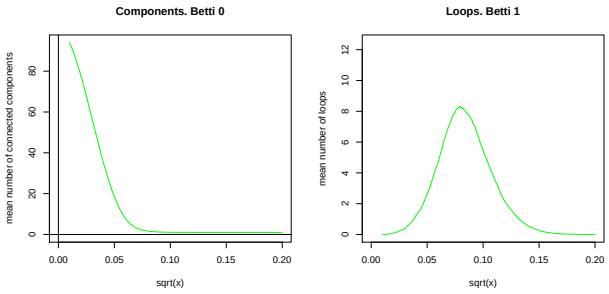

Figura 1.11: Curvas de Betti promedio para las componentes conexas (izquierda) y los ciclos (derecha) de una nube de datos.

1.3 Contenido y Estructura

Esta tesis obedece la siguiente estructura: el Capítulo 2 trata en profundidad el tema de bondad de ajuste en procesos puntuales; el Capítulo 3 aborda cómo clasificar objetos 3D. En ambos capítulos se mencionan los objetivos específicos con respecto a los problemas planteados, esto es, se establece la pregunta de interés particular en

cada caso. Luego, se da una visión general de los conceptos principales de las disciplinas de procesos puntuales y aprendizaje estadístico, respectivamente. En cada caso, se muestra la manera tradicional o convencional de resolver la problemática mencionada. Posteriormente, se plantea una solución alternativa a estas cuestiones basadas en curvas de Betti. Se busca responder a la pregunta de si el ATD es capaz de competir y superar a estas técnicas convencionales. Para esto, se diseña y ejecuta un experimento de simulación que mida el rendimiento de la solución planeada. Se concluye cada capítulo con conclusiones correspondientes a cada caso. Por último, se da una conclusión general del uso de curvas de Betti en problemas de ciencia de datos.

2

Prueba de Hipótesis para CSR

El análisis estadístico de procesos puntuales involucra, entre otras cosas, el estudio de pruebas de hipótesis para verificar supuestos como la aleatoriedad espacial completa (Cressie, 2015), o CSR por sus siglas en inglés (complete spatial randomness). Las pruebas para CSR se abordan principalmente a través de la prueba de Quadrat y la prueba MAD (Cressie, 2015). La prueba de Quadrat consiste en dividir la región de estudio en rectángulos de igual área y realizar conteos sobre cada región. La prueba MAD se basa en la función de Ripley, que estudia las distribución de las distancias entre puntos (Ripley, 1976). Estas pruebas se estudian en la Sección (2.2.2). En este capítulo se propone una nueva prueba para CSR basada en curvas de Betti. Ésta aprovecha la información topológica para discernir entre realizaciones de procesos puntuales.

Los procesos puntuales son conjuntos de puntos distribuidos al azar sobre un espacio métrico, como la recta real o el plano cartesiano (Daley and Vere-Jones, 2003). Éstos se utilizan como modelos estadísticos en el análisis de patrones observados de puntos. Tienen gran importancia para ciencias como la biología y la astronomía.

Pueden modelar eventos de interés que abarcan desde la ubicación de árboles en un bosque hasta estrellas en el cielo (Baddeley et al., 2006). Los procesos puntuales son los objetos de estudio en está sección; es de interés verificar la hipótesis CSR en procesos de distintos tipos.

Los procesos puntuales se pueden clasificar según las interacciones que existen entre sus puntos. Los procesos agregativos tienden a tener sus eventos en racimos (Illian et al., 2008), como cúmulos estelares donde las estrellas se encuentran juntas. Los puntos de los procesos repulsivos se encuentran alejados entre si y es muy raro encontrarlos en grupos (Illian et al., 2008), como en especies de árboles grandes cuyas raíces impiden el crecimiento de árboles semejantes en espacios cercanos. Por último, los procesos aleatorios son aquellos que no presentan un patrón definido (Cressie, 2015). Tienen grupos creados por azar, como los procesos de Poisson. Estos tres tipos de procesos serán de ayuda a la hora de definir los escenarios que utilizaremos para comprobar los beneficios de la prueba que se diseñará para probar CSR.

La importancia de la hipótesis CSR radica en que es un paso primordial en cualquier análisis de procesos puntuales. Para efectos de este trabajo, esta hipótesis se define como en Cressie (2015). Bajo esta definición, un proceso que satisface CSR es equivalente a un proceso de Poisson. Si no se rechaza CSR, no tenemos información para decir que el proceso estudiado tenga un comportamiento distinto al de un proceso de Poisson, y un análisis más hondo carecería de sentido. Si se rechaza CSR, quiere decir que el proceso tiene una estructura distinta a la de un proceso de Poisson, dando lugar a estudios más complejos como indagar sobre la repulsividad o la formación de grupos del proceso. Es de interés obtener pruebas que disciernan con mayor eficiencia cuándo un proceso satisface CSR.

En este capítulo se verá cómo surgen diferentes curvas de Betti a partir de procesos puntuales agregativos, repulsivos y aleatorios. Con esto, se generará una estadística de prueba que puede discernir entre patrones puntuales distintos. En el caso repulsivo la prueba basada en ATD es más potente que las pruebas de Quadrat y MAD, como se verá más adelante. Por lo tanto, tiene sentido pensar que técnicas topológicas pueden ser utilizadas para estudiar el comportamiento de procesos pun-

tuales.

Por último, es pertinente acotar que este capítulo surge como un trabajo en conjunto entre el CIMAT y el CNAM en París. Se trabajó en conjunto con el Dr. Avner Bar-Hen, quien aporto muchos de los detalles relacionados con los procesos puntuales; muchas ideas surgen de su artículo Bar-Hen et al. (2015). Esta colaboración se inicia durante el primer congreso en Ciencia de Datos entre Francia y México en la Ciudad de México en noviembre de 2016 organizada por la Universidad Nacional Autónoma de México. Este congreso dio lugar a una estancia de investigación en París en febrero de 2017.

2.1 Objetivos

Las preguntas que surgen en este contexto son:

1. ¿Cómo se puede plantear una prueba de hipótesis con técnicas de ATD?
2. ¿Cómo mejora la potencia de esta prueba respecto a las pruebas usuales de la literatura?

La primera pregunta requiere proponer una prueba de hipótesis para CSR utilizando técnicas de ATD. En la construcción de la estadística de prueba se estima por simulación estocástica la curva de Betti promedio bajo CSR. Ésta es utilizada para diseñar una medida que sea chica cuando el proceso satisface la hipótesis de aleatoriedad espacial completa y grande en otro caso. La región crítica se estima con técnicas de Monte Carlo (Hope, 1968).

La segunda cuestión requiere cuantificar el rendimiento de la prueba. Para esto se diseña y ejecuta un experimento que mida el efecto sobre la potencia de la prueba, es decir, en la probabilidad de discernir correctamente que un proceso no es CSR. El experimento pretende comparar el desempeño entre las pruebas Quadrat, MAD y la prueba basada en curvas de Betti. Su diseño considera procesos que cambian según las interacciones, es decir, procesos agregativos, repulsivos y aleatorios. En todos los casos el tamaño de muestra es fijo, ya que estamos interesados en el efecto de los procesos en la potencia. Este problema ha sido mencionado por Robins and Turner

(2016) y Biscio and Møller (2016). Estos autores han obtenido, mediante otros resúmenes topológicos, que los procesos de tipo repulsivo son más fáciles de distinguir que los procesos agregativos. En nuestro caso, se espera obtener un resultado similar.

Los objetivos específicos de este capítulo son:

1. Construir una prueba de hipótesis para CSR utilizando curvas de Betti.
2. Diseñar y ejecutar un experimento de simulación para estudiar la potencia de la prueba del objetivo 1 y compararla con las pruebas Quadrat y MAD.

En la siguiente sección se describen algunos detalles técnicos, como conceptos básicos de teoría de procesos puntuales y teoría de pruebas de hipótesis. En la Sección (2.3) se construye la prueba de hipótesis basada en curvas de Betti. En la Sección (2.4) se diseña y ejecuta el experimento de simulación. Por último, se dan algunas conclusiones y trabajo futuro.

2.2 Antecedentes generales

Esta sección resume qué son los procesos puntuales. Se describen los principales patrones puntuales utilizados en este trabajo. Se da una idea de cómo funcionan describiendo sus parámetros. Por último, se describen las nociones básicas de pruebas de hipótesis estadísticas, las pruebas Monte Carlo y las pruebas para CSR de la literatura estándar. Esta sección sirve de referencia al lector para abordar la estadística de prueba, la región crítica y el experimento de simulación. Estos elementos serán descritos en secciones posteriores de este trabajo.

2.2.1 Teoría de procesos puntuales

Un proceso puntual espacial es un patrón aleatorio de puntos en un conjunto d-dimensional. Existen tres principales caracterizaciones para un proceso puntual: a través de las distribuciones finito dimensionales, por las probabilidades de encontrar vacíos en el proceso y a través de la función generadora de probabilidades. Los tres enfoques son equivalentes y pueden consultarse a detalle en (Daley and Vere-Jones, 2007). Comúnmente, las realizaciones de patrones puntuales representan la

ubicación de grupos de objetos de estudio, como árboles en un bosque o estrellas en el Universo (Cressie, 2015). En dimensión uno, la ubicación en el espacio es identificable con la variable tiempo, por lo que los procesos puntuales son extremadamente útiles para estudiar fenómenos que tienen que ver con ocurrencias en el tiempo (fallas, muertes, reclamos de seguros, etc.). En lo general, Los procesos puntuales son útiles como modelos estadísticos en el análisis de nubes de puntos.

Los procesos puntuales pueden agruparse en aleatorios, agregativos y repulsivos, según las interacciones entre sus puntos. Cada modelo tiene una estructura topológica diferente. Se aprovechará esta discrepancia para mostrar las capacidades de la curva de Betti en discernir entre diferentes tipos de patrones puntuales. A continuación, se describen de manera general los patrones puntuales de interés.

Procesos aleatorios

Los procesos aleatorios son aquellos en los que las interacciones ocurren de manera totalmente al azar, es decir, sus puntos no presentan tendencias repulsivas o atractivas. Un proceso aleatorio y un proceso de Poisson son equivalentes, siguiendo el enfoque de Cressie (2015). De esta forma, La prueba de hipótesis CSR consiste en inferir si un proceso puntual es de Poisson. En este sentido, este tipo de proceso es el punto de partida para nuestro estudio.

Durante el desarrollo de este trabajo, se requerirá simular realizaciones de procesos aleatorios. El proceso binomial es un proceso de Poisson homogéneo donde el número de puntos es fijo, digamos n. Éste es llamado así porque la distribución del número de puntos en una región W, subconjunto de la región de estudio S, tiene distribución binomial con n intentos y probabilidad de éxito $p = \text{área}(W)/\text{área}(S)$. Es posible generarlo a través de variables aleatorias uniformes (Streit, 2010). Su rápida simulación permitirá estimar la distribución de la estadística de prueba y la región crítica en el procedimiento de pruebas de hipótesis, descritos más adelante. En la Figura (2.1) se muestran realizaciones de procesos binomiales con $n = 50$ y $n = 100$.

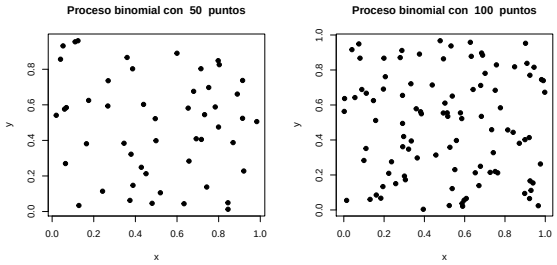

Figura 2.1: Procesos binomiales para $n = 50$ (izquierda) y $n = 100$.

Procesos agregativos

Los procesos agregativos están formados por grupos de puntos. En este tipo de patrones, la densidad puntual no es uniforme y varía significativamente en el región de estudio. En algunas aplicaciones es de interés identificar el centro de los conglomerados. En esta tesis, se utiliza el proceso de Matern por ser sencillo y flexible. Este proceso representa parte de los casos alternativos en los cuales se explora qué tan buena es la prueba de hipótesis que formularemos más adelante. Un catálogo amplio de modelos para procesos agregativos se ofrece en Illian et al. (2008).

El proceso de Matern se forma en dos etapas, de padres y de hijos. Primero, se genera un proceso de Poisson, de padres, de tasa κ. Los puntos de este proceso sirven como centro para el siguiente proceso de Poisson, de hijos, de tasa μ. El proceso de hijos se genera sobre una bola de radio r y centro en los puntos padres. Se genera un proceso diferente para cada punto padre. Finalmente, se retiran los puntos padres y lo que queda es una realización del proceso de Matern con parámetros κ, μ y r. En la Figura (2.2) se muestran realizaciones del proceso de Matern para $r = 0.2$ y $r = 0.3$.

Con el parámetro r es posible tener una idea del efecto del radio en el cambio de la topología. En el comportamiento de las características topológicas de este proceso, se ha observado que las componentes conexas y los ciclos pequeños mueren más rápidamente que en un proceso aleatorio. Por otro lado, se visto que los ciclos largos persisten más. Estas observaciones están basadas en simulación estocástica de procesos de Matern. Las simulaciones fueron realizadas con ayuda del lenguaje R y los paquetes TDA y spatstat. En la Figura (2.4) se muestran las curvas de Betti prome-

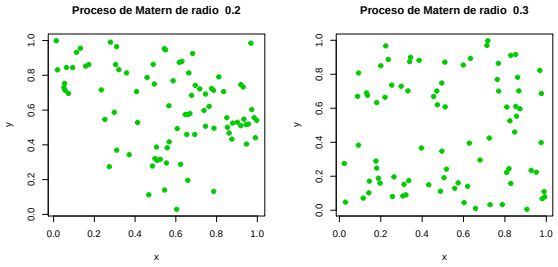

Figura 2.2: Procesos agregativos para radio $r = 0.2$ (izquierda) y radio $r = 0.3$ (derecha).

dio para los tres tipos de procesos trabajados en esta tesis, en verde para el proceso de Matern. Para estudiar con más detalle los procesos agregativos y de Matern se recomienda leer a Illian et al. (2008).

Procesos repulsivos

Los procesos repulsivos son procesos en donde los puntos tienden a alejarse entre si, y también son llamados procesos de inhibición o prohibición Illian et al. (2008). Parte de su densidad está formada por la función de interacción: esta función mide la cercanía de una pareja de puntos al proceso puntual, como la función $\exp\{-||x-y||\}$, donde $x, y \in \mathbb{R}^2$. Mientras x e y estén más alejados, menor peso tendrá la función de interacción. Un ejemplo en el que se da este tipo de procesos es en árboles grandes en un bosque. Sus raíces forman un radio en el cual no puede crecer vegetación semejante. Así, éstos tienden a repelerse entre sí y es muy poco probable encontrar grupos. El proceso más conocido para modelar este tipo de fenómenos es el proceso de Strauss; ver Strauss (1975). Para esta tesis, los procesos repulsivos representan parte de los casos alternativos con los cuales se investiga qué tan efectiva es la prueba de hipótesis que formularemos más adelante.

El proceso de Strauss es atractivo por la interpretación de sus parámetros. El radio de interacción, denotado por R, es una distancia para la cual la probabilidad de encontrar parejas de puntos es baja. Esta incertidumbre se controla a través del nivel de repulsividad, denotado por γ. Éste puede ir de cero a uno: cuando $\gamma = 1$ no existe repulsión, y el proceso es totalmente aleatorio; cuando $\gamma = 0$ no se permite que

ningún punto esté a una distancia menor al radio de interacción de cualquier otro punto. Por último, β es un parámetro de intensidad y controla la cantidad de puntos que genera el proceso. En la Figura (2.3) se muestran realizaciones de procesos repulsivos para $\gamma = 0.5$ y $\gamma = 0.1$.

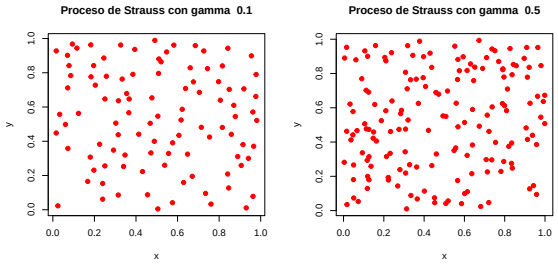

Figura 2.3: Procesos repulsivos para $\gamma = 0.1$ y $\gamma = 0.5$.

Este proceso permite estudiar el efecto de la repulsividad en la topología del proceso puntual a través del parámetro γ. En cuanto al comportamiento de las características topológicas de este proceso, se ha observado que las componentes conexas persisten un poco más que en el proceso aleatorio. Por otro lado, se ha visto que los ciclos se encuentran en mayor cantidad para radios pequeños, pero que no persisten luego de cierto radio. Esto posiblemente esté relacionado al radio de interacción. Estas observaciones están basadas en simulación estocástica de procesos de Strauss. Las simulaciones fueron realizadas en el lenguaje R y los paquetes TDA y spatstat. En la Figura (2.4) se muestran las curvas de Betti promedio para los tres tipos de procesos trabajados en esta tesis, en rojo para el proceso de Strauss.

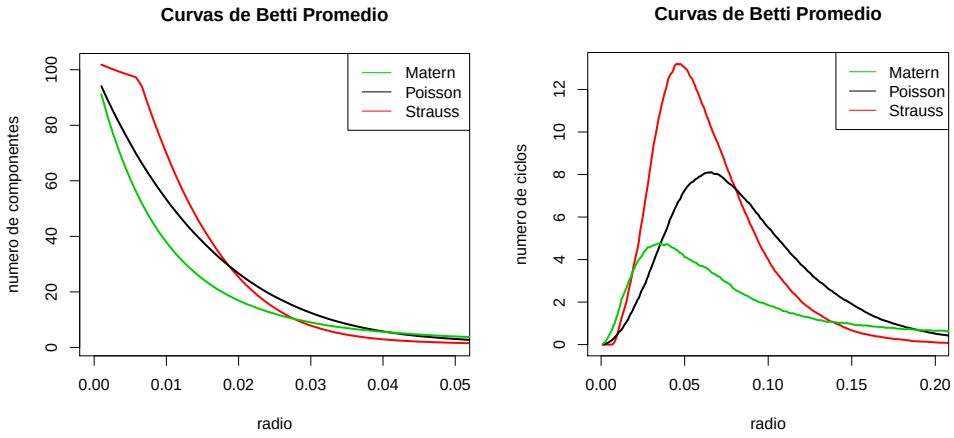

Figura 2.4: Curvas de Betti promedio para los tres tipos de proceso: agregativo (verde), repulsivo (rojo) y aleatorio (negro). A la izquierda se encuentra la curva correspondiente a componentes conexas y a la derecha la curva correspondiente a los ciclos

2.2.2 Teoría de pruebas de hipótesis

El problema de prueba de hipótesis surge de preguntarse acerca de la plausibilidad de que un modelo de probabilidad dado, denotado por H_0, rija el fenómeno de interés, con base en una observación empírica del mismo fenómeno. Una de las formulaciones clásicas que existen para abordar esta pregunta consiste de plantear un contexto binario: con base en la información contenida en la muestra, o se rechaza H_0, o no se rechaza. Esta formulación se debe a Neyman and Pearson (1933), y será el enfoque adoptado en este trabajo. Para una exposición moderna de este enfoque, ver Lehmann and Romano (2006). Las premisas que se adoptan son las siguientes:

- La inferencia estadística tiene como fin una decisión conductual binaria.
- Existe una hipótesis nula, que es la hipótesis que se desea probar, y se denota por H_0.
- Existe una hipótesis alternativa, que es la hipótesis a favor de la cual posiblemente se rechaza H_0 como función de los datos, y que se denota por H_1.
- El contexto binario da lugar a errores de tipo I y II. El error de tipo I se comete cuando se resuelve rechazar H_0 a favor de H_1 siendo que H_0 es cierta. El error

de tipo II se comete cuando se resuelve no rechazar H_0 cuando H_0 es falsa.

- El error tipo I es conceptualmente más grave que el de tipo II.
- El usuario provee la probabilidad tolerable para cometer error de tipo I, denotado por α. En nuestro caso se ilustra con $\alpha = 0.05$. Este α se especifica dependiendo de la aplicación para la cual sea necesario. La teoría de pruebas de hipótesis de Neyman-Pearson procede a minimizar el error de tipo II sujeto a que el error de tipo I ha sido acotado por el usuario.

Siguiendo esta linea de pensamiento, una prueba de hipótesis consiste de dos cosas: (i) elegir una estadística de prueba $T = T(X_1, \ldots, X_n)$, y (ii) determinar un subconjunto de valores posible de T, llamado región crítica, C, de la prueba. En nuestro caso la región crítica es de la forma $C = (c, \infty)$, donde c es llamado valor crítico. La regla a utilizar es: rechazar H_0 si y sólo si $T \in C$, esto es, si $T > c$.

Una prueba se dice significativa si $T > c$; en caso contrario se dice que es no significativa. Esto da lugar al concepto de nivel de la prueba, que se ha definido como la probabilidad de cometer error de tipo I. La potencia de la prueba se define como uno menos la probabilidad de cometer error de tipo II. Una prueba es mejor que otra cuando, una vez determinado un nivel de significancia α, tiene mayor potencia. La prueba de hipótesis queda caracterizada por la estadística de prueba T y el punto crítico c.

Otro enfoque importante es el planteado por Fisher (1925), en donde la inferencia no tiene como fin una decisión conductual sino ponderar la plausibilidad de una hipótesis utilizando medidas de discrepancia como el p-valor, ver Fisher (1925) y Sprott (2008). Para el enfoque bayesiano ver Bernardo and Rueda (2002). Se adopta el enfoque Neyman-Pearson por ser más conocido en la literatura de ecología y biología y porque tiene medidas de cuantificación ampliamente conocidas como lo es la potencia de la prueba.

Pruebas de hipótesis de Monte Carlo

Las pruebas de Monte Carlo se utilizan cuando no es posible determinar la distribución de la estadística de prueba bajo la hipótesis nula de manera teórica. En

el desarrollo de la prueba de hipótesis surgirá este inconveniente. Por lo tanto, las técnicas de Monte Carlo serán útiles para aproximar la región de rechazo. Este procedimiento alternativo recurre a simulación estocástica para estimar el valor crítico. Las pruebas de Monte Carlo fueron descritas por primera vez por Hope (1968). Para el uso de pruebas de Monte Carlo en procesos puntuales se recomienda leer Ripley (1977).

Las pruebas de Monte Carlo requieren dos acciones principales: (i) calcular una estadística de prueba, y (ii) simular datos que satisfagan la hipótesis nula. En nuestro caso los datos son los procesos puntuales y se simulan de tal suerte que satisfagan CSR. El objetivo es utilizar simulación estocástica para estimar el valor crítico c. Al momento de tener un juego de datos a evaluar, se rechaza si el estadístico de prueba evaluado con la muestra observada, digamos T, es mayor que c.

El procedimiento para realizar la prueba de Monte Carlo es el siguiente. Se generan m conjuntos de datos simulados, es decir, m procesos puntuales que satisfagan CSR. Luego de esto, se calcula para cada proceso puntual la misma estadística de prueba T, que en nuestro caso, está relacionada a las curvas de Betti de los datos (se definirá en la Sección 2.3.1). Se obtienen $t_1, \ldots, t_m$ con los cuales se estima la distribución de la estadística T utilizando la función de distribución empírica. Correspondientemente, recurrimos al cuantil empírico de nivel α para estimar el valor crítico c^*. La distribución la estadística de prueba que se desarrolla en este trabajo resulta muy compleja para calcular de manera teórica. Las pruebas de Monte Carlo son un enfoque viable dado que es posible simular con facilidad procesos CSR.

Pruebas de hipótesis estándar para CSR

Entre las pruebas clásicas para discernir la hipótesis CSR se encuentran la prueba de cuadrantes y la prueba MAD. La prueba de cuadrantes, o de Quadrats, cuenta el número de puntos en regiones rectangulares predefinidas (Cressie, 2015). Luego, se calcula una estadística de prueba basada en estos conteos. La prueba MAD calcula el valor máximo entre la función K de Ripley teórica y estimada (Baddeley et al., 2014). Ambas pruebas serán explicadas en esta sección. Éstas sirven de comparación para estudiar el rendimiento de la prueba elaborada en esta tesis.

Prueba de cuadrantes

Esta prueba es la más básica para estudiar CSR (Cressie, 2015). Sea ξ una realización de un proceso puntual en una región de rectangular S. Los cuadrantes son celdas rectangulares congruentes $A_1, \ldots, A_k$ que forman una partición de S. Por ejemplo, la Figura (2.5) muestra un caso con 16 cuadrantes.

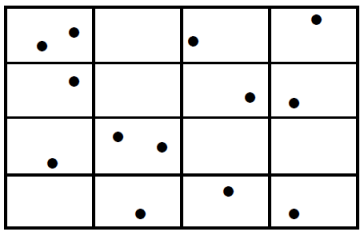

Figura 2.5: Cuadrantes y observaciones empíricas de un proceso puntual.

Un proceso que satisface CSR es equivalente a un proceso homogéneo de Poisson. Dado que los cuadrantes son conjuntos disjuntos y congruentes, la distribución del conteo para cada celda es la misma. Sea N_i la v.a. que cuenta la cantidad de puntos en el cuadrante A_i. Bajo la hipótesis CSR, $N_1, N_2, \ldots N_k$ son v.a.i.i.d. con distribución de Poisson con tasa área$(A_1) \times \lambda^*$, donde λ^* es la tasa esperada del proceso y está dada por

$$\lambda^* = \frac{n}{\text{área}(S)},$$

donde n es la cantidad total de puntos de la realización ξ.

La estadística de prueba se construye con el número esperado de puntos en cada celda, es decir, n/k. Ésta se define como

$$t = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k},$$

donde n_i es el conteo resultante en la celda A_i. Asintóticamente t tiene distribución χ^2 con $k-1$ grados de libertad bajo la hipótesis CSR. El valor critico para esta prueba viene dado por $c^* = \chi^2_{k-1,\alpha}$, donde $\chi^2_{k-1,\alpha}$ es el cuantil en donde una distribución χ^2 con $k-1$ grados de libertad acumula 1-α de probabilidad.

Esta prueba de hipótesis tiene la restricción de todos cuadrantes tienen que tener la misma área. Además, los resultados dependen del tamaño de la partición escogida. No existe una manera natural de escoger esta partición, por lo que los resultados pueden muy sensibles a esta partición. Para más detalles de esta prueba consultar la Sección 8.2.3 de Cressie (2015).

Prueba MAD

La prueba MAD (del inglés, Maximum Absulote Deviation) utiliza la desviación máxima entre la función $\hat{K}$ estimada de los datos observados y de la función K teórica bajo la hipótesis nula. La función K de Ripley es una herramienta para analizar realizaciones de procesos puntuales Dixon (2002). Esta función se define como

$$K(t) = \lambda^{-1}\mathbb{E}\left(\begin{array}{c}\text{número de eventos adicionales a}\\ \text{una distancia } t\text{de un evento arbitrario}\end{array}\right), \tag{2.1}$$

donde λ es la densidad (número por unidad de área) de eventos. La función $K(t)$ puede detectar patrones de agrupamientos en realizaciones de procesos puntuales a diferentes escalas.

El estimador, propuesto por Ripley (1976), para $K(t)$ es

$$\hat{K}(t) = \hat{\lambda}^{-1}\sum_{i}\sum_{i\neq j} w(l_i, l_j)^{-1}\frac{\mathbb{I}(d_{ij} < t)}{N},$$

donde d_{ij} es la distancia euclidiana entre los puntos i y j del proceso e $\mathbb{I}$ es la función indicadora. Los pesos $w(l_i, l_j)$ proporcionan la corrección de borde. Estos valen 1 cuando el círculo con centro l_i que pasa por l_j está totalmente contenido en S, la región de estudio. Si parte de este círculo cae afuera de la región de estudio, entonces $w(l_i, l_j)$ es la proporción de la circunferencia de ese círculo que cae la región de estudio.

La estadística de prueba es

$$T = \max_{0\leq r\leq R} |\hat{K}(r) - K(r)|,$$

donde R es un valor de tolerancia para el cual la diferencia $\hat{K}(r) - K(r)$ empieza a ser pequeña. La estadística T puede verse como la separación vertical máxima entre las

gráficas $\hat{K}(r)$ y $K(r)$ sobre el rango $[0, R]$. El valor crítico para este caso se estima por métodos de Monte Carlo, como en la Sección 2.2.2. Esta prueba es útil para discernir procesos agregativos de procesos CSR dado que el máximo es muy diferente es este caso.

2.3 Prueba para CSR con curvas de Betti

En esta sección se procede a especificar la estadística de prueba y la región crítica de la prueba de hipótesis basada en ATD. La idea de la prueba será utilizar la curva de Betti media para evaluar la discrepancia entre realizaciones de procesos puntuales con respecto a la hipótesis nula, que es la hipótesis CSR. Como se verá, en algunos casos esta prueba tiene mejor potencia que las pruebas usuales de CSR.

2.3.1 Estadística de Prueba

La estadística de prueba busca cuantificar la discrepancia entre un proceso puntual observado y la hipótesis CSR, empleando conceptos de topología. Así, esta discrepancia se cuantifica a través de ciertas integrales sobre las curvas de Betti. La estadística de prueba tiene la siguiente definición.

Definición: Sea ξ una realización de un proceso puntual, sea $\omega \in [0, 1]$ y sean β_0 y β_1 curvas de Betti asociadas a ξ. Se define la estadística de prueba como

$$d(\xi) = (1-\omega)\int_0^\infty (\beta_0(r) - \beta_0^*(r))^2\, dr + \omega \int_0^\infty (\beta_1(r) - \beta_1^*(r))^2\, dr. \quad (2.2)$$

En la práctica se necesitan dos aproximaciones para implementar esta estadística de prueba. La primera de ellas consiste en estimar la curva de Betti media con la curva de Betti promedio. La razón es que no existe un resultado teórico conocido para calcular la curva de Betti media para esta hipótesis. Como se vio en la Sección 1.2.3, la curva de Betti promedio, obtenida bajo realizaciones de un proceso CSR, es una buena estimación de la curva de Betti media. En esta tesis se consideran 1000 repeticiones para su estimación. Este valor, aunque arbitrario, provoca resultados suaves

y que cambian poco al incluir más repeticiones.

La segunda aproximación consiste en evaluar numéricamente las integrales. El límite superior depende de la curva de Betti promedio y de la curva de Betti observada; la curva de Betti observada es la curva asociada a los datos sobre los cuales queremos probar la hipótesis CSR. Para obtener el límite superior, se calcula el primer el primer valor para el cual la curva de Betti promedio se vuelve cero y no crece nuevamente; este limite se denota por $x_{\text{prom,i}}$, donde i es el grado de homología. En notación matemática esto es,

$$x_{\text{prom,i}} = \text{ínf}\{r \in \mathbb{R} : \overline{\beta}_i(x) = 0 \quad \forall x \geq r\}.$$

Repetimos este proceso con la curva de Betti observada para obtener $x_{\text{obs,i}} = \text{ínf}\{r \in \mathbb{R} : \beta_i(x) = 0 \quad \forall x \geq r\}$, donde i en el grado de homología. El límite superior x_i se define entonces como el máximo entre los valores calculados, esto es,

$$x_i = \text{máx}(x_{\text{prom}}, x_{\text{obs}}).$$

La idea es obtener un punto de corte para la evaluación de las integrales. Luego de este punto las integrales no aportan información en el sentido que tanto la curva promedio como la observada son cero.

Las curvas de Betti promedio pueden utilizarse siempre que trabajemos con la misma cantidad de puntos y una región de estudio fija. No es necesario repetir su cálculo para una nueva nube de puntos del mismo tamaño obtenida sobre la región de estudio. En nuestro caso los procesos tienen 100 puntos y suceden sobre $[0, 1] \times [0, 1]$. Por lo tanto, podemos utilizar las mismas curvas de Betti estimadas para hacer pruebas CSR con otros juegos de datos que tengan las mismas restricciones de número de datos y región de estudio. No ocurre lo mismo con los límites de integración: éstos dependen de la realización del proceso puntual que se este analizando, el cual se asume que será diferente para cada prueba.

Resumiendo lo anterior, la estadística de prueba que se utiliza en la práctica puede definirse como se hace a continuación.

Definición: Sea ξ una realización un proceso puntual, sea $\omega \in [0,1]$ y sean β_0 y β_1 curvas de Betti asociadas a ξ. Esto es,

$$\widehat{d}\,(\xi) \;=\; (1-\omega)\int_0^{x_0}\left(\beta_0(r)-\overline{\beta}_0(r)\right)^2 dr + \omega\int_0^{x_1}\left(\beta_1(r)-\overline{\beta}_1(r)\right)^2 dr. \quad (2.3)$$

Para este trabajo se consideran pesos $\omega = 0$ y $\omega = 1$. Estos corresponden a considerar por separado cada grado de homología. La pregunta sobre la ω óptima no se aborda en este trabajo por considerar que el nuestro objetivo puede abordarse con la selección mencionada de ω. Sin embargo, es una pregunta legítima.

La Figura (2.6) ilustra el área entre curvas para cada grado de homología. En la parte superior de la imagen se observa el área entre las curvas para el caso donde el proceso estudiado satisface CSR. En la parte inferior se observa el área entre las curvas para un proceso que no satisface CSR; en este caso el área es mucho mayor. Mientras mayor es el área entre curvas, mayor es el valor de la estadística de prueba. Esto puede considerarse como evidencia en contra de la hipótesis CSR.

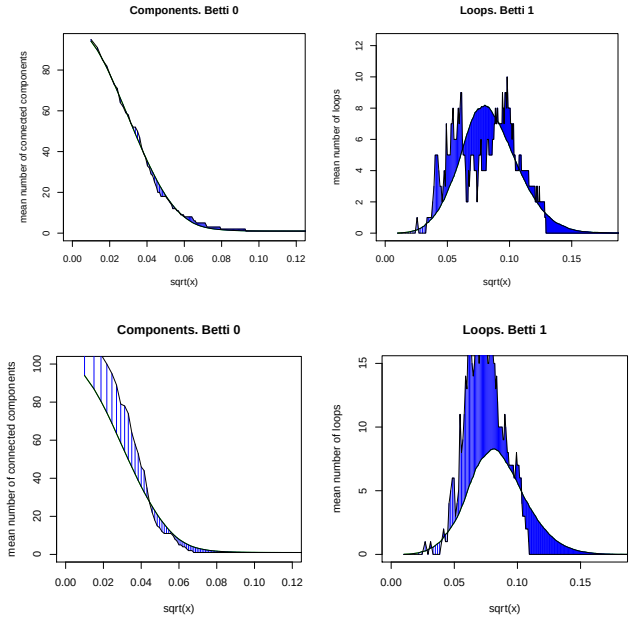

Figura 2.6: Arriba, área entre curvas para un proceso que satisface CSR. Abajo, área entre curvas para un proceso que no satisface CSR. Se muestran los resultados para las dos primeras curvas de Betti.

2.3.2 Región de crítica

La estimación de la región crítica requiere de la distribución de la estadística de prueba (2.3) bajo la hipótesis nula. No es conocido un resultado teórico para distribuciones de las curvas de Betti en procesos puntuales, y tampoco para funciones similares a nuestra estadística de prueba. Una solución es utilizar simulación estocástica para estimar esta distribución. Se emplea la distribución empírica de la estadística de prueba como estimador. El teorema de Glivenko-Cantelli, Teorema 7.4 en Wasserman (2013), garantiza la convergencia de la distribución empírica a la distribución deseada. Un tamaño de réplicas grandes produce una buena estimación de la distribución de la estadística de prueba bajo H_0.

El procedimiento de pruebas de hipótesis basadas en simulación es conocido como pruebas de Monte Carlo, explicadas en la sección (2.2.2). En este trabajo se simularon $m = 1000$ procesos binomiales, es decir, bajo la hipótesis CSR, con $n = 100$ puntos para estimar la distribución de la estadística de prueba bajo H_0. Para cada proceso se calcula la estadística de prueba (2.3), obteniendo $d_1^*, d_2^*, \ldots, d_m^*$. La distribución empírica de estos valores es una buena estimación de la distribución deseada.

El valor crítico c_α^* se define como el cuantil $1 - \alpha$ de esta distribución, es decir, el valor donde la distribución acumula $1 - \alpha$ de probabilidad. Lo anterior puede verse en la Figura (2.7). La región de rechazo queda determinada por todos los valores mayores que c_α^*. Así, si el valor de la estadística de prueba es mayor que c_α^*, entonces rechazamos que el proceso satisface CSR. En caso contrario no podemos rechazar la hipótesis de CSR.

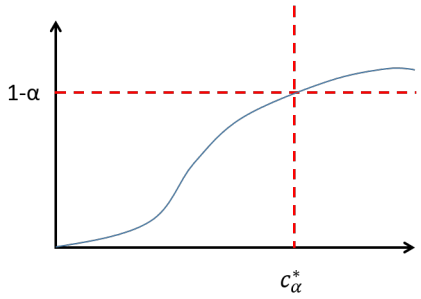

Figura 2.7: Distribución de la estadística de prueba. El valor c^*_α es el cuantil donde la distribución empírica de la estadística de prueba acumula $1-\alpha$ de probabilidad.

2.4 Experimento de simulación

En esta sección se explica el diseño y los resultados de un experimento de simulación para examinar la prueba de hipótesis diseñada en la Sección 2.3. Este experimento se contrasta con las pruebas usuales para procesos puntuales, descritas en la Sección (2.2.2). Se espera que este experimento aporte información sobre el uso de curvas de Betti en el estudio de procesos puntuales.

El objetivo del experimento es estudiar la efectividad de la prueba basada en ATD. Para esto se estima la potencia de esta prueba y de las pruebas usuales. Además, se consideran escenarios basados en los tipos de procesos puntuales. Finalmente, se hacen comparaciones de los resultados para determinar qué prueba tiene mejor rendimiento. Como se verá, la prueba de ATD tiene mayor potencia que las pruebas clásicas en el caso repulsivo, y para procesos agregativos las curvas de Betti no aportan una diferencia significativa con respecto a la prueba MAD o de cuadrantes.

2.4.1 Diseño del experimento

A continuación se describen los tratamientos, el tamaño de muestra y el proceso de análisis del experimento. Estos elementos fueron extraídos de la metodología usual en el diseño de experimento propuesta por Dean et al. (2017). En nuestro caso, las unidades muestrales pueden verse como las realizaciones de los procesos puntuales. Esta metodología se escogió para garantizar que las consideraciones per-

tinentes sean hechas antes de iniciar la simulación de datos y el análisis posterior.

Los tratamientos de este experimento están representados por los tipos de procesos involucrados: aleatorios, repulsivos y agregativos. Los procesos aleatorios se simulan a través de realizaciones de procesos binomiales. Estas relaciones se obtienen simulando dos variables aleatorias uniformes en $[0, 1]$. Con éstas se forma un par ordenado que corresponde a un punto en $[0, 1] \times [0, 1]$. La repetición de este proceso n veces garantiza tener un proceso binomial de n puntos (Robert and Casella, 2004).

Para los procesos repulsivos se simulan procesos de Strauss en dos niveles. El primer nivel es "menos repulsivo". Éste tiene parámetros $\beta = 150$, $\gamma = 0.5$ y $R = 0.05$; se denotará por S_1. El segundo nivel es "más repulsivo". Sus parámetros son $\beta = 250$, $\gamma = 0.1$ y $R = 0.05$, denotado S_2. En el proceso de Strauss, el parámetro γ controla qué tan separados están los puntos entre sí. El valor $\gamma = 1$ representa aleatoriedad completa y $\gamma = 0$ simboliza un proceso totalmente repulsivo; ver Sección (2.2.1).

En el caso de los procesos agregativos se simularon procesos de Matern en dos niveles. El primer nivel es "menos agregativo". Tiene parámetros $\kappa = 10$, $s = 0.3$ y $\mu = 10$, y lo denotaremos con M_1. El segundo nivel es "más agregativo". Sus parámetros son $\kappa = 10$, $s = 0.2$ y $\mu = 10$, y se representará como M_2. En el proceso de Matern, el parámetro s controla el radio donde pueden nacer los hijos, a partir de los padres. Radios más chicos implican procesos con puntos más concentrados; ver Sección (2.2.1).

La Figura (2.8) resume de manera esquemática los procesos considerados y sus parámetros. A la izquierda están los agregativos, y a la derecha los repulsivos. Es importante notar que este "ordenamiento" es estrictamente esquemático. No es posible asignar una orden entre estos tipos de procesos, en el sentido de que un proceso que sea más repulsivo es a la vez menos agregativo.

Para la obtención de las muestras, se simulan 1000 procesos puntuales para cada uno de los cinco tratamientos. Se considera sólo un tamaño de muestra porque es de interés entender cómo cambia la potencia con los diferentes tratamientos y

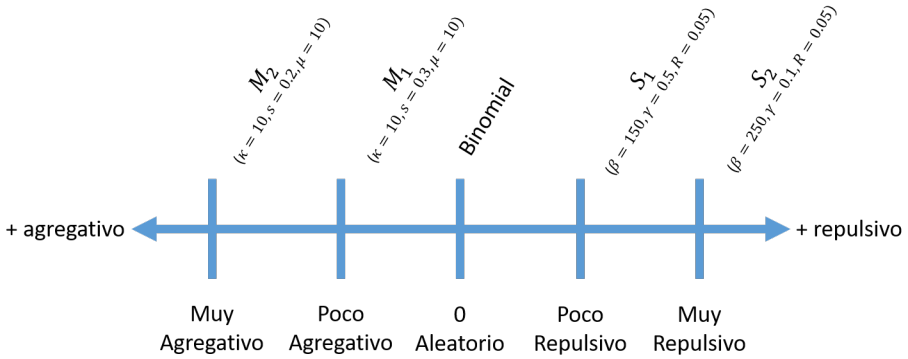

Figura 2.8: Tratamientos del experimento. A la izquierda se encuentran los procesos agregativos. A la derecha se están los procesos Repulsivos. Estos niveles corresponden a los casos a los que se someten las pruebas de hipótesis.

no cómo cambia cuando tenemos más procesos. Esta cantidad, aunque arbitraria, es suficiente para obtener una buena estimación de la potencia de cada prueba. Los procesos puntuales constan de $n = 100$ puntos. Este valor es acorde con la cantidad de observaciones que se obtienen en aplicaciones de biología, y es suficiente para propósitos ilustrativos. Para simular el proceso de Strauss y de Matern se utiliza el paquete "spatstat" de R, el cual esta disponible de manera libre en cran.r-project.org. Las simulaciones se condicionan para que cada proceso puntual tenga 100 puntos. Esta restricción controla la sensibilidad de las curvas de Betti al número de puntos en el proceso.

Para estimar la potencia de cada prueba de hipótesis, se calcula el porcentaje de veces que las realizaciones de los procesos puntuales son rechazadas para cada prueba. Este procedimiento se repite en cada uno de los tratamientos especificados. La potencia de cada prueba y de cada tratamiento se agrupa en una tabla resumen. Cada renglón de la tabla representa una prueba distinta. Los tratamientos se disponen en las columnas. De esta forma, podemos cuantificar y comparar el desempeño de cada prueba.

2.4.2 Resultados del experimento y discusión

A continuación se muestra la tabla resumen con los resultados del experimento. Esta tabla agrupa la potencia estimada para cada tratamiento y para cada prueba de hipótesis. Para el proceso binomial, se observan potencias cercanas a 5 %. Este número es esperado ya que los procesos binomiales satisfacen la hipótesis nula, y $\alpha = 0.05$.

En los procesos de Strauss, columnas dos y tres, se observa que la potencia es superior para las pruebas basadas en curvas de Betti. La prueba MAD tuvo un rendimiento deficiente, siendo indistinguible un proceso aleatorio de uno repulsivo. Esto ocurre por que la prueba MAD considera el máximo de las distancias entre vecinos cercanos. Basado en simulaciones propias, se ha observado que esta distancia es similar en procesos aleatorios y repulsivos. La prueba basada en curvas de Betti con $\omega = 0$, que considera componentes conexas, es superior a la prueba para $\omega = 1$, que toma en cuenta los ciclos, alcanzando cerca de 98 % de rechazos correctos. Esto quiere decir que las componentes conexas son más efectivas para detectar discrepancias en las nubes de datos. Estos resultados resaltan el potencial de las pruebas de ATD con respecto las pruebas tradicionales, en el caso de procesos repulsivos.

	$B(n = 100)$	S_1	S_2	M_1	M_2
Cuad.	5 %	11 %	39 %	56 %	90 %
MAD	3 %	<1 %	<1 %	72 %	96 %
$\omega = 0$ (β_0)	4 %	44 %	98 %	46 %	91 %
$\omega = 1$ (β_1)	3 %	12 %	46 %	32 %	76 %

Tabla 2.1: Resultado del experimento. Cada fila representa un procedimiento de prueba de hipótesis. En cada columna están los distintos tipos de procesos. El porcentaje en cada casilla es la proporción de rechazos correctos, es decir, la potencia estimada de cada prueba.

El rendimiento es diferente para los procesos de Matern. En este caso no se observa mejoría con respecto a las pruebas usuales. De hecho, las pruebas tradicionales

superan a las pruebas basadas en curvas de Betti. La prueba MAD fue muy efectiva para detectar procesos de agrupamiento. En efecto, detectar procesos con grupos fue parte de los objetivos al crear esta prueba (Baddeley et al., 2014). Pese a que las pruebas tradicionales superaron a las pruebas basadas en curvas de Betti, la prueba $\omega = 0$ continua siendo mejor que la prueba $\omega = 1$. Esto es, la prueba que considera las componentes conexas tiene mayor poder de discernimiento. En resumen, con la estadística de prueba escogida, las pruebas basadas en curvas de Betti no son tan eficientes para detectar diferencias entre procesos agregativos y aleatorios como las pruebas tradicionales.

La Figura (2.9) resalta los resultados del experimento recopilados en la Tabla (3.2). En esta figura se separan por colores las pruebas. En negro se resaltan las pruebas clásicas y en rojo las pruebas basadas en curvas de Betti. De esta forma, es más sencillo observar la eficiencia de las pruebas basadas en curvas de Betti para los casos repulsivos, ubicados a la derecha, y la escasa utilidad en los procesos agregativos, ubicados a la izquierda.

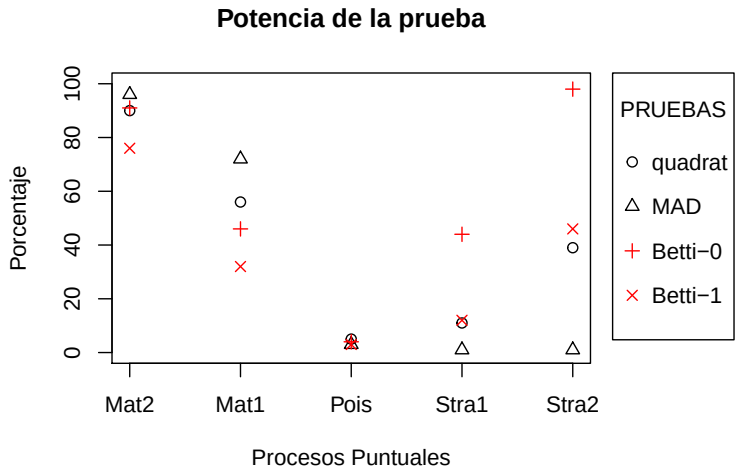

Figura 2.9: Gráfica del resultado del experimento. En rojo se etiquetan las pruebas basadas en ATD. En negro se encuentran coloreadas las pruebas de cuadrantes y MAD. Para los procesos de Strauss, ubicados a la derecha, las pruebas basadas en curvas de Betti son superiores.

Como parte del análisis de los resultados obtenidos, se estudió la variación estimada de las curvas de Betti promedio para cada tipo de proceso, trazadas en la Figura (2.10). La parte izquierda representa los casos para el primer grado de homología, las componentes conexas. La parte derecha corresponde al segundo grado de homología, los ciclos. Estas curvas se construyeron en dos pasos:

1. Se estimaron las curvas de Betti promedio para cada tipo de proceso: en verde el proceso agregativo de Matern, en rojo el proceso binomial, y en azul el proceso repulsivo de Strauss. El procedimiento para calcular las curvas de Betti promedio se encuentra en la Sección (1.2.3). En la Figura (2.10), la curva que está en medio de la región coloreada representa la curva promedio. Éstas se estimaron con 1000 simulaciones de cada proceso.

2. Se estimó la desviación estándar puntual de las curvas de Betti, es decir,

$$\text{SD}(\beta(r)) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\beta_i^{(k)}(r) - \overline{\beta}_i^{(k)}(r)\right)^2}$$

   para cada $r \geq 0$, donde i es el grado de homología y $n = 1000$. Para este cálculo, se utilizaron las simulaciones generadas en el paso anterior. Para representar visualmente esta desviación, se suma y resta la cantidad $\text{SD}(\beta(r))$ a la curva $\overline{\beta}_i^{(k)}(r)$ para cada $r \geq 0$ y se rellena el espacio con colores que se transparenten, como en la Figura (2.10).

De la parte izquierda de la Figura (2.10), se puede dar una idea de por qué la prueba $\omega = 0$ tiene mejor potencia que la prueba $\omega = 1$. A pesar de que parece que las curvas son más cercanas entre sí, existe mucha menor dispersión. Esta menor dispersión implica una mayor precisión. En la parte derecha de la Figura (2.10), las regiones coloreadas se solapan en una región mucho mayor. Esto podría justificar decidir quedarnos sólo con la prueba de curvas de Betti de componentes conexas. Este estudio gráfico, basado en las curvas promedio y desviación estándar, refleja el potencial de utilizar curvas de Betti en el análisis de datos.

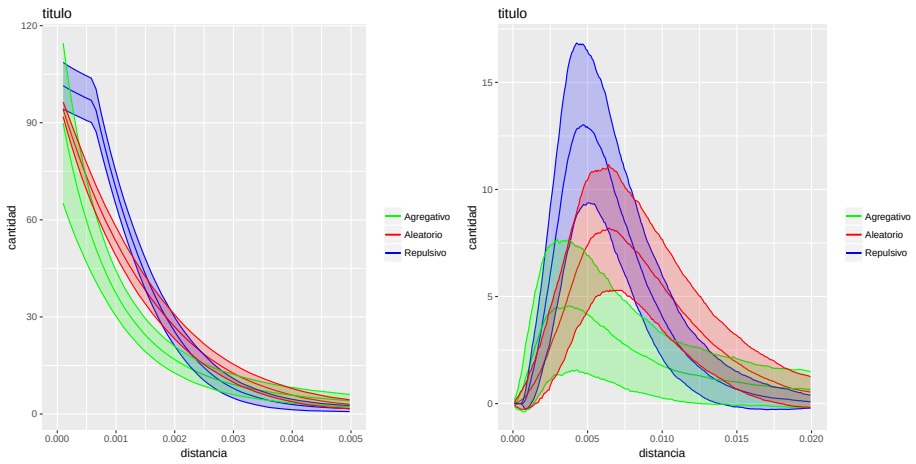

Figura 2.10: Curvas de Betti promedio con desviación estándar. En la parte izquierda, se encuentran las curvas de Betti promedio para las componentes conexas de los tres tipos de procesos puntuales estudiados. En la parte derecha se encuentran las curvas de Betti promedio para los ciclos de estos tres procesos. La región coloreada corresponde la desviación estándar en cada punto de la curva de Betti.

2.5 Conclusiones

En este capítulo se introdujo un nuevo procedimiento de pruebas de hipótesis para procesos puntuales basado en curvas de Betti. Se mostró, apoyados en el resultado del experimento de simulación, que las técnicas de ATD pueden ser efectivas para detectar diferencias en procesos repulsivos y aleatorios. Sin embargo, en el caso de patrones agregativos estas pruebas son superadas por las técnicas usuales. Se recomienda utilizar técnicas de ATD cuando se este en presencia de un fenómeno en que los puntos tiendan a alejarse entre si.

El uso de curvas de Betti promedio permitió la interpretación de la topología del proceso puntual estudiado. Estas curvas permitieron identificar las diferencias, en la formación de componentes conexas y de ciclos, de cada tipo de proceso. De esta forma, se entendió de una manera más acentuada el resultado del experimento de simulación. El empleo de la desviación estándar permitió visualizar detalles como las regiones en donde la topología de los procesos es similar. Estos aspectos de vi-

sualización representan una ventaja en comparación a otros resúmenes topológicos.

Estas pruebas pueden extenderse a mayores dimensiones, tantas como tenga el espacio donde vivan los procesos puntuales. Por ejemplo, si se tiene un conjunto de puntos en $\mathbb{R}^d$ es posible obtener d pruebas diferentes basadas en curvas de Betti. Siguiendo la construcción de la Sección (2.3.1), una estadística de prueba podría ser de la forma:

$$d\left(\xi\right) \quad = \quad \sum_{i=0}^{d-1} \omega_i \int_0^{x_i} \left(\beta_i(r) - \overline{\beta}_i(r)\right)^2 dr, \tag{2.4}$$

donde ξ es la realización del proceso a estudiar y $\sum \omega_i = 1$.

Una posibilidad para complementar este trabajo es hacer un estudio basado en el enfoque de pruebas de hipótesis de Fisher. En este enfoque se estaría interesado en calcular el p-valor. Por lo que, en lugar de tener una respuesta binaria al problema, se mediría la plausibilidad de que la hipótesis nula sea cierta. En este sentido, se trabajaría en un caso más general. Por lo tanto, se tendría una visión más completa del problema de pruebas de hipótesis abordado con técnicas topológicas. Otra posibilidad para extender es estudiar diferentes combinaciones de ω en la estadística de Prueba. La idea sería determinar un método que decida la ω óptima, en el sentido que maximice la potencia. Esta ω optima podría resultar en una prueba que sea significativamente superior a la planteada en esta tesis.

3

Clasificación de objetos 3D

El problema de clasificación consiste en generar un algoritmo que sea capaz de identificar, de entre un grupo de objetos y un grupo de categorías, a que categoría pertenece un objeto. Estos algoritmos basan sus predicciones en un conjunto de datos para los cuales se conoce a qué categoría pertenece cada objeto, llamado conjunto de entrenamiento. Ejemplos comunes son: detectar correo basura, clasificar números escritos a mano y discernir entre pacientes enfermos y sanos. En este capítulo se propone un algoritmo que utiliza las curvas de Betti para generar vectores que, como se verá, son mas eficientes para clasificar que los puntos por sí solos. La relevancia de este nuevo enfoque radica en que con pocos datos es posible obtener buenos resultados.

Uno de los algoritmos principales para abordar este problema es la máquina de soporte vectorial o SVM (por sus siglas en inglés, Support Vector Machine), como se expone en Hastie et al. (2009). La SVM separa las categorías de los objetos construyendo hiperplanos en el espacio donde viven los datos. Los parámetros del hiperplano son estimados mediante la optimización un problema convexo. En nuestro

caso, la SVM será de utilidad ya que es un algoritmo que puede ser utilizado con los vectores que se generan a partir de las curvas de Betti. Además, permite la comparación con el método "tradicional" de clasificación, que se explicará más adelante.

Entenderemos un objeto 3D como una nube de puntos en $\mathbb{R}^3$. Los puntos se toman de manera uniforme sobre la superficie del objeto 3D. En nuestro estudio se simulan de figuras como el cubo, la esfera y la botella de Klein. Éstas son figuras sencillas que permiten ilustrar la implementación del método propuesto. En este sentido, nuestra definición difiere del concepto de imagen en el contexto de reconocimiento de patrones. En lugar de ser una imagen digital compuesta por una matriz de pixeles, un objeto 3D es una matriz con información espacial de los puntos muestreados. Esta definición separa nuestro estudio de la clasificación de imágenes que surgen en visión artificial (computer vision).

La importancia de clasificar objetos 3D puede verse en aplicaciones del análisis estadístico de formas o SSA (Statistical Shape Analysis). Un ejemplo, en 2D, es el estudio de la forma del cráneo de gorilas machos y hembras (Dryden and Mardia, 2016). Ellos toman muestras de algunos puntos específicos del cráneo. Estos puntos son llamados "landmarks", y son fijos. Dryden and Mardia (2016) muestran que es posible detectar diferencias significativas entre cráneos dependiendo del género. Aunque las aplicaciones son similares, nuestro método de recolección de datos y el de Dryden and Mardia (2016) difieren ligeramente. Ellos fijan previamente el lugar donde se localizan los puntos mientras que en este trabajo la ubicación de los puntos es aleatoria sobre la superficie del objeto 3D.

El papel del ATD en este contexto opera bajo la premisa de que objetos 3D diferentes deben tener curvas de Betti diferentes. En la Sección (3.3) se construye un algoritmo para clasificar objetos 3D. Este algoritmo utiliza las curvas de Betti para detectar diferencias topológicas y poder discernir entre los objetos. Como se verá, en algunos casos se obtiene un error de clasificación menor a que si realizáramos el procedimiento "tradicional" de clasificación.

3.1 Objetivos

Las preguntas que surgen en este contexto son:

1. ¿Cómo se construye un algoritmo de clasificación que tome en cuenta curvas de Betti?

2. ¿Cómo mejora el error de clasificación al utilizar las curvas de Betti?

La primera pregunta requiere la implementación de un algoritmo para clasificar objetos 3D que permita utilizar curvas de Betti. En la elaboración de nuestro algoritmo, se trabaja las primeras tres curvas de Betti. Con ellas se crea un vector apto para ser utilizado en el algoritmo de SVM, y que se llamará vector de Betti. Este vector se construye acotando y concatenando las curvas de Betti, como se explica en detalle en la Sección (3.3). Para la comparación de este método, se plantea una solución alternativa utilizando directamente la nube de datos en el algoritmo SVM. Para esto, se transforma la nube de puntos en un vector. Este proceso se verá en la Sección (3.2.3), y será llamado método "tradicional". Es la primera solución natural que se implementaría para abordar el problema de clasificación.

La segunda pregunta requiere una medida de cuantificación del rendimiento del algoritmo. Esto se aborda, de manera análoga al Capítulo 2, mediante el diseño y la ejecución de un experimento de simulación. El diseño del experimento considera un tamaño de muestra fijo ya que es de interés el efecto del uso de herramientas topológicas y no el cambio del error de clasificación al aumentar el tamaño de muestra. Finalmente, el algoritmo propuesto se compara con el método "tradicional" para evaluar la utilidad de las herramientas topológicas.

En el experimento, se mide el rendimiento a través del error de clasificación. Éste es el número esperado de predicciones incorrectas (Hastie et al., 2009). Este error se estima utilizando validación cruzada (VC). La VC toma particiones del conjunto de entrenamiento. Luego, una fracción de la partición es etiquetada como conjunto de validación, y es utilizada para estimar el error de predicción. Este proceso se repetite sucesivas veces variando la fracción que representa al conjunto de validación (Hastie et al., 2009).

El problema de clasificación utilizando técnicas de ATD ha sido abordado por Peter Bubenik. Él diseña un procedimiento basado en vectores construidos a partir de los panoramas de persistencia (Bubenik, 2015). Estos vectores sirven como entrada para la SVM. Este enfoque aprovecha toda la información extraída en la homología persistente. Cabe destacar que no existe una publicación oficial de este algoritmo; fue presentado por él durante la tercera escuela de análisis topológico de datos en el Estado de México en enero de 2017. Los resultados mostrados sugieren que el método es efectivo en el análisis supervisado. Sus ideas son la base para nuestro algoritmo; sin embargo, nuestro algoritmo utiliza curvas de Betti en lugar de panoramas de persistencia.

Los objetivos específicos de este capítulo son:

1. Mostrar la construcción de un algoritmo de clasificación utilizando curvas de Betti.

2. Mostrar los resultados de un experimento de simulación para estudiar el error de clasificación del algoritmo del objetivo anterior.

3.2 Antecedentes generales

El problema de clasificación es parte del aprendizaje estadístico o aprendizaje de máquina (Machine Learning). Esta rama de la estadística y la computación puede separarse en aprendizaje supervisado y no supervisado. En el aprendizaje supervisado el objetivo es predecir el valor de un dato basado en conjunto de entrenamiento. Es este conjunto se conoce el grupo al que pertenece cada dato. Por otro lado, el aprendizaje no supervisado carece de conjunto de entrenamiento. El objetivo en este caso es describir nuevas las relaciones y patrones en un conjunto de datos. Para más detalles, ver Hastie et al. (2009).

3.2.1 Máquina de soporte vectorial

Una manera de separar categorías de puntos es utilizando hiperplanos separadores (Hastie et al., 2009). Pueden darse dos casos: en el primero es posible encontrar un plano que separe perfectamente los puntos en dos clases; en el segundo caso no es posible encontrar dicho hiperplano, como se muestra en la Figura (3.1). Los algoritmos suelen ser diferentes para cada caso. La máquina de soporte vectorial o SVM incluye ambas posibilidades. Este algoritmo considera el hiperplano que tenga la mayor separación o margen entre las clases a separar. Para esto, se resuelve un problema de optimización convexo donde la variable a maximizar es el margen, y se denota por M.

Dado que es posible que los puntos no se puedan separar completamente en cada clase, el margen esta sujeto a variables de holgura. Estas variables permiten que algunos puntos estén en lado opuesto del margen. Se denotan como $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$. El problema de optimización a resolver es:

$$\max_{\beta, \beta_0, ||\beta||=1} M \tag{3.1}$$
$$\text{sujeto a } y_i \left(x_i^T \beta + \beta_0\right) \geq M(1 - \xi_i), \qquad i = 1, \ldots, N.$$

Nuestros datos de entrenamiento consisten en N pares (x_1, y_1),(x_2, y_2),...,(x_N, y_N), con $x_i \in \mathbb{R}^p$. La variable y_i esta en $\{-1, 1\}$ dependiendo de la clase a la que pertenezca el punto i.

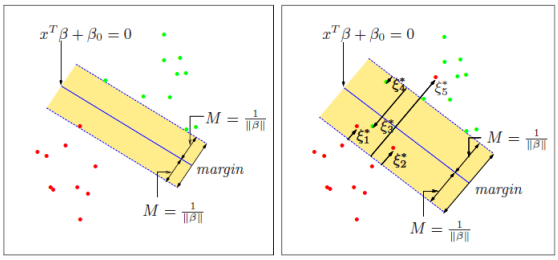

Figura 3.1: Hiperplanos separadores. En el lado izquierdo se aprecia una combinación de puntos para los cuales es posible encontrar un plano que los separe en dos clases. En el lado derecho se tiene el caso contrario. No es posible trazar una linea recta que separe a los puntos rojos y verdes. Imagen tomada de Hastie et al. (2009).

La idea detrás de esta formulación es que el valor ξ_i, en la restricción

$$y_i \left(x_i^T \beta + \beta_0\right) \geq M(1 - \xi_i),$$

es la cantidad proporcional por la cual la predicción $f(x_i) = x_i^T \beta + \beta_0$ está en el lado equivocado de su margen. Por lo tanto, al acotar la suma $\sum \xi_i$, acotamos la cantidad proporcional total por la cual las predicciones caen en el lado incorrecto del margen. Una clasificación incorrecta ocurre cuando $\xi_i > 1$. Si se acota $\sum \xi_i$ a un valor constante, digamos K, se acota el número total de clasificaciones incorrectas. Para una formulación más detallada, consultar Hastie et al. (2009).

3.2.2 Validación cruzada

La validación cruzada es una técnica para evaluar el rendimiento de un modelo. Es recomendable cuando el propósito del modelo es predecir (Hastie et al., 2009). Específicamente, se utilizará el método de validación cruzada k-fold. Éste utiliza parte de los datos disponibles para estimar los parámetros del algoritmo de predicción, y otra parte para probarlo (Hastie et al., 2009). Se escogió este método por ser efectivo con pocas muestras y no requerir supuesto sobre los datos.

Para llevar a cabo la validación cruzada k-fold, se separan los datos de entrenamiento en k partes iguales. En la Figura (3.2) se muestra un esquema con $k = 5$. La k-ésima parte (tercera en la figura) se etiqueta como conjunto de validación. Se estiman los parámetros del algoritmo de clasificación tomando como conjunto de entrenamiento las otras $k - 1$ partes de los datos. Luego, se estima el error de predicción de este algoritmo mediante el pronóstico correcto en el conjunto de validación, es decir, en la k-ésima parte de los datos. Hacemos esto para $1, 2, \ldots, k$. Estas k estimaciones se promedian para calcular el error promedio de predicción.

Sea $\hat{f}$ la función de predicción y $\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, K\}$, una función de etiquetado que indica la partición a la que corresponde la observación i. Sea $\hat{f}^{-k}(x)$

Figura 3.2: Validación cruzada k-fold con $k = 5$. El tercer rectángulo representa la fracción que corresponde al conjunto de validación. En este caso, se repite el procedimiento cinco veces. En cada rotación se cambia la fracción de los datos que sirven de conjunto de validación.

la función ajustada calculada sin la k-ésima parte. La estimación del error por validación cruzada es

$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{L}\left(y_i, \hat{f}^{-\kappa(i)}(x_i)\right),$$

donde L es una función de perdida que es igual a cero si la predicción es correcta ($y_i = \hat{f}^{-\kappa(i)}(x_i)$) y es igual a 1 si la predicción no es correcta ($y_i \neq \hat{f}^{-\kappa(i)}(x_i)$). En este trabajo se utiliza $K = 10$ con fines ilustrativos del método.

3.2.3 Clasificación de objetos 3D con SVM

Se describe el algoritmo de clasificación que utiliza directamente los puntos de los objetos 3D, y que se denota por "tradicional". Este enfoque surge de manera natural como primera aproximación a una solución al problema de clasificación. Otros métodos más complejos involucran kernelización (Hastie et al., 2009). La kernelización transforma los datos para resaltar sus diferencias y obtener un menor error de clasificación. Este enfoque requiere habilidad por parte del modelador para encontrar la función tipo kernel que destaque las diferencias en los datos. Si no se realiza con cuidado, la kernelización puede resultar en un sobreajuste. En este trabajo, se prefiere el método "tradicional" por motivos ilustrativos. En este sentido, la comparación realizada es una medida del aporte de usar técnicas topológicas con respecto a la solución trivial.

El algoritmo de clasificación "tradicional" tiene los siguientes pasos:

1. Transformar la matriz de datos a un vector de dimensión $3 \times n$ concatenando

cada columna, esto es,

$$\begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{pmatrix} \Rightarrow (x_n, \ldots, x_n, y_1, \ldots, y_n, z_1, \ldots, z_n)\,.$$

Esta transformación es necesaria para poder utilizar la SVM. De esta forma, se vuelven comparables los resultados de esta algoritmo y el que se describirá en la Sección (3.3).

2. Utilizar estos vectores como entrada de la SVM.

3. Correr el algoritmo de SVM.

Los resultados del algoritmo SVM suelen presentarse en una matriz de confusión. La matriz de confusión es una tabla de contingencia que muestra cómo fueron clasificados los objetos; ver Tabla (3.1). Cada columna representa el número de predicciones de cada clase. Cada renglón es la verdadera etiqueta del objeto. Esta tabla tiene como objetivo la validación del algoritmo.

Para entender el valor de las entradas de la matriz de confusión, supongamos que tenemos N objetos tipo A y N objetos de tipo B. De haber obtenido un resultado perfecto en el algoritmo, la matriz de confusión tendría N en la diagonal y cero en el resto de sus entradas. Es usual llamar a los elementos de la matriz como Verdadero Positivo y Verdadero Negativo cuando el modelo acierta, y Falso Negativo y Falso Positivo cuando el modelo no acierta. El nombre preciso de cada caso se refleja en las entradas de la Tabla (3.1). Los nombres Positivo y Negativo no tienen interpretación matemática; son etiquetas. Esta matriz es importante para calcular el error de clasificación. El error de clasificación se estima de la siguiente forma:

$$\mathrm{EP} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{VP} + \mathrm{FP} + \mathrm{FN} + \mathrm{VN}}.$$

realidad\predicción	Tipo A	Tipo B
Tipo A	Verdadero Positivo (VP)	Falso Positivo (FP)
Tipo B	Falso Negativo (FN)	Verdadero Negativo (VN)

Tabla 3.1: Matriz de confusión. Las columnas representan el número de predicciones de cada clase. Los renglones son las etiqueta reales del objeto. Un objeto de tipo A que sea predicho como de tipo A estaría en el primer renglón y en la primera fila.

3.3 Clasificación utilizando curvas de Betti

En esta sección se explica el algoritmo de clasificación que involucra técnicas topológicas. Este algoritmo consiste en la transformación de los objetos 3D en vectores construidos a partir de las curvas de Betti. Estos vectores se utilizan como entrada de la SVM, y en este trabajo serán llamados vectores de Betti. Se presenta una definición de los vectores de Betti y la estructura del algoritmo. Cabe destacar que estos vectores son una aportación original de este trabajo en la búsqueda de resolver el problema de clasificación con las curvas de Betti.

3.3.1 Vectores de Betti y algoritmo de clasificación

Los vectores de Betti resumen la información topológica de una nube de datos en un formato compatible con SVM, la cual requiere observaciones que sean vectores. La idea general es concatenar las primeras tres curvas de Betti en un solo vector. Para explicar la construcción de los vectores de Betti, se ha separado el proceso en dos etapas: inicial y de construcción. En la etapa inicial se define la cantidad de entradas del vector y el límite hasta donde se evalúan las curvas de Betti. Estos parámetros pueden ser modificados por el usuario dependiendo de las necesidades. En la etapa de construcción se crea el vector de Betti basado en los valores obtenidos en la etapa anterior. De cada nube de datos se genera un vector de Betti. El vector resultante del proceso se ve graficado como la Figura (3.4).

En la etapa inicial se utilizan las nubes de datos del conjunto de entrenamiento para generar curvas de Betti promedio. Las curvas promedio ayudan a definir los

límites hasta donde se evaluará cada curva de Betti, denotados por x_0, x_1, x_2. El límite x_i se define como

$$x_i = \text{ínf}\left\{x : \beta_i^*(y) = 0 \qquad \forall y \geq x\right\},$$

donde β_i^* es la i-esima curva de Betti media. De manera similar al Capítulo 2, β_i^* se estima utilizando la curva de Betti promedio, $\overline{\beta}_i$. La idea es aprovechar el menor intervalo, empezando en cero, para el cual cada curva de Betti en cuestión aporta información, en el sentido de tener valores positivos.

Después de tener el límite de evaluación para las curvas de Betti, se procede a construir el vector. Para esto, se define la cantidad de entradas que tendrá el vector, denotada por l. Luego, se crean rejillas de l elementos en los intervalos $[0, x_0]$, $[0, x_1]$ y $[0, x_2]$. Las rejillas son vectores de la forma,

$$P_i = \left(p_{i1} = 0, p_{i2} = \frac{x_i}{l}, \ldots, p_{il} = x_i\right)^T, \qquad i = 0, 1, 2. \tag{3.2}$$

El proceso se ilustra en la Figura (3.3).

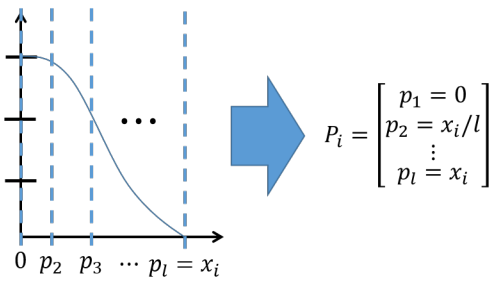

Figura 3.3: Etapa inicial. A la izquierda se muestran los límites x_i. A la derecha se observa la rejilla P_i, donde $i = 0, 1, 2$.

En la etapa de construcción se evalúan los puntos p_{ij} del vector P_i de la Ecuación (3.2) en sus respectivas curvas de Betti, como en la Ecuación (3.3). De esto, se obtienen tres vectores de longitud l cuyas entradas son los valores de las curvas de Betti para cada elemento de la partición. Estos vectores tienen la forma

$$\beta_i(P_i) = \left(\beta_i\left(p_{i1}\right), \beta_i\left(p_{i2}\right), \ldots, \beta_i\left(p_{il}\right)\right), \qquad i = 0, 1, 2, \tag{3.3}$$

donde i es el grado o dimensión de la curva de Betti que se este evaluando. Finalmente, se genera un vector de longitud $3 \times l$. Éste vector tiene la forma

$$\beta = (\beta_0(P_0), \beta_1(P_1), \beta_2(P_2)) .$$

El vector β es el vector de Betti, ilustrado en la Figura (3.4).

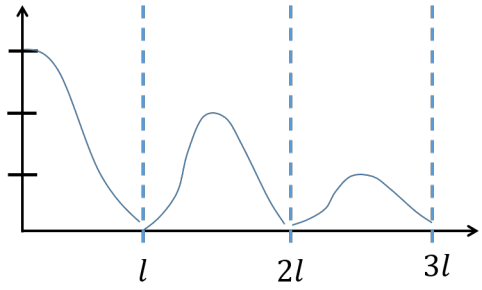

Figura 3.4: El vector de Betti se obtiene concatenando y evaluando las primeras tres curvas de Betti.

Los pasos para el algoritmo basado en curvas de Betti son los siguientes:

1. Transformar las nubes de datos en diagramas de persistencia.
2. Transformar diagramas de persistencia en curvas de Betti.
3. Construir los vectores de curvas de Betti, como se explicó en la sección anterior.
4. Utilizar estos vectores como entrada de la SVM.
5. Correr el algoritmo de SVM. Ver Figura (3.5).

3.4 Experimento de simulación

El objetivo del experimento es estimar los errores promedio de predicción para el algoritmo de clasificación con curvas de Betti y el método "tradicional". Se consideran distintos tratamientos basados en comparaciones dos a dos de los diferentes tipos figuras: la esfera, el cubo y la botella de Klein. Luego se determina la efectividad de cada modelo a través de la estimación del error esperado, el cual es obtenido con validación cruzada. Con este experimento, se pretende analizar y cuantificar el aporte del ATD en la clasificación de objetos 3D.

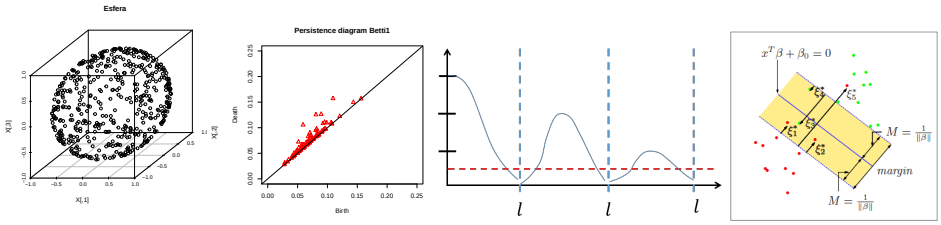

Figura 3.5: Algoritmo de clasificación con vectores de Betti. De izquierda a derecha, se observa primero la nube de puntos. Luego, el diagrama de persistencia seguido de el vector de Betti. Por último, el resultado de una corrida del algoritmo SVM.

3.4.1 Diseño del experimento

Se describen elementos importantes del experimento de simulación. Estos elementos abarcan las unidades muestrales, los tratamientos, el tamaño de muestra y el proceso de análisis. Estos elementos forman parte de la metodología usual en el diseño de experimentos propuesta por Dean et al. (2017). De manera análoga al Capítulo 2, esta metodología se escogió para garantizar que las consideraciones pertinentes sean hechas antes de iniciar la simulación de datos y el análisis posterior.

En este experimento, las unidades muestrales pueden verse como las nubes de puntos. Estas nubes son muestreadas de manera aleatoria y uniforme sobre esferas, cubos y botellas de Klein. La esfera es simulada con el paquete TDA en el lenguaje de programación R. Este algoritmo simula variables aleatorias gaussianas con media cero y varianza uno. Luego, las estandariza a través de la norma espectral de matrices o norma "2" (Fasy et al., 2014). De esta forma, se garantiza obtener puntos distribuidos de manera uniforme sobre la esfera de radio r (Biscay et al., 2016). En este trabajo, el radio de la esfera se tomó como uno por razones ilustrativas.

El cubo se simuló a través de un algoritmo diseñado en esta tesis. En este algoritmo, primero se escoge de manera aleatoria y uniforme alguna de las seis caras del cubo en donde aparecerá el primer punto. Sobre esta cara se simulan dos variables aleatorias uniformes. Con estos dos valores resultantes y la cara escogida, se genera un valor en $\mathbb{R}^3$ que corresponde al punto simulado. Este proceso se repite tantas veces como puntos se requieran. El lado del cubo se escogió de tal forma que el cubo

resultante tenga el mismo volumen que la esfera de radio 1.

Los datos para la banda de Moebius y la botella de Klein fueron simulados a través del código descrito en Flores et al. (2017). Su procedimiento tiene la ventaja de producir puntos uniformemente distribuidos sobre toda la superficie. Otros métodos presentan concentraciones en la parte más angosta de la botella. En la Figura (3.6) se muestra una visualización de estos objetos en 3D.

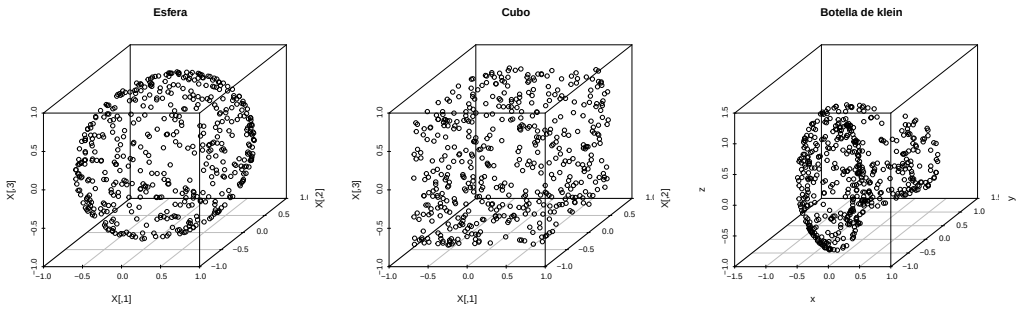

Figura 3.6: De izquierda a derecha se encuentran la esfera, el cubo y la botella de Klein. Por motivos ilustrativos, en esta imagen cada objeto tiene 500 puntos. En el experimento de simulación se toman 50 puntos.

Los tratamientos pueden verse como las comparaciones dos a dos de los tipos de figuras. Esto da lugar a los escenarios Esfera-Cubo (EC), Esfera-Botella de Klein (EK) y Cubo-Botella de Klein (CK). Se anticipa que el caso EC sea difícil dado que ambas figuras son topológicamente equivalentes. En los otros dos tratamientos, se esperan porcentajes bajos de error de clasificación. La razón de esto es que los objetos difieren en topología. Por ejemplo, los tres primeros números de Betti (ordenados según componentes conexas, ciclos, y agujeros de dos dimensiones) son 1, 0 y 1 para la esfera, igual que para el cubo. Por otro lado, para la Botella de Klein son 1, 2 y 1. En este sentido, se espera que estas diferencias teóricas se reflejen en el experimento de simulación.

Se generan 10 nubes de datos por cada tipo de figura. Esta cantidad, aunque relativamente pequeña, resulta suficiente para obtener errores de predicción bajos en

el algoritmo basado en curvas de Betti. La cantidad de puntos en cada objeto 3D es de 50 puntos. Se considera sólo un tamaño de muestra porque estamos interesados en entender cómo cambia el error de predicción en cada objeto, y no cómo cambia cuando tenemos más muestras de objetos. Intuitivamente, se esperaría que mientras más objetos tuviésemos, mejor capacidad de discernir entre objetos.

De cada tratamiento, se calcula el error promedio de clasificación para el algoritmo basado en curvas de Betti y para el método "tradicional". Los resultados se resumen en matrices de confusión. Esta tabla da lugar a una visualización gráfica para comparar de manera más efectiva los resultados, los cuales se muestran en la siguiente sección.

3.4.2 Resultados del experimento y discusión

Se muestra el resultado del error de clasificación para cada método y para cada escenario comparativo en la Tabla (3.2). El método de clasificación que utiliza curvas de Betti es superior en todos los casos. El resumen gráfico de esta tabla puede verse en la Figura (3.7).

	Esfera-Cubo	Esfera-Klein	Cubo-Klein
Clasificación con sólo SVM	50 %	50 %	30 %
Clasificación con curvas de Betti	10 %	5 %	10 %

Tabla 3.2: Resultado del experimento. Error de clasificación promedio, basado en validación cruzada.

El método "tradicional" resulta en 50 % de error de clasificación para los casos Esfera-Cubo y Esfera-Klein. Para este caso, es igual utilizar este método o lanzar un volado para decidir el tipo de objeto. Esto se debe a que sólo estamos utilizando 10 datos para crear el modelo. Con tan poca información, la SVM aun no da lugar a un modelo efectivo. Experimentalmente se ha observado que para obtener un error de clasificación con menos de 25 % se requieren al menos 250 nubes de datos, en lugar de 10. Por lo tanto, las curvas de Betti aportan información para clasificar efectiva-

mente los objetos 3D.

El método basado en curvas de Betti se comporta razonablemente bien. A pesar de tener poco datos, se obtiene un error menor o igual a 10 % para todos los casos. A diferencia de lo esperado, el caso Esfera-Cubo tiene un desempeño tan bueno como el caso Cubo-Klein. Una razón por la que el método clasificó bien el caso EC es las esquinas del cubo. Esto puede ser indicio de que la homología persistente recoge información de la ubicación y espacio que ocupan los objetos 3D, además de las características topológicas. Los resultados obtenidos se pueden apreciar en la Figura (3.7). Un análisis detallado de las curvas de Betti promedio de los resultados se encuentra en la próxima sección.

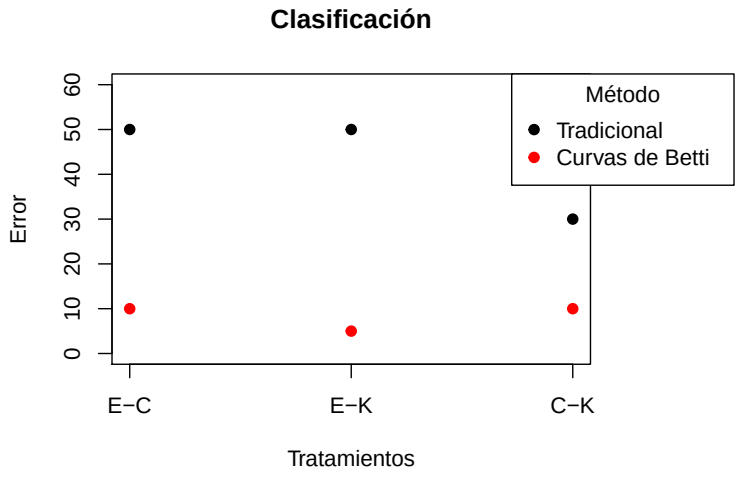

Figura 3.7: Resumen gráfico de experimento de simulación. En negro se muestran los resultados para el método "tradicional". En rojo se presentan los porcentajes de error para el método basado en curvas de Betti, y que tiene menor error para todos los tratamientos.

Como parte del análisis de los resultados obtenidos, se estudió la variación estimada de las curvas de Betti promedio para cada tratamiento propuesto. Este análisis es similar al que se hizo en la Sección 2.4.2. En cada una de las Figuras (3.8), (3.9) y (3.10), la curva de Betti promedio es la línea que está en el centro de cada banda, en verde para un tipo de figura y rojo para otro. Estas curvas están acompañadas de

una región coloreada que representa una desviación estándar para cada valor en r de la curva de Betti, es decir, para ese radio del complejo. Esta desviación es obtenida puntualmente con las 10 nubes de datos simuladas, de manera similar a la Sección 2.4.2.

En el caso del Esfera-Cubo, se obtuvo un resultado diferente al esperado. Aunque ambos objetos son topológicamente equivalente, el algoritmo logró predecir correctamente con error bajo. Las curvas de Betti promedio para este caso se aprecian en la Figura (3.8). Las componentes conexas, a la izquierda de la figura, son ligeramente diferentes en ambos casos. Para la segunda curva de Betti, en el centro de la figura, no se aprecian diferencias significativas para los ciclos. Las primeras dos características topológicas aportan poca información al algoritmo. En este caso, las diferencias se encuentran en la tercera curva de Betti. Por lo tanto, todas las características topológicas permiten que el algoritmo pueda discernir efectivamente.

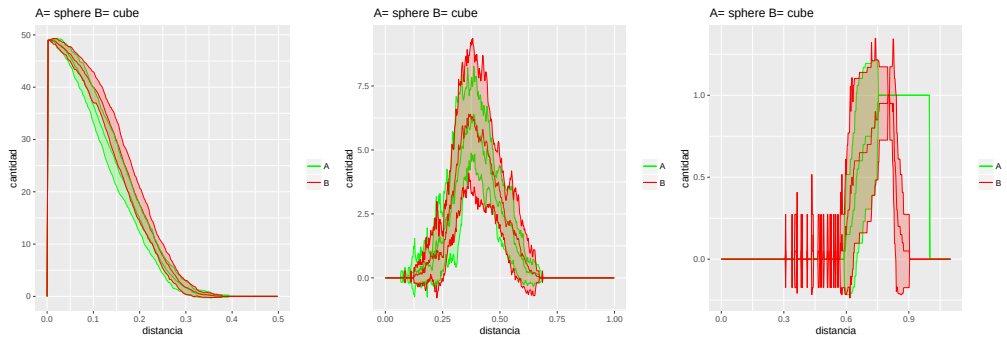

Figura 3.8: Resultados para el caso Esfera-Cubo. De izquierda a derecha se encuentran la primera, segunda y tercera curva de Betti promedio respectivamente, con la desviación estándar obtenida.

En el caso Esfera-Klein, el algoritmo se comportó como se esperaba. Las curvas de Betti promedio para este caso se aprecian en la Figura (3.9). De manera similar al caso Esfera-Cubo, las primeras dos curvas de Betti no aportan mucha información. De la figura puede observarse que las curvas de Betti promedio, junto con su desviación, son muy similares entre sí para las componentes conexas y los ciclos. Nuevamente, las diferencias se encuentran en la tercera curva de Betti. Podemos concluir, para es-

te caso, que se requieren todas las características topológicas para que el algoritmo pueda clasificar correctamente.

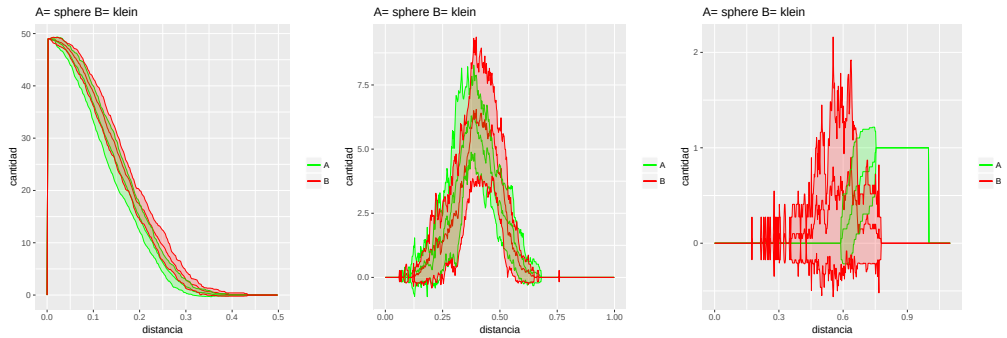

Figura 3.9: Resultados para el caso Esfera-Klein. De izquierda a derecha se encuentran la primera, segunda y tercera curva de Betti promedio respectivamente, con la desviación estándar obtenida.

Las curvas de Betti promedio para el caso Cubo-Klein pueden observarse en la Figura (3.10). Este caso tuvo un rendimiento ligeramente menor a lo que se pensaba. Dadas las diferencias entre las nubes de puntos, Figura (3.6), se esperaba que el porcentaje de error fuese tan bajo como el caso Esfera-Klein. Sin embargo, su porcentaje fue tan alto como el caso Esfera-Cubo. Una explicación de este resultado puede ser que la variación en la curva de Betti para las componentes conexas (la más izquierda en la Figura (3.10)) son prácticamente iguales. De manera similar a los casos anteriores, las diferencias se encuentran en la tercera curva de Betti.

3.5 Conclusiones

En este capítulo se construyó un algoritmo de clasificación basado en técnicas de ATD. Este algoritmo se comparó con métodos tradicionales. Se mostró que las curvas de Betti pueden ser efectivas para discernir entre diferentes figuras, incluso con pocos datos. Esto indica que la homología persistente es un descriptor provechoso de la forma de los datos.

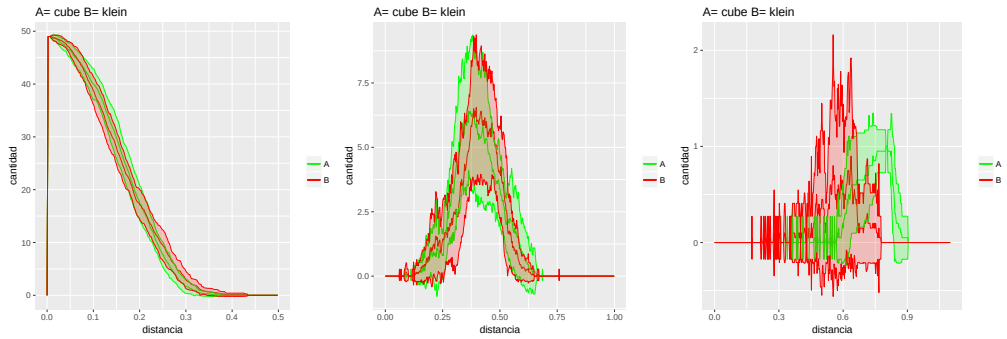

Figura 3.10: Resultados para el caso Cubo-Klein. De izquierda a derecha se encuentran la primera, segunda y tercera curva de Betti promedio respectivamente, con la desviación estándar obtenida.

En todos los casos analizados, las primeras dos curvas de Betti no bastan para clasificar las nubes de datos. En caso de haber utilizado sólo las componentes conexas y los ciclos, hubiese sido necesario tener una muestra más grande para poder catalogar correctamente cada nube de datos. En este sentido, fue relevante aprovechar todas las curvas de Betti.

Las técnicas de TDA son sensibles a transformaciones de los datos. Como en todo análisis de datos, es necesario procesar y limpiar los datos antes de introducirlos en los algoritmos. Se recomienda definir el marco de estudio antes de realizar cálculos de homología persistente. De esta forma, no tendremos discrepancias relacionadas con una inconsistencia en las magnitudes de los datos.

De manera similar al capítulo anterior, las curvas de Betti promedio permitieron hacer un análisis más profundo de los resultados. De esta forma, se pudo inspeccionar qué grado de homología ayuda más a clasificar los objetos. Esto permite tener una concepción mas completa de la topología de las nubes de puntos, y detectar posible rangos en los que falle el método de clasificación. De esta forma, se muestra el potencial que tienen las curvas de Betti para analizar y visualizar resultados.

4

Comentarios finales

En esta tesis se compararon técnicas de análisis topológico y ciencia de datos para dos casos de estudio. En el primero de ellos se analizó la potencia de una prueba de hipótesis basada en curvas de Betti para CSR. Esta prueba se comparó con las pruebas usuales de la literatura de procesos puntuales. Entre las principales conclusiones, se observo que para procesos con características repulsivas las pruebas de ATD resultan efectivas. Por otro lado, para los procesos agregativos se observó que la prueba construida no aporta una ventaja significativa. En el caso del problema de clasificación, se diseñó un algoritmo que etiqueta nubes de puntos en esferas, cubos y botellas de Klein. Nuestro procedimiento se basó en los vectores de Betti de cada tipo de figura. El algoritmo propuesto produjo un error de clasificación bajo con pocas nubes de datos. Con base en los resultados de ambos experimentos, las técnicas topológicas pueden ser útiles para el análisis de datos.

La curva de Betti es una herramienta muy versátil. En ambos casos de estudio se encontró una manera de utilizarla para abordar el problema en cuestión. La visualización de estas curvas permitió el análisis de resultados de los experimentos

de simulación. En el caso de procesos puntuales, las curvas promedio ayudaron a entender el comportamiento de la topología en cada tipo de patrón. En el caso de clasificación se determinó la importancia de evaluar todos los grupos de homología en los vectores de Betti. En la literatura las curvas de Betti son poco utilizadas; esta tesis es un aporte original por descubrir nuevas formas de hacer ATD.

Aunque el ATD funciona bien en algunos casos, tiene algunas desventajas. Un problema es la interpretación de los resúmenes topológicos en altas dimensiones. En nuestro caso, las nubes de puntos representan eventos en dos o tres dimensiones. Sin embargo, de tener más variables, es complicado determinar el significado de la homología persistente. En este sentido, resulta relevante buscar conjuntos de datos reales que puedan ser analizados con técnicas de ATD. Esto con la finalidad de encontrar relaciones que ayuden a entender el papel de las características topológicas en casos más generales.

Es primordial la limpieza y transformación de los datos para que sean comparables. Por ejemplo, todos los procesos puntuales se generaron en la región $[0,1]\times[0,1]$. Si hubiésemos tenido dos procesos repulsivos en regiones de diferente área, las curvas de Betti hubiesen detectado diferencias. De manera similar para el caso de clasificación, de haber tenido dos nubes de puntos muestreadas de esferas de radios diferentes, se hubiesen diferenciado aunque ambas fuesen el mismo tipo de objeto. Para solventar esto es necesario definir, desde el principio, la región de estudio para las nubes de datos antes de que éstos sean analizados. El objetivo de esto es poder comparar los resultados.

Bibliografía

Baddeley, A., Bárány, I., and Schneider, R. (2006). *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13-18, 2004*. Springer.

Baddeley, A., Diggle, P. J., Hardegen, A., Lawrence, T., Milne, R. K., and Nair, G. (2014). On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, 84(3):477–489.

Bar-Hen, A., Emily, M., and Picard, N. (2015). Spatial cluster detection using nearest neighbor distance. *spatial statistics*, 14:400–411.

Bartle, R. G. (1975). *The elements of real analysis*, volume 2. Wiley New York.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70(3):351–372.

Biscay, R., Nakamura, M., Pérez-Abreu, V., and Reveles, F. (2016). *Persistencia, probabilidad e inferencia para análisis topológico de datos*. CIMAT.

Biscio, C. and Møller, J. (2016). The accumulated persistence function, a new useful functional summary statistic for topological data analysis, with a view to brain artery trees and spatial point process applications. *arXiv preprint arXiv:1611.00630*.

Bobrowski, O. and Kahle, M. (2014). Topology of random geometric complexes: a survey. *arXiv preprint arXiv:1409.4734*.

Brockwell, P. J. and Davis, R. A. (2013). *Time series: theory and methods*. Springer Science & Business Media.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Cattell, R. B. (1943). The description of personality: basic traits resolved into clusters. *The journal of abnormal and social psychology*, 38(4):476.

Chowdhury, S. and Mémoli, F. (2016). Persistent homology of asymmetric networks: An approach based on dowker filtrations. *arXiv preprint arXiv:1608.05432*.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer Science & Business Media.

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

Dean, A. M., Voss, D., et al. (2017). *Design and analysis of experiments 2nd Edition*. Springer.

Dixon, P. M. (2002). Ripley's k function. *Encyclopedia of environmetrics*.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.

Edelsbrunner, H., Facello, M., Fu, P., and Liang, J. (1995). Measuring proteins and voids in proteins. In *System Sciences, 1995. Proceedings of the Twenty-Eighth Hawaii International Conference on*, volume 5, pages 256–264. IEEE.

Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Soc.

Fasy, B. T., Kim, J., Lecci, F., and Maria, C. (2014). Introduction to the r package tda. *arXiv preprint arXiv:1411.1830*.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Flores, G., Hernández, Y., and Pérez-Abreu, V. (2017). Simulaci'on en variedades y análisis topológico de datos. *Manuscrito en preparación*.

Giusti, C., Pastalkova, E., Curto, C., and Itskov, V. (2015). Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460.

González-Cucurachi, V. (2016). Aspectos estadísticos en análisis topológico de datos y una aplicación en ecología.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer-Verlag New York.

Hatcher, A. (2002). *Algebraic topology*. Cambridge University Press.

Hope, A. C. (1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 582–598.

Ibarra-Rodríguez, J. (2015). Modelos de homología persistente en filogenética.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

Kahle, M. and Meckes, E. (2010). Limit theorems for betti numbers of random simplicial complexes. *arXiv preprint arXiv:1009.4130*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Michel, B. (2015). *A Statistical Approach to Topological Data Analysis*. PhD thesis, UPMC Université Paris VI.

Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

Møller, J. and Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684.

Munkres, J. R. (1984). *Elements of algebraic topology*, volume 2. Addison-Wesley Menlo Park.

Murray, W., Gill, P., and Wright, M. (1981). Practical optimization.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in statistics*, pages 73–108. Springer.

Park, C., Pranav, P., Chingangbam, P., Van De Weygaert, R., Jones, B., Vegter, G., Kim, I., Hidding, J., and Hellwing, W. A. (2013). Betti numbers of gaussian fields. *arXiv preprint arXiv:1307.2384*.

Penrose, M. (2003). *Random geometric graphs*. Oxford University Press.

Perea, J. A. and Carlsson, G. (2014). A klein-bottle-based dictionary for texture representation. *International journal of computer vision*, 107(1):75–97.

Peréz-Angulo, J. (2016). Análisis topológico de datos: Robusticidad y análisis de sensibilidad de algoritmos. Master's thesis, Centro de Investigaciones en Matemáticas (CIMAT).

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Robins, V. and Turner, K. (2016). Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D: Nonlinear Phenomena*, 334:99–117.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Sprott, D. A. (2008). *Statistical inference in science*. Springer Science & Business Media.

Strauss, D. J. (1975). A model for clustering. *Biometrika*, 62(2):467–475.

Streit, R. (2010). *Poisson Point Processes*. Springer.

Tryon, R. C. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Edwards brother, Incorporated, lithoprinters and publishers.

Van de Weygaert, R., Vegter, G., Edelsbrunner, H., Jones, B. J., Pranav, P., Park, C., Hellwing, W. A., Eldering, B., Kruithof, N., Bos, E., et al. (2011). Alpha, betti and the megaparsec universe: on the topology of the cosmic web. In *Transactions on Computational Science XIV*, pages 60–101. Springer-Verlag.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Wright, M. L. (2016). Introduction to persistent homology. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 51. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Zomorodian, A. J. (2005). *Topology for computing*, volume 16. Cambridge university press.