UNIVERSIDAD DE GUANAJUATO

Análisis topológico de datos y aprendizaje estadístico para clasificación de galaxias

Tesis presentada como requisito parcial para obtener el título de:

Licenciado en Matemáticas

Victor Andres Amaya Carvajal

Universidad de Guanajuato
Departamento de Matemáticas
Guanajuato, México
2018

Análisis topológico de datos y aprendizaje estadístico para clasificación de galaxias

Victor Andres Amaya Carvajal

Tesis presentada como requisito parcial para obtener el título de:
Licenciado en Matemáticas

Director de Tesis:
Dr. Abraham Martín del Campo Sánchez

Co-director de Tesis:
Dr. Solai Jeyakumar

Lector Especial:
Dr. Víctor Manuel Perez Abreu Carrión

Universidad de Guanajuato
Departamento de Matemáticas
Guanajuato, México
2018

Dedicado a toda mi familia en Honduras...

Agradecimientos

A mis padres, muchas gracias por todo el amor y consejos que me han brindado. Gracias por haberme dado su apoyo en estos 5 años que estuve fuera de casa. Gracias por mostrarme que el trabajo duro tiene su recompensa, ya que ello ha sido una de mis motivaciones a lo largo de la carrera.

Quiero agradecerle a mis asesores. Muchas gracias por su paciencia conmigo, por su amistad y sus valiosos consejos, tanto académicos como personales. Gracias por mostrarme el valor que tiene la interacción de las matemáticas con otras ramas del conocimiento. Gracias por ser un ejemplo a seguir y guiarme en mi formación académica.

Estoy muy agradecido con el CIMAT y la Universidad de Guanajuato por la excelente y rigurosa formación en matemáticas que me han brindado. Estoy especialmente agradecido con el programa de becas del CIMAT, ya que dicho apoyo económico me permitió realizar mis estudios en matemáticas. Quiero agradecer al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca de ayudante de investigador con expediente 4337.

Quiero agradecerle a todos mis compañeros de la generación del 2013 del DEMAT, gracias por estos 5 años de amistad.

Quiero agradecerle a mis sinodales por tomarse el tiempo de leer esta tesis y por sus valiosos comentarios que me permitieron lograr una mejor redacción de este trabajo.

Por último, quiero agradecerle a todas las personas en México que me han ayudado a lo largo de mi travesía en este hermoso país que me ha sabido tratar muy bien.

Desiderata

Camina plácido entre el ruido y la prisa,
y recuerda la paz que se puede encontrar en el silencio.
En cuanto te sea posible y sin rendirte,
mantén buenas relaciones con todas las personas.
Enuncia tu verdad de una manera serena y clara,
y escucha a los demás,
incluso al torpe e ignorante,
también ellos tienen su propia historia.
Evita a las personas ruidosas y agresivas,
ya que son un fastidio para el espíritu.
Si te comparas con los demás,
te volverás vano o amargado
pues siempre habrá personas más grandes y más pequeñas que tú.
Disfruta de tus éxitos, lo mismo que de tus planes.
Mantén el interés en tu propia carrera,
por humilde que sea,
ella es un verdadero tesoro en el fortuito cambiar de los tiempos.
Sé cauto en tus negocios,
pues el mundo está lleno de engaños.
Pero no dejes que esto te vuelva ciego para la virtud que existe,
hay muchas personas que se esfuerzan por alcanzar nobles ideales,
la vida está llena de heroísmo.
Sé tú mismo,
y en especial no finjas el afecto,
y no seas cínico en el amor,
pues en medio de todas las arideces y desengaños,
es perenne como la hierba.
Acata dócilmente el consejo de los años,
abandonando con donaire las cosas de la juventud.
Cultiva la firmeza del espíritu

para que te proteja de las adversidades repentinas,
mas no te agotes con pensamientos oscuros,
muchos temores nacen de la fatiga y la soledad.
Sobre una sana disciplina,
sé benigno contigo mismo.
Tú eres una criatura del universo,
no menos que los árboles y las estrellas,
tienes derecho a existir,
y sea que te resulte claro o no,
indudablemente el universo marcha como debiera.
Por eso debes estar en paz con Dios,
cualquiera que sea tu idea de Él,
y sean cualesquiera tus trabajos y aspiraciones,
conserva la paz con tu alma
en la bulliciosa confusión de la vida.
Aún con todas sus farsas, penalidades y sueños fallidos,
el mundo es todavía hermoso.
Sé alegre. Esfuérzate por ser feliz.

Desiderata por Max Ehrmann.

Introducción

Un problema actual en astronomía es saber cuál es la relación entre la morfología de una galaxia y sus propiedades físicas observables. Es decir, encontrar las cualidades físicas que definen o determinan la forma geométrica de las galaxias, aspecto que todavía es usual hacer mediante una clasificación morfológica visual en el contexto actual de bases de datos complejas.

El contar con herramientas estadísticas, matemáticas y computacionales modernas permite analizar y clasificar nuevas galaxias sin necesidad de recurrir a realizar observaciones visuales para cada una de ellas. Esto implica un ahorro significativo de tiempo así como precisar la metodología con la cual se clasifican las galaxias.

El objetivo de este trabajo es proponer un enfoque del análisis topológico datos (ATD) y el aprendizaje estadístico para la clasificación de galaxias de acuerdo a su brillo en diferentes ondas del espectro electromagnético.

La base de datos analizada en este trabajo es de reciente creación; la cual resulta de una combinación de los catálogos SDSS, FIRST y NVSS [36, 37, 38]. Fue compilada por Amy Kimball y Željko Ivezić [2, 5], del Observatorio Radioastronómico Nacional, en Socorro, Nuevo México, iniciando en 2008 y actualizado en el 2018. Para cada una de las observaciones (galaxias) se tiene un total de 9 mediciones físicas, correspondientes a los *brillos* en radio y algunas bandas de colores provenientes de su espectro óptico. Cabe destacar que esta base de datos contiene más mediciones para cada una de las galaxias pero en este trabajo nos restringimos al uso de 9 de ellas. Se tiene un total de medio millón de observaciones de las cuales 1058 ya fueron clasificadas visualmente [5]. Es este trabajo proponemos una metodología para clasificar galaxias la cuál se contrasta con los resultados obtenidos de manera visual.

Como es usual, es primero necesario limpiar y adecuar esta base de datos para contar con una nube de puntos para su estudio. Se eliminaron y estimaron datos faltantes y se transformaron algunas de las mediciones para que todas estuviesen en las mismas unidades físicas.

Asimismo, se procedió a realizar un análisis exploratorio de los datos mediante una triangulación vía complejos simpliciales, con el objetivo de indagar sobre la topología de la nube de puntos. Se observó que la forma de los datos no presentaba grupos de homología de dimensión mayor que cero, es decir, solamente se observaron componentes conexas y no una estructura topológica más compleja.

En el marco del análisis topológico de datos y el aprendizaje estadístico, el problema de clasificación corresponde al de encontrar el número de componentes conexas de una nube de puntos — la dimensión (β_0) del grupo de homología cero.

Se empleó el algoritmo Mapper para realizar un análisis de la estructura topológica de los datos. Este algoritmo presenta ventajas con respecto a otros métodos como el ser muy rápido, eficiente en el uso de memoria de cómputo y, bajo las condiciones adecuadas a describirse más adelante, capaz de resumir información topológica de la nube de puntos bajo estudio [1, 28, 42]. Otro motivo por el cual se decidió utilizar Mapper se debe a su cómoda implementación que nos permitió realizar una modificación del mismo el cual nos permite mantener un registro de la ubicación exacta de todas las galaxias en las diferentes componentes conexas.

Uno de los requerimientos para el uso del algoritmo Mapper es la elección de una función que ayude a discriminar entre las diferentes observaciones, llamada función de filtro. Esta elección no es automática y representa un reto — y arte — en el uso apropiado de Mapper de acuerdo a la naturaleza del problema. Una parte importante de este trabajo fue inspeccionar y analizar una variedad de funciones de filtro para discriminar entre las galaxias desde diferentes *ángulos*. Con estos análisis se obtenían varias componentes conexas de las cuales fue común que una de ellas fuese significativamente más grande que las demás. Al analizar con más detenimiento dicha componente se pudo observar que habían dos tipos de galaxias que se localizaban en diferentes extremos de la componente mayor, lo cual nos da como información que estos grupos son esencialmente distintos.

Debido a lo anterior, optamos por emplear un algoritmo de reducción de dimensión con el objetivo de visualizar los datos así como

su estructura y forma. Se optó por el algoritmo t-stochastic neighbor embedding (t-SNE), el cual es no lineal, que por la complejidad de nuestro datos, brinda una mejor separación entre los diferentes tipos de radio galaxias. El Análisis de Componentes Principales — el cuál es un algoritmo de reducción de dimencionalidad lineal — no brindó una separación clara de ellos.

Existen otras métodos para clasificar galaxias de manera automática. Entre los más exitosos se encuentran las redes neuronales y el aprendizaje profundo (*deep learning*) [15]. La ventaja que ofrece t-SNE con respecto a estos otros métodos es que éste nos da un valor numérico que está relacionado, de alguna manera, con las características físicas observables de las galaxias, lo cual resulta en una buena función de filtro para el problema bajo estudio.

Dos conclusiones importantes de nuestro trabajo son las siguientes. La primera, con la metodología propuesta se pudieron clasificar de manera clara y utilizando únicamente propiedades físicas dos de los seis tipos de radio galaxias existentes catalogadas por su morfología visual. Estos dos tipos, corresponden a los que tienen el mayor número de observaciones, se subdividen a su vez en dos grupos cada uno. La segunda, la cual corresponde a una conclusión metodológica, es que nuestra propuesta es eficiente en el uso de memoria y tiempo de cálculos, debido a que ambos, Mapper y t-SNE, lo son.

Existen varias implementaciones libres de estos algoritmos. Nosotros trabajaremos con las versiones hechas para `R` por nuestra mayor familiaridad con el lenguaje. Empleamos la versión de Mapper disponible en la librería `TDAmapper` elaborada por Paul Pearson profesor del Hope College [12]. Para t-SNE, se empleó el código creado por el autor original del algoritmo [30] el cual se encuentra disponible en la librería `Rtsne`.

Este trabajo consta de 5 capítulos. El Capítulo 1 presenta elementos de astronomía en el contexto de datos complejos. Se mencionan algunas de las herramientas estadísticas que se utilizan actualmente y cómo el análisis topológico de datos se ha empleado en otros problemas astronómicos como el estudio de la geometría de la red cósmica de galaxias [16, 17]. También realizamos una breve descripción sobre las

diferentes maneras con las cuales los astrónomos clasifican y catalogan las galaxias.

El Capítulo 2 describe la teoría relacionada con la topología computacional. Se presentan de manera breve los conceptos de complejos y homología simplicial al igual que las nociones básicas de homología persistente. Asimismo, se mencionan los códigos de barras, la cual es una herramienta valiosa que nos muestra, de manera cualitativa, información sobre los grupos de homología persistente encontrados en la estructura subyacente de los datos.

El Capítulo 3 da una descripción detallada y rigurosa del algoritmo Mapper. Mostramos la construcción topológica del mismo y su implementación en la práctica. Se hace énfasis en la importancia de la elección de la función de filtro y la cubierta abierta que se escoge, así como algunas sugerencias de cómo se pueden escoger estos parámetros. También se hace mención sobre algunas de las ventajas y desventajas que presenta este algoritmo.

En el Capítulo 4 se hace una descripción matemática del algoritmo de reducción de dimensión t-SNE. Mostramos la modelación del problema de reducción de dimensión desde un punto de vista probabilista. En particular, se enfatiza el que t-SNE emplea una distribución normal para los puntos de dimensión alta y una distribución t-student de colas pesadas para los puntos de dimensión baja. Se explica la razón de utilizar una distribución con colas pesadas y la elección de la t-student.

El Capítulo 5 hace una descripción de la base de datos empleada y su análisis en base a la metodología propuesta. Detallamos el trabajo de la *limpieza* de los datos, cambios de magnitudes y de escalas, i.e., como adecuamos los datos para poder trabajar con ellos. Mostramos los resultados de las exploraciones preliminares realizadas así como los resultamos obtenidos mediante el algoritmo t-SNE y los resúmenes topológicos que nos brinda Mapper.

Contenido

Capítulo 1

Análisis estadístico de datos en astronomía

1.1. Introducción

Varias ciencias naturales están experimentando el fenómeno de *big data*, y la astronomía es uno de los principales participantes de este fenómeno. Los telescopios modernos son capaces de producir terabytes de datos en cada observación, y las simulaciones requeridas para modelar el universo observable empuja a las computadoras a sus límites. Con el objetivo de extraer información importante de estos datos, los científicos necesitan ser capaces de pensar de forma computacional. Este capítulo tiene como propósito resumir brevemente algunas de las herramientas del aprendizaje estadístico empleadas actualmente en la interpretación y análisis de datos, asimismo mencionar una aplicación del análisis topológico de datos hecha en astronomía [16, 17]. Expondremos y motivaremos por la pregunta en la cual se basa este trabajo de tesis.

Con el pasar del Siglo XXI, se ha llevado a cabo un cambio en la manera de realizar investigación en astronomía. En la actualidad, la investigación está basada en el análisis estadístico y geométrico de los datos recolectados por los telescopios. El proceso se puede describir, a grandes rasgos, de la siguiente manera: se recolectan datos mediante observaciones – ya sea en la forma de imágenes, mediciones de inten-

sidad de luz, brillo, etc – de una sector específico del cielo. Luego, el astrónomo se encarga de analizar los datos mediante técnicas estadísticas y computacionales, así como su conocimiento de física e intuición, la geometría y evolución del universo.

Se han logrado muchos avances en esta nueva era de la astronomía con el análisis comprensivo de las base de datos. Entre algunos de los progresos logrados está el entendimiento de la formación y evolución del universo [27].

Hoy en día, existen una gran cantidad de catálogos los cuales fueron recabados utilizando datos de diferentes bandas del espectro de la luz, los cuales van desde los rayos-γ, rayos-X, ultravioleta (UV), óptico, infrarrojo hasta las ondas de radio. En este trabajo utilizaremos una base de datos que combina, mediante la técnica de *cross-matching*, datos en óptico y radio [5]. En el capítulo 5 detallaremos y describiremos las mediciones que vamos a emplear, así como su análisis.

Como ejemplo de los datos en astronomía, en las imágenes en [**1-1**] se muestra en la parte superior una representación de distribución de objetos en el *WISE Galaxy Catalog* y en la parte inferior una imagen de la distribución de las fuentes del catálogo NVSS, el cual muestra fuentes con densidad de flujo en un rango superior a los 2.5 mJy. Parte de los datos a trabajar provienen de este último catálogo.

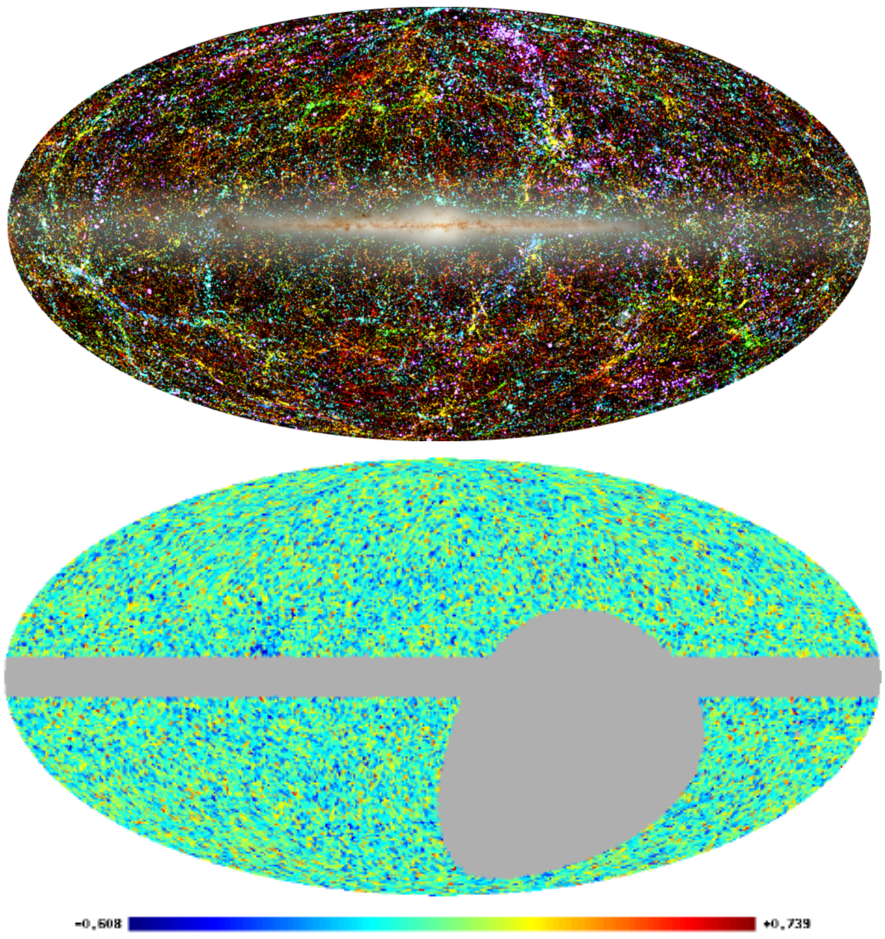


Figura 1-1
Distribución de galaxias en diferentes catálogos, [24] y [25]. Los colores representan la densidad de galaxias en la bóveda celeste.

1.2. Minería de datos en astroestadística

Con el objetivo de aprovechar y afrontar los obstáculos que estos datos masivos nos ofrecen, han nacido nuevas disciplinas como lo son la astroestadística así como la astroinformática. Estas ramas del conocimiento, altamente multidisciplinarias, buscan afrontar los problemas al con trabajar con bases de datos inmensas con el objetivo de encontrar patrones y explicaciones que yacen dentro de ellos.

Un paso importante cuando se está trabajando con datos reales es transformarlos de tal manera que puedan ser empleadas las herramientas ya conocidas para su análisis; ya que en muchas ocasiones estos vienen en la forma de imágenes, texto o binarios. La *minería de datos* en astronomía nos ayuda a encontrar las relaciones físicas y matemáticas en las diferentes poblaciones, al igual que ajustar y verificar las hipótesis de los modelos físicos que se aplican sobre ellos.

Entre algunas de las tareas que la minería de datos aborda está la clasificación, resumen, regresión, asociación, agrupamiento y detección de datos atípicos o anomalías (también llamados *outliers* por su nombre en inglés).

Además de las herramientas ya desarrolladas en el análisis de datos, surge en matemáticas una nueva llamada *análisis topológico de datos*. Este combina herramientas de la topología algebraica, ciencias de la computación, probabilidad y estadística para el estudio de los invariantes topológicos, bajo transformaciones homotópicas, de la estructura subyacente en los datos.

1.3. Algunas herramientas estadísticas empleadas en astronomía

En astronomía se emplean varias herramientas pertenecientes al área de aprendizaje estadístico. Entre algunas de las tareas y herramientas se encuentran las siguientes:

- Clasificación.
  Redes Neuronales, máquina de soporte vectorial (SVM), árboles de decisión, K-vecinos más cercanos y *random forests*.

- Regresión.
  Redes Neuronales, *Support Vector Regression* (SVR), árboles de decisión, *random forests*, regresión k-vecinos más cercanos, kernel de regresión, regresión con componentes principales.

- Agrupamiento.
  Análisis de componentes principales, K-medias, agrupamiento jerárquicos, modelos con mezclas Gausianas.

Las redes neuronales y el aprendizaje profundo (*deep learning*) están sido cada vez más empleadas en astronomía. En un estudio realizado por Marc Huertas-Company [15], astrónomo del observatorio de París, se utilizan técnicas de aprendizaje profundo para la identificación de galaxias en sus diferentes etapas de formación. Dicho trabajo ha resultado muy exitoso y se están realizando más investigaciones para tratar de mejorar los resultados obtenidos.

Observación 1.1. *Cabe resaltar que en dicho estudio se utilizan estas técnicas como cajas negras ya que no se sabe con exactitud qué elementos utiliza el algoritmo para dar su respuesta. Por otro lado, matemáticamente, no se entiende de manera exacta por qué las redes neuronales y el aprendizaje profundo son tan efectivos.*

1.3.1. Análisis topológico de datos

Entre las herramientas más novedosas utilizadas para el estudio de datos astronómicos se encuentran las provenientes del *análisis topológico de datos* ATD. Estas técnicas han sido empleadas para estudiar la geometría de la llamada *red cósmica de galaxias*. Algunos de estos trabajos pueden ser consultados en [16] y [17]. En la figura [**1-2**] vemos un ejemplo de α-complejos que se utilizan para hacer una representación de la red cósmica, en donde cada punto negro representa una galaxia.

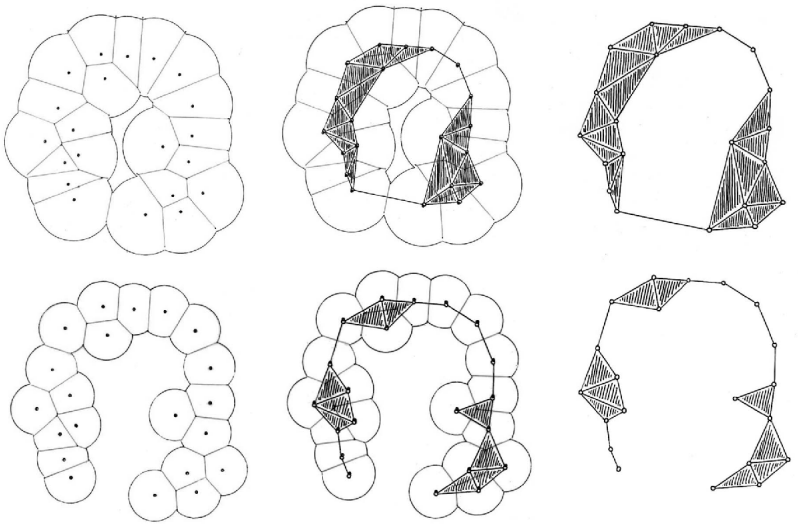


Figura 1-2: Presentación de la red cósmica de galaxias mediante los α-complejos [26]

También se pueden hacer otras conexiones de acuerdo a ciertos criterios predeterminados por el usuario (entre algunos de ellos tenemos: tamaño, distancia al vecino más cercano, forma, etc).

En este trabajo se utilizará una combinación de técnicas provenientes del aprendizaje estadístico y análisis topológico de datos. Nuestro enfoque y problema estudiado son distintos a los trabajos descritos arriba. Nosotros nos enfocaremos en un problema de clasificación y no al estudio geométrico de la distribución de galaxias en el espacio.

1.4. Tipos de clasificaciones de galaxias y radio galaxias

Entre los diferentes tipos de clasificaciones podemos resaltar aquellas basadas en la morfología. Entre las más famosas está la definida por Edwin Hubble, mejor conocida como el *esquema de Hubble* o el *diagrama de diapasón*, la cual fue desarrollada en 1936 [10]. En esta clasificación las galaxias son catalogadas como: elípticas, espirales, leticulares, espirales barradas, espirales intermedias e irregulares.

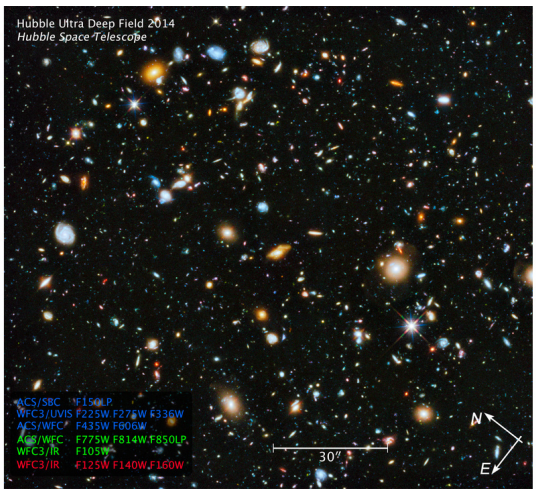


Figura 1-3: Galaxias observadas por el telescopio Hubble (Hubble Ultra Deep Field 2014 [40])

En la figura [**1-3**] podemos ver imágenes reales tomadas por el telescopio Hubble en donde podemos observar diferentes galaxias y su respectiva morfología.

Otra forma de hacer la clasificación de las galaxias es de acuerdo a la intensidad de luz (brillo) que estas emiten en diferentes bandas. Así, podemos clasificar las galaxias cercanas a la nuestra de acuerdo a que tanto brillan con respecto a las otras. El brillo que emiten las galaxias abarca todo el espectro electromagnético, en particular, nuestro interés es estudiar el brillo de las galaxias tanto en el espectro óptico como en el de radio. Los objetos con los cuales vamos a trabajar son llamados radio galaxias y se definen a continuación.

Definición 1.2. *Una Radio Galaxia es una galaxia activa que es muy luminosa en ondas de radio, con luminosidades de hasta* 10^{39} *W.*

Geométricamente, la estructura a gran escala más común de las radio galaxias son los denominados lóbulos: son estructuras dobles, aproximadamente elipsoidales y a veces simétricas situadas en cada lado del núcleo activo. El núcleo de las fuentes de radio está localizado dentro del objeto visible en óptico, ya sea una galaxia o un quasar. El tamaño de estas fuentes de radio va desde los pocos *parsec* hasta unos

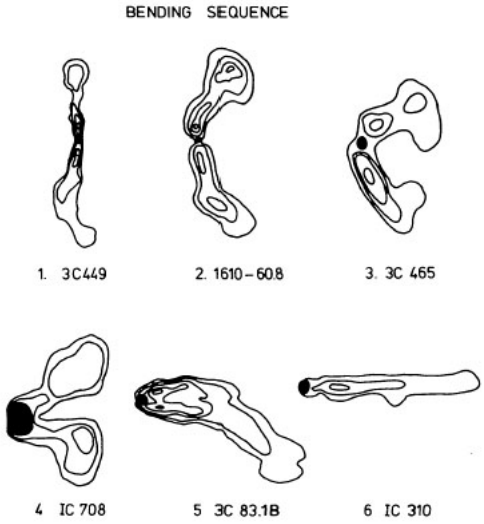


Figura 1-4: Morfología de radio galaxias. Normalmente la galaxia óptica se encuentra en el centro o dentro del componente más brillante (como en el caso 6) [32]

cuantos Mpc ($1pc = 3 \times 10^{13} km$). Las fuentes de radio más pequeñas son conocidas como *fuentes radio compactas* y las más grandes son llamadas *fuentes de radio gigantes*. En la Figura [**1-4**] podemos ver la morfología algunas de estas radio galaxias.

Sabemos que la cantidad de lóbulos que podemos observar puede depender del ángulo con el cual éstas son observadas desde la Tierra. Pero ésto no es algo que nos interese por el momento; lo que nos importa es el flujo total y los tamaños observados.

1.5. El problema astronómico

Actualmente en astronomía, una pregunta de interés es la siguiente:

¿Cuál es la relación entre las mediciones físicas y la morfología de las galaxias?

Nuestro objetivo consiste en abordar esta pregunta mediante herramientas del análisis topológico de datos y el aprendizaje estadístico.

Para ello, vamos a trabajar con una base de datos que contiene el flujo en radio de muchas radio galaxias, datos fotométricos y espectrales en óptico, así como las distancias para cada una de ellas.

Como primer paso, antes de tratar de dar respuesta a la pregunta anterior, nos preguntamos:

¿Es posible encontrar una combinación de variables, basada en las propiedades físicas observables, capaz de reproducir la clasificación morfológica conocida de las galaxias?

En otras palabras, se quiere una manera *precisa* de automatizar el proceso de clasificación; un método capaz de clasificar galaxias sin necesidad de recurrir a la clasificación visual, pero sí utilizando propiedades físicas medibles. Cabe recalcar que en este trabajo no haremos uso de las imágenes mas sí las propiedades físicas de cada galaxias.

En los capítulos 2, 3 y 4 vamos a desarrollar la teoría matemática y computacional necesaria para este trabajo de tesis. En el capítulo 5, vamos a mostrar los resultamos obtenidos y el avance que se obtuvo en la resolución de la pregunta planteada en este capítulo.

Capítulo 2

Topología computacional

2.1. Introducción

La disciplina de *topología computacional* tiene una historia relativamente corta en comparación con otras ramas de las matemáticas como lo es el análisis, el álgebra y la geometría. La primera publicación, relacionada con el área, fue hecha por Robins [19] en el año 1998 y luego, de manera independiente, fue definida la noción de *homología persistente* por Edelsbrunner, Letscher y Zomorodian [20] en 2002.

La topología computacional es el estudio, mediante el uso de algoritmos, de las propiedades topológicas subyacentes en espacios discretos. Para ello, siempre se tiene en mente la implementación práctica de dichos algoritmos por medio de computadoras. La topología computacional es una rama interdiciplinaria que combina elementos de la topología algebraica, computación y recientemente probabilidad y estadística.

Esta disciplina tiene como uno de sus objetivos el estudiar los invariantes topológicos latentes dentro de una nube de puntos $D = \{x_1, x_2, \ldots, x_n\}$ y de esta manera poder inferir propiedades de ella, como puede ser el deducir la variedad en donde están encajados dichos puntos. Estas técnicas puedan emplearse para analizar datos provenientes de la vida real, así como haremos en este trabajo al analizar datos provenientes de radio galaxias. El *análisis topológico de datos* se ha posicionado en los últimos años como una herramienta muy prome-

tedora en el área de matemáticas aplicadas, ya que ha sido empleada en diversos sectores como lo es la biología, química, finanzas e ingeniería [21, 22].

En una plática con Edelsbrunner, nos relató como había surgido la idea de homología persistente. En los libros la idea aparece de manera muy elegante y formal, pero no se habla sobre el proceso de maduración y creación de la idea. *A nosotros nos interesaba saber más sobre el lado humando del proceso.* Nos comentó que él se encontraba trabajando con biólogos y deseaban crear un modelo computacional para cierta proteína. Se quería probar si dicha proteína contenía un agujero que pasaba a través de ella o no, lo cuál era importante para saber el proceso de sinterización de cierta enzima. Cuando fijaban diferentes radios el agujero aparecía y desaparecía. Por ello se tenía que pensar en cómo llevar un seguimiento del proceso, de esta manera, establecer cómo *persistían* las características de interés. Poco a poco esta idea, y gracias a varias pláticas con otros colegas matemáticos, se fue madurando y formalizando para finalmente lograr definir lo que conocemos hoy como homología persistente.

Este capítulo está basado en las notas introductorias escritas por F. Chazal, B. Michel M. Carrière, S. Oudot [6, 7]; al igual que la parte III del libro de Edelsbrunner [8].

La división del presente capítulo es la siguiente. Se empezará por definir las nociones básicas de topología algebraica, como lo es el concepto de complejo simplicial. Luego se definirán varias formas con las que se pueden construir dichos objetos para posteriormente definir la noción de homología simplecial. Finalmente, se definirá el concepto de *homología persistente.*

2.2. Complejos simpliciales

Una de los conceptos fundamentales que en el análisis topológico de datos es sin duda el de simplejos y complejos simpliciales abstractos. Los simplejos pueden pensarse como generalizaciones de triangulos. Un complejo simplicial abstracto es nada más una colección de simplejos que satisfacen ciertas propiedades, de las cuales hablaremos más adelante. El porqué estos objetos matemáticos resultan tan importantes yace en la idea de poder triangular conjuntos discretos de puntos (una nube de datos) para luego estudiar su estructura topológica y de esta manera poder inferir información de la misma.

Un (k-)simplejo estándar es la generalización k-dimensional de un triángulo. De manera más precisa, un simplejo es la envolvente conexa de un conjunto de $(k+1)$ puntos independientes afines en un espacio Euclidiano de dimención $d \geq k$. Veamos a continuación la definición formal matemática:

Definición 2.1. *Un* ***k-simplejo estándar*** Δ^k *es el subconjunto de* $\mathbb{R}^{k+1}$ *definido como:*

$$\Delta^k \coloneqq \left\{ (t_0, t_1, \ldots, t_k) \in \mathbb{R}^{k+1} \colon \sum_{i=0}^{k} t_i = 1, t_i \geq 0, \textit{para toda } i \right\}.$$

También se puede definir un simplejo para puntos arbitrarios del espacio.

Definición 2.2 (Simplejo k-dimensional.)**.** *Dado un conjunto de* $(k+1)$ *puntos* $X = \{x_0, x_1, \ldots, x_k\} \subseteq \mathbb{R}^d$ *afines independientes, el simplejo* $\sigma = [x_0, x_1, \ldots, x_n]$ *generado por* $\mathbb{X}$*, es la envolvente conexa de* $\mathbb{X}$*. Los puntos de* X *son llamados los vértices de* σ *y los simplejos generados por los subconjuntos de* X *son llamados las caras de* $\mathbb{X}$*.*

Para un complejo simplicial σ generado por X, la *dimensión* del simplejo es igual a la cardinalidad de X menos 1, i.e., $\dim(\sigma) = |X| - 1$. Los simplejos de dimensión $0, 1, 2, 3$ son comúnmente llamados vértices, aristas, triángulo y tetaedros, respectivamente.

A la hora de querer triangular una nube de puntos, para tratar de tener una buena aproximación de la misma, se necesitan emplear

muchos triángulos. Para dichos triángulos, nos gustaría definir cierta estructura para mantener un orden y poder diferenciar unos de otros. Es por ello que definiremos a continuación una estructura para un conjunto de simplejos.

Definición 2.3 (Complejo Simplicial Abstracto)**.** *Un complejo simplicial K es un conjunto de simplejos que satisfacen las siguientes condiciones:*

i. Cualquier cara de un simplejo de K está también en K.

ii. La intersección de cualesquiera dos simplejos $\sigma_i, \sigma_j \in K$ es el conjunto vacío $\emptyset$ o una cara de tanto σ_i como de σ_j.

Se le llama complejo simplicial abstracto ya que no estamos pensando en su realización geométrica en el espacio Euclidiano, sino en sí como un conjunto con la propiedad de ser cerrado bajo contensiones.

Definición 2.4 (Dimensión)**.** *Los diferentes subconjuntos de K son los simplejos de K. La dimensión de un simplejo σ se define como* $\dim(\sigma) = \text{card}(\sigma) - 1$. *La dimensión de un complejo simplicial abstracto está dada por:*

$$\dim(K) \coloneqq \max_{\sigma \in K} \dim(\sigma).$$

Para cada simplejo σ nos podemos fijar en todas las caras cuya dimensión sea igual a p. Dichas caras son conocidas como las p-caras de σ. De la definición anterior, si σ es un simplejo de un complejo simplicial abstracto K entonces todas las p-caras de σ también deben de estar contenidas en K. De esta manera, podemos entonces considerar el conjunto de p-caras. La unión del conjunto de p-caras se conoce como el p-esqueleto de K. Formalmente podemos escribir lo anterior como sigue:

Definición 2.5 (Esqueleto)**.** *Sea K un complejo simplicial abstracto y $p \leq \dim(K)$. El p-esqueleto de K se define como:*

$$\text{sk}_p(K) \coloneqq \bigcup_{\sigma \in \mathcal{H}_p} \sigma,$$

donde $\mathcal{H}_p \coloneqq \{\sigma \in K \colon \dim(\sigma) = p\}$.

Ejemplo 2.6. *El conjunto de vértices de K es el 0-esqueleto de K.*

Definición 2.7 (Cara maximal)**.** *Un simplejo σ de un complejo simplicial abstracto K se la llama cara maximal de K si*

$$\dim(\sigma) \geq \dim(\tau), \ \ \forall \tau \in K.$$

Algo importante que observar es lo siguiente: basta con saber cuáles son las caras maximales de un complejo simplicial K ya que, de manera automática por definición, todas las caras de éstos también deberán de estar en el complejo simplicial. Esta observación es importante para nosotros, ya que en la computadora sólo necesitamos almacenar en memoria las caras maximales de los complejos simpliciales.

Por otro lado, ahora que ya tenemos la definición de complejo simplicial, queremos estudiar de qué maneras podemos utilizarlos para la triangulación de una nube de puntos $\mathcal{D} \subseteq \mathbb{R}^n$. Para ello, podemos hacer uso de que los puntos se encuentran encajados en un espacio Euclidiano y que podemos definir diferentes métricas en dicho espacio. Sabemos que toda métrica define una topología en el espacio donde ésta está definida. Por esta razón podemos aventajarnos de ello y utilizarlo para la construcción de complejos simpliciales.

Comenzaremos recordando la definición de un espacio métrico.

Definición 2.8 (Espacio Métrico)**.** *Un espacio métrico (M,ρ) es un conjunto M asociado a una función $\rho\colon M \times M \to \mathbb{R}_+$, llamada distancia, de tal forma que, para cualesquiera $x, y, z \in M$:*

1. *$\rho(x, y) \geq 0$ y $\rho(x, y) = 0$, si y sólo si, $x = y$,*
2. *$\rho(x, y) = \rho(y, y)$,*
3. *$\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.*

Ejemplo. Un ejemplo de métrica es que conocida *distancia Euclidiana*, la cual se define para cualesquiera dos puntos $x, y \in \mathbb{R}^n$ de la siguiente manera:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2)},$$

en donde (x_i, y_i) representan la i-ésima coordenada canónica del los vectores x y y, respectivamente. Es fácil verificar que la función d anterior satisface los axiomas de un métrica.

Existe muchas métricas, en nuestro caso al analizar datos algunas métricas resultan de mayor utilidad. Entre algunas de ellas se encuentran las siguientes:

- L^1**.** $\|x, y\|_1 \coloneqq \sum_{i=1}^{n} |x_i - y_i|$.
- L^2**.** $\|x, y\|_2 \coloneqq \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$.
- **Chebyshev.** $d(x, y) \coloneqq \max_{1 \leq i \leq n} |x_i - y_i|$.

Todas las métricas anteriores definen el mismo espacio topoógico en $\mathbb{R}^n$ y llamado Espacio Euclidiano.

Ahora, ya con la noción de distancia, estamos en la capacidad de definir un tipo particular de complejo simplicial que nos serán de mucha utilidad más adelante. Los complejos de Čech nos ayudan a construir, de forma intuitiva, un simplejo sobre una nube de puntos dada.

Definición 2.9 (Complejo de Čech.)**.** *Sea ρ una métrica definida en $\mathbb{R}^n$. Para $r \geq 0$, Sea $B_r(x) = \{y \in \mathbb{R}^n \colon \rho(x, y) \leq r\}$, la bola cerrada de radio r con centro en $x \in X$. El complejo de Čech para un conjunto finito $X \subseteq \mathbb{R}^n$ de radio r, se define como:*

$$\check{C}ech(X) = \left\{ Q \subseteq X \colon \bigcap_{x \in Q} B_r(x) \neq \emptyset \right\}.$$

Para todo subconjunto $Q \subseteq X$ de cardinalidad finita. El gran inconveniente con los complejos del tipo Čech es su costo computacional para ser calculados. La razón es que se debe revisar, para cada $x \in X$, si la intersección con las bolas centradas en cada uno de puntos en $X \setminus \{x\}$ es no vacía. Hacer esto en dimensión dos y tres no es muy difícil, pero ya en dimensión mayor esto se complica al no tener intuición geométrica que nos pueda ayudar a hacer los cálculos.

Una forma de solucionar este problema, es modificando un poco la definición anterior y en lugar de verificar intersecciones entre cada uno

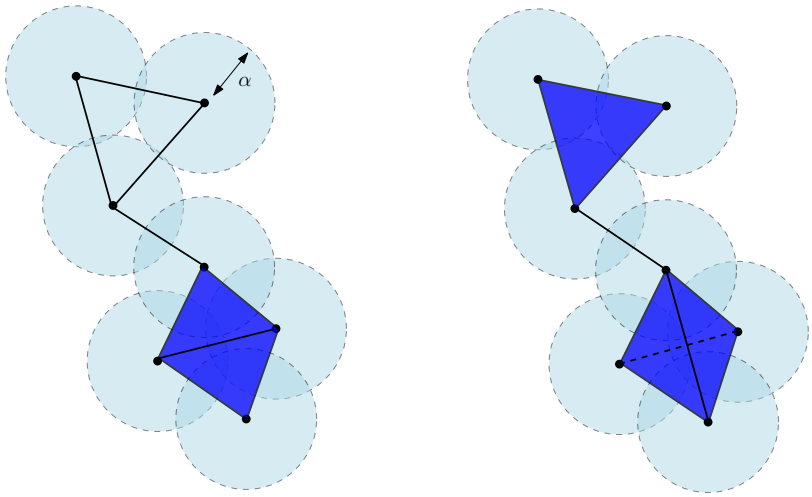


Figura 2-1: A la izquierda vemos un complejo de Čech. A la derecha vemos un complejo de Vietoris-Rips. (Chazal et al. [6])

de los puntos, se analizan las intersecciones por parejas. Estos tipos de simplejos se conocen como del tipo *Vietoris-Rips*, los cuales se definen a continuación.

Definición 2.10 (Complejo de Vietoris-Rips.)**.** *Sea X un conjunto de puntos en un espacio métrico (M, ρ) y α un número real no negativo. Un complejo simplicial de Vietoris-Rips, denotado por* $\text{Rips}_\alpha(X)$*, es un conjunto de simplejos $[x_0, x_1, \ldots, x_k]$ tales que $\rho(x_i, x_j) \leq \alpha$, para toda $i, j \in \{1, 2, \ldots, k\}$.*

Se sigue de forma inmediata de la definición que los complejos de Čech y Vietoris-Rips son, en efecto, complejos simpliciales abstractos. Sin embargo aún cuando X es un subconjunto finito de $\mathbb{R}^n$, $Rips_\alpha(\mathbb{X})$ no admite una realización geométrica en $\mathbb{R}^n$, con la misma n; en general, ésta puede ser de dimensión mayor que n.

En la figura [**2-1**] mostramos, para un mismo conjunto de puntos, los complejos de Čech y Vietoris-Rips que se construyen con el mismo radio $\alpha > 0$.

Otra forma interesante (y computacionalmente útil) de construir complejos simpliciales, es el conocido como *nervio* de una cubierta. En este caso no hacemos uso de ninguna métrica, mas sí se usa la noción

de conjunto abierto dentro un espacio topológico.

Definición 2.11 (Nervio de una Cubierta)**.** *Dada una cubierta* $\mathcal{U} = \{U_i\}_{i\in I}$ *(con* $|I| < \infty$*) de un conjunto* $\mathbb{M}$*, el nervio de* $\mathcal{U}$ *es el complejo simplicial* $N(\mathcal{U})$ *cuyos vértices están dados por los* U_i*'s tales que:*

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

Observación 2.12. *El nervio de una cubierta es un complejo simplicial abstracto. Para ver esto basta notar que si* $\sigma_1 = [U_1, \ldots, U_m] \in N(\mathcal{U})$*, con* $\sigma_2 \subseteq \sigma_1$*, entonces existe* $k \leq m$ *tal que*

$$\sigma_2 = \bigcap_{j=0}^{k} U_{i_j} \subseteq \sigma_1 = \bigcap_{i=0}^{m} U_i \neq \emptyset.$$

Para ver lo anterior, haremos uso de la definición de complejos simpliciales; queremos saber de que forma podemos comparar dos espacios topológicos, así poder discernir de unos y otros. Una de las maneras de hacer esto es comprobar si podemos deformar un espacio en otro de manera continua. Esto se conoce en topología como homeomorfismo, i.e., espacios que tienen la misma forma bajo deformaciones continuas.

Definición 2.13 (Homeomorfismo)**.** *Dos espacios topológicos* X, Y *se dicen homeomorfos si existen funciones continuas biyectivas* $f\colon X \to Y$ *y* $g\colon Y \to X$ *tales que* $f \circ g$ *y* $g \circ f$ *son las identidades en* Y *y* X*, respectivamente.*

Notar que en la definición anterior $g = f^{-1}$.

Resulta que pedir que dos espacios sean homeomorfos es una propiedad muy fuerte, en especial en el contexto de `TDA`, por lo que necesitamos otra manera de poder verificar si dos espacios topológicos comparten características similares. Resulta más sencillo verificar si dos espacios topológicos son homotópicos o no, cosa que definiremos a continuación, ésto junto con el nervio de una cubierta nos ayudarán a estudiar, más adelante, la estructura de un conjuntos de datos.

Definición 2.14 (Homotopía)**.** *Sean $f_0, f_1 \colon X \to Y$, dos funciones continuas. Dichas funciones se dicen homotópicas si existe una función continua $H \colon X \times [0,1] \to Y$ tal que, $\forall x \in X$,*

$$H(x,0) = f_0(x) \qquad y \qquad H(x,1) = f_1(x). \tag{2-1}$$

Dos espacios topológicos X y Y, se dicen ser *homotópicamente equivalentes* o simplemente *homotópicos*, si existen funciones continuas $f \colon X \to Y$ y $g \colon Y \to X$, tales que $f \circ g$ y $g \circ f$ son homotópicas a la función identidad de Y y X, respectivamente.

Teorema 2.15 (Teorema del Nervio)**.** *Sea $\mathcal{U} = \{U_i\}_{i \in I}$ una cubierta abierta de un espacio topológico X tal que, cualquier $H \subseteq I$ la intersección $\bigcap_{j \in H} U_j$ es vacía o contraible a un punto. Entonces, X y $N(\mathcal{U})$ son homotópicamente equivalentes.*

Comentario. El complejo de Čech es isomorfo al nervio de una colección de bolas cerradas. Por tanto, por el Teorema del Nervio, su homotopía es equivalente a la unión de bolas, $B_r(X) = \bigcup_{x \in X} B_r(x)$.

2.3. Homología simplicial

La teoría de homología simplicial está definida de forma general para campos conmutativos. Sin embargo, nosotros en `TDA` usualmente trabajamos con el grupo cíclico $\mathbb{F}_2 := \{0, 1\}$, el cual es también un campo conmutativo. La razón de utilizar $\mathbb{F}_2$ yace en que las cuentas se simplifican y se puede hacer uso de herramientas computacionales ya existentes para realizar todos los cálculos. Cabe recalcar que en la teoría de homología no importa el campo conmutativo con el cual se trabaje. Por esto, trabajar con un campo en particular no implica ningún problema o limitación en la teoría.

Para determinar si dos espacios topológicos son homeomorfos basta con dar el homeomorfismo entre ellos. Sin embargo, probar lo contrario no es tarea fácil. Una forma de hacer es encontrar invariantes topológicos, bajo homeomorfismos, que no sean los mismos en ambos espacios.

Una de las técnicas utilizadas para diferenciar espacios es la de estudiar los *agujeros* en diferentes dimensiones. La homología se encarga de formalizar el concepto de agujeros multidimensionales. Así, en esta sección la definiremos para nuestros objetos de estudio: los complejos simpliciales. Primero debemos definir funciones entre los espacios de complejos simpliciales.

Sea K un complejo simplicial (finito) y $k \in \mathbb{N}$. El espacio de k-cadenas de K, denotado por $C_k(K)$, es el conjunto de sumas (formales) finitas de k-simplejos de K. Es decir, si $\{\sigma_1, \ldots, \sigma_n\}$ es el conjunto de k-simplejos de K, las k-cadenas de K se pueden escribir de la forma:

$$c = \sum_{i=1}^{n} t_i \sigma_i, \text{ para } t_i \in \mathbb{F}_2.$$

Podemos también definir la suma entre cadenas de la misma forma que la suma entre polinomios, es decir, sumando término a término los coeficientes de los k-simplejos.

Lema 2.16. *$C_k(K)$ es un grupo conmutativo.*

Volviendo con los complejos simpliciales, para formalizar la idea de *detección de agujeros* necesitamos las siguientes definiciones de topología algebraica.

Definición 2.17. *La frontera de un k-simplejo $\sigma = [v_0, v_1, \ldots, v_k]$ es la $(k-1)$-cadena definida por:*

$$\partial_k(\sigma) = \sum_{i=1}^{k} (-1)^i [v_1, \ldots, \hat{v_i}, \ldots, v_k],$$

donde $[v_1, \ldots, \hat{v_i}, \ldots, v_k]$ es la $(k-1)$-cadena generada por todos los vértices a excepción de v_i.

Comúnmente, cuando no hay riesgo de confusión, a $C_k(K)$ es denotado simplemente por C_p.

Observación 2.18. *La definición anterior de ∂_k es valida para cualquier campo conmutativo; pero nosotros al estar trabajando en $\mathbb{F}_2$ se sigue que $(-1) = 1$, lo cual simplifica la expresión anterior.*

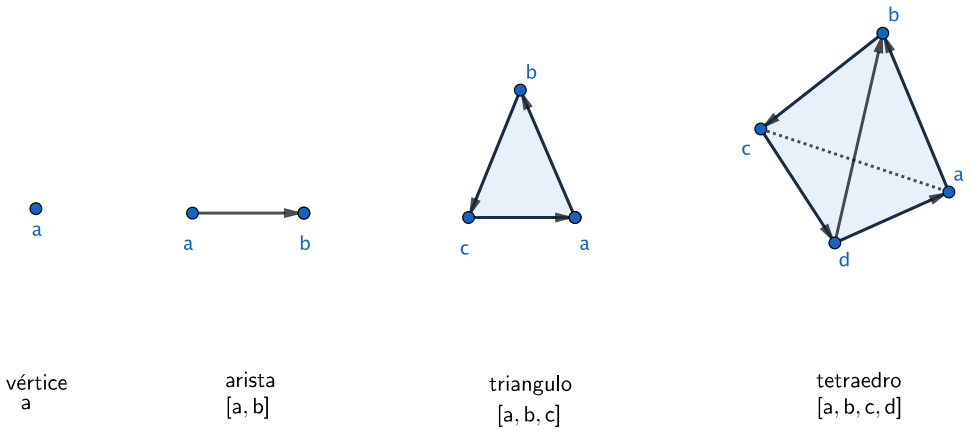


Figura 2-2: Ejemplos de complejos simpliciales.

Ejemplo. Vamos a aplicar operador frontera a los simplejos mostrados en la figura [**2-2**],

- $\partial_1[a,b] = b - a$.
- $\partial_2[a,b,b] = [b,c] - [a,c] + [a,b] = [b,c] + [c,a] + [a,b]$.
- $\partial_3[a,b,c,d] = [b,c,d] - [a,c,d] + [a,b,d] - [a,b,c]$.

Notemos que si tomamos cualquier simplejo de la figura [**2-2**] y aplicamos el operador frontera dos veces obtenemos como resultado el valor cero. Por ejemplo, si tomamos el triangulo $[a,b,c]$

$$\partial_1\partial_2[a,b,c] = [c] - [b] - [c] + [a] + [b] - [a] = 0.$$

Esta observación del operador frontera puede ser generalizada para cualquier complejo simplicial.

Teorema 2.19. $\partial_{k-1} \circ \partial_k \equiv 0$, *para cualquier* $k \geq 1$.

Demostración. Recordemos que un complejo simplicial es una coleeción de k-simplejos la cual es cerrada bajo intersecciones; por la linealidad de ∂_k, basta demostrar el teorema para cualquier k-simplejo $\sigma = [v_0, \ldots, v_k]$,

$$\begin{aligned}
\partial_k \circ \partial_{k+1}(\sigma) &= \partial_{k-1}(\partial_{k+1}(\sigma)) = \partial_k(\partial_{k+1}([v_0, v_1, \ldots, v_k])) \\
&= \partial_{k-1}\left(\sum_{i=1}^{k+1}(-1)^i[v_0, v_1, \ldots, \hat{v_i}, \ldots, v_k]\right) \\
&= \sum_{j<i}(-1)^i(-1)^j[v_0, \ldots, \hat{v_j}, \ldots, \hat{v_i}, \ldots, v_n] \\
&\quad + \sum_{j>i}(-1)^i(-1)^{j-1}[v_0, \ldots, \hat{v_i}, \ldots, \hat{v_j}, \ldots, v_n] \\
&= 0,
\end{aligned}$$

intercambiando i por j en la segunda suma vemos que la segunda cancela a la primera. □

Notemos que el operador frontera induce un operador $\partial_k \colon C_p \to C_{p-1}$ que a cada k-simplejo c le asigna su frontera. A este mapeo se le llama *mapa frontera* y no es difícil observar que dicho operador es un homeomorfismo de grupos.

Definición 2.20 (Ciclos.)**.** *El conjunto*

$$Z_k(K) \coloneqq \{c \in C_k(K) \colon \partial_k(c) = 0\},$$

es llamado el espacio de k-ciclos de K.

Definición 2.21 (Fronteras.)**.** *El conjunto*

$$B_k(K) \coloneqq \{c \in C_k(K) \colon \exists t \in C_{k+1}(K) \text{ tal que } \partial_{k+1}(t) = c\},$$

es llamado el espacio de k-fronteras de K.

Al igual que con C_p, cuando no hay riegos de confusión, escribimos simplemente Z_p y B_p. Observemos que los grupos Z_p y B_p son subgrupos de C_p. Ya que Z_p es el *kernel* de ∂_k y B_p la imagen de ∂_{k+1}.

En otras palabras, cualquier k-frontera es un k-ciclo, i.e. tenemos las siguientes contensiones de grupos, para cualquier complejo simplicial K,

$$B_k(K) \subseteq Z_k(K) \subseteq C_k(K).$$

Lo anterior permite definir un cociente entre grupos y consecuentemente definir los grupos de homología.

Definición 2.22 (Grupos de homología simplicial y números de Betti.)**.** *El k-ésimo grupo de homología de K es el espacio vectorial cociente*

$$H_k(K) \coloneqq Z_k(K)/B_k(K).$$

El k-ésimo número de Betti $\beta_k(K)$ de K es la dimensión del espacio vectorial $H_k(K)$, i.e., $\beta_k(K) \coloneqq \dim H_k(K)$.

A el cociente de grupos H_k se le conoce como clases de homología de dimensión k. Se dice que dos grupos son homólogos si están en la misma clase de homología. En otras palabras, decimos que $z_1, z_2 \in Z_k$ son homólogos si existe $b \in B_k$ tal que $z_1 = z_2 + b$.

Lema 2.23. *Si dos complejos simpliciales tienen realizaciones geométricas homotópicamente equivalentes, entonces sus grupos de homología son isomorfos y sus números de Betti coinciden.*

Observación 2.24. *Si se trabaja con coeficientes en $\mathbb{F}_2$, entonces los números de Betti tienen la siguiente interpretación. $\beta_0(K, \mathbb{F}_2)$, $\beta_1(K, \mathbb{F}_2)$ y $\beta_2(K, \mathbb{F}_2)$ significan, respectivamente, el número de componentes conexas, el número de agujeros (hoyos 2-dimensionales) y el número de vacios (hoyos 3-dimensionales) del complejo simplicial K.*

Teorema 2.25. *Sean X, Y dos espacios homotópicamente equivalentes. Entonces $H_*(X; \mathbb{F}) \simeq H_*(Y; \mathbb{F})$.*

2.4. Homología persistente

2.4.1. Introducción

Homología persistente es una herramienta que nos permite estudiar los invariantes toplógicos de un conjuto discreto $\mathcal{D}$ (nube de puntos) mediante una filtración dada por complejos simpliciales. La homología persistente estudia los grupos de homología de la nube de puntos a diferentes escalas. De forma más precisa, permite el estudio de la evolución

de las características topológicas de un conjunto de puntos, que posiblemente contengan algo de ruido, mediante una familia de espacios topológicos anidados.

Una manera intuitiva de entender la homología persistente es la siguiente. Imaginemos un conjunto de datos $\mathcal{D}$ y nos hacemos las siguientes preguntas: ¿será que los datos pertenecen o tienen alguna estructura topológica subyacente?. Si este es es el caso, ¿cómo podríamos calcular la estructura topológica de la nube de puntos?.

Claramente, al ser $\mathcal{D}$ un conjunto discreto no podemos aplicar directamente las herramientas de topología algebraica que ya conocemos. Lo que sí podemos hacer es triangular los datos vía complejos simpliciales y luego estudiar la estructura de estos últimos como una forma de aproximar los de $\mathcal{D}$. De aquí surge otra pregunta importante: ¿cuál es la triangulación que nos permitirá tener la *mejor* aproximación posible? Lo que podemos hacer es crear una filtración de complejos simpliciales y calcular los grupos de homología de dicha filtración a cada nivel. Luego, el siguiente paso consiste en estudiar los grupos que *persisten* con el tiempo ya que, de manera intuitiva, son estos los que nos proporcionan información importante de la nube de puntos $\mathcal{D}$.

2.4.2. Filtraciones y módulos de persistencia

Definición 2.26. *Sea K un complejo simplicial abstracto. Una **filtración** $\mathscr{F}$ del complejo K es una colección de subcomplejos $K_0, K_1, \ldots, K_n$ tales que*

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K.$$

Observemos que, de acuerdo a la definición anterior, que un simplejo $\sigma_i \in K_i$ no puede aparecer antes que lo hagan todas sus caras. Otra manera de ver a las filtraciones en la siguiente. Consideremos un complejo simplicial K cualquiera y una función $f\colon K \to \mathbb{R}$. A la función f se le pide ser monótona creciente sobre las caras de simplejo, es decir, si σ es una cara de τ entonces $f(\sigma) \leq f(\tau)$. La monotomía de f implica que los subcojuntos de nivel, $K(a) = f^{-1}(-\infty, a]$, es un subcomplejo de K para toda $a \in \mathbb{R}$. De esta manera, la definición anterior se obtiene tomando a n como el número de simplejos contenidos

en el complejo simplicial K y definiendo $K_m \coloneqq f^{-1}(-\infty, m]$ para toda $m \in \{0, 1, 2, \ldots, n\}$.

Observación 2.27. *La idea intuitiva detrás de homología persistente es la de estudiar la evolución de los grupos de homología a través de una filtración.*

Notemos que para cada $i \leq j$ se tiene una una inclusión natural entre el simplejo K_i y el K_j, por lo cual, existe también un homomorfismo inducido en los grupos de homología $H_p(K_i)$ y el $H_p(K_j)$, para cada dimensión p. Con lo cual obtenemos sucesión de grupos de homología inducida por homomorfismos:

$$0 = H_p(K_0) \to H_p(K_1) \to \cdots \to H_p(K_n) = H_p(K).$$

Formalizando lo anterior, los grupos de homología persistente se definen a continuación:

Definición 2.28. *Sea $\mathscr{F}$ una filtración. El (i, j)-ésimo grupo de homología persistente de nivel p de la filtración es la imagen de*

$$f_p^{i,i+1} \colon H_p(K_i) \to H_p(K_{i+1})$$

el homomorfismo inducido por la inclusión entre los grupos de homología,

$$H_p^{i,j} \coloneqq \operatorname{Im}\left(f_p^{i,j}\right).$$

Se define también el número de Betti persistente como $\beta_p^{i,j} \coloneqq rank(H_p^{i,j})$.

En topología algebraica, el p-ésimo número de Betti (β_p) de un complejo simplicial K tiene un significado geométrico muy interesante. Informalmente, β_p cuenta en número de superficies disconexas de dimensión p. De esta manera, los primeros 3 tres números de Betti tiene las siguientes interpretaciones.

- β_0 cuenta el número de componentes conexas.
- β_1 cuenta el número de agujeros (o círculos) dos-dimencionales.
- β_2 cuenta el número de hoyos (o vacíos) tri-dimencionales.

En la figura [**2-3**] Podemos observar los números de Betti para algunas variedades topológicas conocidas.

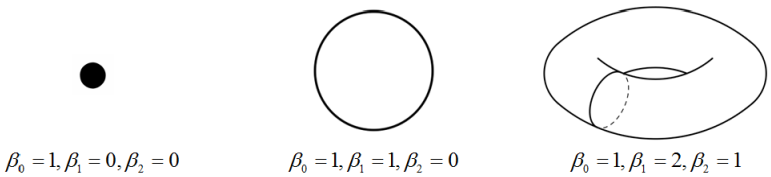


Figura 2-3: Números de Betti en variedades (Kraft [23])

Módulos de persistencia

El objeto más importante de homología persistente es el llamado módulo de persistencia, el cual definimos a continuación. Esa sección estará basada en la tesis doctoral de Carrière [9].

Definición 2.29. *Sea $\mathbb{K}$ un campo. Un módulo de persistencia U es un conjunto de $\mathbb{K}$-espacios vectoriales indexados por $\mathbb{R}$, denotado por $\{U_\alpha\}_{\alpha\in\mathbb{R}}$, y una familia de funciones lineales $\{u_\alpha^\beta \colon U_\alpha \to U_\beta\}_{\alpha,\beta\in\mathbb{R},\alpha\leq\beta}$ tales que:*

i. $\forall \alpha \in \mathbb{R},\ u_\alpha^\alpha = \mathrm{id}_{U_\alpha}$,
ii. $\forall \alpha \leq \beta \leq \gamma \in \mathbb{R},\ u_\beta^\gamma \circ u_\alpha^\beta = u_\alpha^\gamma$.

Como siempre que se da una nueva definición, veamos un ejemplo.

Ejemplo. Este ejemplo es conocido como *modulo de un intervalo* $\mathbf{I}_I$ para un intervalo $I \subseteq \mathbb{R}$. Como el conjunto de K-espacios vectoriales, definimos $(\mathbf{I}_I)_\alpha = K$ si $\alpha \in I$, en otro caso es $\{0_K\}$; para el conjunto de funciones lineales tomamos la función identidad en K, $(i_I)_\alpha^\beta = \mathrm{id}_K$ si $[\alpha, \beta] \subseteq I$, sino $(i_I)_\alpha^\beta$ se define cómo la función idénticamente cero.

Para poder trabajar de forma más general necesitamos tener una definición más general de filtración, ya que ahora estamos trabajando con espacios indexamos por los reales. Veamos la definición a continuación.

Definición 2.30. *Una filtración es una familia de espacios topológicos $\{X_\alpha\}_{\alpha\in\mathbb{R}}$ que están ordenamos con respecto a contenciones. Esto es, si $s \leq t$ entonces $X_s \subseteq X_t$.*

Definición 2.31 (Módulo de persistencia de una filtración.)**.** *Sea* $\mathbb{K}$ *un campo,* X *es una espacio topológico y* $\{X_\alpha\}_{\alpha\in\mathbb{R}}$ *una filtración de* X*. Sean* $H_p(X_s;\mathbb{K})$ *y* $H_p(X_t;\mathbb{K})$ *los p-ésimos grupos de homología de* X_s *y* X_t *para* $s \leq t$*. Definimos la inclusión* $\iota_s^t\colon H_p(X_s;\mathbb{K}) \to H_p(X_t;\mathbb{K})$ *inducido por la inclusión canónica* $X_s \hookrightarrow X_t$*. Notemos que estos mapeos dependen de la homología de dimensión p y pudieran no ser inyectivos. El p-ésimo módulo de persistencia de* X *asociado a la filtración* $\{X_s\}_{s\in\mathbb{R}}$ *es el módulo de persistencia* $\{H_p(X_\alpha;\mathbb{K}), \{\iota_s^t\}_{t\geq s}\}_{s\in\mathbb{R}}$*.*

Para poder hablar de manera más general, empezaremos por recordar la definición de una categoría en matemáticas:

Definición 2.32. *Una categoría* $\mathcal{C}$ *consiste en:*

- *Una colección de objetos Ob(*$\mathcal{C}$*) (e.g. conjuntos, grupos, espacios vectoriales, etc.)*
- *Para cada par de objetos* $X, Y \in Ob(\mathcal{C})$ *un conjunto* $\mathrm{Mor}_{\mathcal{C}}(X,Y)$ *los morfismos de* X *a* Y *(las funciones que van de* X *a* Y*, los homomorfismos entre los grupos* X *e* Y*, las transformaciones lineales entre* X *e* Y*, etc.)*
- *Composición de morfismos, i.e., una función*

  $$\circ\colon \mathrm{Mor}_{\mathcal{C}}(X,Y) \times \mathrm{Mor}_{\mathcal{C}}(Y,Z) \to \mathrm{Mor}_{\mathcal{C}}(X,Z)$$

  (lo que corresponde a la composición de funciones entre conjuntos, homomorfismos de grupos, transformaciones lineales, etc.)
- *Para cada objeto* $X \in Ob(\mathcal{C}$ *existe el elemento identidad* $id_X \in \mathrm{Mor}_{\mathcal{C}}(X,X)$*.*

La composición se supone que es asociativa en el sentido natural. Para cada $f \in \mathrm{Mor}_{\mathcal{C}}(X,Y)$ *se supone que* $id_Y \circ f = f$ *y que* $f \circ id_X = f$*.*

Observación 2.33 (Morfismos)**.** *Los módulos de persistencia pueden ser vistos como objetos en una categoría conmutativa, cuyos morfismos definiremos a continuación.*

Definición 2.34. *Sean* $\mathbf{U} = \{U_\alpha, u_\alpha^\beta\}$ *y* $\mathbf{V} = \{V_\alpha, v_\alpha^\beta\}$ *dos módulos de persistencia. Un morfismo* Ψ *entre* $\mathbf{U}$ *y* $\mathbf{V}$ *es una familia de morfismos* $\Psi = \{\psi_\alpha\colon U_\alpha \to V_\alpha, \alpha \in \mathbb{R}\}$ *tal que, para todo* $\alpha, \beta \in \mathbb{R}$ *que cumplen que* $\alpha \leq \beta$*, se tiene que* $\psi_\beta \circ u_\alpha^\beta = v_\alpha^\beta \circ \psi_\alpha$*.*

Si en la definición anterior cada uno de los ψ_α es un isomorfismo, entonces Ψ es llamado un isomorfismo y $\mathbf{U}, \mathbf{V}$ son isomorfos, en cuyo caso escribimos como $\mathbf{U} \simeq \mathbf{V}$.

2.4.3. Código de barras

Una manera de observar de forma cualitativa la evolución de los grupos de homología es mediante los llamados *códigos de barra* (ver capítulo 11.4 del libro de Edelsbrunner [8]). Un código de barra consiste en un conjunto de puntos en el plano que nos muestran el intervalo (a, b), con $a \in \mathbb{R}_+, b \in \mathbb{R}_+ \cup \{\infty\}$, en cuál apareció y desapareció un grupo de homología. A continuación presentaremos una serie de definiciones que nos ayudaran a formalizar la idea anterior.

Definición 2.35 (Nacimiento y Muerte)**.** *Sea $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$ una filtración y $f_p^{i,i+1}\colon H_p(K_i) \to H_p(K_{i+1})$, para cada $p \in \{0, 1, 2, \ldots, n\}$ la función entre los grupos de homología inducida por inclusión natural entre los simplejos. Sea $[\alpha] \in H_p(K_i)$ un clase de homología no trivial:*

- *Decimos que $[\alpha]$ nace en el tiempo t_i, al nivel i-ésimo de la filtración, si $[\alpha] \notin H_p^{i-1,i}$.*
- *Si $[\alpha]$ nace en K_i (i.e. a tiempo t_i) decimos que ésta muere entrando a K_j (a tiempo t_j) si se une con alguna clase mientras pasa de K_{j-1} a K_j, esto es, $f_p^{i,j-1}([\alpha]) \notin H_p^{i-1,j-1}$, pero si se cumple que $f_p^{i,j}([\alpha]) \in H_p^{i-1,j}$.*
- *Si una clase de homología $[\alpha]$ que nace a tiempo t_i muere a tiempo t_j, con $j \geq i$, su persistencia se define como $t_j - t_i$ y el índice de persistencia como $j - i$. Para las clases que no mueren, su persistencia se define como ∞.*

Ejemplo. En la figura [**2-4**] podemos observar en la primera imagen un conjunto de puntos con aparente forma de *ocho*. Se procede a construir paso a paso una filtración por medio de complejos de Vietoris-, junto con su respectivo código de barras. Se observa como persiste una de las componentes conexas al igual que uno de los agujeros de dimensión 1.

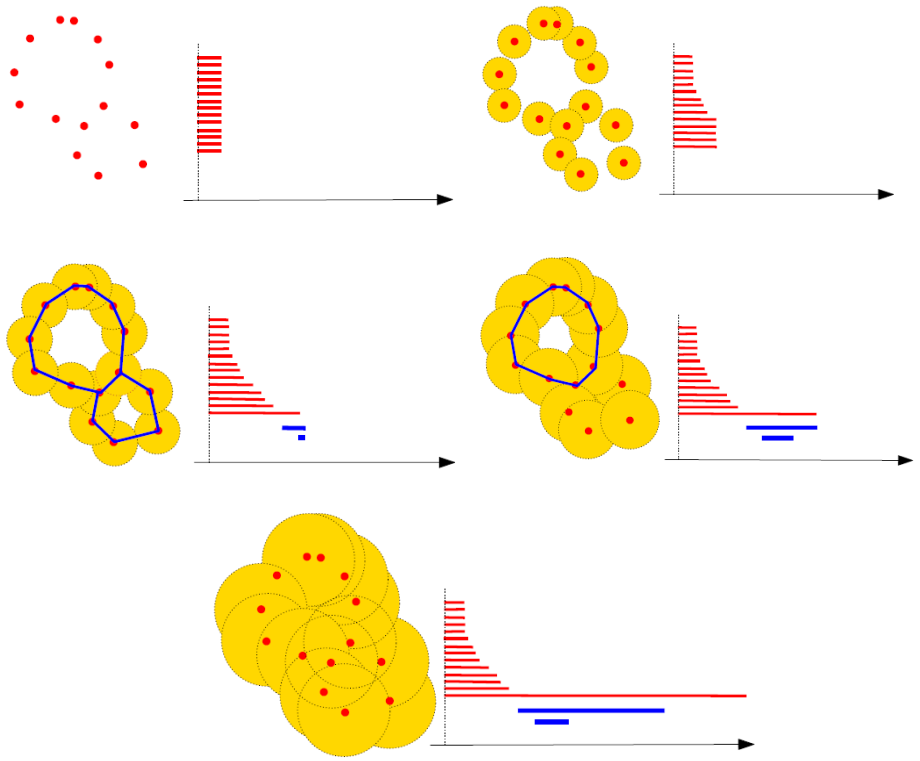


Figura 2-4: Ejemplo de un código de berras proveniente de una filtración vía complejos de Vietoris- (Chazal et al [6])

Capítulo 3

El algoritmo Mapper

3.1. Introducción

Mapper es un algoritmo que nos permite resumir la estructura topológica de una nube de puntos mediante un agrupamiento parcial de los datos y el uso de funciones que ayuden a discriminar entre ellos – conocidas como funciones de filtro. El algoritmo nos proporciona dicho resumen en la forma de un grafo (gráfica).

El algoritmo Mapper [1] fue desarrollado por G. Singh, F. Memoli, G. Carlsson en el 2007 en el Instituto de Matemáticas Computacionales de la Universidad de Stanford[1]. Mapper es una aplicación de la teoría discreta de Morse para el análisis de datos multivariados (High-Dimensional Data sets) y el reconocimiento de objetos tridimensionales. Este realiza un agrupamiento parcial de los datos mediante funciones que están definidas sobre ellos. Esto, sin ser dependiente de ningún algoritmo de agrupamiento específico.

La importancia de este algoritmo radica en la capacidad de poder realizar un análisis cualitativo, simplificación y visualización de los datos de manera eficiente en tiempo y memoria. En muchas ocasiones, los datos provenientes de observaciones reales que están encajados en un espacio de dimensión $d > 3$. Para poder visualizarlos, se busca una *buena* manera de cómo representarlos en un espacio de dimensión dos

[1] `https://icme.stanford.edu/`

o tres. En algunos casos, inclusive con un representación de dimensión baja de los datos, no es posible discernir la estructura geométrica de los mismos. Mapper es un propuesta para reducir la dimensión de los datos dando una representación simplicial de los mismos; la cual captura información geométrica y topológica a una resolución específica.

Existen dos versiones de este algoritmo. Una de ellas es la construcción topológica y la otra es la versión empírica (en la cuál se trabaja con datos reales). En la versión topológica, se trabaja con alguna función continua $f\colon X \to Z$, en donde X, Z son espacios topológicos. Para cualquier cubierta abierta $\mathcal{U} = \{U_\alpha\}_{\alpha\in A}$ de Z, lo que Mapper devuelve como resultado es el nervio del pull-back refinado de la cubierta abierta anterior, con esto queremos decir que $M(\mathcal{U}, f) \coloneqq N(f^{-1}(\mathcal{U}))$, en donde cada preimagen es divida en sus componentes conexas. Discutiremos esto con más detalle en la siguiente sección.

Notemos que Mapper resulta de gran utilidad cuando se requiere construir un complejo simplicial en algún espacio X muy complicado; ya que podemos trabajar en otro espacio Z más sencillo y luego tomar el pull-back de alguna cubierta abierta de Z bajo una función continua f definida entre estos dos espacios.

A continuación, estudiaremos con más detalle la versión topológica para luego mostrar una forma de como podemos llevar ésta idea a la práctica.

3.2. Versión topológica

Primero empezaremos definiendo el *nervio* de una cubierta. Esta definición es importante ya que será la forma en la cual se construirán los complejos simpliciales.

Definición 3.1. *Dada una cubierta finita $\mathcal{U} = \{\mathcal{U}_\alpha\}_{\alpha\in A}$ de un espacio topológico X, se define el* ***nervio*** *de la cubierta $\mathcal{U}$ como el complejo simplicial $N(\mathcal{U})$ cuyos vértices están dados por los elementos del conjunto A y la familia $\{\alpha_0, \alpha_1, \ldots, \alpha_k\}$ generan un k-simplejo en $N(\mathcal{U})$ si y sólo si $\mathcal{U}_{\alpha_0} \cap \mathcal{U}_{\alpha_1} \cap \cdots \cap \mathcal{U}_{\alpha_k} \neq \emptyset$.*

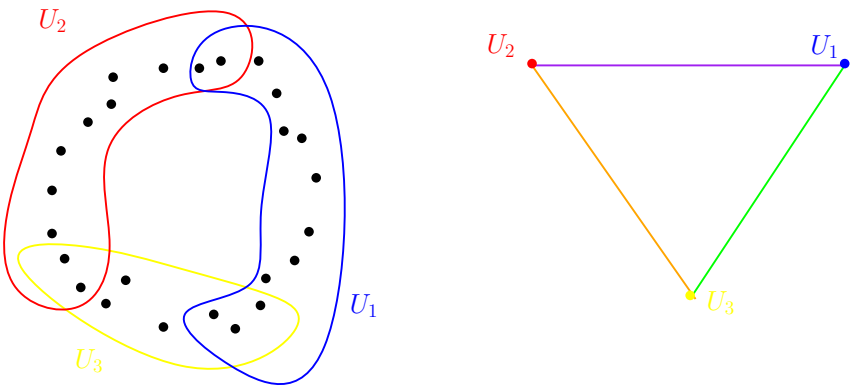


Figura 3-1: Ejemplo del nervio de una cubierta con puntos tomados de la esfera S^1. (Chazal et al. [6])

Como estamos dando una definición nueva de un concepto abstracto, es conveniente dar un ejemplo del mismo para ayudarnos a entender la idea.

Ejemplo. Tomemos 25 puntos aleatorios de manera uniforme sobre la esfera unitaria S^1 y una cubierta abierta de estos puntos dada por $\mathcal{U} = \{U_1, U_2, U_3\}$ así como se muestra en la figura [**3-1**]. Podemos ver a la izquierda de la figura los puntos tomados de la esfera así como la cubierta $\mathcal{U}$ y a la derecha el nervio de la cubierta $\mathcal{U}$. La imagen siguiente fue tomada de [6].

La siguiente definición juega un papel importante es la construcción del complejo simplicia que devuelve Mapper.

Definición 3.2 (Pullback de una cubierta.)**.** *Sea $f\colon X \to Y$, donde X, Y son espacios topológicos, una función continua y sea $\mathcal{U} = \{U_\alpha\}_{\alpha\in A}$ una cubierta abierta de Y. La cubierta pull-back de X inducida por $(f,\mathcal{U})$ es la colección de conjuntos abiertos $\{f^{-1}(U_\alpha)\}_{\alpha\in A}$. La cubierta pull-back refinada es la colección de componentes conexas de los conjuntos abiertos $f^{-1}(U_\alpha)$, para toda $\alpha \in A$.*

Nótese que la cubierta pull-back refinada de $\mathcal{U} = \{U_\alpha\}_{\alpha\in A}$ puede verse de la siguiente manera: definamos $V_\alpha = f^{-1}(U_\alpha)$, notemos que cada V_α es un conjunto abierto por ser la preimagen de un conjunto

abierto bajo una función continua. Entonces, para cada $\alpha \in A$, V_α puede ser escrita como

$$V_\alpha = \coprod_{i \in \{1,\ldots,c(\alpha)\}} V_\alpha^i,$$

en donde $c(\alpha)$ es el número de componentes conexas en V_α. Entonces,

$$\mathcal{V} = \{V_\alpha^i\}_{\alpha,i}$$

para toda $\alpha \in A$ y toda $i \in \{1, \ldots, c(\alpha)\}$. Por ello $\mathcal{V}$ es una cubierta conexa de X. Por tanto, $\mathcal{V}$ es la cubierta pull-back refinada de $\mathcal{U}$ de f.

Ya tenemos las herramientas necesarias para poder definir formalmente a Mapper, la defínción es la siguiente:

Definición 3.3 (Mapper)**.** *Sea $f\colon X \to Z$ una función continua, con X, Z espacios topológicos. Sea $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ alguna cubierta abierta de Z. Sea $\mathcal{V}$ la cubierta pull-back refinada de $\mathcal{U}$ bajo f. Mapper es el conjunto $M(f,\mathcal{U})$ conformado por el nervio de $\mathcal{V}$, i.e.*

$$M(\mathcal{U}, f) \coloneqq N(\mathcal{V}).$$

Observación 3.4. *Nótese que Mapper depende de tanto la función $f\colon X \to Z$, así como de la cubierta $\mathcal{U}$ del conjunto Z. Por eso se denota de esta manera: $M(f,\mathcal{U})$.*

En resumen, la idea principal de detrás Mapper es la siguiente. Dada una nube de puntos X, un espacio topológico Z y una *conveniente* función continua $f\colon X \to Z$, poder resumir X a través de la cubierta pull-back refinada de una cubierta $\mathcal{U}$ de $f(X)$. Si se escoge una buena cubierta abierta de X y función f, podremos ser capaces de resumir con una gráfica (grafo) información de X.

3.2.1. Funciones de filtro

La función continua $f\colon X \to Z$, también llamada *función de filtro*, que relaciona a los espacios topológicos con los que estamos trabajando juega un papel muy importante en la construcción del complejo simplicial. Por ejemplo, es fácil observar que la función constante $f(x) = c$,

para todo $x \in X$ y algún $c \in Z$, no nos está proporcionando ninguna información sobre X, ya que la preimagen de $f^{-1}(I) = X$, para todo $I \subseteq Z$ tal que $c \in I$.

La función de filtro debe escogerse de tal manera que se realcen las propiedades que se están estudiando en cada situación específica. Se buscan funciones que resalten las propiedades que se quieren estudiar.

Como se menciona en [6], las funciones de filtro que se emplean usualmente son las siguientes:

I. Estimación de densidades.
II. Análisis de componentes principales.
III. Funciones de centralidad y excentricidad, las cuales son respectivamente las siguientes:

$$f(x) = \sum_{y \in X} d(x, y), \qquad f(x) = \max_{y \in X} d(x, y),$$

donde $d(\cdot, \cdot)$ es una métrica sobre el espacio X.
IV. Alguna función que describa a lo datos.

Veamos un ejemplo en donde se resalta la importancia de la función de filtro.

Ejemplo. Supongamos que tenemos una función continua f, definida de la siguiente manera:

$$\begin{gathered} f \colon \mathbb{S}^1 \subseteq \mathbb{R} \to [-1, 1], \\ (x, y) \mapsto y \end{gathered}$$

para todo punto $(x, y) \in \mathbb{S}^1$. Esta función manda el círculo unitario al intervalo $[-1, 1] \subseteq \mathbb{R}$. Dada la cubierta abierta del intervalo cerrado [-1,1] (bajo la topología de subespacio)

$$\mathcal{U} = \{[-1, -1/3), (-1/2, 1/2), (1/3, 1]\},$$

Observamos que la preimagen de $f^{-1}([-1, -1/3))$ y $f^{-1}([1/3, 1))$, geométricamente, están conformadas por una componente conexa cada una. Por otro lado, la preimagen de $f^{-1}([-1/2, 1/2))$ está compuesta

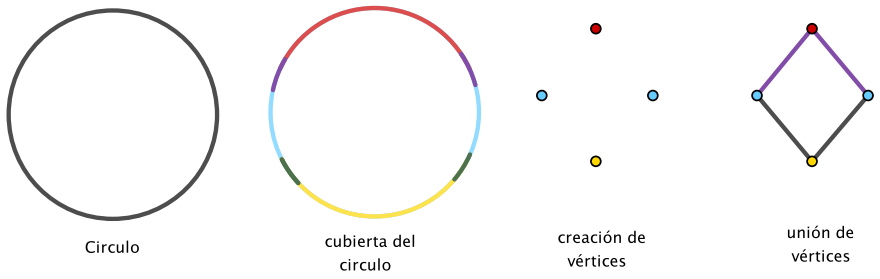


Figura 3-2: Complejo simplicial construido con el pull-back refinado de una cubierta del intervalo [-1,1].

por dos componentes conexas. Recordemos que el siguiente paso es subdividir cada una de las preimagenes en sus componentes conexas. Cada una de las componentes es un vértice del simplejo. Ponemos una arista entre cualesquiera dos vértices si, y sólo si, la intersección entre las componentes correspondientes es no vacía. Notemos que con la función y cubierta anterior obtenemos el complejo simplicial mostrado en la figura [**3-2**].

Observación 3.5. *Notemos que tanto la cubierta $\mathcal{U}$ como la función de filtro f que fueron empleadas en este ejemplo funcionaron muy bien ya que sabíamos de antemano información sobre la geometría de los datos.*

La función de filtro juega el papel de *discriminante* en la base de datos. Esta nos sirve para, sin importar el orden en la cual se encuentran las observaciones, tratar de acercar las funciones que tienen un valor de filtro similiar y alejar a las que no. De esta manera, se garantiza que elementos que están en una misma componente conexa es porque comparten entre sí características en común. Es de aquí que proviene el nombre de *filtro*. Esta función es la que reorganiza las observaciones de tal manera que observaciones similares (cualquiera que sea nuestro criterio de similitud) se encuentren cercanas una de otras y alejadas de las que lo son.

3.2.2. Cubiertas abiertas

Dada la función de filtro f, una forma estándar en la cual podemos escoger la cubierta $\mathcal{U}$ de $f(X)$ es: escogiendo un real $r \geq 0$ y un porcentaje $p \in (0, 1)$ para así generar un conjunto de bolas abiertas de radio r, este valor es conocido como *resolución*, y con porcentaje de intersección entre pares de las bolas consecutivas, de tal forma que la unión de todas las bolas abiertas cubran al conjunto X.

Observación 3.6. *Cabe mencionar que Mapper es también sensible a cambios en las cubiertas. Cambiar el tamaño de la resolución o porcentaje puede dar lugar a salidas diferentes. Por ejemplo, para un conjunto acotado X, para $x \in X$ existe $r > 0$ tal que $X \subseteq B_r(x)$; pero la preimagen de esta cubierta, claramente, no nos está dando información relevante de X. Una forma de como proceder es: probar con varios valores de r y p y quedarse con aquellos que nos den más información, de acuerdo a las características que se estén buscando.*

Supongamos que tenemos dos cubiertas distintas de X. ¿Qué relación tienen los complejos que son construidos, como se dijo antes, con sus repectivos nervios? Para responder esta pregunta, procedamos a tomar dos cubiertas y estudiar lo que sucede. Sean $\mathcal{U} = \{\mathcal{U}_\alpha\}_{\alpha \in A}$ y $\mathcal{V} = \{\mathcal{V}_\beta\}_{\beta \in B}$ dos cubiertas del espacio X. Un mapa entre cubiertas es una función $f\colon A \to B$ tal que $\mathcal{U}_\alpha \subseteq \mathcal{V}_{f(\alpha)}$, $\forall \alpha \in A$.

Observemos que, si f es un mapa entre dos cubiertas $\mathcal{U} = \{\mathcal{U}_\alpha\}_{\alpha \in A}$ y $\mathcal{V} = \{\mathcal{V}_\beta\}_{\beta \in B}$, entonces f induce una mapeo simplicial

$$N(f)\colon N(\mathcal{U}) \to N(\mathcal{V}).$$

Ya que, por la forma en que fue defininda el nervio de una cubierta, f manda vértices de $\mathcal{U}$ a vértices en $\mathcal{V}$. Consecuentemente, si tenemos una familia de cubiertas $\mathcal{U}_1, \mathcal{U}_2 \ldots, \mathcal{U}_N$ y mapas entre cubiertas $f_i\colon \mathcal{U}_i \to \mathcal{U}_{i+1}$, para $i = 1, \ldots, N$-1, obtenemos un diagrama de complejos simpliciales y un mapeo simplicial:

$$N(\mathcal{U}_0) \xrightarrow{N(f_0)} N(\mathcal{U}_1) \xrightarrow{N(f_1)} \ldots \xrightarrow{N(f_{N\text{-}1})} N(\mathcal{U}_N).$$

Empecemos suponiendo que contamos con un espacio topológico X y una función continua $f\colon X \to Z$, donde Z es un espacio de referencia

(también llamado espacio de parámetros), digamos el intervalo $[a, b] \subseteq \mathbb{R}$. Supongamos también que Z cuenta con una cubierta abierta $\mathcal{U} = \{U_i\}_{i \in I}$, donde $|I| < \infty$, es decir, $\mathcal{U}$ es una cubierta finita de Z. Ya que f es una función continua, se tiene que $f^{-1}(U_i)$ es un conjunto abierto en X para toda $i \in I$. Por lo tanto, el conjunto $\{f^{-1}(U_i)\}_{i \in I}$ conforma una cubierta abierta para X, i.e.,

$$X \subseteq \bigcup_{i \in I} f^{-1}(U_i).$$

Ahora, notemos que para cada $i \in I$, podemos considerar la descomposición del conjunto $f^{-1}(U_i)$ en sus componentes conexas de tal manera que lo podamos escribir de la siguiente manera:

$$f^{-1}(U_i) = \bigcup_{j=1}^{c(i)} V_i^j,$$

en donde $c(i)$ representa el número de componentes conexas en $f^{-1}(U_i)$. Realizando este proceso obtenemos una cubierta abierta para el conjunto X, a esta cubierta la denoramos por $\overline{\mathcal{U}}$.

Ahora, consideramos dos espacios topológicos X y Z, así como una función continua $f\colon X \to Z$. Supongamos que se tienen dos cubiertas finitas $\mathcal{U}, \mathcal{V}$ de Z. Si se tiene un mapeo de cubiertas $\mathcal{U} \to \mathcal{V}$ de Z, entonces existe un mapeo correspondiente entre las cubiertas $\overline{\mathcal{U}} \to \overline{\mathcal{V}}$. Para ver esto, notemos que si $U \subseteq V$ entonces $f^{-1}(U) \subseteq f^{-1}(V)$ y es claro que cada componente conexa de $f^{-1}(U)$ está contenida en exactamente una de las componentes conexas de $f^{-1}(V)$. Así, el mapa entre cubiertas de $\overline{U}$ a $\overline{V}$ requiere que el conjunto $U_\alpha(i)$ sea enviado a un único conjunto de la forma $V_{f(\beta)}(j)$ tal que $U_\alpha(i) \subseteq V_{f(\beta)}(j)$.

3.3. Agrupamientos jerárquicos

En general, el objetivo de un algoritmo de agrupamiento es el de particionar el espacio de observaciones en un número de grupos de tal manera que si dos observaciones están en un mismo grupo significa que ambas tiene características similares.

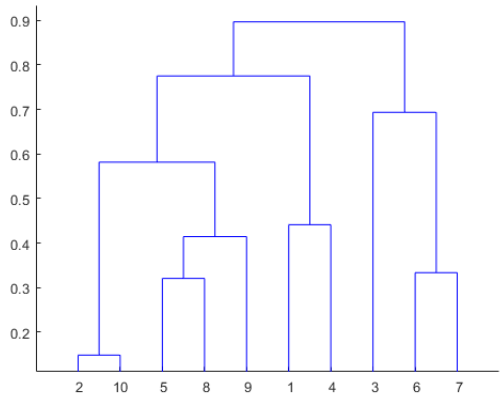


Figura 3-3: Ejemplo de un dendograma

Se necesitan dos ingredientes para poder aplicar un algoritmo de agrupamiento(*clustering*): un conjunto finito de puntos y una medida de disimilitud entre ellos. El algoritmo [1] nos muestra los pasos generales utilizados en cualquier algoritmo de agrupamiento.

Algoritmo 1 El algoritmo de aglomeramiento

1. Se inicia con tanto grupos como puntos haya, en donde cada punto se encuentra en un único grupo. La medida de similitud entre los grupos se toma como la distancia entre cada uno de los puntos.
2. Buscar las parejas de grupos más cercanos (similares) y luego unirlos en un mismo grupo.
3. Recalcular las distancias entre cada uno de los grupos formados, tanto de los nuevos formados como de los viejos.
4. Repetir los pasis 2 y 3 hasta que se alcance el número deseado de grupos o, en su defecto, que todos los puntos se encuentren en un mismo grupo.

Estos métodos también son conocidos como *algoritmos de aglomeramiento jerárquicos* ya que es posible construir un árbol de clasificación llamado *dendograma* [**3-3**] el cual nos indica la cantidad de grupos existentes así como el nivel al cual dos grupos diferentes se unen.

Es importante notar que la medida de similitud juega un papel

muy importante, ya que es posible que dos medidas distintas nos den como resultado dos agrupamientos distintos. Por ejemplo, si nuestro conjunto de datos consiste de expresiones genéticas de un conjunto de individuos, uno puede definir como función de similitud a semejanza de los patrones genéticos de estos mismos; pero otro función de similitud podría bien ser la edad de los individuos. Estas dos opciones pueden dar como resultado dos agrupamientos distintos no habiando finalmente una *mejor* medida de similitud.

A continuación vamos a mostrar 3 de las variantes más importantes. En todos los casos siguientes supondremos que D es un conjunto finito de n puntos y

$$\varphi\colon \mathrm{D} \times \mathrm{D} \to \mathbb{R}_+,$$

en donde φ es una medida de similitud entre los puntos contenidos en D. Una posible φ puede ser la función distancia entre los puntos, notando que dos puntos son similares si distancia es muy *pequeña*.

3.3.1. Distancia mínima o similitud máxima (Single linkage)

Este método consiste en que la medida de similitud entre los grupos está dada por la similitud mínima entre sus componentes. De esta manera, tras efectuarse k pasos se habrán formado $n - k$ grupos.

La distancia entre dos grupos G_i (con n_i puntos) y G_j (con n_j puntos) está definida como:

$$d(G_i, G_j) \coloneqq \min_{x_l \in G_i, x_m \in G_j} \{d(x_l, x_m)\},$$

en donde $l = 1, \ldots, n_i; \quad m = 1, \ldots, n_j$.

Si se emplea una medida de similitud, la distancia entre dos grupos se define como:

$$\varphi(G_i, G_j) \coloneqq \max_{x_l \in G_i, x_m \in G_j} \{\varphi(x_l, x_m)\},$$

en donde $l = 1, \ldots, n_i; \quad m = 1, \ldots, n_j$.

3.3.2. Distancia máxima o similitud mínima (Complete linkage)

En este método la medida de similitud entre los grupos consiste tomar la distancia máxima entre sus componentes. Al igual que en el caso anterior, luego de k pasos se habrán formado $n-k$ grupos.

La distancia entre dos grupos G_i (con n_i puntos) y G_j (con n_j puntos) está definida como:

$$d(G_i, G_j) \coloneqq \underset{x_l \in G_i, x_m \in G_j}{\text{máx}} \{d(x_l, x_m)\},$$

en donde $l = 1, \ldots, n_i; \quad m = 1, \ldots, n_j$.

Si se emplea una medida de similitud, la distancia entre dos grupos se define como:

$$\varphi(G_i, G_j) \coloneqq \underset{x_l \in G_i, x_m \in G_j}{\text{mín}} \{\varphi(x_l, x_m)\},$$

en donde $l = 1, \ldots, n_i; \quad m = 1, \ldots, n_j$.

3.3.3. Distancia o similitud promedio ponderada (Average distance)

La distancia o similitud entre dos grupos, está definida por el promedio ponderado de las distancias o similitudes de los componentes de un grupo respecto a otro. Sean G_i (con n_i elementos) y G_j (con n_j elementos), supongamos que G_i está formado por otros dos grupos G_{i_1} y G_{i_2} con n_{i_1} y n_{i_2} elementos respectivamente, de tal modo que $n_i = n_{i_1} + n_{i_2}$.

Así, en términos de distancias (o similitudes), la distancia promedio ponderada es:

$$d(G_i, G_j) \coloneqq \frac{n_{i_1} d(G_{i_1}, g_j) + n_{i_2} d(G_{i_2}, g_j)}{n_{i_1} + n_{i_2}}.$$

Comentario 3.7. *Para una descripción más detallada recomendamos consultar la tesis Análisis Topológico de Datos: Robusticidad y análisis de sensibilidad de algoritmos [28].*

3.4. Versión empírica

Vamos a mostrar una manera en la cual se puede hacer la construcción anterior, i.e. la construcción topológica, pero en esta ocasión para una nube de puntos $\mathcal{D}$ dada. En este caso, un algoritmo de agrupamiento (*clustering*) será el que tome el papel de la noción de dividir el espacio en componentes conexas.

Supongamos que tenemos un conjunto X con N puntos, y una función $f\colon X \to Z$ que está definida en todo $x \in X$, con Z un espacio métrico (comúnmente tomamos $Z = \mathbb{R}, \mathbb{R}^2, S^1$). Es decir, se conoce el valor bajo f de todos los elementos de X. Se supone también que X cuenta con una medida de *disimilitud*. Se pide esto último para poder medir la diferencia entre los puntos de X y así poder emplear algún algoritmo de agrupamiento. Nótese que no se está imponiendo ninguna restricción sobre el algoritmo de agrupamiento que se quiera usar. El siguiente paso consiste en encontrar un intervalo que cubra a el rango de la función f, el cual llamaremos I. Nótese que esto se puede hacer la siguiente manera, definamos α y β como:

$$\alpha = \min_{x \in X} f(x), \qquad \beta = \max_{x \in X} f(x), \qquad I \coloneqq [\alpha, \beta].$$

Ahora, se divide I en intervalos más pequeños $\mathcal{S} \coloneqq \{I_i\}$, en donde los intervalos adyacentes tienen cierto nivel de traslape. Esto nos da dos parámetros para controlar la *resolución* del complejo simplicial que vamos a construir; los cuales son: la longitud de los intervalos más pequeños (l) y el porcentaje de traslape que tienen los intervalos adyacentes (p).

Luego, para cada uno de los intervalos $I_j \in \mathcal{S}$, definimos

$$X_j \coloneqq \{x \in X\colon f(x) \in I_j\}.$$

Claramente, los $\{X_j\}$ forma una cubierta de X i.e. $X \subseteq \bigcup_j X_j$. Cada uno de los X_j es dividido en grupos usando algún algoritmo de agrupamiento, obteniendo así conjuntos de la forma X_{jk}. Cada uno de estos grupos representa un vértice de la gráfica que estamos construyendo. Dibujamos una arista entre dos vértices si $X_{jk} \cap X_{lm} \neq \emptyset$. Es decir, los

grupos que corresponden a vértices adyacentes no tienen intersección vacía.

Así pues, podemos escribir el algoritmo Mapper como se muestra en el algoritmo 2.

Algoritmo 2 El algoritmo Mapper

Input: Un conjunto de puntos X con una medida de disimilitud entre sus puntos. Una función $f\colon X \to Z$, con $(Z = \mathbb{R}, R^2, S^1)$ y una cubierta $\mathcal{U}$ de $f(X)$.

1. Para cada $U \in \mathcal{U}$, descomponer $f^{-1}(U)$ en sus componentes conexas $C_{U,1}, \cdots, C_{U,k_U}$.

2. Calcular el nervio de la cubierta de X definida por $C_{U,1}, \cdots, C_{U,k_U}$, para cada $U \in \mathcal{U}$

Output: Un complejo simplicial:

- un vértice $V_{U,i}$ por cada uno de los grupos $C_{U,i}$,
- una arista entre $V_{U,i}$ and $V_{U',j}$ sí y sólo sí $C_{U,i} \cap C_{U',j} \neq \emptyset$

Observación 3.8. *Vemos que la idea geométrica de dividir las preimágenes de la cubierta de I es reemplazada por la idea de agrupamiento en cada uno de los X_j.*

Ventajas de Mapper

Entre algunas de las ventajas que tiene el utilizar Mapper en un análisis topológico exploratorio de una nube de puntos se encuentran las siguientes:

- Rapidez y eficiencia computacional en la construcción del complejo simplicial
- Salidas fáciles de entender en forma de una gráfica
- Con una modificación hecha del algoritmo, podemos llevar registro de la ubicación de cada galaxia.

Mapper es una herramienta útil al hacer una exploración inicial de los datos. Sirve para estudiar la estructura topológica de los mismos. De esta manera, poder extraer o intuir información sobre los diferentes grupos existentes en los datos.

Desventajas de Mapper

Entre algunas de las desventajas al utilizar Mapper se encuentran:

- Dependencia en la función de filtro y la cubierta abierta tomada.
- Rigidez con la cual se escoge la cubierta abierta.
- Dependencia de los parámetros.

Cómo escoger la función de filtro y la cubierta abierta son dos los convenientes más grandes que se tienen al emplear Mapper. La función de filtro debe ser los suficientemente *buena* para poder detectar características importantes de la topología de los datos. La forma en como se toma la cubierta abierta también es importante, ya que se deben definir con cuántos intervalos se cubrirá el conjunto y el nivel de empalme que tendrán los abiertos que estén ubicados de manera contigua. Con las implementaciones con que actualmente se cuenta sólo es posible únicamente escoger una cubierta abierta dada por rectángulos uniformes.

Con la dependencia a los parámetros nos referimos a que si cambiamos un parámetro en el algoritmo, como ser el porcentaje de traslape que tienen los abiertos, la gráfica puede cambiar drásticamente. Esto puede ser también considerado como una ventaja del algoritmo ya que, dada la misma función de filtro, podemos analizar la base de datos mediante diferentes *resoluciones*.

Capítulo 4

Aprendizaje estadístico

4.1. Introducción

Dada una colección finita de puntos n-dimensionales, los algoritmos de reducción de dimensión permiten encontrar una representación de éstos en un espacio de dimensión menor, usualmente de dimensión dos o tres, de tal manera que dicha representación preserve algunas nociones geométricas, como ser distancias o vecinos más cercanos entre los puntos. Al tener una representación de dimensión dos o tres de los datos, podemos visualizarlos y estudiar su forma. De esta manera, podemos ser capaces de identificar patrones en ellos. En el invierno del 2017, tuve la fortuna de poder participar en el verano de investigación científica organizado por la Universidad de Auckland, Nueva Zelanda (recordemos que, por debajo del ecuador, es verano en los meses de diciembre, enero y febrero). Ahí aprendí sobre la versatilidad del algotimo t-SNE y surgió la idea de hacer una combinación entre el análisis topológico de datos y el aprendizaje estadístico para nuestro problema de clasificación de galaxias. En el verano se trabajó mejorando la implementación del algoritmo en el lenguaje de programación R, bajo la dirección del Dr. James M. Curran. En este capítulo describiremos los fundamentos matemáticos del algoritmo. Para un estudio más detallado y ejemplos recomendamos [29].

4.2. t-distributed Stochastic Neighbor Embedding (t-SNE)

Sea $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, con cada punto $x_i \in \mathbb{R}^n$ (espacio de dimensión alta). Queremos encontrar un conjunto de puntos $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ de tal forma que los puntos $y_i \in \mathbb{R}^k$ (espacio de dimensión baja), con $k = 2, 3$, preserven, en sentido geométrico, la estructura de los puntos en el espacio de dimensión alta. Es decir, queremos que puntos que están cerca (lejos) en $\mathbb{R}^n$ permanezcan cerca (lejos) en su representación de dimensión baja.

Para poder encontrar esta representación de dimensión baja, a los puntos del conjunto $\mathcal{X}$ se les asignará una medida de probabilidad. De esta manera, poder discernir entre cualesquiera dos puntos x_i y x_j. Se comienza definiendo la siguiente probabilidad condicional para cada punto:

$$p_{i|j} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq j} \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}.$$

Luego, como queremos medir diferencias de forma simétrica, se define:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}.$$

Para alguna $\sigma_i > 0$. Como estamos pensado a las distancias como una forma de definir similitud hacemos $p_{i|i} \coloneqq 0$.

El algoritmo t-SNE es una modifición del algoritmo SNE. La diferencia radica que en lugar de asignar una distrución Gaussiana a los puntos de dimensión baja, se les asigna una distribución t-student con un grado de libertad. Esto se debe a que cuando nos movemos de un espacio de dimensión m a uno de dimensión n, con $m > n$, *perdemos* espacio. Es por ello que necesitamos de una distribución con colas pesadas para poder dispersar los puntos. A las diferencias entre los puntos y_i e y_j de $\mathcal{Y}$ se les asigna la siguiente probabilidad,

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_{k \neq l}(1 + \left\|y_k - y_l\right\|^2)^{-1}}.$$

Observemos la expresión anterior es simétrica con respecto a ambas i, j por lo que no es necesario ninguna simetrización como se hizo para los puntos x_i's de $\mathcal{X}$. Por un argumento análogo, se define $q_{ii} \coloneqq 0$.

Si los puntos y_i e y_j estuviesen correctamente localizados, la diferencia de los puntos $p_{ij} - q_{ij}$ debiera de ser cero. Motivamos por este argumento, recordamos que la divergencia de Kullback-Leibler es una forma de medir qué tan *diferentes* son dos distribuciones de probabilidad. Como ya hemos asignado una probabilidad para cada uno de los puntos en los conjuntos $\mathcal{X}$ e $\mathcal{Y}$, podemos minimizar la divergencia de Kullback-Leibler para así aprender cuál es la ubicación de los puntos en $\mathcal{Y}$. De este modo, t-SNE intenta minimizar la función de costo C dada por:

$$C = \sum_i \mathrm{KL}(P_i||Q_i) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4-1}$$

donde $P_i = \{p_{i1}, p_{i2}, \ldots, p_{iN}\}$ y $Q_i = \{q_{i1}, q_{i2}, \ldots, q_{iN}\}$, para cada $i \in \{1, 2, \ldots, N\}$.

Desaforadamente, la función de costo C es no conexa.Por ello es necesario utilizar un método especial para su minimización. Se utilizará el método de gradiente descendiente para ello.

4.3. Gradiente descendiente

Teorema 4.1. *El gradiente de la ecuación 4-1 tiene la forma siguiente:*

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}.$$

Demostración. Esta demostración fue tomada de [29]; se corrigieron algunos errores que contenía la original. Recordemos que las probabilidades asignadas a las diferencias entre los puntos estaban dadas por:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}, \qquad q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}},$$

y $p_{ii} = q_{ii} = 0$. La divergencia de Kullback-Leibler entre las distribuciones P_i y Q_i está dada por

$$\begin{aligned} C = \sum_i \mathrm{KL}(P_i||Q_i) &= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \\ &= \sum_i \sum_j \left(p_{ij} \log(p_{ij}) - p_{ij} \log(q_{ij}) \right). \end{aligned} \tag{4-2}$$

Ahora, para poder trabajar de manera menos engorrosa, vamos a introducir las siguientes variables auxiliares,

$$d_{ij} = \|y_i - y_j\|, \qquad Z = \sum_{k \neq l} \left(1 + d_{kl}^2\right)^{-1}.$$

Notemos que si y_i cambia, las únicas parejas de distancias que cambian son d_{ij} y d_{ji}, para toda j. Por tanto, usando la regla de la cadena, podemos ver el que gradiente de la función de costo C con respecto a y_i está dado por

$$\begin{aligned} \frac{\partial C}{\partial y_i} &= \sum_j \left(\frac{\partial C}{\partial d_{ij}} + \frac{\partial C}{\partial d_{ji}} \right) \frac{(y_i - y_j)}{\|y_i - y_j\|} \\ &= 2 \sum_j \frac{\partial C}{\partial d_{ij}} \frac{(y_i - y_j)}{\|y_i - y_j\|} \\ &= 2 \sum_j \frac{\partial C}{\partial d_{ij}} \frac{(y_i - y_j)}{d_{ij}}. \end{aligned} \tag{4-3}$$

Para poder calcular el gradiente de $\frac{\partial C}{\partial d_{ij}}$ vamos hacer uso de la ecuación (4-2). Notemos que la primera parte no depende de $d_{ij} = \|y_i - y_j\|$, tan solo depende de los valores x_i's:

$$\begin{aligned}
\frac{\partial C}{\partial d_{ij}} &= -\sum_{k\neq l} p_{kl} \frac{\partial \log(q_{kl}))}{\partial d_{ij}} \\
&= -\sum_{k\neq l} \frac{\partial \log(q_{kl}Z/Z)}{\partial d_{ij}} \qquad \text{(ya que } Z \neq 0\text{)} \\
&= -\sum_{k\neq l} p_{kl} \frac{\partial \left(\log(q_{kl}Z) - \log(Z)\right)}{\partial d_{ij}} \\
&= -\sum_{k\neq l} p_{kl} \left(\frac{1}{q_{kl}Z} \frac{\partial (1+d_{kl}^2)^{-1}}{\partial d_{ij}} - \frac{1}{Z}\frac{\partial Z}{\partial d_{ij}} \right).
\end{aligned}$$

En esta expresión hicimos uso del hecho que $q_{kl} = (1+d_{kl}^2)^{-1}/Z$. Ahora, notemos que $\frac{\partial (1+d_{kl}^2)^{-1}}{\partial d_{ij}}$ es cero solamente cuando $k = i$ and $l = j$. De este modo, el gradiante de $\frac{\partial C}{\partial d_{ij}}$ está dado por:

$$\frac{\partial C}{\partial d_{ij}} = 2\frac{p_{ij}}{q_{ij}Z}(1+d_{ij}^2)^{-2}d_{ij} - 2\sum_{k\neq l} p_{kl}\frac{(1+d_{ij}^2)^{-2}}{Z}d_{ij}.$$

Ahora, recordemos que $q_{ij} = (1+d_{kl}^2)^{-1}/Z$, y que $\sum_{k\neq l} p_{kl} = 1$, por lo que podemos reescribir la última ecuación de la siguiente manera:

$$\begin{aligned}
\frac{\partial C}{\partial d_{ij}} &= 2p_{ij}(1+d_{ij})^{-1}d_{ij} - 2q_{ij}(1+d_{ij})^{-1}d_{ij} \\
&= 2(p_{ij}-q_{ij})(1+d_{ij})^{-1}d_{ij}.
\end{aligned}$$

Sustituyendo este término en la ecuación (4-3) se obtiene

$$\begin{aligned}
\frac{\partial C}{\partial y_i} &= 2\sum_j \frac{\partial C}{\partial d_{ij}} \frac{(y_i - y_j)}{d_{ij}} \\
&= 2\sum_j 2(p_{ij}-q_{ij})(1+d_{ij})^{-1}d_{ij}\frac{(y_i-y_j)}{d_{ij}} \\
&= 4\sum_j (p_{ij}-q_{ij})(1+d_{ij})^{-1}(y_i-y_j).
\end{aligned}$$

Lo anterior prueba el resultado.

□

Ya que la función de costo es una suma que corre sobre todos los x_i's e x_j's vemos que éste es un algoritmo cuya complejidad es $\mathcal{O}(N^2)$, donde N es el número total de puntos en la nube de puntos $\mathcal{X}$.

Aún nos hace falta describir cómo encontrar el valor de σ_i para poder calcular el valor de $p_{i|j}$. La distribución de P_i tiene una entropía que decrece a medida que lo hace el valor de σ_i, t-SNE realiza una búsqueda binaria para encontrar el valor de σ_i que produce una distribución P_i con una perplexidad definida por el usuario. La perplexidad está definida por:

$$\text{Perp}(P_i) = 2^{H(P_i)},$$

donde $H(P_i)$ representa la de entropía de Shannon de P_i. Recordemos que la entropía está definida como:

$$H(P_i) = -\sum_j p_{ij} \log_2 p_{ij}.$$

Intuitivamente, la perplexidad indica el número de grupos que se espera encontrar en los datos. Este valor puede irse variando y de esta manera encontrar el valor que nos genera la mejor separación de los grupos encontrados.

4.4. Representación de dimensión baja

Ya tenemos todos los ingredientes necesarios para calcular la representación de dimensión baja de los puntos en el conjunto $\mathcal{X}$. Como ya se mencionó anteriormente, utilizaremos el método de gradiente descendiente. Se inicializan los puntos de $\mathcal{Y}$ generando una muestra aleatoria de una distribución Gaussiana isotrópica con varianza pequeña y centrada alrededor del origen. Para acelerar el tiempo del cálculo de la optimización y evitar caer en algún mínimo local, se agrega al gradiente un término relativamente grande llamado momentum. Matemáticamente, la actualización del gradiente con el término de momentum está dado por:

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\partial C}{\partial \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right),$$

donde $\mathcal{Y}^{(t)}$ indica la solución en la t-ésima iteración, η representa la tasa de aprendizaje y $\alpha(t)$ representa el momentum al tiempo de el t-ésima iteración.

Capítulo 5

Clasificación y resultados

5.1. Introducción

A una distancia considerable de nuestra galaxia, se encuentra una gran variedad de objectos llamados objetos extragalácticos. Estos tienen diferentes formas y tamaños; entre ellos los quasares y galaxias. Muchos de estos objetos emiten radiación en diferentes longitudes del espectro electromagnético como son radio, infrarrojo (IR), óptico o rayos-X, etc.

Los objetos que emiten en radio tienen una estructura típica (conformada por lóbulos, jets y un núcleo) como la que se muestra en la figura [**5-4**]. Existe una gran variedad de radio galaxias en el universo que difieren unas con otras en, por ejemplo, en sus diferentes formas geométricas, su color, luminosidad, espectro, etc. Al existir muchas variantes nos atrae la idea de poder clasificar estos objetos de acuerdo a sus similitudes.

Comúnmente, las radio galaxias se clasifican por su morfología mediante observaciones visuales. Sin embargo, no se sabe nada de la relación, si es que existe, entre esta clasificación visual y las propiedades físicas observables y medibles. Por ello es que surge la siguiente pregunta:

¿Es posible encontrar una combinación de variables, basada en las propiedades físicas observables, capaz de reproducir la clasificación morfológica conocida de las galaxias?

Es decir, queremos saber si es posible encontrar una clasificación, utilizando propiedades que podemos medir en diferentes bandas, de tal forma que esta clasificación coincida con la ya conocida de radio galaxias.

De esta manera, precisar la metodología con el cuál se clasifican las galaxias, con lo que podremos analizar y clasificar nuevas galaxias sin necesidad de recurrir a la observación visual. Esto implicaría una avance en la investigación astronómica ya que implica una ahorro sustancial en el tiempo de clasificación de objetos nuevos.

5.2. Base de datos

Los datos con los que se cuentan proviene de una compilación de 3 catálogos astronómicos distintos. Dichos datos fueron recolectados por Kimball & Ivezić [5] en 2014, del National Radio Astronomy Observatory [2]. A simple vista, hacer una recompilación de catálogos no parece una tarea tan difícil. Pero un momento de reflexión nos hacer pensar en la siguiente pregunta: ¿cómo se está seguro de que una galaxia que tomaron de uno de los catálogos es la misma que la que tomaron de otro? Habiendo tantas galaxias en $\mathcal{G}$ (el conjunto de todas las galaxias observables), no es trivial poder emparejarlas de tal manera que se pueda estar seguro, bajo cierta incertidumbre, de que estamos hablando de los mismos objetos. La metodología de *cross-matching* explicada en Kimball & Ivezić [5] en 2014, nos ofrece unos parámetros que miden la certeza del emparejamiento de estas galaxias. Además, este trabajo incluye una clasificación visual hecha por Kimball et al. [35] en 2011.

Los catálogos individuales son los siguientes:

• **Sloan Digital Sky Survey (SDSS):** Es un proyecto de investigación el cual recolecta imágenes en la banda visible, utilizando un sistema fotométrico de 5 filtros. Estas imágenes contienen el espectro óptico para cada una de estas fuentes. Ello permite extraer el corrimiento rojo y poder conocer sus distancias, con respecto a la Tierra. Esta operación es realizada con un telescopio con ángulo de campo

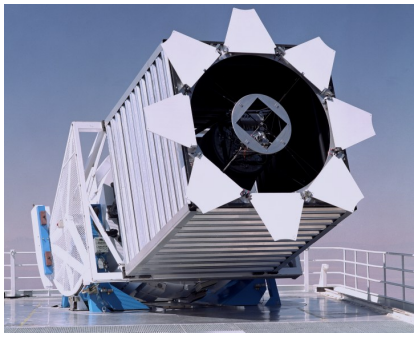


Figura 5-1: SDSS [36]

Figura 5-2: Conjunto VLA [39]

de visión de 2.5 metros de diámetro localizado el observatorio *Apache Point* de Nuevo México. La cámara del telescopio está formada por treinta CCD cada uno con una resolución de 2048×2048 pixeles, totalizando aproximadamente 120 Megapixeles. La fotografía [**5-1**] es del telescopio SDSS.

Los siguientes dos catálagos astronómicos fueron creados mediante el mismo radio telescopio. El *Very Large Array* (VLA) [33] es un observatorio radio astronómico localizado en Nuevo México. Este conjunto de radio antenas se encuentra sobre un sistema de rieles que permite acercarlos a alejarlos entre sí (ver imagen [**5-2**]). Entre más se alejan entre ellos mejor será la resolución y podrán ver con mas detalle los objectos que están dentro de su rango de visión. Si están muy cerca puedan capturar estructuras cuya emisión es más débil.

• **FIRST (VLA alta resolución):** Son las siglas en inglés de *Faint Images of the Radio Sky at Twenty-cm*. El catálogo cubre alrededor del 51 % del cielo, con regiones centradas en los polos galácticos norte y sur. Las regiones fueron escogidas de tal modo que cubrieran las mismas que el **SSDS**.

• **NVSS (VLA baja resolución):** Son las siglas en inglés de *NRAO VLA Sky Survey*. Este catálogo del hemisferio norte que cubre alrededor del 82 % del cielo.

Denotaremos por $\mathcal{B}$ a la base de datos de galaxias con la que con-

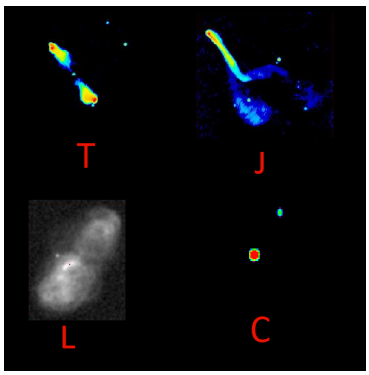


Figura 5-3: Ejemplos de diferentes clases de radio galaxias [34]

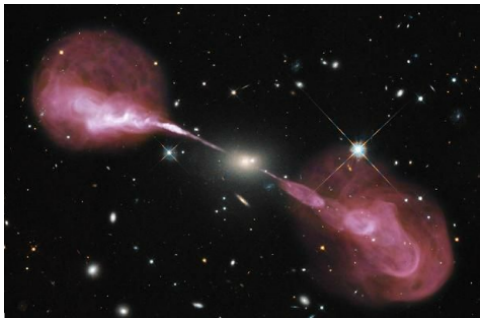


Figura 5-4: Sobre-posición de imagen en radio y óptico de la radio galaxia Hercules-A [41]

tamos. Ésta es una matriz de tamaño 581648×9; con 581,648 observaciones y 9 características, las cuales describiremos más adelante.

También se cuenta con una segunda base de datos $\mathcal{D} \subseteq \mathcal{B}$ en la cual, además de las mismas características que se tienen en $\mathcal{B}$, en ésta se incluye la clasificación de las galaxias contenidas. En $\mathcal{D}$ hay un total de 1058 observaciones ya clasificadas. De este modo la matriz $\mathcal{D}$ es de tamaño 1058×10. La base de datos $\mathcal{D}$ fue clasificada mediante observaciones visuales.

A continución, presentamos las características físicas que se tienen para cada una de las galaxias, así como una breve descripción del significado físico de cada uno de ellos.

First flux (*fflux*):
Densidad de flujo de las galaxias, este dato proviene del catálogo FIRST. La unidad de medida es el Jansky (W/m^2Hz). La densidad de flujo es la cantidad de brillo que podemos detectar de la radio galaxia.

NVSS flux (*nflux*):
Densidad de flujo de galaxias (las mismas galaxias anteriores). Éste dato proviene del catálogo de NVSS. Al igual que antes, esta unidad representa que tan brillante en radio la galaxia vista desde la Tierra.

Flujo en diferentes filtros ópticos (Colores: *u,g,r,i,z*):
Brillo de las galaxias ópticas, en cada uno de los colores del espectro visible. Los filtros de color del SDSS, Slogan Digital Sky Survey, están diseñados para bloquear la luz en todas las otras frecuencias y sólo dejar pasar aquellas que son cercanas a la longitud de onda que se desea observar. En la tabla [**5-1**] podemos encontrar la longitud de onda que corresponde a cada uno de los filtros.

Estas observaciones están dadas en una escala logarítmica (magnitudes). Se construyeron los colores mediante la sustracción de valores en diferentes bandas, i.e, *u-g, g-r, r-i, i-z*.

Filtro	Longitud de onda
Ultravioleta(u)	$3543 \times 10^{-10} m$
Verde(g)	$4770 \times 10^{-10} m$
Rojo(r)	$6231 \times 10^{-10} m$
Infrarojo Cercano(i)	$7625 \times 10^{-10} m$
Infrarojo(z)	$9134 \times 10^{-10} m$

Tabla 5-1: Filtros SDSS.

Corrimiento al rojo (z):
El corrimiento al rojo está relacionado con la velocidad con que se mueve un objeto con respecto a la Tierra. Su cálculo se realiza midiendo la longitud de onda emitida por algún elemento químico, por ejemplo el hidrógeno, en la Tierra y la observada del objeto en cuestión. Si λ denota la longitud de onda y por f a la frecuencia del elemento químico escogido. El corrimiento al rojo se escribir de la siguiente manera:

$$\begin{aligned} z &= \frac{\lambda_{observada} - \lambda_{emitida}}{\lambda_{emitida}} \\ &= \frac{f_{emitida} - f_{observada}}{f_{observada}}. \end{aligned}$$

Recordemos que la relación entre frecuencia y longitud de onda es la siguiente: $f = c/\lambda$, donde c denota la velocidad de la luz. Mediante

modelos cosmológicos el corrimiento al rojo se relaciona con la distancia entre el objeto y la Tierra.

5.3. Manejo y limpieza de datos

Un inconveniente que se tiene con la base de datos en cuestión es que existen algunas galaxias cuya información está incompleta. Esto es un inconveniente ya que Mapper y t-SNE no son capases de discernir cuando esto sucede. No es buena idea colocar el valor de 0 en las entradas faltantes ya que esto introduciría un sesgo en la base de datos. Por lo cual se tomó el subconjunto cuya información está completa para cada una de las galaxias.

Por otro lado, los colores colectados por el SDSS están en una escala logarítimica, por que hicimos un cambio de unidades mediante una transformación T, dada por:

$$\begin{aligned} \mathrm{T}\colon \mathcal{D}|_{\text{colores}} &\to \mathcal{D}|_{\text{colores}}, \\ \mathrm{T}(d) &\mapsto (10^{-0.4})^d \end{aligned}$$

Luego, al tener todos los valores de la base de datos $\mathcal{D}$ en las mismas unidades, se procedió a realizar un reescalamiento de los datos. Se emplearon tres tipos de re-escalamientos, los cuales son:

Re-escalamiento 1. Re-escalamiento al intervalo [0,1]:

El re-escalamiento lo hacemos de la siguiente manera. Supongamos que $X = \{x_1, x_2, \ldots, x_N\} \subseteq \mathbb{R}^n$. Este conjunto es transformado en otro $Y = \{y_1, y_2, \ldots, y_N\} \subseteq \mathbb{R}^n$, con $y_i \in [0, 1]$ vía la siguiente transformación.

$$y_i = \frac{x_i - \text{mín}(X)}{\text{máx}(X) - \text{mín}(X)}.$$

Esta transformación está bien definida siempre y cuando el conjunto X no sea constante (para evitar que $\text{máx}(X) = \text{mín}(X)$). Recordemos que en un conjunto finito el máximo y el mínimo coinciden sí, y sólo sí, el el conjunto es constante. Denotaremos a esta transformación como $t_1\colon \mathcal{D} \to \mathcal{D}$.

Re-escalamiento 2. Estandarización normal:

Utilizando la misma notación anterior, definimos el conjunto Y tomando

$$y_i = \frac{x_i - \mu_i}{\sigma_i},$$

en donde $\mu_i = \mathbb{E}[x_i]$ y $\sigma_i = \text{var}(x_i)$. Denotaremos a esta transformación como $t_2 \colon \mathcal{D} \to \mathcal{D}$.

Re-escalamiento 3. Estandarización por columnas:

Otros de los re-escalamientos realizados fue la estandarización de cada una de las columnas. Al estar las columnas (pero no necesariamente todas) en diferentes unidades de medición, es necesario que cada columna esté estandarizada. De este modo, no perder la formación que nos proporcionan cada una de ellas.

Es decir, supongamos que la matriz que contiene los datos está dada por $A = (a_{ij})_{i,j=1}^{n}$. Definimos la varianza muestral como

$$\sigma_i = \frac{1}{n-1} \sum_{k=1}^{n} (a_{ki} - \overline{x}_i)(a_{kj} - \overline{x}_j),$$

en donde

$$\overline{x}_k = \frac{1}{n} \sum_{i=1}^{n} a_{ik}.$$

Para hacer la estandarización dividimos la columna i de A entre $\sigma_i \neq 0$. Esto es, para cada una de las columnas hacemos

$$A_i \leftarrow A_i / \sigma_i,$$

donde A_i representa la columna i-ésima de A. Denotaremos a esta transformación como $t_3 \colon \mathcal{D} \to \mathcal{D}$.

5.4. Exploración topológica de los datos

Se llevó a cabo un análisis topológico de datos mediante tres algoritmos. Detallamos a continuación los resultados obtenidos con cada uno de ellos.

5.4.1. Complejos Vietoris-Rips

El primero paso realizado fue un análisis por medio de complejos del tipo Viertoris-Rips [43] con el objetivo de analizar la geometría subyacente en la base de datos. Al ser el cálculo de este tipo de complejos simpliciales computacionalmente muy costoso, se realizaron los cálculo en el servidor de `ATD` del CIMAT.

Tomamos la base de datos $\mathcal{D}$ y aplicamos el re-escalamiento 1 (mapear todos los valores al intervalo [0,1]). Es decir, tomanos el conjunto

$$\mathcal{D}' \coloneqq t_1(\mathcal{D}).$$

Luego, tomamos una muestra aleatoria de tamaño 300 del conjunto $\mathcal{D}'$ y lo analizamos mediante los complejos de Vietoris-Rips. Podemos ver el resultado en la imagen [**5-5**]. Hacemos lo mismo pero para una muestra de tamaño 500 y para la base completa con 1058 puntos; los resultados se encuentran en las figuras [**5-6**] y [**5-7**], respectivamente.

Podemos observar que en los datos, su estructura subyacente cuenta sólamente con componentes conexas, que los los grupos de homología de dimensión cero son los únicos que persisten con el tiempo.

5.4.2. Ripser

El algoritmo Ripser fue creado por Ulrich Bauer, profesor de la Universidad Tecnológica de Munich (TUM) [13]. Esta es una implementación muy eficiente para el cálculo de una filtración utilizando complejos de Vietoris-Rips. Mediante la aplicación online disponible en `http://live.ripser.org/` se corrieron varios subconjuntos de $\mathcal{D}'$. Mediante este ejercicio se verificó que se contaba únicamente con componentes conexas en la base de datos de radio galaxias.

El siguiente paso consiste en analizar con más detalle cada una de las componente conexas, de esta manera estudiar las galaxias que se encuentran en las mismas componentes y la relación que estas comparten entre sí.

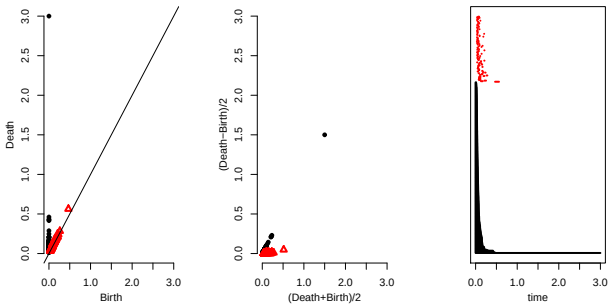


Figura 5-5: 300 puntos

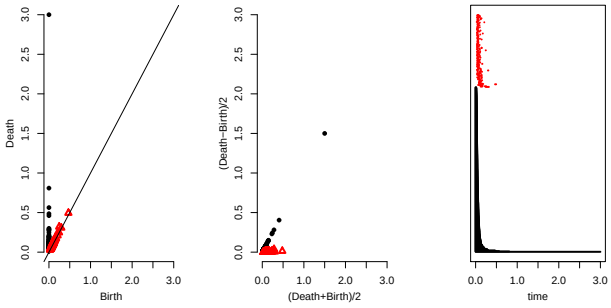


Figura 5-6: 500 puntos

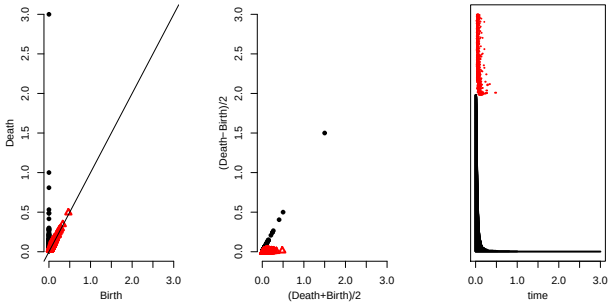


Figura 5-7: 1058 puntos

5.4.3. Exploración de los datos mediante Mapper

Existen varias implementaciones computaciones del algoritmo Mapper, de las cuales podemos resaltar la versión en `Python` creada por Daniel Müllner en la Universidad de Stanford [11]. Otra implementación está hecha en `R`, bajo el nombre de `TDAmapper`, y fue desarrollada por Paul Pearson, profesor en Hope College [12]. Nosotros trabajaremos con el algoritmo en `R`. Ambas implementaciones son muy rápidas y eficientes en el uso de memoria.

A diferencia del cálculo de complejos simpliciales de tipo Vietoris-Rips y Čech, que son computacionalmente muy costosos, Mapper nos contruye una gráfica en una cantidad mucho menor de tiempo. Esto es conveniente, especialmente cuando se trabaja con una base de datos muy grande, ya que nos permite encontrar rápidamente características topológicas importantes en los datos.

En el capítulo 3 se enfatizó la necesidad de los siguientes elementos al hacer uso del algoritmo Mapper en una nube de datos $\mathcal{D}'$:

1. Una función de filtro. Es decir, una función real que esté definida sobre todos los puntos en $\mathcal{D}'$.
2. Una cubierta abierta del espacio de parámetros, la cual es la imagen de $\mathcal{D}'$ bajo la función de filtro.

Con nuestra base de datos, no es claro cuál función de filtro emplear de tal forma que se pueda extraer información importante de los datos. Si uno no tiene idea sobre cuál función utilizar, es común emplear la proyección de los datos sobre las primeras componentes principales (dadas por PCA). La razón es que al realizar dicha proyección, estamos resumiendo los datos de alguna manera, tomando las proyecciones que maximizan la varianza.

La implementación de Mapper con la cual se está trabajando solamente permite lidiar con cubiertas abierta dadas por rectángulos uniformes. Podemos controlar el porcentaje de traslape que se tienen de los rectángulos y el la longitud de los lados. En la imagen [**5-8**] se encuentra un ejemplo de juguete con una representación visual.

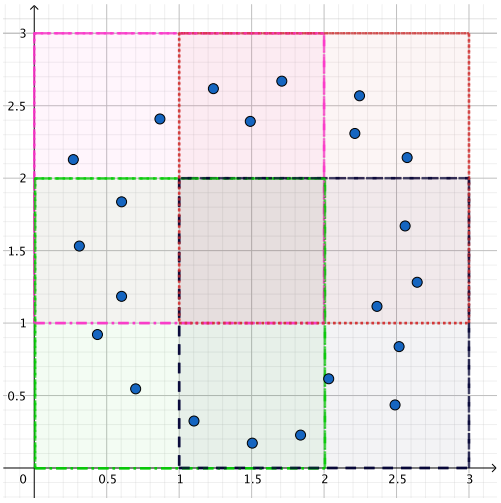


Figura 5-8: Ejemplo de una cubierta abierta uniforme sobre puntos aleatorios de S^1

Se realizó un análisis exploratorio mediante las siguientes funciones de filtro,

$$f : \mathcal{D}' \to \text{im}\,(f(\mathcal{D}')) \subseteq \mathbb{R}^k,$$

definidas como:

1. Flujo dado por los dos radio telescopios *fflux, nflux*.
2. Transformaciones logarítmicas y combinaciones de las características físicas provenientes de cada una de las galaxias.
3. Cociente de los radios de la elipse que mejor aproxima a el área de la radio galaxia.

En la figura [**5-9**] se encuentra una típica salida de Mapper. Cabe mencionar que los números en los vértices no tienen ningún significado. Es nada más una forma arbitraría de darle nombre a los vértices.

Modificación del algoritmo Mapper

En la implementación de Mapper en el lenguaje `R` se realizó una modificación del algoritmo para poder llevar un registro sobre la ubicación de las galaxias en cada momento. Esto conviene porque no sólo

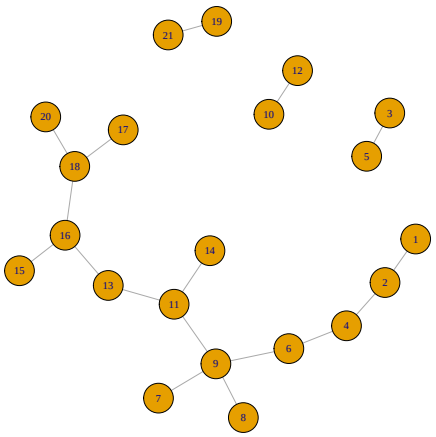


Figura 5-9: Ejemplo de una salida de Mapper en `R`

estamos interesados en saber la cantidad de componentes conexas (grupos) existentes en los datos, sino también saber exactamente dónde se encuentran cada una de las galaxias. Así, se puede analizar de las galaxias de las cuales aún no sabe su clasificación.

5.5. Reducción de dimensión mediante t-SNE

Luego de realizar varias pruebas exploratorias con Mapper, se decidió utilizar un algoritmo de reducción de dimensión para poder visualizar los datos y de esta manera poder encontrar lo parámetros adecuados para Mapper que mejor nos resuma la información topológica. En las figuras [**5-10**,**5-11**,**5-12**] se encuentran tres reprentaciones bajo los distintos re-escalamientos, mencionados en la sección 5.3.

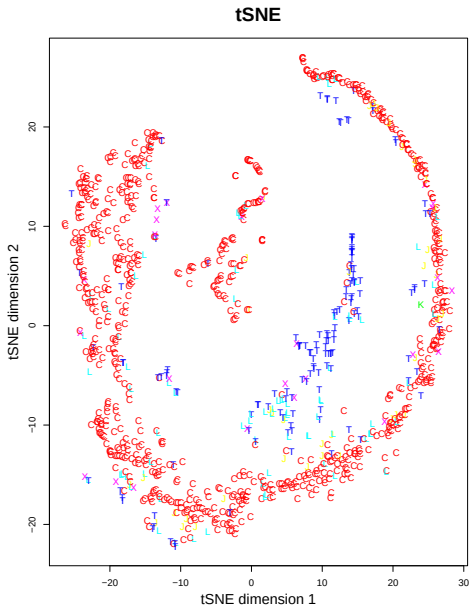


Figura 5-10: Sin reescalamiento

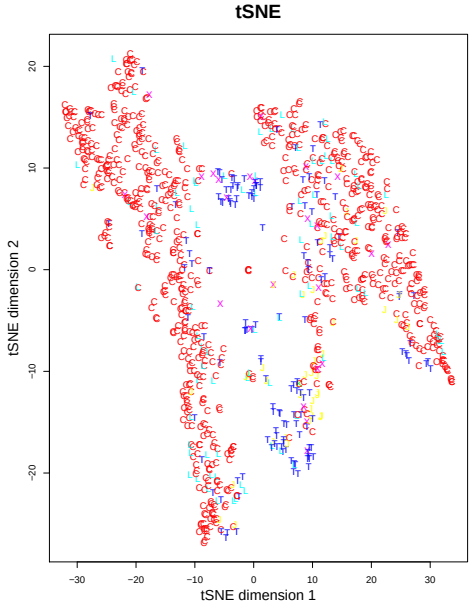


Figura 5-11: Re-escalamiento 1

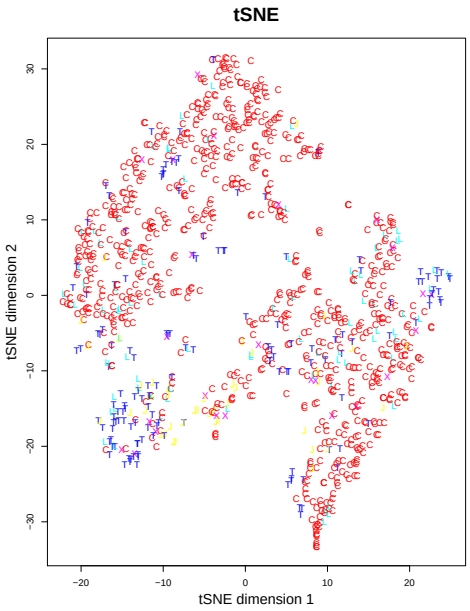


Figura 5-12: Re-escalamiento 2

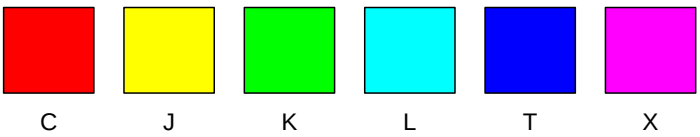
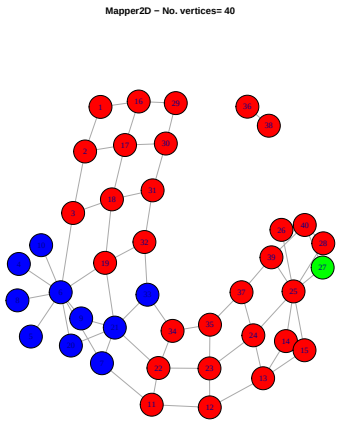


Figura 5-13: Tabla de colores

En las tres representaciones de dimensión dos dadas en las figuras [**5-10**,**5-11**,**5-12**], cada grupo de galaxias es representando mediante un color distinto según su tipo. Los colores están definidos para cada grupo según la figura [**5-13**].

Observación 5.1. *Dada la heurística del algoritmo t-SNE, este inicia con puntos escogidos al azar en el espacio de dimensión menor, las salidas no son exactamente las mismas en diferentes corridas del algoritmo, aún con la misma nube de puntos. Lo bueno es que la estructura de la representación de dimensión baja se mantiene siempre similar, salvo una rotación o reflexión de lo mismos.*

Como nuestra implementación de Mapper nos permite únicamente tomar rectángulos uniformes como función de filtro, el re-escalamiento número 1 es el que nos conviene utilizar ya que, a simple vista, vemos en la figura [**5-11**] que esta es la más fácil de las tres de dividir por medio de cuadrados uniformes. En la tabla [**5-2**] vemos los parámetros utilizados así como su respectiva salida.



Característica	Valor usado
Función de filtro	(t-SNE.1, t-SNE.2)
No. de intervalos	(10,3)
Overlap	50 %
No. Bins	7
Reescaling	1

Tabla 5-2: t-SNE como función de filtro

En donde t-SNE.1 y t-SNE.2 representan las primeras dos coordenadas dadas por t-SNE en la reducción de dimensión de la base de datos de

dimensión 9 a dimensión 2. Estamos cubriendo con 10 intervalos uniformes el eje horizontal y con 3 el vertical de la figura [**5-11**], ambos con nivel de traslape del 50 %. Estos parámetros pueden ser modificados con el objetivo de, dada la misma función de filtro, estudiar la base de datos desde distintas *resoluciones*.

Significado de los colores:

Por otro lado, en el lado derecho de la tabla [**5-2**], los colores de los vértices en la gráfica significan que hay una mayoría de galaxias del tipo correspondiente. Para su coloración, seguimos el algoritmo 3.

Observación 5.2. *Es importante tener claro que los colores en la gráfica mostrada en la figura derecha de [**5-2**] indican una mayoría de galaxias del tipo correspondiente y no que hay únicamente galaxias de ese tipo. Esta coloración la hicimos con la idea de poder identificar con más facilidad donde se estaban concentrando los diferentes tipos de galaxias. Esto es importante cuando se están realizando una primera exploración de los datos con diferentes funciones de filtro y parámetros y se quiere encontrar aquellos que extraen de una manera más clara la información relevante que se está buscando.*

Algoritmo 3 Coloración de vértices

1. Para cada uno de los vértices, contamos la cantidad de galaxias de cada tipo en cada uno de ellos.

2. Identificamos, para cada uno de los vértices, el tipo de galaxia con mayor cantidad de miembros.

3. Pintamos el vértice de acuerdo a la mayor cantidad de galaxias del tipo correspondiente.

Vemos como Mapper nos está identificando correctamente que en la parte derecha e izquierda de los datos se encuentra una mayoría de galaxias del tipo C (rojas). Por otro lado, en el centro se tiene una mayor concentración de galaxias del tipo T (azules).

5.6. Conclusiones

Dos conclusiones importantes sobre nuestro trabajo en el problema de clasificación de radio galaxias son las siguientes. Primero, nuestro método ha mostrado ser congruente con la clasificación visual existente. Esto sugiere la posible existencia de una relación entre las mediciones físicas de las (radio) galaxias con su morfología. Resta encontrar con exactitud cuál es dicha relación. Segundo, fuimos capaces de detectar los dos tipos de radio galaxias con el mayor número de observaciones (los tipos C y T) empleando únicamente sus mediciones físicas — sin necesidad de recurrir a la observación visual para cada una de ellas. Más trabajo es requerido para poder detectar correctamente a los diferentes tipos restantes de radio galaxias. El método propuesto aquí es un primer paso, uno prometedor, en la formalización y automatización del proceso de clasificación de radio galaxias. Además, el resumen topológico que brinda Mapper, de la clasificación hecha por t-SNE, nos brinda una posible vía con la cual podamos medir la certidumbre que se tiene sobre la asignación hecha para cada una de las nuevas radio galaxias que están aún sin clasificar.

Apéndice A

Anexo: Tutorial, librería TDA en R

Ejemplo. Vamos a generar en R un conjunto de 25 puntos de $\mathbb{S}^1$ puntos en R con el código que mostraremos abajo. En la figura [**A-1**] podemos ver los puntos generados y en la figura [**A-2**].

```
#Ejemplo 1.
library(TDA)
library(rgl)

#generamos 25 puntos de la esfera S^2:

ptos <- sphereUnif(n=35, 2, r = 1)
#plot3d(ptos) #para ver la esfera de forma tridimencional.

maxscale <- 5 # limite para la filtracion.
maxdimension <- 1 # componentes y loops.

DiagRips <- ripsDiag(X=ptos,maxdimension, maxscale)

plot(DiagRips[["diagram"]], barcode = TRUE, main = "Barcode")
```

En la figura [**A-2**] las barras negras representan las componentes conexas y las barras negras los lazos.

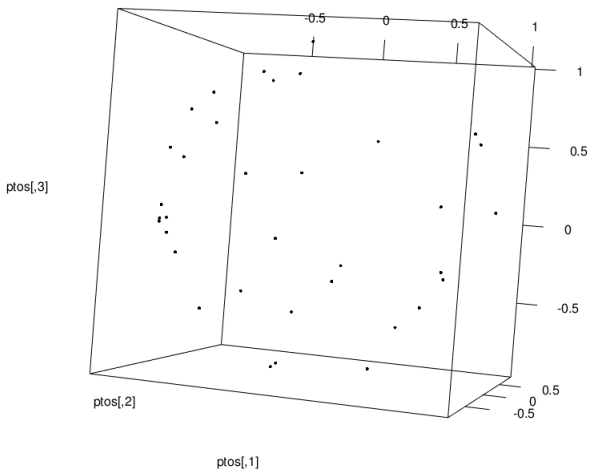


Figura A-1: 25 puntos aleatorios en la esfera $\mathcal{S}^1$.

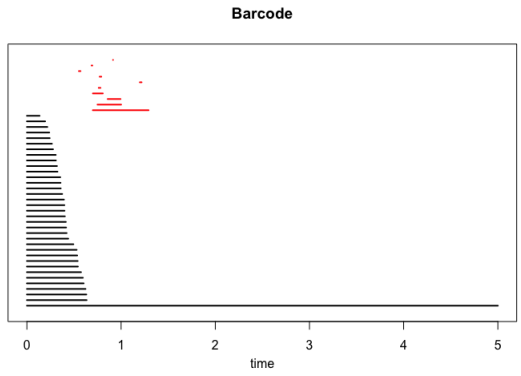


Figura A-2: Código de barra de los 25 puntos antes generados.

Bibliografía

[1] G. Singh, F. Mémoli, G. Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *Eurographics Symposium on Point-Based Graphics (2007).* `https://research.math.osu.edu/tgda/mapperPBG.pdf`

[2] Amy E. Kimball, Željko Ivezić. An updated version of the unified radio catalog: a multi-wavelength radio and optical catalog of quasars and radio galaxies. *Conference Proceedings for IAU Symposium 304 "Multi-wavelength AGN Surveys and Studies"*, 2014.

[3] Michael R. Blanton, et al. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. *The Astronomical Journal,* 2017.

[4] Bradley W. Carroll, Dale A. Ostlie. *An introduction to modern astrophysics.* Pearson; 2 edition, 2006.

[5] Amy E. Kimball. National Radio Astronomy Observatory. `http://www.aoc.nrao.edu/~akimball/`

[6] Frédéric Chazal, Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *Journal de la Société Française de Statistique,* 2017.

[7] Carriere, Mathieu and Michel, Bertrand and Oudot, Steve Y. Statistical analysis and parameter selection for Mapper. *arXiv preprint arXiv:1706.00204,* 2017.

[8] Herbert Edelsbrunner. *A Short Course in Computational Geometry and Topology.* Springer, 2014.

[9] Mathieu Carriere. On Metric and Statistical Properties of Topological Descriptors for geometric Data. Computational Geometry [cs.CG]. Université Paris-Saclay, 2017. English.

[10] Ronald J. Buta. Galaxy Morphology. *Planets, Stars, and Stellar Systems,* ***Vol.*** *6, 2011.*

[11] Daniel Müllner. Mapper Algorithm implementation in Python. Stanford University. `http://danifold.net/mapper/`

[12] Paul Pearson. Mapper Algorithm implementation in R. Hope College. `https://github.com/paultpearson/TDAmapper`

[13] Ulrich Bauer. Ripser: a lean C++ code for the computation of Vietoris–Rips persistence barcodes (2015–2018). `https://github.com/Ripser/ripser`

[14] Kleinmann, S. G, Lysaght, M. G., Pughe, W. L et al. The Two Micron All Sky Surve. *Science with Astronomical Near-Infrared Sky Surveys. Springer Netherlands, 1994.*

[15] Tim Stephens. Face recognition for galaxies: Artificial intelligence brings new tools to astronomy. *Newscenter UC Santa Cruz, 2018.*

[16] Pratyush Pranav, Herbert Edelsbrunner, Rien van de Weygaert et al. The Topology of the Cosmic Web in Terms of Persistent Betti Numbers. *arXiv preprint Monthly Notices of the Royal Astronomical Society, no.4, 4281-4310* 2017.

[17] Rien van de Weygaert, Gert Vegter, Herbert Edelsbrunner, Bernard J.T. Jones et al. Alpha, Betti and the Megaparsec Universe: on the Topology of the Cosmic Web. *Transactions on Computational Science XIV, páginas 60-101. Springer-Verlag, 2011*

[18] Thierry Sousbie. The persistent cosmic web and its filamentary structure I: Theory and implementation. *arXiv preprint arXiv:1009.4015, 2010.*

[19] Vanessa Robinss. Towards computing homology from finite approximations. *Topology Proceedings, volume 24, pages 503–532, 1999.*

[20] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological per- sistence and simplification. *Discrete and Computational Geometry, 28(4):511–533, 2002.*

[21] Marian Gidea. Topology data analysis of critical transitions in financial networks. *3rd International Winter School and Conference on Network Science, pag: 47–59. Springer, 2017.*

[22] Marc Offroy, Ludovic Duponchel. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta,* ***vol.*** *910, 2016.*

[23] Rami Kraft.*Illustrations of Data Analysis Using the Mapper Algorithm and Persistent Homology.* Master's Thesis, KTH Royal Institute of Technology, Sweden, 2016

[24] T. Jarrett. Large Scale Structure in the Local Universe: The WISE Galaxy Catalog. *To appear in Publications of the Astronomical Society of Australia,* 2004.

[25] A. Marcos-Caballero, P. Vielva, E. Martinez-Gonzalez et al. A new model for NVSS galaxy catalogue using the redshift distribution and halo minimum mass. *arXiv preprint arXiv:1312.0530,* 2013.

[26] Rien van de Weygaert, Erwin Platen, Gert Vegter, Bob Eldering, Nico Kruithof. Alpha Shape Topology of the Cosmic Web. *International Symposium on Voronoi Diagrams, 2010.*

[27] Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel. Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. *arXiv preprint arXiv:1704.04650v1,* 2017.

[28] Jesús Manuel Pérez Angulo. *Análisis Topológico de Datos: Robusticidad y análisis de sensibilidad de algoritmos.* Tesis de maestría, CIMAT 2016.

[29] Laurens van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research 9(Nov):2579-2605,* 2008.

[30] Laurens van der Maaten. Package 'Rtsne', 2017. https://github.com/lvdmaaten/bhtsne/

[31] Jeff Jauregui. *Principal component analysis with linear algebra.* http://www.math.union.edu/~jaureguj/PCA.pdf

[32] More Information on Tailed Radio Galaxies (Part 2). https://blog.galaxyzoo.org/author/vrooje/page/2/

[33] National Radio Astronomy Observatory. Verry Large Array. https://science.nrao.edu/facilities/vla/

[34] J. P. Leahy, A . H. Bridle, and R. G. Strom. An Atlas of DRAGNs. http://www.jb.man.ac.uk/atlas/

[35] Amy E. Kimball, Željko Ivezić, Paul J. Wiita, Donald P. Schneider. Correlations of quasar optical spectra with radio morphology. *arXiv preprint arXiv:1103.4791,* 2011.

[36] The Sloan Foundation 2.5m Telescope at Apache Point Observatory, in south east New Mexico. Telescopes and Instruments. http://www.sdss.org/instruments/

[37] Robert H. Becker, David J. Helfand, Richard L. White, Michael D. Gregg, Sally A. Laurent-Muehleisen. Faint Images of the Radio Sky at Twenty-Centimeters. *Astronomical Data Analysis Software and Systems III, A.S.P. Conference Series,* ***Vol.*** *61,* 1994. http://sundog.stsci.edu/

[38] Condon, J. J.; Cotton, W. D.; Greisen, E. W.; Yin, Q. F.; Perley, R. A.; Taylor, G. B.; Broderick, J. J. The NRAO VLA Sky Survey. *The Astronomical Journal,* ***Vol.*** *115, Issue 5, pp. 1693-1716.* https://www.cv.nrao.edu/nvss/

[39] Fred Espenak. Very Large Array, photo Gallery A. 2007. www.mreclipse.com/Observatory/VLA/VLAgalleryA.html

[40] Fotografía de galaxias. Hubble Ultra Deep Field, 2014. http://hubblesite.org/image/3380/news_release/2014-27

[41] Baum, O'Dea, Perley and Cotton. The Curious Lives Of Radio Galaxies – Part One.
https://blog.galaxyzoo.org/author/vrooje/page/2/

[42] Lum, P.Y. et al. Extracting insights from the shape of complex data using topology. *Scientific Reports 3, 1236; DOI:10.1038/srep01236,* 2013.

[43] Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria. TDA: Statistical Tools for Topological Data Analysis. https://cran.r-project.org/web/packages/TDA/index.html.