Centro de Investigación en Matemáticas A.C.

Estadísticos Topológicos en el Estudio de Electrocardiogramas

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

MAESTRO EN CIENCIAS

CON ORIENTACIÓN EN:

PROBABILIDAD Y ESTADÍSTICA

PRESENTA:

Marcos Torres Vivanco

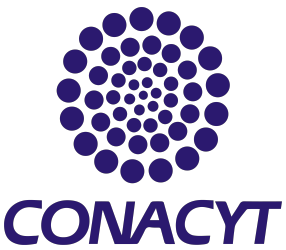
Asesor

Dr. Carlos Vargas Obieta

Co-Asesor

Dr. Miguel Nakamura Savoy

Guanajuato, Guanajuato
Junio, 2020



Índice general

Agradecimientos

A mis padres y hermano por todo el apoyo que me han dado y por siempre creer en mi.

A mis amigos Andrea, Carmen, Javier y Pardo por no dejar que me diera por vencido en ningún momento de la maestría y llenar estos dos años de buenos momentos.

A CONACyT por apoyarme en muchos aspectos durante mi maestría y licenciatura.

A mis asesores de tesis y sinodales por ayudarme a completar este proyecto.

A Eduardo, Rafa y Ricardo por sus consejos y buenas ideas.

A todos los compañeros, profesores e investigadores que me han hecho amar la topología y estadística desde la licenciatura hasta la maestría.

Capítulo 1

Introducción

El Análisis Topológico de Datos, o simplemente ATD, es una disciplina de origen reciente, que combina herramientas de distintas áreas como la computación y las matemáticas. La topología es la rama de las matemáticas que estudia la forma de los conjuntos, de hecho en esta rama la palabra *topología* es usada como sinónimo de *forma*. Por lo tanto, la idea principal del ATD es obtener información de la forma en que se encuentran estructurados nuestros datos. En esta tesis se estudia el potencial del ATD para poder clasificar electrocardiogramas.

A lo largo de los últimos años han surgido cada vez más aplicaciones del ATD. Algunas de estas aplicaciones las podemos encontrar en el área de la biología y la salud, como las descritas en Chazal (2017) [**1**], donde se aplican técnicas de ATD para comparar la estructura de las proteínas y para clasificar datos de sensores de movimiento. En el trabajo de De Silva (2012) [**2**] se propuso usar ATD para el estudio de series de tiempo. Investigaciones posteriores retomaron esta idea, por ejemplo en Perea et al. (2015) [**3**], donde se cuantifica la periodicidad de una serie de tiempo usando teoría de topología. Este trabajo está inspirado en estos trabajos, pues se extrae información de un tipo particular de series de tiempo: los electrocardiogramas.

Siguiendo la filosofía del ATD, nuestro objetivo es analizar la forma de los ECG, por lo cual necesitamos considerar alguna estructura geométrica asociada a la serie de tiempo del corazón. En esta parte nos auxiliamos del teorema de Takens, el cual nos dice que es posible reconstruir la topología del atractor del sistema dinámico asociado a la serie de tiempo. En nuestro caso, bajo algunos supuestos, el teorema de Takens nos garantiza que podemos reconstruir parte de la dinámica del corazón usando un ECG.

La técnica más usual del ATD es la homología persistente, que nos ofrece un resumen de los nacimientos y muertes de componentes conexas, ciclos y huecos de una nube de puntos al inflar esferas alrededor de cada dato. Es con esta técnica que se estudian los ECG a lo largo de este trabajo.

1.1. Problema

Para esta tesis se trabajará con ECG de dos ritmos distintos: ritmo normal y ritmo con fibrilación atrial (una cardiopatía común). Asi mismo, los datos tienen dos orígenes: unos serán simulados a partir de un modelo y otros se obtienen de la base de datos del sitio web Physionet [**4**].

El sitio Physionet organizó un concurso en 2017 donde se pidió construir un algoritmo para poder clasificar los diferentes ritmos de la base de datos. En Ignacio (2019) [**5**] se propone un enfoque a este problema usando ATD. Es sobre esta idea que se construye la presente tesis, explorando la posibilidad de usar ATD para obtener información relevante en series de tiempo.

El objetivo de este trabajo es estudiar el potencial del análisis topológico de datos en la clasificación de electrocardiogramas. Para esto construiremos un algoritmo que separe los ECG en las clases de ritmos con fibrilación atrial y ritmo normal, usando información topológica de las series de tiempo. También construiremos otro método de clasificación de ECG, basado únicamente en información médica, y lo compararemos con los métodos que se han desarrollado a partir de ATD.

1.2. Contribución

La primera contribución de este trabajo es la creación de un nuevo modelo para simular ECG's y usarlos en los algoritmos de clasificación. Dicho modelo está basado en el trabajo de Kubicek (2014) [**6**], pero se agrega un ruido que toma en cuenta el contexto de los electrocardiogramas para volverlos más parecidos a los que se obtienen en la vida real. Con esto es posible crear bases de datos sintéticas de ECG y trabajar con ellas.

La segunda contribución es el ajuste a datos de ECG del método presentado en Chazal (2017) [**1**] y que fue usado originalmente para la clasificación en sensores de movimiento. Con esto se exhibe una nueva aplicación al uso de los panoramas de persistencia en el área de ciencia de datos.

La tercer contribcuión de esta tesis es la propuesta de un método conceptual para estudiar los códigos de barras obtenidos con la filtración de Čech, de tal forma que se pueda diferenciar las barras que describan características locales de aquellas que describan características globales, sin recurrir a la metodología usual de dividir el código en barras cortas y barras largas.

1.3. Organización de la tesis

En el Capítulo 2 se resumen los antecedentes técnicos necesarios para poder abordar la tesis. Comenzamos con la teoría de datos funcionales y series de tiempo necesarios para describir los ECG y los panoramas de persistencia. Luego, se da un repaso breve de las herramientas que usaremos de la teoría de análisis topológico de datos. También se exponen algunos conceptos del campo médico usados en el trabajo, así como los componentes de un ECG. Concluimos el capítulo abordando el tema de clasificación estadítica, en donde mencionamos los algoritmos que usaremos en este trabajo.

En el Capítulo 3 se señala la forma de trabajar con series de tiempo usando topología, comenzando por el método de la ventana deslizante, la cual nos permite obtener una nube de puntos a partir de una serie temporal. Este resultado está basado en el teorema de Takens, el cual enunciamos. Posteriormente, se presentan los estadísticos topológicos que serán usados como características en los clasificadores. Al final del capítulo presentamos un ejemplo de la obtención de dichos estadísticos sobre un ECG.

En el Capítulo 4 se presenta una nueva propuesta conceptual para diferenciar características topológicas locales y globales de una nube de puntos. Para esto primero se definen los conceptos de filtración de Čech y de triangulación de Delone y posteriormente se enuncia la propuesta para encontrar características locales. También se expone un ejemplo en donde la metodología usual para separar características locales y globales, basanda en considerar longitudes en el código de barras, llega a fallar.

En el Capítulo 5 se presentan los resultados obtenidos de clasificar los datos simulados y los datos reales con los estadísticos obtenidos en el Capítulo 3. También se explica el modelo que se creó para construir los datos simulados.

Finalmente, en el Capítulo 6 se presentan las conclusiones a las que se llegó después de realizar las clasificaciones en el Capítulo 5, así como algunos comentarios finales sobre los aspectos computacionales de los algoritmos estudiados.

Capítulo 2

Antecedentes

La motivación de esta tesis surge del problema de diferenciar los electrocardiogramas (ECG) de pacientes con alguna aflicción cardiaca de pacientes con ritmo sano. Este es un caso particular del problema de clasificar señales biomédicas, con dos componentes importantes: la teoría de clasificación estadística y el campo médico. Describiremos ambas componentes para entrar en el contexto del problema.

Para abordar lo relativo a los métodos de clasificación, recurrimos a diversas teorías. En primer lugar enunciamos y revisamos propiedades de las series de tiempo y datos funcionales, los cuales proporcionan un marco natural para trabajar con electrocardiogramas, permitiendo abordarlos más sistemáticamente, y son necesarios para definir una conexión entre series de tiempo y topología, a través del Teorema de Takens, el cual enunciamos en un capítulo posterior. En segundo lugar, se utilizan conceptos del análisis topológico de datos (ATD). Se incluye entonces una breve reseña de las ideas principales y de las herramientas propias a la teoría.

Para poder entender el contexto del campo médico en el que se desarrolla este trabajo, se define un ECG y sus componentes, así como la forma de identificarla gráficamente.

En este capítulo se presenta un resumen de los temas recién enumerados, con el objetivo de proporcionar una idea general, sin entrar en muchos detalles. Para profundizar en el tema de datos funcionales se recomienda consultar Ramsay (2005) [**7**], para el área de series de tiempo se recomienda Shumway (2005) [**8**] y Chazal (2017) [**1**] para consultar detalles sobre las principales herramientas del TDA.

Como complemento del estudio de las señales del corazón se recomienda Kubícek et al. 2014 [**6**]. Para profundizar en materia de clasificación se sugiere Gareth (2013) [**9**], donde se contemplan otros métodos de clasificación.

Si se desea conocer sobre otros métodos de clasificación de ECG se puede consultar los métodos presentados en Xiong (2017) [**10**], Yazdani (2017) [**11**] y Smolen (2017) [**12**].

2.1. Datos funcionales y series de tiempo

2.1.1. Datos funcionales. Con la tecnología reciente es posible obtener información que toma multiples valores para una única observación o incluso datos cuyo dominio es un intervalo o región. Con esto, podemos definir datos que son funciones sobre un espacio continuo, conocidos como *datos funcionales*. Los ECG con los que trabajamos a lo largo de esta tesis son un ejemplo de datos funcionales, pues son funciones que toman como dominio un intervalo continuo de tiempo. Las principales contribuciones a la teoría de datos funcionales son gracias a James O. Ramsay quien, además de bautizarlos, da una exposición de estos en su libro Ramsay (2005) [**7**]. En esta sección se resumen los aspectos teóricos que usaremos de datos funcionales, tomando como guía el libro de Ramsay.

En la teoría, los datos funcionales son modelados como procesos estocásticos $X = X(t)_{t\in[0,T]}$ cuyas curvas pertenecen al espacio de Hilbert $L^2[0,T]$. Por lo tanto, el conjunto de datos debe tener un dominio y un rango perfectamente especificados. Por ejemplo, cuando consideramos nuestro conjunto de ECG, estos viven en el espacio de funciones continuas con dominio el intervalo de tiempo y rango los reales.

En la práctica, se tienen algunas opciones para poder trabajar con los datos funcionales aprovechando las propiedades de las funciones y el espacio donde viven. La primera es considerar una base del espacio de funciones y estimar los coeficientes de la base usando la información disponible. La segunda opción es considerar una versión discretizada de las funciones. Para esto tomamos observaciones de la función en una rejilla de valores del dominio, usualmente a la misma distancia, esto nos da un vector que será tan grande como la rejilla elegida. Para esta tesis optaremos por usar esta segunda opción pues nos resulta útil al momento de realizar el trabajo estadístico.

2.1.2. Series de tiempo. Un caso particular de datos funcionales son las *series de tiempo*, que son funciones medidas a través de un periodo de tiempo o en un conjunto de valores de tiempo ordenados. Por ejemplo, los ECG con los que trabajamos son series de tiempo, medidos en intervalos de 30 segundos y en cada segundo se toman 300 observaciones a la misma distancia. La teoría de series de tiempo se ha estudiado extensamente. A continuación se resume la teoría necesaria para este trabajo tomando como guía Shumway (2005) [**8**].

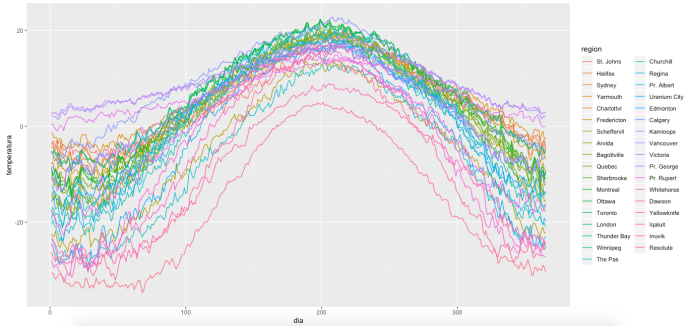


FIGURA 2.1. La temperatura promedio por día en las diferentes regiones de Canadá es un ejemplo de datos funcionales.

Las series de tiempo periódicas, son aquellas que presentan una naturaleza cíclica, es decir que para un determinado periodo de tiempo los valores de la serie se vuelven a repetir. Esta naturaleza se ve reflejada en los ECG pues de manera periódica se repiten los patrones correspondientes a un pulso del corazón.

Una componente que aparece comunmente en las series de tiempo que son tomadas de observaciones reales es *el ruido*, el cual aparece ya sea por errores de medición o factores en el entorno que afectan directamente los valores de la serie de tiempo. El ruido es una de las componentes más importantes al momento de estudiar y modelar las series de tiempo como los ECG con los que trabajamos.

Una forma de modelar el ruido es usando variables sin ninguna correlación entre ellas, de media 0 y varianza σ, x_t. A este modelo se le conoce como ruido blanco y es usual generar valores usando una variable aleatoria gaussiana de media 0 y varianza σ^2, es decir, $x_t \sim N(0, \sigma^2)$.

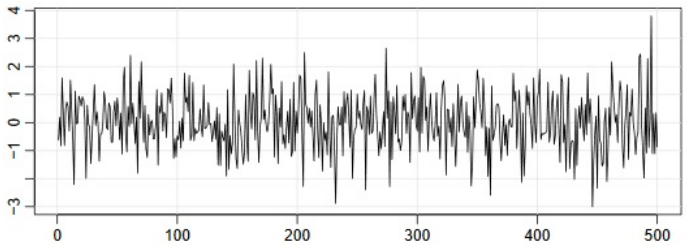


FIGURA 2.2. Ejemplo de ruido blanco.

En ocasiones cuando se trabaja con aparatos de medición, el ruido se encuentra influenciado por el tiempo. Por ejemplo, entre más tiempo se use una báscula presenta más fallas o los valores empiezan a presentar un sesgo. Es por esto que en ocasiones es necesario considerar que las observaciones de ruido tienen una correlación con otros valores del ruido pasado. Esto se puede modelar con una *caminata aleatoria*

$$x_t = \delta + x_{t-1} + r,$$

donde δ es una constante conocida como deriva y r es una variable aleatoria de ruido, generalmente una variable gaussiana $N(0, \sigma)$.

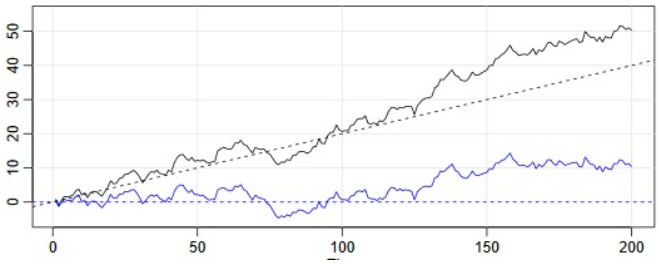


FIGURA 2.3. Ejemplo de caminata aleatoria con y sin deriva.

Una caminata aleatoria con deriva nula y donde sus valores no se salen de un determinado intervalo se conoce como caminata aleatoria con barreras reflejantes. Los extremos del intervalo de la barrera reflejante pueden tomar el valor de $\pm\infty$. Más adelante usaremos caminatas aleatorias con barreras reflejantes para poder construir modelos más precisos de ECG.

2.2. Análisis topológico de datos

El *Análisis Topológico de Datos* (abreviado ATD) es un campo de estudio que surgió de manera reciente con los trabajos de Edelsbrunner et al. (2002) [**13**] y Zomodian et al. (2005) [**14**]. Toma ideas de otros campos ampliamente estudiados como topología algebraica, topología computacional y geometría computacional. La motivación principal del ATD es obtener información de un conjunto de datos usando teoría de topología. En esta tesis exploramos el uso de ATD en el estudio de ECG, esta aplicación en particular fue propuesta por primera vez en Ignacio (2017) [**5**].

Supongamos que tenemos dos nubes de puntos y queremos obtener información para diferenciarlas. Con herramientas de ATD esto es posible, pues nos proporciona una descripción cualitativa y cuantitativa sobre la forma de las nubes. Esto es la que usaremos para estudiar ECG en un futuro una vez que le asociemos una nube de puntos.

El primer paso para estudiar una nube de puntos es asociarle una estructura topológica con más propiedades. Para esto recurrimos a los *complejos simpliciales*, que son son estructuras topológicas formadas por bloques conocidos como *simplejos*, los cuales son puntos, segmentos, triángulos, tetraedros y sus generalizaciones a dimensiones mayores. El simplejo de dimensión k es conocido como *k-simplejo*.

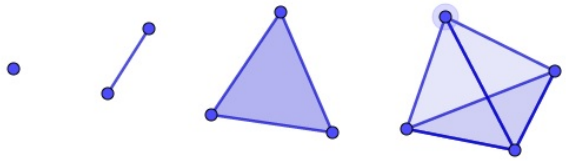


FIGURA 2.4. De izquierda a derecha vemos ejemplos de simplejos de dimensión 0, 1, 2 y 3.

La principal ventaja de trabajar con complejos simpliciales es que además de su estructura topológica, poseen una estructura combinatoria. Esto permite utilizar algoritmos para obtener invariantes algebraicos. Para consultar más sobre estos algoritmos se puede consultar Otter (2015) [**15**].

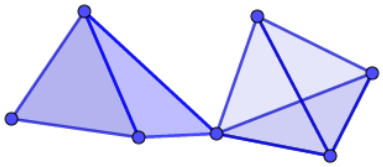


FIGURA 2.5. Ejemplo de un complejo simplicial.

Uno de los invariantes topológicos más importantes es la *homologia simplicial*, la cual consiste en una familia de espacios vectoriales, con índice en los naturales, a los que se les conoce como *grupos de homología*. Se les suele denotar H_k donde

$k \in \mathbb{N}$ y se le llama el grupo de homología de dimensión k. El tamaño de la base de estos espacios vectoriales se les conoce como *números de Betti* y se denota β_k. Estos valores nos dan información sobre la forma del complejo simplicial, por ejemplo los números de Betti de dimensión 0, 1 y 2 cuentan las componentes conexas, hoyos y huecos, respectivamente, en el complejo simplicial.

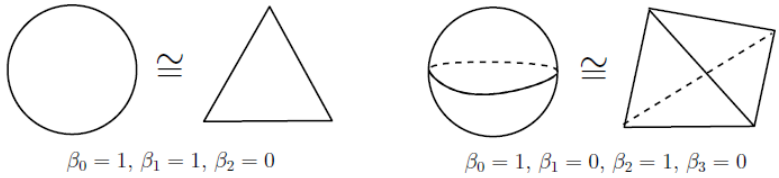


FIGURA 2.6. Imagen expuesta en Chazal (2017) [**1**] con los números de Betti de dos espacios distintos, el número de Betti β_0 cuenta componentes conexas, β_1 ciclos y β_2 los huecos.

Existen muchos complejos simpliciales que se le pueden asociar a una misma nube de puntos. Para la labor de clasificación de este trabajo usaremos el complejo de Vietoris-Rips por sus ventajas computacionales, pues es uno de los más rápidos de calcular ya que sólo se necesita conocer los puntos y las aristas para construir el resto de los simplejos que lo conforman. Formalmente, si X es una nube de puntos y $\mathcal{P}$ es su conjunto potencia, entonces el *complejo de Vietoris-Rips* de la nube X con radio ε es

$$\mathrm{VR}_\varepsilon(X) = \{\sigma \subset \mathcal{P}(X) \ : \ d(x,y) \leq 2\varepsilon \, , \forall x, y \in \sigma\}.$$

Notemos que en la definición del complejo de Vietoris-Rips se necesita de un parámetro ε. Dicho parámetro se puede interpretar como la distancia necesaria para poder unir dos puntos de la nube con una arista.

Una pregunta natural es: ¿Cuál es el mejor ε para poder asociar un complejo a nuestra nube de datos? Para tener un estudio más completo, en ATD se trabaja con todo un intervalo de valores de ε, con lo que se obtiene una familia de complejos simpliciales, llamada *filtración*. Calculando la homología simplicial a cada uno de los elementos de la filtración se obtiene la *homología persistente*, que describe la evolución de los grupos de homología simplicial a través de la escala ε. La homología persistente es la principal herramienta de trabajo del análisis topológico de datos y la usaremos para caracterizar nuestros datos.

En Edelsbrunner (2002) [**13**] y Zomorodian y Carlsson (2005) [**14**], se presentan las primeras ideas para resumir la información de homología persistente: *el código de barras y el diagrama de persistencia*. Ambas gráficas son equivalentes y representan de manera visual los valores de ε en donde nacen y mueren los generadores de los grupos de homología. Son una herramienta poderosa, pues si un generador permanece vivo por mucho tiempo, es decir es persistente, entonces proporciona una descripción de la forma en la nube de puntos. Por ejemplo, en la Figura 2.7 existe una barra de dimensión 1, representada con color azul, que es muy persistente, por lo que sabemos que la nube de puntos tiene al menos un ciclo grande.

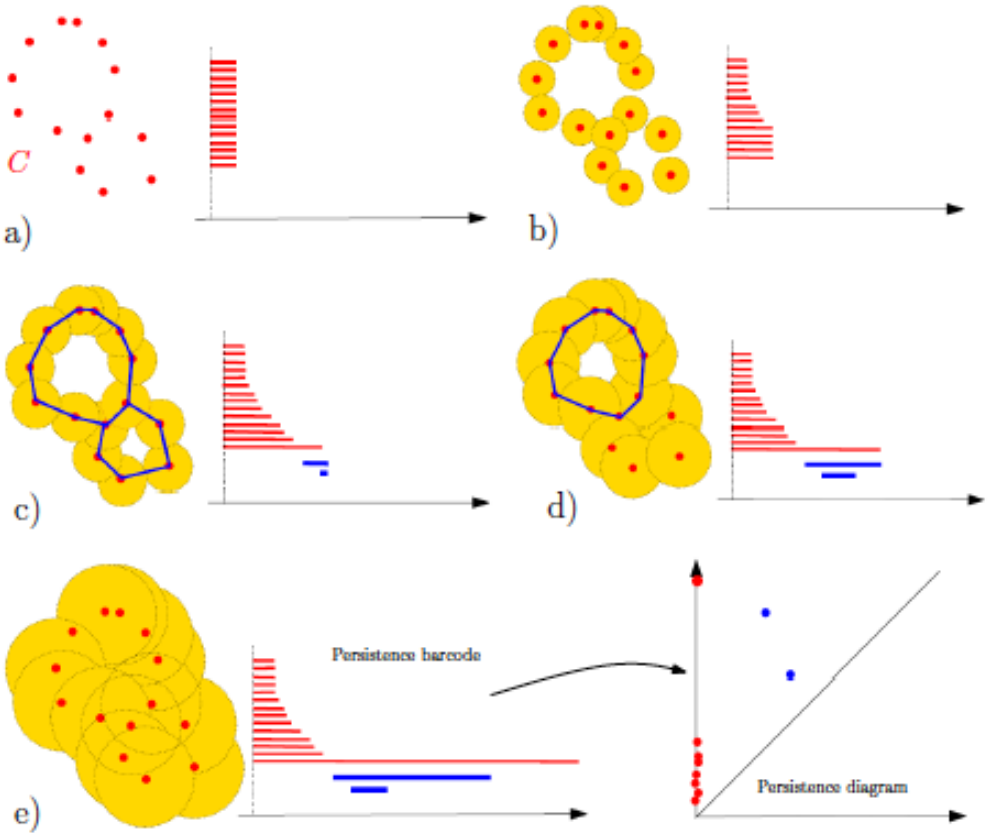


FIGURA 2.7. Calculando la homología persistente en varios pasos sobre una nube de puntos, al final se puede ver la equivalencia entre el código de barras y el diagrama de persistencia, Chazal (2017) [**1**].

2.2.1. Panoramas de Persistencia. La importancia del diagrama de persistencia no sólo se encuentra en la busqueda de elementos persistentes, también es útil en la obtención de estadísticos, veamos como obtener un nuevo estadístico que usaremos a lo largo de la tesis. Primero, notemos que toda pareja de puntos de nacimientos y muertes (b, d) cumple que $b < d$, por lo que los puntos del diagrama de persistencia se encuentran por encima de la recta identidad. Por esto consideramos la transformación

$$T(x, y) = (\frac{x+y}{2}, \frac{x-y}{2}) \tag{1}$$

que mueve la recta identidad al eje de las x.

Para una dimensión de homología fija, consideremos los puntos de nacimientos/muertes y para cada punto en el diagrama de persistencia construimos una función

$$f_{(b,d)} = \begin{cases} 0 & \text{si} \quad x \notin (b, d) \\ x - b & \text{si} \quad x \in \left(b, \frac{b+d}{2}\right] \\ -x + d & \text{si} \quad x \in \left(\frac{b+d}{2}, d\right) \end{cases} .$$

Con estas funciones podemos definir otra familia de funciones $\{\lambda_k\}_{k\in\mathcal{N}}$, donde $\lambda_k(x)$ es el k-ésimo valor más grande de $\{f_{(b_i,d_i)}(x)\}_{i=1}^n$. Dichas funciones son conocidas como *panoramas de persistencia* y fueron definidas por Peter Bubenik (2015) [**16**].

Si tomamos todos los panoramas de persistencia asociados a cada dimensión de homología es posible reconstruir el diagrama de persistencia. Estos serán los estadísticos de naturaleza topológica que usaremos para obtener información de los ECG.

2.3. Campo médico

En el campo médico, al registro de la actividad del cuerpo se le denomina *señal médica*. El estudio de estas señales juega un rol importante pues con ellas es posible detectar alguna enfermedad o dar seguimiento a la condición de una persona. En esta tesis trabajaremos con un tipo de señal particular: *los electrocardiogramas*.

La cardiología es la rama de la medicina dedicada al estudio y diagnóstico del corazón. Una de las herramientas más usadas por los cardiólogos es el electrocardiograma o ECG, que es una representación visual de la actividad eléctrica del corazón.

Existen diversas variantes de ECG, divididas principalmente en la cantidad de derivaciones o componentes de la señal (ver Figura 2.9) pero en esta tesis trabajaremos con ECG de una derivación o simples.

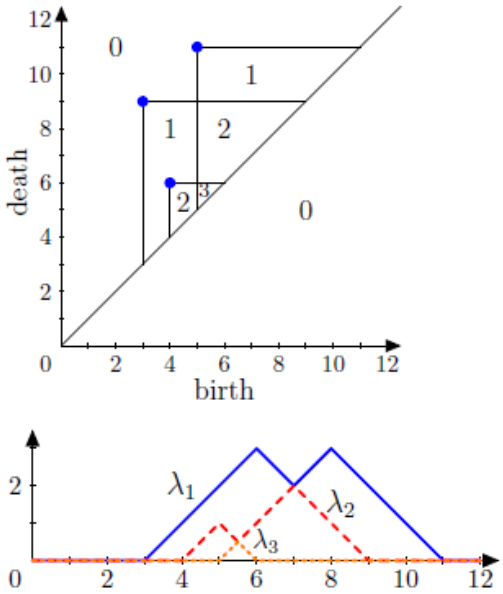


FIGURA 2.8. Representación gráfica de los panoramas de persistencia. Podemos notar que se obtienen de aplicar la transformación 1 al diagrama de persistencia.

Un ECG es una serie de tiempo que registra la actividad eléctrica, en miliVoltios (mV), de latidos consecutivos del corazón a lo largo de un intervalo de tiempo, usualmente medido en segundos o milisegundos. Dado que los latidos son similares entre ellos, la serie de tiempo es periódica. Existen parámetros y componentes en un ECG, por ejemplo la frecuencia con la que se mide la señal medida en Hertz (Hz) o el ruido presente en las mediciones. Conocer todos estos componentes es importante al momento de usar un algoritmo de clasificación sobre los datos.

Un latido normal se compone de diversas ondas, denotadas por las letras de la P a la T, e intervalos entre las dichas ondas. En la Figura 2.10 se puede apreciar un diagrama de un latido con sus componentes. La amplitud de las ondas y la longitud de los intervalos representan información valiosa sobre la actividad cardiaca y es usada en la clasificación de los ECG. En este trabajo se usaran los intervalos RR, es decir, la distancia entre dos ondas R consecutivas. En la Tabla 2.1 se encuentran resumidos los rangos de amplitud y longitud de las componentes de un latido normal.

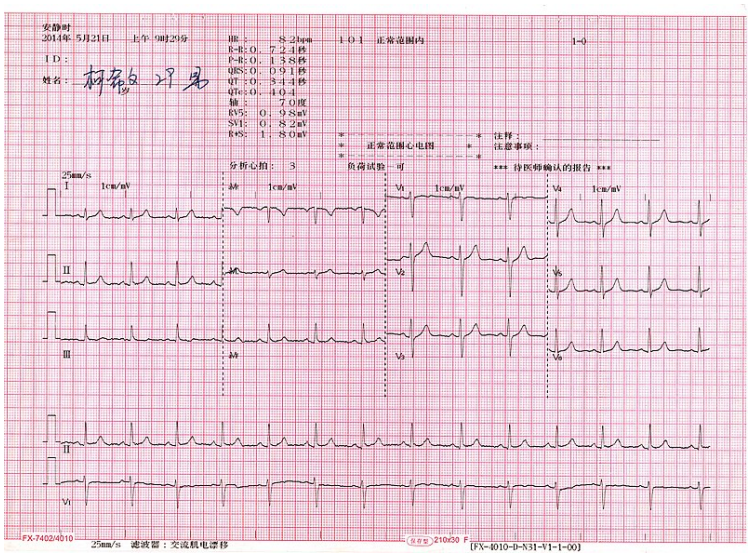


FIGURA 2.9. El ECG de 12 derivaciones está compuesto por 12 series de tiempo simultáneas midiendo señales en diferentes zonas, tomado de [**17**].

Nombre	Duración	Amplitud
Onda P	0-160 ms	0-0.3 mV
Intervalo PR	120-200 ms	0 mV
Complejo QRS	50-120 ms	0-0.5 mV
Segmento ST	0-320 ms	0-0.5 mV
Intervalo QT	50-440 ms	0 mV
Onda T	160 ms	0-0.8 mV

CUADRO 2.1. Componentes del ECG de un latido. [**6**]

Los ECG son usados principalmente en el diagnóstico de cardiopatías o enfermedades del corazón. Esto se hace identificando diferencias entre los ECG obtenidos de personas sanas y aquellos con cardiopatías. Algunas de estas diferencias se expresan en algunos de los componentes de los ECG. Por ejemplo: bajo ciertas cardiopatías algunos intervalos tienden a ser más largos o cortos de lo normal. La información de estas cardiopatías es la que nos permite realizar clasificación sobre conjuntos de ECG.

La *fibrilación auricular*, o FA, es la cardiopatía más común a nivel mundial, teniendo gran impacto en los sistemas de salud de varios países, Le Heuzey (2004) [**18**]. Es ocasionada por latidos irregulares en las auriculas del corazón y se pueden

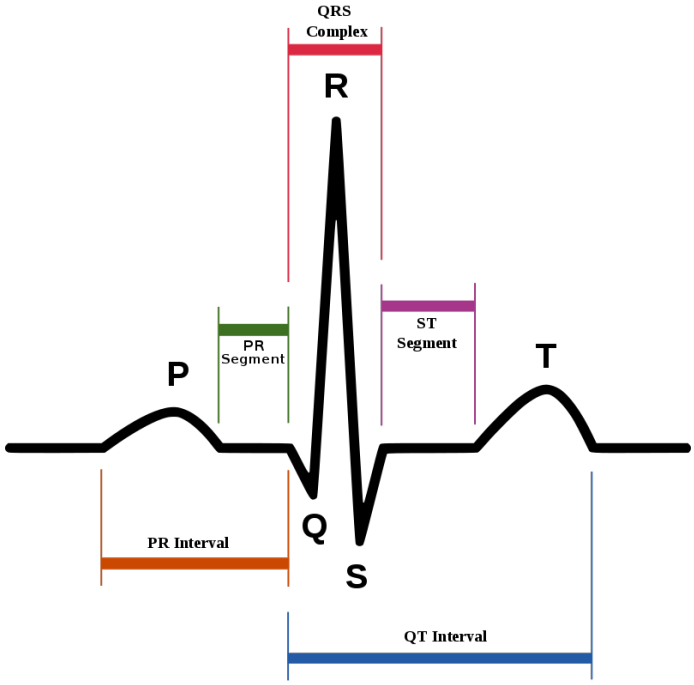


FIGURA 2.10. Un ECG está compuesto por ondas e intervalos entre ellas, tomado de [**17**].

detectar con el ECG del paciente, pues se expresa borrando la onda P, causando una lata varianza en los intervalos RR y generando zigzags en el segmento ST. En esta tesis trabajaremos con el problema de clasificar entre ECG que tienen ritmo con fibrilación auricular y los que tienen ritmo normal.

2.4. Algoritmos de clasificación

Un problema de *clasificación* consiste en asignar una etiqueta a un dato nuevo, después de haber aprendido una regla de asignación usando otros datos. Para poder entrenar un modelo de clasificación es necesario contar con un conjunto de datos previamente etiquetados y que cuenten con características o variables que ayuden a identificar los rasgos en común de una clase. Generalmente se trabaja con dos conjuntos de datos: los de entrenamiento, con los que se aprende la regla de asignación y los de prueba, con los que se verifica la eficacia del clasificador. En nuestro caso, contaremos con datos de ECG, los cuales previamente fueron identificados por cardiólogos dentro de las clases de ritmo normal o ritmo con fibrilación auricular y el objetivo es determinar a que clase pertenece un dato nuevo, para así desarrollar un

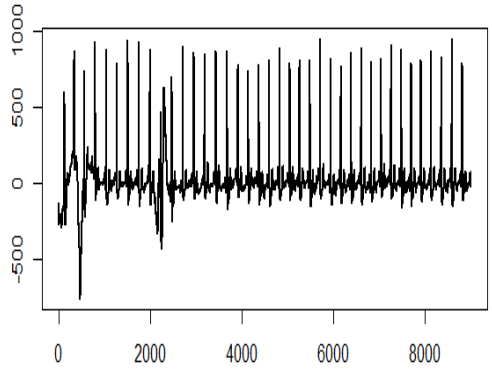
diagnóstico.

Existen diversos algoritmos de clasificación, como regresión logística, k-vecinos más cercanos y bosques aleatorios. Para esta tesis se trabajará con el algoritmo de *bosques aleatorios*, pues además de tener buenos resultados es un método robusto a la dimensionalidad de las características de los datos. El método de bosques aleatorios consiste en someter a votación la elección de varios árboles de decisión elegidos de manera aleatoria.

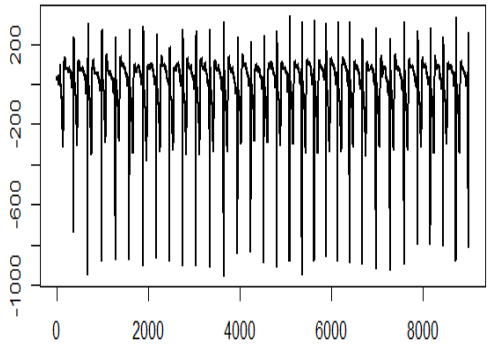
En esta tesis, usaremos dos tipos distintos de características para clasificar: las de origen médico y las de origen topológico. Las características médicas son aquellas que se obtienen directamente de los ECG y que tienen una justificación en el campo médico, por ejemplo, los intervalos RR y su varianza relativamente grande ocasionada por la irregularidad de los pulsos. Las características topológicas son aquellas que obtendremos después de una transformación de los ECG y describen la forma de los ECG, por ejemplo, el diagrama de persistencia. Las características médicas son las más típicas y su uso es tradicional en la clasificación de ECG. Las características topológicas surgieron en años recientes y son una propuesta nueva para el estudio de este tipo de señales Perea (2015) [**3**].

Las características médicas presentan diversas ventajas y desventajas. Entre las ventajas podemos mencionar la facilidad para interpretarlas, pues están relacionadas con el efecto directo de la cardiopatía sobre el ritmo. Además las características médicas nos proporcionan mejores resultados en la clasificación. Entre sus desventajas podemos mencionar el hecho de que dependen de la interpretación médica y de conocimiento específico del efecto de cada enfermedad sobre el ECG, lo que significa que las características usadas pueden a cambiar con cada enfermedad, para lo que cada clasificación necesita un estudio diferente.

De la misma manera podemos hablar de las ventajas y desventajas de las características topológicas. Dentro de sus principales ventajas tenemos que al estudiar los datos usando únicamente su forma, las características son más robustas a bases de datos que contengas ECG invertidos, lo cual ocurre por un ajuste erróneo en el aparato de medición, ver Figura 2.11. Otra ventaja es su potencial en aplicaciones a otro tipo de señales tanto dentro como fuera del área médica. Dentro de sus desventajas tenemos una interpretabilidad muy pobre, pues la topología resulta una herramienta un tanto abstracta.



(A) ECG en el sentido usual.



(B) ECG ordenado en el sentido opuesto.

FIGURA 2.11. Ejemplo de dos ECG donde uno presenta un error que hace que aparezca en sentido opuesto.

Capítulo 3

Topología en series de tiempo

En este capítulo se estudia la forma en que se puede usar la teoría de ATD para extraer información de las series de tiempo. Primero se expone el método de la ventana deslizante con retardo, que nos proporciona una representación topológica de una serie de tiempo. Posteriormente, se revisaran algunos de los métodos que se han desarrollado recientemente para la clasificación usando TDA, así como de algunas de sus propiedades y ventajas para su posterior aplicación en este trabajo. Terminamos presentando un ejemplo donde se obtiene los estadísticos topológicos de un ECG.

Se recomienda consultar el artículo original donde se presenta el método de la ventana deslizante con retardo en Takens (1981) [**19**], así como en algunos artículos donde se usa este método para estudiar series de tiempo desde un enfoque de sistemas dinámicos como Stam (2005) [**20**] y Hundewale (2012) [**21**]. En Perea (2015) [**3**] se usa este método junto con técnicas de TDA para estudiar la periodicidad en series de tiempo y en Ignacio (2019) [**5**] es usado junto con la información del diagrama de persistencia para realizar clasificación de ECG.

3.1. Método de ventana deslizante con retardo

El *método de la ventana deslizante* es un procedimiento que se aplica a una serie de tiempo para obtener una nube de puntos que tenga información sobre la forma de la serie original. Esto se hace agrupando valores consecutivos en vectores de la misma dimensión. La nube de puntos vive en un espacio euclidiano, que generalmente tiene una dimensión alta. La información obtenida se usa para caracterizar la serie de tiempo y poder obtener estadísticos con los que se puedan trabajar.

Para entender el método de la ventana deslizante con retardo primero debemos entender el resultado en el que está basado: *el Teorema de Takens*. Una *variedad* es un espacio topológico que localmente se parece mucho a un espacio euclidiano. Un *sistema dinámico* sobre una variedad es una función que nos indica, dado un tiempo t y punto sobre la variedad x, la nueva ubicación de dicho punto en el tiempo t. Un *atractor* es un conjunto de puntos hacia los que un sistema dinámico tiende a evolucionar. El Teorema de Takens nos dice que si tenemos una única coordenada

de la evolución de un punto, es decir su *órbita*, es posible recuperar la topología del atractor. De hecho si tenemos una coordenada de todas las órbitas es posible recuperar toda la variedad.

La órbita de un punto es un conjunto de coordenadas que va cambiando con el tiempo, por lo que si tomamos una única coordenada tenemos una serie de tiempo. Es aquí donde estra en juego este trabajo, pues podemos suponer que nuestras series de tiempo son coordenadas de algún sistema dinámico más complejo. En particular podemos suponer que los ECG son solo una coordenada de un sistema dinámico más complejo que describe la dinámica del corazón. Por lo tanto, queremos recuperar información topológica de este atractor, usando el Teorema de Takens.

La forma de recuperar la estructura topológica de un sistema dinámico a partir de una serie de tiempo $S(t),\ t \in \mathbb{R}^+$ es la siguiente: primero necesitamos dos parámetros el tamaño de ventana $m \in \mathbb{B}$ y el tiempo de retardo $\tau \in \mathbb{R}^+$. Entonces la ventana deslizante con retardo es el conjunto de puntos

$$\mathrm{VD}(S) = \{(S(t), S(t+\tau), \ldots, S(t+m\tau)) \ : \ \forall\, t \in \mathbb{R}^+\}.$$

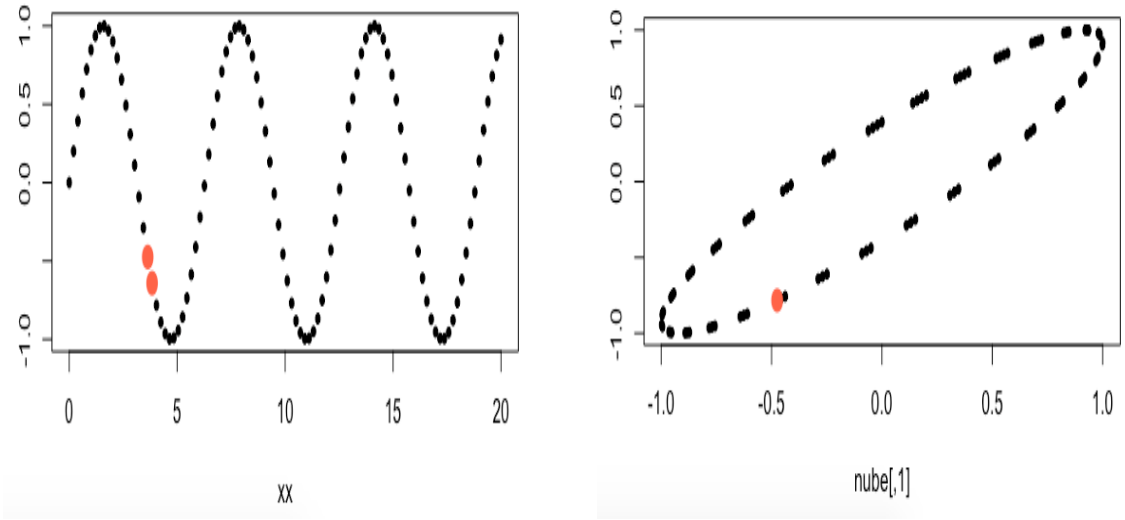
Sin embargo, en la práctica no es posible tener observaciones de la serie de tiempo en todos los tiempos t, generalmente se tiene un conjunto discreto de tiempos en los que se realiza la observación $\{t_0, t_1, \ldots, t_n\}$, por lo que nuestro parámetro τ ahora es un número natural que nos indica cuantos tiempo dejamos de espacio y la ventana deslizante es la nube de puntos

$$\mathrm{VD}(S) = \{(S(t_i), S(t_{i+\tau}), \ldots, S(t_{i+m\tau})) \ : \ \forall\, i \in \{0, 1, \ldots, n\}\}.$$

Para tener un mejor resultado con nuestros datos se debe tener en cuenta las siguientes consideraciones sobre los parámetros del tamaño de la ventana y el tiempo de retardo. Primero, es desconocida la dimensión de la variedad original en la que se encuentra el sistema dinámico asociado, por lo que el tamaño de la ventana debe ser lo suficientemente grande para que la nube de puntos tenga una dimensión igual o mayor que la variedad original, esto nos puede generar un problema de alta dimensionalidad cuya solución presentamos más adelante. Segundo, el tiempo de retardo debe no se recomienda que sea cercano a uno, pues valores adyacentes en una serie de tiempo suelen ser cercanos, por lo que los puntos de la nube de puntos se pueden agrupar mucho alrededor de la recta identidad en el caso que se toma un retardo muy pequeño.

A continuación presentamos un ejemplo que ilustrará las ideas que acabamos de exponer. Comenzamos con la variedad $S^1 = \{(\cos(\theta), \sin(\theta)) \ : \ \theta \in (0, 2\pi]\}$ y con

el sistema dinámico $f(t,(\cos(\theta),\sin(\theta))) = (\cos(\theta+t),\sin(\theta+t)), \forall t \in \mathbb{R}$, que hace que los puntos del círculo se roten a la izquierda o derecha según el signo de t. Si nos quedamos sólo con la segunda coordenada del sistema dinámico, esta seria de la forma $\sin(x)$. Por lo tanto, aplicando el método de la ventana deslizante con tamaño de ventana $m = 2$ y $\tau = 1/5$ tenemos la nube de la Figura 3.1. Podemos notar que dado que tomamos un valor de τ muy pequeño la nube de puntos aparece cerca de la recta identidad $y = x$.



(A) Valores en la función seno. (B) Nube formada por la función seno.

FIGURA 3.1. Método de la ventana deslizante aplicado a la función seno.

En De Silva (2012) [**2**], se menciona que para series periódicas la variedad del atractor serán circulos que se puede enredar en varios ciclos. Dado que los ECG son series de tiempo periódicas, al momento de estudiarlos consideraremos la homología persistente de índice 1, pues es la que cuenta los ciclos.

Si la serie de tiempo es de la forma $f(t) = \sin(at) + \cos(bt)$, entonces la serie de tiempo pertenece a la órbita de un punto sobre el toro, De Silva (2012) [**2**]. Sabemos que si a/b es racional dicha órbita corresponde a un nudo toroidal Murasugi (2008) [**22**]. Pero en el caso en que a/b, la órbita corresponde al sistema dinámico conocido como toro irracional, el cual tiene como atractor al toro mismo pues el ciclo nunca se cierra. Por lo tanto, aplicando el método de la ventana deslizante sobre la serie $f(t)$ nos da una nube de puntos como se puede apreciar en la Figura 3.2. Este resultado nos muestra que es viable el uso de la topología para el estudio de otro tipo de series

de tiempo asociados a sistemas más complejos.

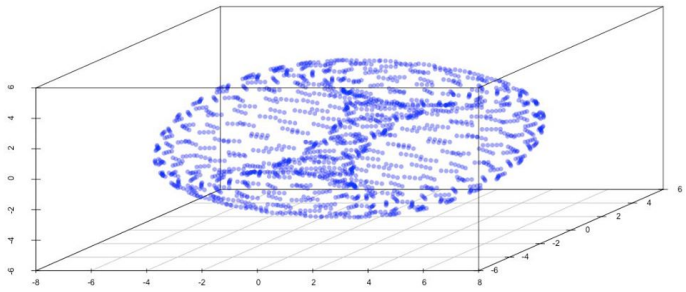


FIGURA 3.2. Órbita del toro irracional obtenida con el método de la ventana deslizante.

Existen varias formas de mejorar la calidad de la información obtenida con el método de la ventana deslizante al aplicarlos a ECG. La primera de ellas se plantea en Kim (2018) [**23**], y consiste en realizar una proyección sobre las componentes principales de la nube de puntos obtenida en el último paso. Esto tiene una doble finalidad: por un lado reduce la dimensión de los datos, por lo que se disminuye la complejidad en los cálculos computacionales y también reduce el efecto del ruido presente en la serie de tiempo. Otra forma de aminorar el ruido consiste en aplicar el método de la ventana deslizante sólo sobre una parte de la serie de tiempo con una buena calidad de señal.

Otra forma de mejorar la calidad de la información es haciendo que los ciclos de las órbitas obtenidas sean lo más redondos posibles. En Perea (2015) [**3**] se demuestra que los ciclos son más redondos cuando el tamaño de la ventana se aproxima al tamaño del periodo de la serie de tiempo, donde redondes se define como el radio del círculo más grande que cabe dentro de un ciclo en la órbita obtenida. Tomaremos en cuenta todas estos aspectos para mejorar la información al momento de obtener nuestros estadísticos sobre los ECG.

3.2. Estadísticos topológicos en series de tiempo

Para poder caracterizar la serie de tiempo original usando la nube de puntos obtenida con el método de la ventana deslizante con retardo existen diferentes opciones. Dado que el método surgió de la teoría de sistemas dinámicos es usual usar elementos de esta teoría para caracterizar a la serie de tiempo. Por ejemplo, en Hundewale (2012) [**21**], se usa el exponente de Lyapunov y la entropia de Kolmogorov

para estudiar ECG. En De Silva (2012) [**2**] se propuso por primera vez el uso de homología persistente para estudiar series de tiempo.

En esta sección estudiaremos dos propuestas para la obtención de características topológicas para posteriormente usarlas en la labor de clasificación de series de tiempo. Ambas propuestas se usan con nuestros datos de ECG, para poder compararlas.

3.2.1. Método de Chazal. En [**1**] se presenta un ejemplo de clasificación de datos pertenecientes a sensores de movimiento usando los panoramas de persistencia, los cuales vimos enteriormente, y aplicando un algoritmo de bosques aleatorios. Inspirados en esta idea se propone calcular los panoramas de persistencia obtenidos de la nube de puntos del método de la ventana deslizante con retardo. Para poder usar estos datos en un algoritmo de clasificación primero debemos considerar una versión discreta de los panoramas. Tomamos 100 observaciones igualmente espaciadas para cada uno de ellos. Para la homología de dimensión 0 y 1 tomamos los primeros 3 panoramas y los agrupamos juntos. Por lo tanto, tenemos como características un vector de dimensión 600 que representa a los panoramas.

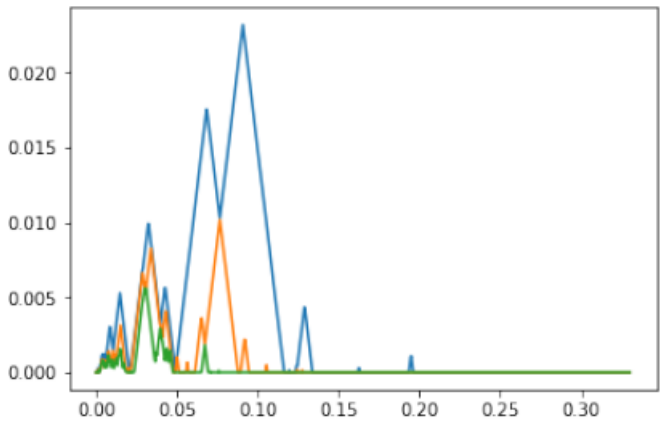


FIGURA 3.3. Panoramas de uno de los datos clasificados en Chazal (2017) [**1**].

Posteriormente, usaremos estas características para realizar la clasificación usando bosques aleatorios, pues al ser un método robusto a la alta dimensionalidad no se verá afectado por el gran tamaño de nuestro vector de características.

3.2.2. Método de Ignacio. Este método fue propuesto en Ignacio (2015) [**5**] y consiste en obtener la nube de puntos asociada al ECG usando el método de la ventana deslizante. Posteriormente, se calcula el código de barras de la nube de puntos correspondiente. Usando la información de las barras se calcula un total de 17 estadísticos relacionados con la distribución de los nacimientos, muertes y persistencias.

Algunos de estos estadísticos implican excluir las barras más persistentes. En el área de ATD, se considera que los datos más persistentes proporcionan características globales de la nube de puntos, mientras que los datos menos persistentes dan descripciones locales. En la siguiente sección comentaremos más sobre este tema.

En la Tabla 3.1 se muestran los estadísticos que se calculan.

Característica	Media	Desv. Estándar	Skewness	Curtosis	Suma
Dimensión 0					
Muertes	Sí		Sí	Sí	
Dimensión 0 $*$					
Muertes	Sí	Sí	Sí		
Dimensión 1					
Nacimientos		Sí	Sí		
Muertes				Sí	
Persistencia	Sí				Sí
Dimensión 1 $*$					
Persistencia				Sí	Sí
Dimensión 2					
Nacimientos				Sí	
Muertes	Sí				
Dimensión 2 $*$					
Nacimientos			Sí		

CUADRO 3.1. Estadísticos del código de barras obtenidos en Ignacio (2019) [**5**] para la clasificación de ECG. Las barras con $*$ no incluyen las 5 % más persistentes. [**6**]

Posteriormente, se usan estos 17 estadísticos para realizar la clasificación usando el algoritmo de bosques aleatorios.

3.3. Ejemplo estadísticos topológicos de un ECG

Describiremos la forma en que se obtienen los estadísticos topológicos en un ECG, que usaremos en el siguiente capítulo para realizar la tarea de clasificación. Comenzaremos con un ECG de 30 segundos, obtendido con una frecuencia de 300 Hz, es decir se toman 300 muestras cada segundo. Nuestro ECG es un vector de longitud 9000 que gráficamente se ve como en la Figura 3.4. Para el método de ATD es muy costoso computacionalmente trabajar con una frecuencia tan elevada, por lo que agrupamos cuatro puntos consecutivos, calculando su promedio, y obtenemos un vector de longitud 2250.

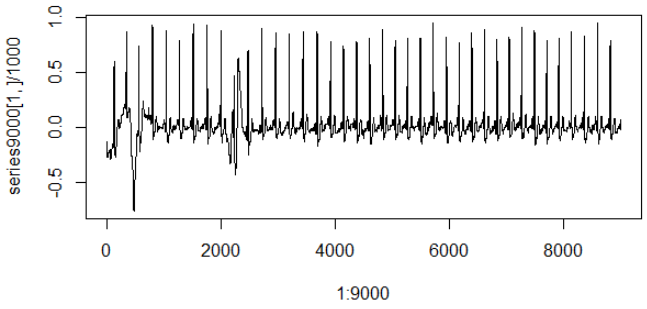


FIGURA 3.4. Electrocardiograma de 30 segundos con ritmo normal [**4**].

Siguiendo la recomendación de trabajar con un intervalo del ECG que tenga buena calidad, dividimos el ECG en nuevo partes iguales y usando el paquete `Neurokit` de Python se calcula la calidad de la señal de cada uno de los segmentos. Seleccionamos el que cuenta con mejor calidad para seguir trabajando. El periodo de los ECG es de un pulso cada 0.8 segundos aproximadamente, por lo que siguiendo el consejo para maximizar la redondes tomamos un tamaño de ventana de 0.8 segundos.

Usando un tiempo de retardo de 2 puntos en el vector del ECG, se obtiene la nube generada por el método de la ventana deslizante. La dimensión de la nube es $m = 60$, para reducir la dimensión de los puntos de la nube y para reducir el ruido aplicamos PCA, obteniendo los primeros 3 componentes de la nube, lo que nos da una nube de puntos en $\mathbb{R}^3$. Esta será la nube de la que obtendremos la información topológica.

Finalmente, usando la librería GUDHI, disponible para R y Python calculamos, el diagrama de persistencia de la nube de puntos y los correspondientes panoramas

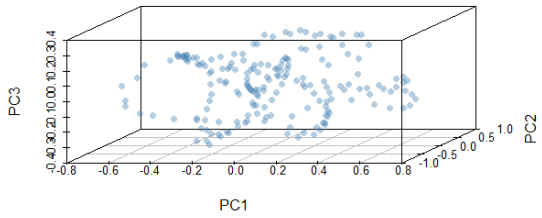
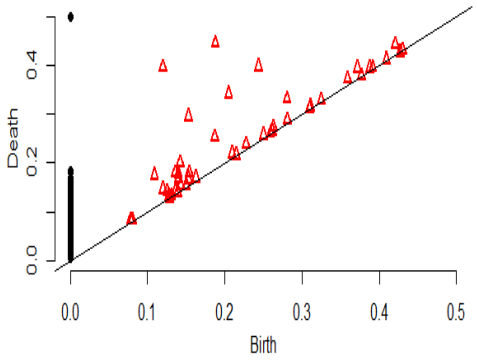
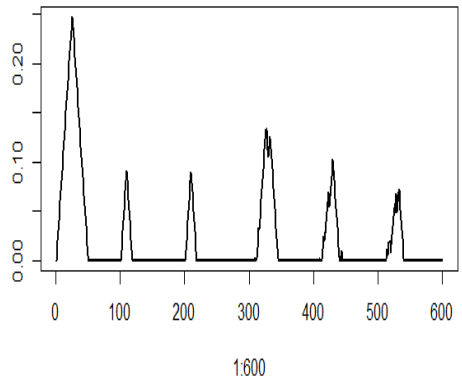


FIGURA 3.5. Nube de puntos asociada al ECG de la Figura 3.4.

de persistencia. En la Figura 3.6 podemos ver el diagrama de persistencia donde los puntos negros corresponden a los nacimientos y muertes de las componentes conexas, es decir la homología de dimensión 0, y en color rojo los nacimientos y muertes de los ciclos, homología de dimensión 1. Finalmente, se obtienen los primeros 3 panoramas de dimensión 0 y los primeros 3 panoramas de dimensión 1 agrupados en una misma función. Con los versión vectorizada de los panoramas y con los estadísticos del método de Ignacio tenemos los estadísticos topológicos que buscamos.



(A) Diagrama de persistencia de la nube de la Figura 3.5.



(B) Algunos panoramas de la nube de la Figura 3.5.

FIGURA 3.6. Estadísticos topológicos del ECG 3.4.

Capítulo 4

Particiones del código de barras para obtención de estadísticos

En el método de Ignacio se hace una separación de las barras del código de barras en dos categorías: las 5 % más persistentes y las 95 % menos persistentes. Se construyen estadísticos de distinta naturaleza con estas barras, como se aprecia en la Tabla (2.1). Esta división surge de la idea fundamental del ATD, que intuye a las barras más persistentes como descriptoras de características globales de la forma de la nube de puntos.

En sus inicios, el análisis topológico de datos simplemente excluía a una porción importante de barras pequeñas al momento de realizar el análisis estadístico. Por ejemplo, en Edelsbrunner (2002) [**13**] solo se considera la información de los ciclos que se encuentran por encima de cierto umbral. Posteriormente se comenzaron a incorporar a las barras pequeñas en el análisis, para construir estadísticos que aportaran nueva información. Un ejemplo concreto son las siluetas Chazal et al. (2013) [**24**].

Las opciones para elegir estadísticos que ayuden en la clasificación son grandes y la selección de los más significativos depende altamente de los datos con los que se esté trabajando. El proceso de optimización hace un barrido, más o menos empírico, de posibles características clasificadoras, ponderando su relevancia en la clasificación contra su complejidad computacional.

Hasta el momento, los estadísticos del ATD basados en homologías de dimensión mayor o igual que dos, difícilmente terminan dentro de la lista de principales características clasificadoras. Por otro lado, para las herramientas del ATD que se obtienen de la homología de dimensión cero, se podría argumentar que se encuentran incluídas dentro la teoría clásica de análisis de grupos, en particular en algoritmos de agrupamiento jerárquico. En consecuencia, las mejoras experimentadas en los problemas de clasificación (al incorporar las herramientas del ATD) se concentran hasta ahora, en estadísticos derivados de la homología de dimensión 1.

En general, para cualquier dimensión, la consideración de separar el 5 % o alguna proporción fija de las barras más grandes nos resulta un tanto arbitraria. En esta dirección, se abordan en este capítulo las siguientes preguntas:

1. ¿Cómo diferenciar entre una "barra chica"(que puede usarse para generar estadísticos que describan la geometría local) de una "barra grande"(inherente a la forma global de la nube)? ¿Cómo decidir la proporción óptima de barras que se considerarán como persistentes?

2. ¿Puede hacerse esta diferenciación de una manera mas conceptual?

Responderemos primero de forma positiva a la segunda pregunta, para la que introduciremos algunos conceptos importantes, como los complejos de Čech y las triangulaciones de Delone. Mostramos que, cuando se consideran filtraciones de Čech, existe una correspondencia biyectiva natural entre el conjunto de triángulos agudos de la triangulación de Delone y un subconjunto especial de barras, que llamaremos "de Delone", que consiste típicamente de las barras más pequeñas.

Esto nos ofrece una separación conceptual entre barras chicas (de Delone) y grandes (el resto). Lo anterior aporta también a nuestra comprensión del tratamiento diferenciado que se le da a barras pequeñas (vs. grandes) en la tabla de características presentada al final del capítulo anterior.

Una vez que se remueven todas las barras correspondientes a triangulos de Delone, nos quedamos con las barras correspondientes a ciclos que describen características globales. Aquí cabe enfatizar que no nos quedaremos con una única barra grande representando a cada generador del grupo de homología persistente. Por el contrario, varias de las barras no-Delone pueden corresponder a un mismo generador. Consideramos de relevancia para trabajo a futuro la discusión en torno a un tratamiento más conceptual (implementable) de las barras no-Delone.

Por lo pronto, nos conformamos con una mejora conceptual de la partición del conjunto barras, separando las barras de Delone de las demás. Un problema que nos ocupa actualmente es determinar la implementabilidad de tal partición (en particular, para la clasficación de ECG's). Como hemos mencionado, para estos problemas de clasificación se han usado siempre filtraciones de Vietoris-Rips.

Las barras correspondientes a triángulos agudos de Delone generalmente son pequeñas al compararse con el resto. Sin embargo, en caso de que no se consideren

suficientes puntos en la nube, o bien si éstos se encuentran distribuidos de forma poco regular, es posible que las barras de Delone resulten comparables en tamaño con algunas barras en el complemento. En estos casos, resulta conceptualmente incorrecto asumir que basta con elegir la proporción adecuada de barras chicas vs grandes, como plantea la Pregunta 1. Por fortuna, nuestra partición conceptual evade esta problemática, pues determina cuáles (y no únicamente cuántas) son las barras de Delone.

En la segunda parte de este capítulo, ejemplificamos la problemática anterior comparando nubes de puntos en el toro con distintas distribuciones. Para producir nubes aleatorias de puntos más regulares hacemos uso de ley del círculo de Ginibre. Mostraremos que para nubes de puntos menos regulares las barras correspondientes a características globales no necesariamente se distinguen de las barras pequeñas.

4.1. Čech y Delone

4.1.1. Filtración de Čech. El complejo de Vietoris-Rips tiene la propiedad de ser fácil y rápido de calcular, por lo que presenta ventajas computacionales. Sin embargo, se pierde información topológica de las nubes de puntos. Por ejemplo, si tenemos tres puntos equidistantes el complejo de Vietoris Rips nunca detecta el ciclo que hay dentro de los puntos, pues el simplejo del triángulo aparece en el momento que aparecen las tres aristas. Esta información no se pierde cuando consideramos *el complejo de Čech*, del cual hablaremos a continuación basandonos en Ghrist (2010) [**25**].

Dada una nube de puntos X y un parámetro de longitud $\epsilon > 0$, el complejo de Čech de radio ϵ, denotado $\check{C}_\epsilon$, X con radio ε es

$$\check{\mathrm{C}}_\varepsilon(X) = \{\sigma \subset \mathcal{P}(X) \ : \ \cap_{x\in\sigma} B(x,\epsilon) \neq \emptyset\}.$$

El complejo de Čech es más tardado de calcular, pues requiere verificar varias intersecciones, pero tiene algunas propiedades topológicas de las que carece el complejo de Vietoris-Rips. El mejor ejemplo de esto es el Lema del Nervio, Ghrist (2010) [**25**], el cual nos dice que dada una nube de puntos X y un radio ε el espacio topológico obtenido de la unión de las bolas con centros en X y radio ε, también llamado conjunto de Čech, es homotópicamente equivalente a $\check{\mathrm{C}}_\varepsilon(X)$. Esto nos dice que el complejo de Čech tiene propiedades topológicas parecidas a un espacio topológico que describe mejor la forma de la nube de puntos, como en la Figura 4.1.

Además el complejo de Čech se puede aproximar con dos complejos de Vietoris-Rips relativamente cercanos. Si X es una nube de puntos, para cualquier $\varepsilon > 0$

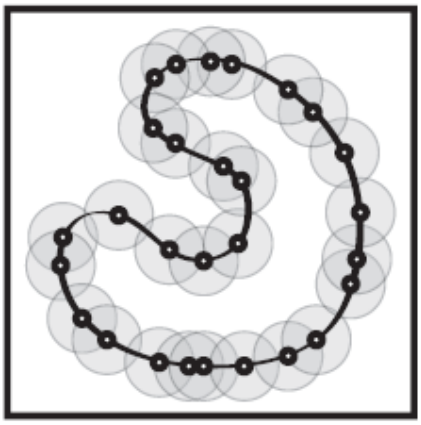


FIGURA 4.1. Nube de puntos con bolas de radio ε y su correspondiente complejo de Čech. Ghrist (2010) [**25**]

tenemos que

$$VR_{\varepsilon}(X) \subset \check{C}_{\varepsilon'} \subset VR_{\varepsilon'}(X),$$

$\forall \varepsilon' \geq \varepsilon\sqrt{2}$.

4.1.2. Triangulación de Delone. En el contexto de asociarle estructuras mas complejas a nubes de puntos sobre espacios topológicos surge el concepto de *triangulación*, la cual es un conjunto de triángulos que tiene como vértices los puntos de la nube. Un tipo particular de triangulación que nos resulta útil por sus propiedades y por aparecer de manera natural en diversas estructuras es *la triangulación de Delone*. Su definición precisa y aplicaciones se puede consultar en Edelsbrunner (2010) [**26**]. A continuación daremos una breve descripción de la triangulación de Delone, así como de algunas de sus propiedades más importantes y en la siguiente sección mencionaremos la importancia de dicha estructura en el contexto de este trabajo.

Una nube de k puntos en $\mathbb{R}^2$ se encuentra en *posición general* si no hay tres puntos colineales, cuatro concíclicos, etc. La Triangulación de Delone para esta nube de puntos se obtiene tomando la envolvente convexa de los puntos y triangulando el interior de manera que para cada triangulo se cumpla la propiedad de Delone: que en el interior o frontera de su circunferencia circunscrita no haya ningun punto de la nube.

Es posible ver que la triangulación de Delone es única, Aurenhammer (2013) [**27**]. Uno de los algoritmos para encontrar la triangulación de Delone consiste en partir

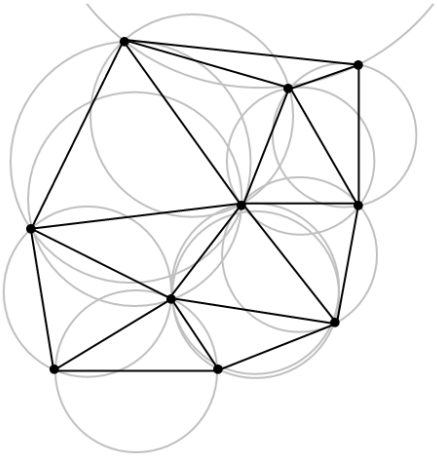


FIGURA 4.2. Triangulación de Delone asociada a una nube de puntos.

de una triangulación arbitraria de los puntos y a través de cambios de arista verificar si se cumple la propiedad de Delone. La triangulación de Delone se puede extender a dimensiones mayores de la manera natural, pero con tetraedros y sus generalizaciones de dimensión mayor. Para los fines de este capítulo basta con trabajar en triangulaciones planas.

4.1.3. Barras de Delone. La relevancia de los conceptos que acabamos de enunciar radica en la siguiente proposición.

Proposición 4.1.1. *Sea X una nube de puntos en $\mathbb{R}^2$, considerese su filtración de Čech con su respectivo codigo de barras, y la triangulación de Delone. Entonces*

i) Existe una barra pequeña por cada triángulo agudo de la triangulación de Delone.

ii) La barra correspondiente a cada uno de dichos triángulos muere cuando $r = R$, el circuncirculo del triángulo, y nace en algun momento $r_0 \leq l/2$, donde l es el lado máximo del triángulo agudo.

DEMOSTRACIÓN. Sea ABC un triángulo agudo de la triangulación de Delone. Consideremos el conjunto de Čech inflado con radio $r = R - \varepsilon$, donde R es el circunradio y $\varepsilon > 0$ es un valor pequeño. Entonces el circuncentro O de ABC se incluye en el conjunto de Čech sólo cuando $r = R$ y no antes, pues por la propiedad de Delone, los vértices de la nube son exteriores al circuncírculo de ABC, por lo que los discos de radio $r < R$ y centro en los puntos de la nube tampoco cubren a O. Por ejemplo en la Figura 4.3, si el punto D pertenece a la nube de puntos por la propiedad de Delone D está a una distancia mayor que R de O.

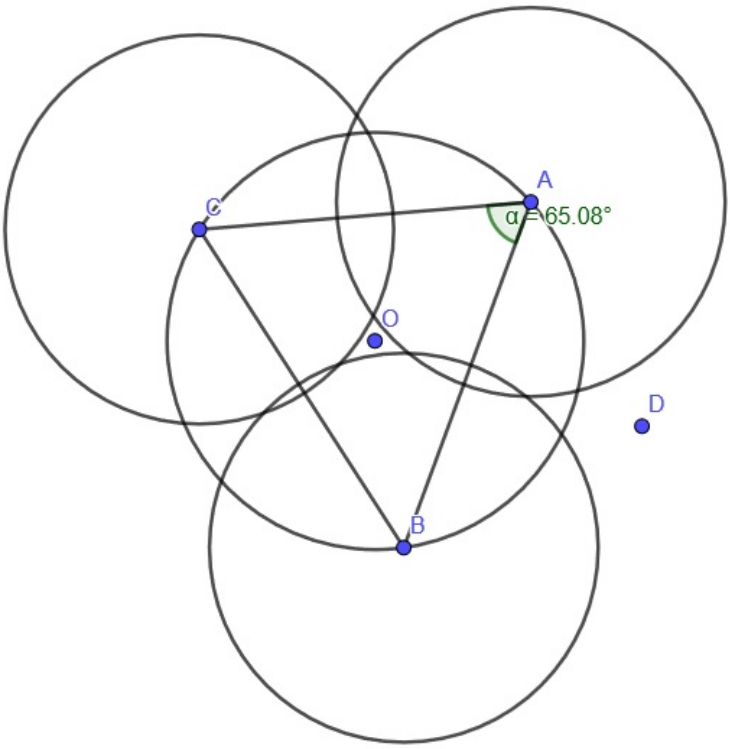


FIGURA 4.3. Triangulo agudo ABC de una triangulación de Delone.

Por otro lado, siempre y cuando $R > r \geq l/2$, habra un ciclo correspondiente al centro O de ABC. Puede ser que este ciclo se forme con anterioridad si muchos puntos estan dispuestos cercanos a la circunferencia, como se ilustra en la Figura 4.4.

Entonces a cada triángulo agudo de Delone le corresponde una barra que termina en R y comienza en algun valor $r_0 \leq l/2 = R(\sin(\alpha))$, donde α es el ángulo opuesto al lado mayor de ABC. □

De la proposición anterior podemos concluir que los triángulos de Delone agudos nos indican el número y la descripción de las barras pequeñas. En las aplicaciones se suelen usar filtraciones diferentes a las de Čech, pero con este resultado obtenemos un conjunto de barras chicas que nos describen propiedades locales de la nube de puntos y que podría aportar información relevante sobre la misma. Este resultado también aporta una diferenciación más conceptual entre barras pequeñas y barras persistentes evadiendo algunos problemas como los que se exponen en la siguiente sección.

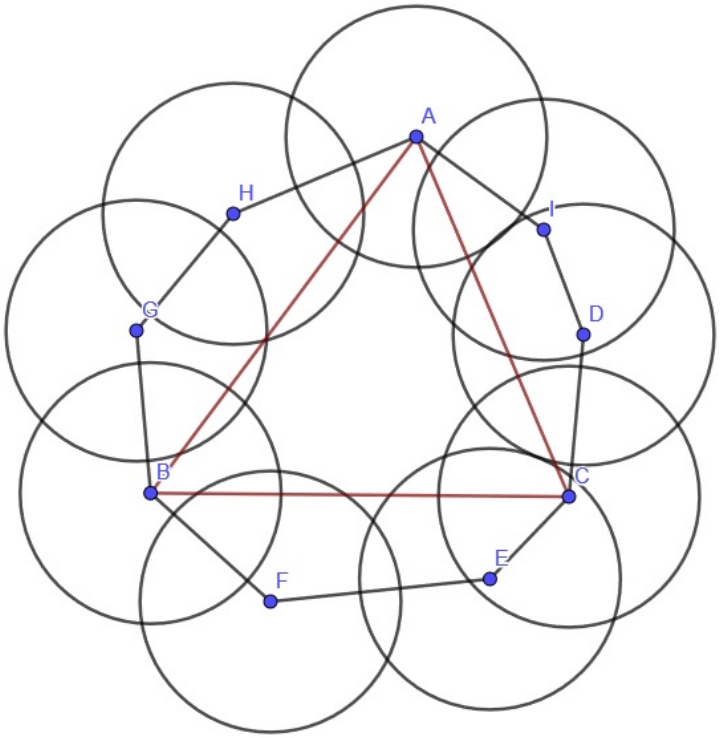


FIGURA 4.4. Ciclo asociado a un triángulo agudo.

4.2. Regularidad de distribuciones en superficies

Recordemos que la motivación de este capítulo es encontrar particiones de las barras que nos aporten información complementaria sobre la nube de puntos. La metodología más usual es separar las barras por su tamaño entre los grupos de persistentes y no persistentes, con la esperanza de que nos aporte información local y global. Pero esto puede llevar a errores cuando la nube de puntos se encuentra distribuida de forma poco regular o cuando no contamos con suficientes puntos. Para mostrar esto, a continuación exponemos un ejemplo con dos distribuciones diferentes sobre la misma variedad con resultados muy distintos en el diagrama de persistencia.

Para este ejemplo tomamos dos nubes con la misma cantidad de puntos sobre el toro. Cada nube será obtenida con una distribución diferente. La primera a partir de una distribución uniforme de puntos independientes y la segunda de una muestra uniforme de puntos que se repelen entre sí.

Nuestro método se basa en un pequeño refinamiento del expuesto en Hernández-Flores [**28**]. Ahí se muestra que los códigos de barras para nubes de puntos más regulares presentan diagramas de persistencia de mejor calidad (al distinguirse más los elementos persistentes del resto). Sin embargo, la diferencia entre ambos modelos no resultó tan drástica como se anticipaba. Observamos entonces que en [**28**], al utilizarse el método de aceptación/rechazo para transferir muestras uniformes del disco al toro, el rechazo de puntos cercanos afecta de manera importante a la regularidad de la nube en la superficie. Esto se manifiesta en los diagramas de persistencia.

Para generar una nube más regular sobre el toro, generamos primero una muestra regular sobre el disco unitario y la transferimos al toro con una parametrización que respete la proporción de las áreas (es decir, se aceptan todos los puntos, transfiriendo efectivamente sin romper la regularidad, la nube de Ginibre al toro).

4.2.1. Matrices Aleatorias de Ginibre. Para generar muestras uniformes en el disco unitario nos basamos en un resultado de matrices aleatorias conocido como *Ley Circular*

Teorema 4.2.1. *Sea* $(X_n)_{n=1}^{\infty}$ *una sucesión de matrices de* $n \times n$ *cuyas entradas son variables i.i.d. gaussianas complejas de media* 0 *y varianza* 1 *(matriz de Ginibre). entonces la distribución de los eigenvalores de* X_n *converge en distribución a la distribución uniforme en el disco unitario.*

Es posible probar que la distribución obtenida por este método tiene un efecto de repulsión entre sus puntos. Este efecto se puede ver más claro cuando vemos los dos ejemplos de distribuciones sobre el disco:

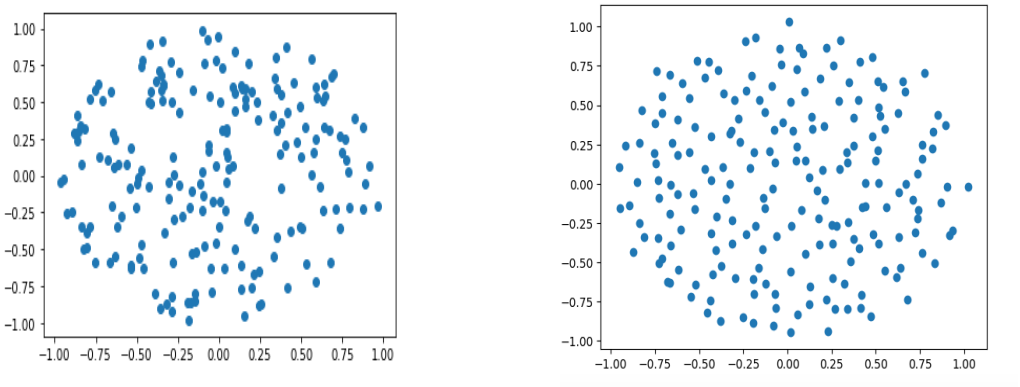
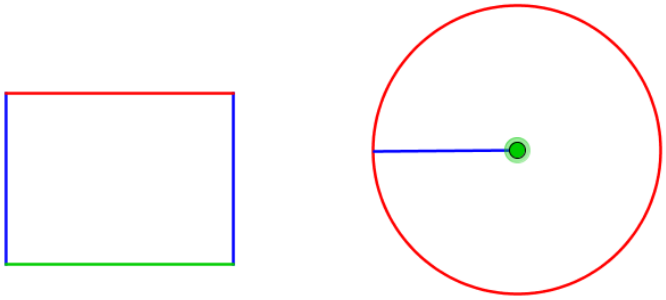


FIGURA 4.5. Distribuciones uniformes en el disco (con la misma cantidad de puntos), obtenidas con una muestra de puntos independientes y con la matriz de Ginibre, respectivamente.

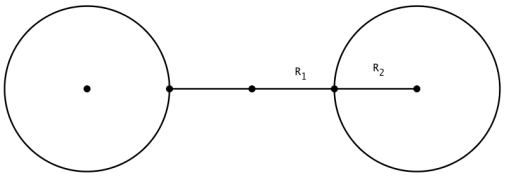
4.2.2. Muestra uniforme independiente. Generamos una una muestra uniforme en el disco unitario, a partir de una muestra uniforme e independiente en el rectángulo $R = [0, 1] \times [0, 2\pi]$. Con este dominio definimos una parametrización del disco $D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ de forma que se preserve la proporción de las áreas.



$$p_1 : R \to D,$$
$$(t, \theta) \mapsto (\sqrt{t}\cos(\theta), \sqrt{r}\sin(\theta)).$$

4.2.3. Muestra sobre el toro. Al toro de radio exterior R_1 y radio interior R_2 lo podemos ver como un sólido de revolución con segmentos:

$$x(t) = R_1 + R_2 + R_2\cos(t)$$
$$y(t) = R_2\sin(t)$$



Consideramos una parametrización del toro, tal que mande a cada radio del disco a un ecuador del toro. Notemos que con esta parametrización el centro del

disco y la frontera del disco estń identificadas en el mismo ecuador, y que cada radio corresponde a un meridiano del toro, observar Figura 4.6.

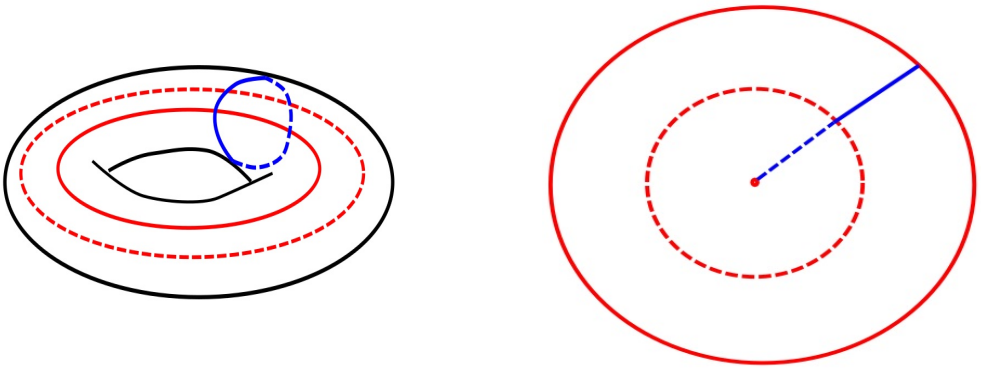


FIGURA 4.6. Parametrización del disco unitario al toro.

Para que las proporciones de las áreas sean las mismas, debemos escoger el ecuador indicado para cada radio del disco. El área de un segmento en el toro es

$$A(t_0) = 2\pi \int_{t_0}^{t_1} x(t) \sqrt{\left(\frac{\partial x}{\partial t}\right)^2 + \left(\frac{\partial y}{\partial t}\right)^2}\, dt.$$

Despejando de esta ecuación se tiene que la parametrización del toro $T(R_1, R_2)$ con dominio el disco asigna el punto $(r, \theta) \in D$ al punto del toro con ángulo exterior θ y ángulo interior ψ, donde ψ es la solución de la ecuación:

$$(R_1 + R_2)\psi + R_2 \sin(\psi) = 2(R_1 + R_2)r^2,$$

la cual se calcula de manera numérica.

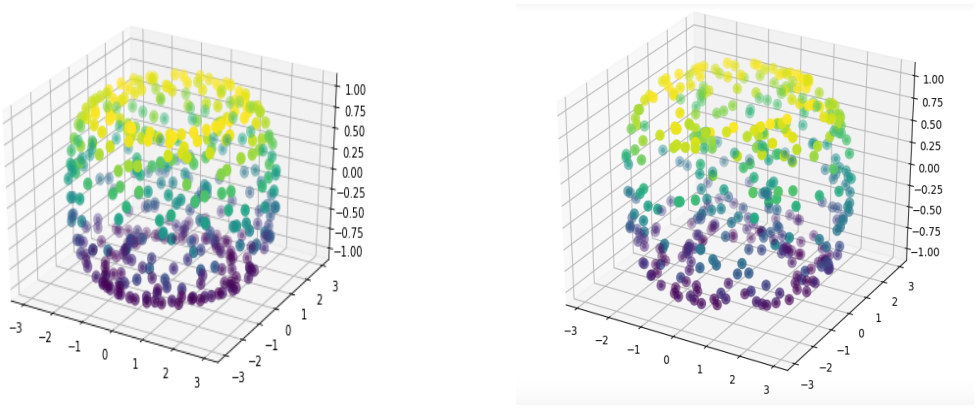


FIGURA 4.7. Muestra uniforme del toro (1,1) con repulsión y sin repulsión.

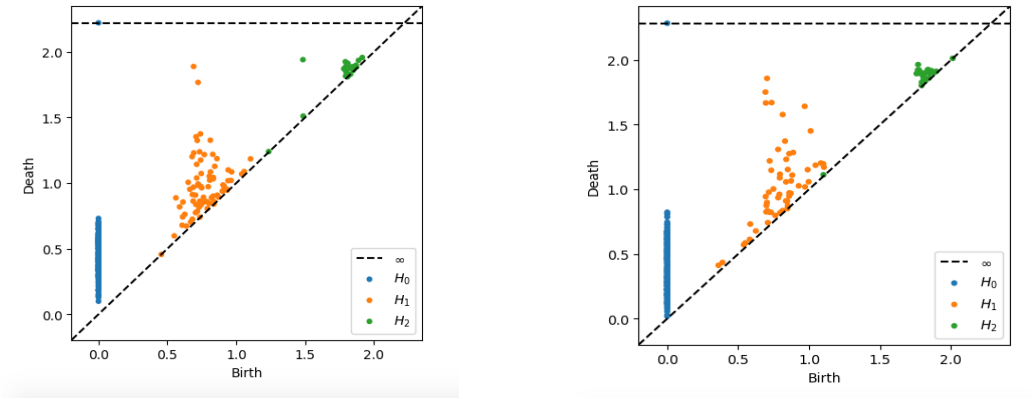


FIGURA 4.8. Diagrama de persistencia del toro (1,1) con repulsión y sin repulsión.

En el diagrama de persistencia de la muestra sobre el toro uniforme con repulsión se pueden identificar claramente los puntos más persistentes que corresponden a los dos ciclos del toro, también se puede distinguir un punto persistente en el grupo de homología de dimensión 2 correspondiente al hueco del toro.

En cambio en la distribución uniforme independiente existen puntos persistentes, por lo que no se puede identificar con claridad a los que representan a las características globales del toro. De la misma forma, en dimension 2, ya no existe un punto distinguido que corresponda al generador del segundo grupo de homología.

Por esto, es necesario investigar nuevas formas de diferenciar la información global y local, que no esten basadas exclusivamente en la persistencia de las barras.

Capítulo 5

Clasificación de ECG

En este capítulo se muestra los resultados de ECG con estadísticos topológicos. En la primer sección se presenta como construir datos sintéticos usando modelos de ECG con ritmos normales y con fibrilación auricular. Estos datos son creados a partir del modelo presentado en Kubiček et al. (2014).

En la siguiente sección, se desea comprobar que tan efectivos son los estadísticos topoólogicos para clasificar. Para esto clasificaremos los ECG simulados usando el método de Chazal, para lo que primero se calculan los panoramas de persistencia. Se tienen dos modelos de ECG, uno sin ruido y otro que simula el efecto de ruido con una caminata aleatoria. Se compara la efectividad de la clasificación topológica en ambos casos.

En la sección subsecuente, clasificamos los datos abiertos del concurso Physionet del año 2017 [**4**], de donde tomamos ECG reales con ritmos normales y con FA. Aquí se presenta un método para clasificar ECG considerando únicamente información médica, la cual está basada en obtener estadísticos de los intervalos RR. Posteriormente, se obtienen las clasificaciones de los datos usando los métodos de Ignacio y de Chazal y se compara con el método médico. Finalmente, se realiza una clasificación usando la información médica y topológica juntas.

El trabajo computacional se realizó en los lenguajes de programación R y Python. Con Python se extrajeron y manipularon los datos de la base de datos original usando el paquete `Numpy`, además se usó el paquete `Neurokit` para obtener información médica de los ECG, como los intervalos RR y la calidad de la señal. Con R se obtuvieron los estadísticos topológicos usando la librería `GUDHI` disponible en el paquete `TDA`, también se realizó la clasifciación de los datos usando el paquete `randomforest`.

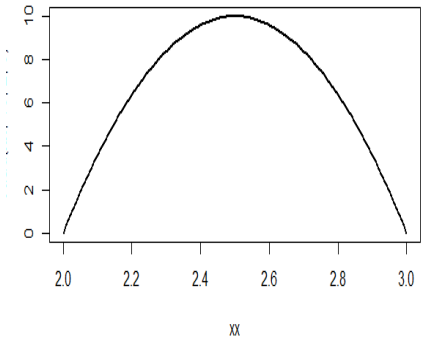
5.1. Clasificación de datos simulados

5.1.1. Construyendo los datos simulados. Construimos un conjunto de datos simulados a partir del modelo de ECG presentado en Kubiček et al. (2014).

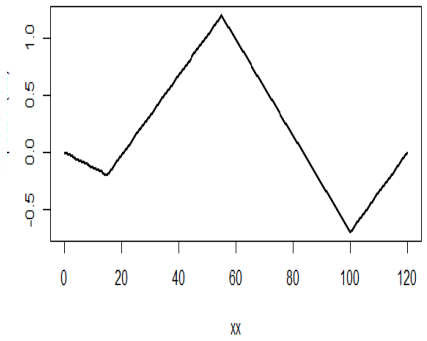
Para construir un ECG artificial debemos simular cada uno de los componentes de un ECG. En la Tabla 2.1 se presentaron los componentes de un ECG, los cuales se pueden dividir en tres tipos:

- Los segmentos. Corresponden a intervalos en que la señal tiene un valor de cero. Los modelamos con un segmento de recta.
- Ondas. Tenemos dos, la onda P y la onda T. Las modelamos con un segmento de una parábola. Ver Figura 5.1.
- Complejo QRS. Corresponde a un complejo formado por las ondas QRS, se modela con una función linear a partes. Ver Figura 5.1.

Usando los rangos de valores descritos en la Tabla 2.1 es posible generar varios pulsos con diferentes parámetros, con lo que se puede construir un ECG simulado. En la Figura 5.2 se muestra un ECG de datos simulados. De esta forma se puede construir una base de datos sintética de ECG con ritmo normal.



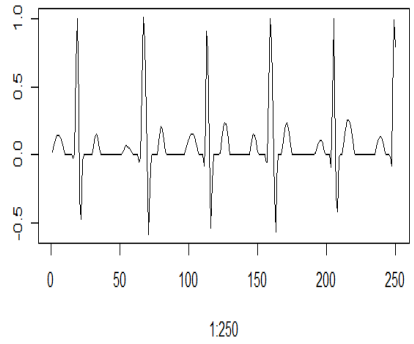
(A) Las onda P y T se modelan con un segmento de parábola.



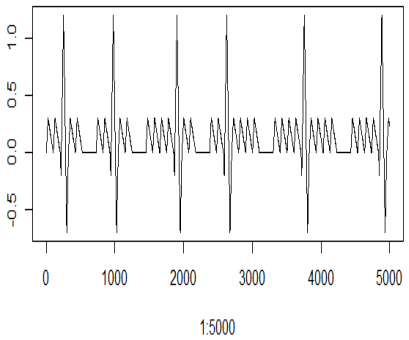
(B) Modelo del complejo QRS.

FIGURA 5.1. Modelos para ondas y segmentos QRS.

Pero necesitamos más de un tipo de dato para poder realizar clasificación. Simularemos ECG de ritmos que presenten fibrilación auricular. Este padecimiento ocasiona que la onda P desaparezca y que en todo el intervalo ST la señal sea errática, es por esto que simulamos los ritmos con fibrilación auricular con zigzags en los intervalos ST, como se aprecia en la Figura 5.2.



(A) ECG simulado de ritmo normal.



(B) ECG simulado de ritmo con fibrilación auricular.

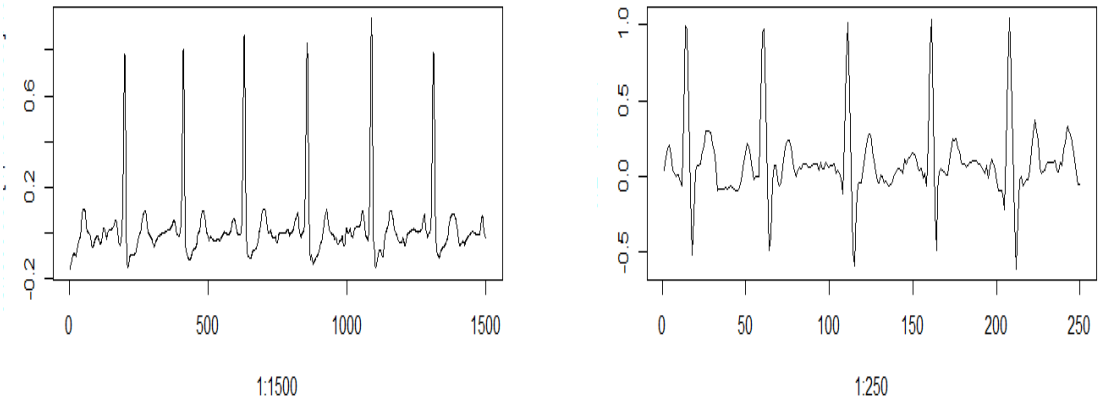
FIGURA 5.2. Simulaciones de ECG.

En capítulos anteriores mencionamos que las series de tiempo presentan un componente de ruido. Entonces para realizar simulaciones de ECG que presenten ruido, a los ECG obtenidos se les agrega aditivamente una caminata aleatoria con barreras reflejantes. En la Figura 5.3 se puede vislumbrar una comparación entre un ECG real y uno simulado al que se le agregó ruido. Se puede apreciar que con esta propuesta se obtienen datos parecidos a los de la vida real.

5.1.2. Clasificando los datos simulados. Simulamos un total de 1200 ECG, de los cuales 600 se obtienen con el modelo de ritmo normal y 600 se obtienen con el modelo de ritmo de fibrilación atrial, todos sin considerar el ruido. Dividimos los datos en dos grupos, tomamos 500 de cada clase para formar el conjunto de entrenamiento y los 100 restantes de cada clase forma el conjunto de prueba.

Realizamos la clasificación con el método de Chazal, para lo cual calculamos los panoramas de la nube que se obtiene del método de la ventana deslizante para cada ECG. Como vimos en el capítulo 3, discretizamos cada función. Obtenemos los primeros 3 panoramas de dimensión 0 y 1, por lo que tenemos 6 vectores de dimensión 100 que agrupamos una tras el otro, obteniendo un vector de dimensión 600.

Finalmente, aplicamos el algoritmo de bosques aleatorios, obteniendo una clasificación correcta en el conjunto de entrenamiento de 96 %. En la Figura 5.4 se muestra una comparación de algunos panoramas así como de la proyección sobre las primeras



(A) ECG real de la base de Physionet [**4**]. (B) ECG simulado.

FIGURA 5.3. Comparación de un ECG real de la base de Physionet y un ECG simulado con ruido.

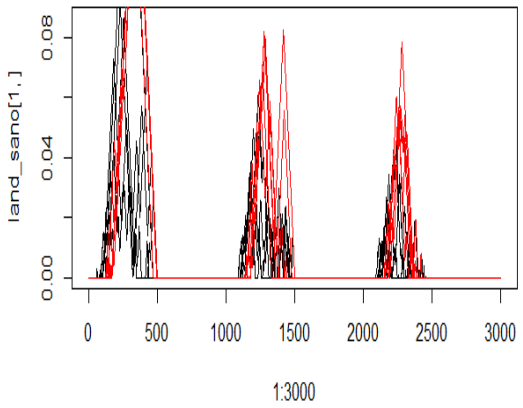
dos componentes de los panoramas.

Repetimos el proceso de clasificación pero ahora simulamos datos que contengan ruido. Los resultados de clasificación fueron de 84 % de clasificación correcta en el conjunto de prueba. La reducción en la clasificación correcta se debe a que el ruido dificulta la tarea de clasificación. Observando en la Figura 5.5 podemos ver el efecto sobre los panoramas y la proyección de los mismos en sus componentes principales.

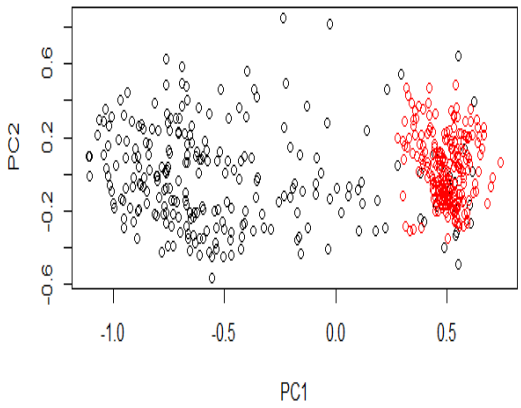
5.2. Clasificación de datos reales

Usaremos la base de datos disponible en la página de Physionet para el concurso de 2017 [**4**]. La base se encuentra dividida en 4 categorías de ECG: ritmo normal, ritmo con fibrilación auricular, otro tipo de ritmo y señales ruidosas. Para los fines de esta tesis sólo usaremos los datos de las clases de ritmos normales y con fibrilación auricular. Algo que caracteriza a los datos reales es que siempre cuentan con ruido por lo que se espera que sea un reto poder clasificar los datos.

De la base de datos tomamos 504 ECG de cada una de las clases de ritmo normal y ritmo con fibrilación auricular, esto para evitar enfrentarnos al problema de datos desbalanceados que no es relevante para este trabajo. Estos 1008 ECG son divididos



(A) Comparación de los panoramas de las dos clases de ECG.



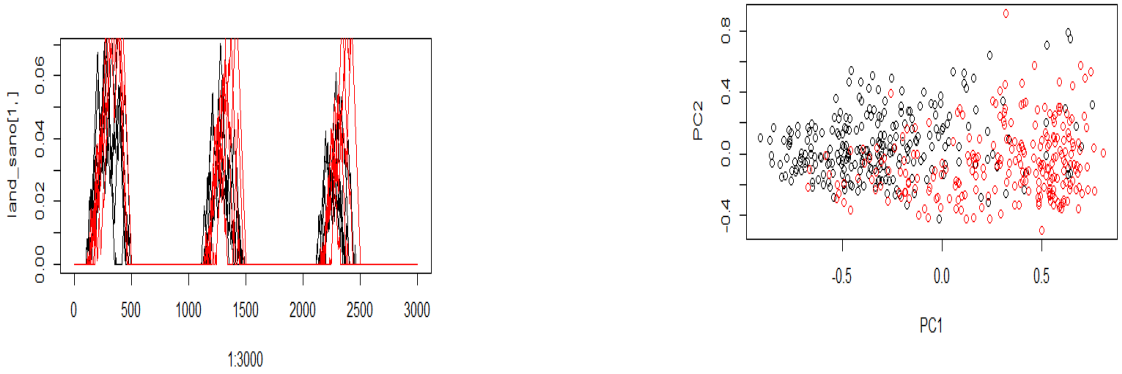
(B) Visualización de las dos primeras componentes principales de los panoramas.

FIGURA 5.4. Comparación visual de las dos clases de ECG, de color negro se muestran los ritmos normales y de rojo los ritmos con fibrilación atrial.

en dos grupos: el conjunto de entrenamiento que se compone de 450 ECG de cada una de las dos clases de datos y el conjunto de prueba que se compone de 54 ECG de cada una de las dos clases.

Primero veamos los resultados de clasificar ECG usando información médica. Este método consiste en obtener, para cada ECG, los valores de sus intervalos RR. Posteriormente, calculamos su media, varianza, curtosis y asimetría estadística. Posteriormente, se usan estás cuatro características para realizar la clasificiación con el algoritmo de bosques aleatorios. Usando el clasificador de los intervalos RR se obtiene una clasificación correcta en el conjunto de prueba de 90.8 %.

Ahora obtendremos información topológica para clasificar los mismos ECG. Para esto primero calculamos la nube asociada a los ECG con el método de la ventana deslizante. Para el método de Ignacio se calcula el código de barras de dicha nube y se obtienen los estadísticos presentados en la tabla 3.1, a estos estadísticos se les aplica el algoritmo de bosques aleatorios para clasificar. Para el método de Chazal se obtienen los primeros 3 panoramas de persistencia de dimensión 0 y 1 de la nube,



(A) Comparación de los panoramas de las dos clases de ECG.

(B) Visualización de las dos primeras componentes principales de los panoramas.

FIGURA 5.5. Comparación visual de las dos clases de ECG, de color negro se muestran los ritmos normales y de rojo los ritmos con fibrilación atrial.

dichos panoramas se discretizan como vectores de dimensión 100 y se concatenan para obtener un vector de dimensión 600 al que se le aplica el método de random forest.

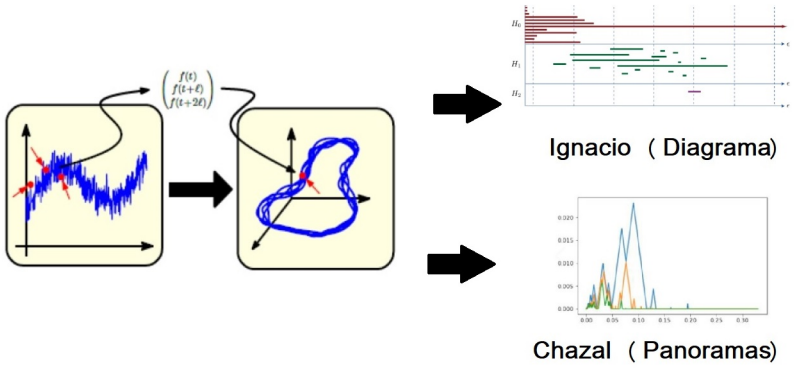


FIGURA 5.6. Procedimiento para obtener los estadísticos topológicos. El método de Chazal usa los panoramas de persistencia, mientras que el método de Ignacio obtiene estadísticos del código de barras.

Al realizar este procedimiento se obtuvo que la clasificación correcta en el conjunto de prueba fue de 61 % usando el método de Ignacio y de 72 % usando el método de Chazal. Estos resultados muestran que la clasificación topológica no es tan buena como la clasificación médica.

Para hacer un estudio más completo, en Ignacio (2019)[**5**] también se combinan las características del clasificador médico con las del topológico para ver si se obtienen mejores resultados. Al combinar los estadísticos de los intervalos RR con los que se obtienen en el método de Ignacio (2019) [**5**] se obtuvo una clasificación correcta de 91.8 %. Por lo tanto, podemos afirmar que el clasificador médico se ve mejorado al combinarlo con el método de Ignacio, aunque la mejoría no es tan grande al ser de sólo 1 %.

Cuando se combinan los panoramas con los datos médicos se obtiene una mala clasificación, lo cual puede ocurrir debido a que los panoramas son representados por vectores grandes por lo que el modelo se sobreajusta. Es por esto que se realizó una reducción de dimensión en los panoramas, se consideraron las primeras 20 componentes principales y fueron usadas como las nuevas características. Al usar estas características junto con los datos médicos se obtuvo una clasificación correcta del 91.8 %, al igual que con el método de Ignacio. Los resultados de estas clasificaciones se pueden encontrar en la Tabla 5.1.

Estadístico o método	Clasificación
Intervalo RR	90.8 %
Ignacio	61 %
Chazal	72 %
Intervalo RR y Ignacio	91.8 %
Intervalo RR y Chazal	91.8 %

CUADRO 5.1. Resultados obtenidos de las clasificaciones en los datos reales.

Capítulo 6

Conclusiones y comentarios finales

6.1. Conclusiones

Los métodos de clasificación basados en ATD, por si solos, no son tan efectivos como los métodos de clasificación médicos, pero al no depender de la interpretación médica tienen un alto potencial para ser estudiados en otro tipo de señales.

Al combinar los estadísticos topológicos con los médicos fue posible mejorar la clasificación de ambos. Esta mejora no fue tan grande, pero abre las puertas a futuras investigaciones con más datos para probar si este incremento es significativo. Además es posible que la clasificación topológica admita mejoras adicionales, usando algún método de clasificiación basado en combinar otros clasificadores o explorando otros métodos, como redes neuronales.

Para el experimento que se hizo con los datos simulados se encontró que la clasificación usando estadísticos topológicos es buena, pero al momento de agregar ruido a los datos se tiene un error de clasificación en el conjunto de prueba, muy parecido al de los datos reales. Debido a esto y al hecho de que al sumar ruido a los datos simulados estos se parecen a los datos reales, se piensa que los métodos basados es topología son muy sensibles al ruido, por lo que se plantea estudiar alguna forma de reducir el ruido o de reconstruir los ECG estimando los parámetros de las ondas para aplicar los métodos sobre series de tiempo menos ruidosas.

El método de la ventana deslizante se aplica a series de tiempo de una única señal, pero existen muchos tipos de sistemas descritos por multiples series de tiempo simultaneas. Por ejemplo, los ECG de 12 derivaciones son 12 series de tiempo que transcurren simultaneamente. Si se puede encontrar una generalización del teorema de Takens para reconstruir el sistema asociado a múltiples señales, puede ser posible aplicar una versión generalizada de la ventana deslizante para estudiar este tipo de datos. El concurso de Physionet de 2020 consiste en clasificar ECG de 12 derivaciones por lo que existe un posible campo de estudio con datos disponibles para trabajar.

6.2. Comentarios finales

Los métodos de Ignacio y de Chazal están basados en el mismo principio obtener la información topológica de la nube de puntos obtenida con el método de la ventana deslizante, la única diferencia son las características usadas para clasificar. El método de Chazal es más tardado pues el cálculo de los panoramas requiere adicional, pero por si sólo supera en precisión al de Ignacio. Por otro lado, el método de Paul es más rápido aunque no es tan preciso por si sólo.

A lo largo de la tesis el tiempo computacional fue un problema que se encontraba con frecuencia, es por eso que se fueron tomando diversas medidas que reducen el tiempo de cómputo. Los tres principales son: la reducción de la frecuencia en los ECG originales, trabajar sólo con una fracción del ECG que tuviera buena calidad de señal y el uso de PCA sobre la nube de puntos obtenida con la ventana deslizante.

Hay muchos pasos que pueden ser modificados en todo el camino desde el momento en que tenemos los datos crudos hasta el momento en que clasificamos. Por ejemplo, el método de la ventana deslizante puede ser modificado por otra forma de asociar una serie de tiempo a un espacio topológico. En lugar de complejos de Vietoris-Rips se pueden usar complejos alpha. Existen más herramientas en el análisis topológico de datos, además de la homología persistente, por lo que se pueden obtener otros estadísticos para usarlos en la clasificación.

Bibliografía

[1] Chazal F. and Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Arxiv*, page 38, 10 2017.

[2] de Silva V., Skraba P., and Vejdemo-Johansson M. Topological analysis of recurrent systems. *Workshop on Algebraic Topology and Machine Learning*, 2012.

[3] Perea J. and Harer J. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838, 2015.

[4] Af classification from a short single lead ecg recording. *The PhysioNet Computing in Cardiology Challenge 2017*, Consultado: 17-feb-19. https://physionet.org/content/challenge-2017/1.0.0/.

[5] Ignacio P., Uminsky D., Dunstan C., Escobar E., and Trujillo L. Classification of single-lead electrocardiograms: Tda informed machine learning. *arXiv:1911.12253*, 2019.

[6] Kubícek J., Penhaker M., and Kahankova R. Design of a synthetic ecg signal based on the fourier series. *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1881–1885, 2014.

[7] Ramsay J. and Silverman B. Functional data analysis. *Springer*, 2005.

[8] Shumway R. and Stoffer D. Time series analysis and its applications (springer texts in statistics). *Springer-Verlag*, 2005.

[9] Tibshirani R., James G., Witten D., and Hastie T. An introduction to statistical learning: with applications in r. *Springer*, 2013.

[10] Xiong Z., Stiles M., and Zhao J. Robust ecg signal classification for detection of atrial fibrillation using a novel neural network. *2017 Computing in Cardiology (CinC)*, pages 1–4, 2017.

[11] Yazdani S., Laub P., Luca A., and Vesin J. Heart rhythm classification using short-term ecg atrial and ventricular activity analysis. *2017 Computing in Cardiology (CinC)*, pages 1–4, 2017.

[12] Smolen D. Atrial fibrillation detection using boosting and stacking ensemble. *2017 Computing in Cardiology (CinC)*, pages 1–4, 2017.

[13] Edelsbrunner H., Letscher D., and Zomorodian A. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.

[14] Zomorodian A. and Carlsson G. Computing persistent homology. *Discrete and Computational Geometry*, 33:249–274, 02 2005.

[15] Otter N., Porter M., Tillmann U., Grindrod P., and Harrington H. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6, 06 2015.

[16] Bubenik P. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1), 2015.

[17] *Electrocardiography*, Consultado 15 nov 2020. https://en.wikipedia.org/wiki/Electrocardiography.

[18] Heuzey L. and Paziaud J. Cost of care distribution in atrial fibrillation patients: the cocaf study. *Am Heart J*, 147:121-6:1881–1885, 2004.

[19] Takens F. Detecting strange attractors in turbulence. *Springer Berlin Heidelberg*, 898:366–381, 1981.

[20] Stam C.J. Nonlinear dynamical analysis of eeg and meg: Review of an emerging field. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 116:2266–301, 11 2005.
[21] Hundewale N. The application of methods of nonlinear dynamics for ecg in normal sinus rhythm. *International Journal of Computer Science Issues*, 9, 01 2012.
[22] Murasugi K. Knot theory and its applications. 01 2008.
[23] Kim K., Kim J., and Rinaldo A. Time series featurization via topological data analysis: an application to cryptocurrency trend forecasting. *arXiv:1812.02987*, 2018.
[24] Chazal F., Fasy B., Lecci F., Rinaldo A., and Wasserman L. Stochastic convergence of persistence landscapes and silhouettes. 2013. https://arxiv.org/abs/1312.0308.
[25] Ghrist R. Elementary applied topology. *University of Pennsylvania*, 2010.
[26] Edelsbrunner H. and Harer J. Computational topology: An introduction. page 282, 01 2010.
[27] Aurenhammer F., Klein R., and Lee D. Voronoi diagrams and delaunay triangulations. page 348, 2013.
[28] Hernández Y. Simulation of point cloud data with various probability distributions on stratified spaces. *Tesis licenciatura DEMAT-CIMAT*, Marzo 2019.