UNIVERSIDAD
COMPLUTENSE
MADRID

Trabajo Fin de Master
Técnicas matemáticas para datos médicos

Alumna: Alicia Torres García

Directora: Ana Maria Carpio Rodriguez

Máster Ingeniería Matemática

2020-2021

Madrid, 1 de julio 2021

Resumen

Los profesionales médicos se enfrentan habitualmente a problemas complejos y difíciles de tratar, como puede ser el diagnóstico de un paciente. En este trabajo se presentan varios recursos basados en la teoría matemática, que podrían ser útiles para identificar patrones y anticiparse a las necesidades de los pacientes.
Los datos médicos suelen aparecer en forma de matrices numéricas o secuencias temporales. Desarrollamos herramientas matemáticas para el cribado automático de dichos datos en dos contextos médicos: el diagnóstico de pacientes con un cuadro clínico de pruebas de laboratorio y la identificación de anomalías cardíacas.
La idea es primero implementar una normalización adecuada de los datos y, a continuación, utilizar las distancias adecuadas para identificar y clasificar los patrones relevantes.
Mientras que las distancias de Hamming están bien adaptadas para el estudio de patrones en matrices numéricas normalizadas que representan datos de pruebas de laboratorio, las distancias de *Time Warping* proporcionan herramientas robustas para el estudio de las señales cardíacas. Las técnicas aquí desarrolladas pueden sentar las bases para el cribado automático de información médica basada en la comparación de patrones.

Palabras clave

Distancia Hamming, distancia del movimiento de la tierra, distancia de la deformación dinámica del tiempo, modelo bayesiano, rankings, probabilidad, patrón, diagnóstico médico, electrocadiograma.

Abstract

Medical professionals are often faced with complex and difficult problems to deal with, such as the diagnosis of a patient. This paper presents several resources based on mathematical theory, which could be useful to identify patterns and anticipate patients' needs.
Medical data are usually in the form of numerical matrices or time sequences. time sequences. We develop mathematical tools for the automatic screening of such data in two medical contexts: the diagnosis of patients with a clinical laboratory data in two medical contexts: the diagnosis of patients with a clinical picture from laboratory tests and the identification of cardiac anomalies.
The idea is to first implement a proper normalisation of the data and then use appropriate distances to identify and classify the relevant patterns. and classify relevant patterns.
While Hamming distances are well suited for the study of patterns in normalised numerical arrays representing laboratory test data, Time Warping distances provide robust tools for the study of cardiac signals.
The techniques developed here can lay the foundation for automatic screening of medical information based on pattern matching.

Keywords

Hamming distance, Earth Mover's distance, Time warping distance, Bayesian model, rankings, likelihood, pattern, medical diagnosis, electrocardiogram.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este primer capítulo se establece un marco contextual y las motivaciones que han llevado a realizar el presenta trabajo, a continuación se enumeran los objetivos principales del mismo y se explica brevemente los capítulos de la memoria.

1.1. Motivación y contexto

El avance de las tecnologías no solo está revolucionando los aspectos de negocio, sino también los relacionados con la vida cotidiana. Actualmente, se están explorando las aplicaciones de herramientas de *machine learning* o tecnologías de *deep learning* en algunas tareas relacionadas con la medicina.
La Ciencia de datos podría ayudar a los médicos a extraer información clínicamente relevante de conjuntos de datos grandes. Además, los profesionales de la asistencia sanitaria podrían abordar problemas complejos que, de otro modo, resultarían difíciles de tratar y consumirían mucho tiempo. La Inteligencia Artificial (IA) podría ser un recurso valioso para los profesionales médicos, que les permitiría utilizar mejor su experiencia y proporcionar valor en todo el mundo sanitario. Podría ser útil para identificar y anticiparse a las necesidades de los pacientes en los centros sanitarios o los laboratorios clínicos. Podría identificar patrones y ayudar a los investigadores a crear grupos dinámicos de pacientes para estudios y ensayos clínicos.
Algunos de las múltiples líneas de trabajo existentes son algoritmos de clasificación de imágenes, modelos basados en datos para diagnosticar ritmos cardíacos irregulares a partir de electrocardiogramas (ECG), detección de anomalías en un paciente, identificación de patrones en los datos y detección de enfermedades a partir de ello, modelos de predicción, ...
En gran medida, estos modelos están basados en una teoría estadística y matématica muy robusta, lo que refleja la importancia de las técnicas matemáticas para el campo de la medicina.

En este trabajo se intenta abordar algunas líneas de trabajo mencionadas anteriormente, por un lado se expone un algoritmo para detección de patrones en las variables clínicas de un paciente, además se incluye un procedimiento bayesiano para comparar el rendimiento de 3 algortimos distintos. Por otro lado se desarrolla un modelo para determinar la clasificación del ritmo cardiaco a partir de electrocardiogramas (ECG).

1.2. Objetivos

Los objetivos abarcados en el presente Trabajo de Fin de Máster son los siguientes:

1. Desarrollo de métodos matemáticos y computacionales para identificar patrones en matrices de datos.

2. Desarrollo de métodos matemáticos y computacionales para clasificar gráficas representando la evolución en tiempo de una variable monitorizada.

1.3. Estructura del documento

A fin de organizar de una manera correcta la memoria, se procede a describir la estructura que presenta este documento.

- En el presente capítulo 1 se contextualiza el problema, se enumeran los objetivos y se describe la estructura de la memoria.

- En el capítulo 2 se hace un estudio de cómo identificar un perfil a partir de variables médicas numéricas y cómo identificar periodos anómalos a partir de ellas, además se incluye un análisis bayesiano basado en el modelo de Plackett-Luce sobre rankings de algoritmos, permitiendo la comparación del rendimiento de varios algoritmos al mismo tiempo.

- En el capítulo 3 se hace un estudio de cómo identificar un perfil a partir de cuadros clínicos gráficos, como los electrocardiogramas.

- Para finalizar, en el capítulo 4 se exponen las conclusiones extraídas a partir de la elaboración del trabajo.

Capítulo 2

Patrones en matrices de datos

Una de las tareas más difíciles de los profesionales de la salud es determinar el diágnóstico médico de un paciente.
El diagnóstico clínico es el medio por el cual se identifica una enfermedad o el estado del paciente, en función de síntomas, signos, la historia clínica y exploraciones complementarias, que permiten definir el cuadro clínico del paciente.

En la búsqueda del diagnósitco nos podemos auxiliar de distintos procedimientos, pero nosotros nos centraremos en el diagnóstico por comparación de patrones.

El diagnóstico por comparación, consiste en construir un síndrome a partir de síntomas y signos del paciente con posibles enfermedades que el paciente pueda presentar. Se realiza el diagnóstico comparando el cuadro clínico que presenta el paciente con el de las distintas enfermedades propuestas. El diagnóstico se realiza por mayor semejanza del cuadro clínico del enfermo con el descrito para cada enfermedad.
Se debe tener en cuenta que, aunque un paciente tenga una enfermedad, no precisa tener todos los síntomas y signos de la misma, y que puede padecer síntomas y signos similares a los de otra enfermedad.

El cuadro clínico para cada enfermedad lo vamos a determinar como patrón de dicha enfermedad. Consideraremos una colección P de K *patrones*. Cada patrón P_k se refiere a un subconjunto de i_n variables médicas. Los distintos patrones P_k se van a comparar con el cuadro clínico del paciente i, T_i, diagnosticando con el que tenga una menor distancia, es decir

$$D_i = min\left\{d\left(T_i, P_1\right), \ldots, d\left(T_i, P_k\right)\right\}$$

donde D_i nos indica el diagnóstico del paciente i, y d es la función distancia.
La función distancia d, es una función que define una distancia entre cada par de elementos de un conjunto. Recordemos la definición de una función distancia.
Una métrica sobre un conjunto X es una función

$$d : X \times X \to [0, \infty)$$

tal que, para cualesquiera que sean $x, y, z \in X$ se satisfacen las siguientes propiedades:

1. $d(x, y) \geq 0$

2. $d(x, y) = d(y, x)$ (Simetría)

3. $d(x, y) = 0 \Leftrightarrow x = y$ (Axioma de coincidencia)

4. $d(x, z) \leq d(x, y) + d(y, z)$ (Desigualdad triangular)

La función distancia, d, va a estar definida en función de cómo consideremos el formato de los patrones. Se van a distinguir dos formas de definir los patrones:

- Patrones como magnitudes numéricas:

  En este caso, los patrones P_k van a presentar los valores numéricos de las variables médicas. Algunas medidas que se pueden introducir en estos patrones pueden ser *la Glucosa, el Colesterol, los Triglicéridos, el Calcio, el Hierro, la VitaminaB12, las Proteinas, ...*
  Para este tipo de patrones, vamos a considerar como función distancia, d, tres funciones distintas:

  - Distancia Hamming con un alfabeto de 3 letras
  - Distancia euclídea
  - Distancia EMD *(Eart Mover's distance)*

- Patrones como magnitudes gráficas:

  En este otro caso, los patrones P_k se presentan como imágenes, como es el caso de *las radiografías, rayos X, ecografías, resonancias magnéticas, tomografías, electrocardiogramas, escanografía nuclear, ...*
  Para estos patrones , la función distancia d que vamos a considerar es:

  - *Time warping distance*
  - Distancia EMD *(Eart Mover's distance)*
  - Wasserstein - 1

En este Capítulo nos centraremos en el estudio de patrones numéricos, dejando el estudio de patrones gráficos para el Capítulo 3.

2.1. Patrones como magnitudes numéricas

Consideramos una matriz $\mathbf{M}$ que recoge los resultados de las pruebas de laboratorio de un paciente durante una serie de días, con magnitudes numéricas, como pueden ser el valor de *la Glucosa o la Bilirrubina*. La matriz $\mathbf{M}$ está formada por N_v filas que contienen los valores de un conjunto de variables (medidas médicas que se han medido en el paciente) y N_c columnas correspondientes a los diferentes tiempos de observación.
Para extraer información de los datos del laboratorio es conveniente normalizarlos los mismos. Cada variable tiene unos rangos de normalidad de referencia, inferior y superior, RI y RS. Comparando con estos valores de referencia, se actualizan los patrones y las tablas de medidas, de manera que ahora tendrán los siguientes valores:

- 'Alto', por encima del valor máximo de referencia, RS.

- 'Normal', entre los dos valores de referencia.
- 'Bajo', por debajo del valor mínimo de referencia, RI.

Para hacer aún mas manejables los datos se utilizará la siguiente notación:

$$\hat{M}_{j,i} = \begin{cases} 1 & \text{si} \quad M_{j,i} \geq RS & \text{(Alto)} \\ 0 & \text{si} \quad RI < M_{j,i} < RS & \text{(Normal)} \\ -1 & \text{si} \quad M_{j,i} \leq RI & \text{(Bajo)} \end{cases}$$

Es decir, se construye una matriz normalizada $\hat{M}_{j,i}$ reemplanzdo $M_{j,i}$ con 1,0 o -1 dependiendo de los valores originales están por encima, entre o debajo de los rangos de normalidad.
Debemos tener en cuenta que a veces se miden variables binarias, cómo puede ser ¿ha sentido mareos?, en el caso de estas variables, la información almacenada en la matriz original M es 'positivo/verdadero/sí' o 'negativo/falso/no', que se reemplaza en la matriz normalizada $\hat{M}$ por 1 en caso de ser positivo y con un 0 en caso contrario.

Tal como se dijo anteriormente, vamos a considerar como funciones distancia, d, tres distancias distintas. Posteriormente, se realizará un *Test de Bayes* basado en el modelo de Placett-Luce para determinar cual de ellas es la mejor para nuestro caso.

2.2. Distancias

Dadas las cohortes de pacientes que muestran síntomas similares, podemos agruparlos por similitud en los resultados de las pruebas de laboratorio utilizando las distancias adecuadas.
Si deseamos utilizar estas distancias de conjuntos de datos específicos para un determinado objetivo, podemos comparar su rendimiento, clasificarlas y realizar un análisis bayesiano para seleccionar la mejor como en [1] para algoritmos de clustering.

2.2.1. Distancia de Hamming

La distancia dependiente del número de diferencias es la conocida como la distancia de Hamming. La distancia de Hamming ([2] y [3]) entre dos cadenas de igual longitud es el número de posiciones en las que los símbolos correspondientes son diferentes, es decir, describe el número de cambios en los simbolos de una cadena necesarios para reducirla a la otra. Sea A un alfabeto de símbolos, sea $C \subset A^n$, el conjunto de cadenas de longitud n sobre A. Sean $u = (u_1, \ldots, u_n)$ y $v = (v_1, \ldots, v_n)$ palabras de C. La *distancia de Hamming* $d_H(u, v)$ es definida como el número de lugares en los que u y v difieren, esto es,

$$d_H(u, v) = \#\left\{i : u_i \neq v_i, i = 1, \ldots, n\right\}.$$

La distancia de Hamming es una métrica de C, por tanto verifica las propiedades de la misma.

Por ejemplo, supongamos que se tiene estos 5 patrones:

$$\begin{array}{l} P_1 = [\ 1,\ 0,\ 0,\ 0,\ -1,\ 0] \\ P_2 = [\ 0,\ -1,\ 0,\ 0,\ 1,\ 0] \\ P_3 = [\ 0,\ 0,\ 1,\ 0,\ 0,\ 1] \\ P_4 = [\ 1,\ 1,\ -1,\ 0,\ -1,\ 0] \\ P_5 = [\ 1,\ 0,\ 0,\ -1,\ -1,\ -1] \end{array}$$

y supongamos que el cuadro clínico para un paciente, es el siguiente:

$$T_1 = [\ 1,\ 0,\ 0,\ 1,\ 0,\ -1]$$

Nuestra función distancia, d, va venir determinada por el número de diferencias. Comparamos cada paciente con cada uno de los patrones y tomamos como distancia el número de diferencias.

$$\begin{array}{lll} d_H(T_1, P_1) = & d([1,0,0,1,0,-1] & ,[1,0,0,0,-1,0]) & = 3 \\ d_H(T_1, P_2) = & d([1,0,0,1,0,-1] & ,[0,-1,0,0,1,0]) & = 5 \\ d_H(T_1, P_3) = & d([1,0,0,1,0,-1] & ,[0,0,1,0,0,1]) & = 4 \\ d_H(T_1, P_4) = & d([1,0,0,1,0,-1] & ,[1,1,-1,0,-1,0]) & = 5 \\ d_H(T_1, P_5) = & d([1,0,0,1,0,-1] & ,[1,0,0,-1,-1,-1]) & = 2 \end{array}$$

Obtenemos así un vector de distancias, para el paciente 1, $d = [3, 5, 4, 5, 2]$. Tenemos que la distancia mínima se obtiene cuando el paciente se compara con el patrón 5, por lo que el diagnóstico final para este paciente será el que indique el patrón 5.

2.2.2. Distancia Euclídea

En matemáticas, la distancia euclidiana o euclídea, es la distancia 'ordinaria' entre dos puntos de un espacio euclídeo. En general, la distancia euclidiana entre dos puntos $P = (p_1, p_2, ..., p_n)$ y $Q = (q_1, q_2, ..., q_n)$, se define como:

$$d_E(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

2.2.3. Distancia EMD *(Earth Mover's distance)*

En estadística, la distancia del movimiento de la tierra (EMD), ([4] y [5]), es una medida de la distancia entre dos distribuciones de probabilidad sobre una región D. En matemáticas esto es conocido como la distancia de *Wasserstein*. De manera coloquial, si interpretásemos las distribuciones como dos formas diferentes de aglomerar una cierta cantidad de tierra sobre la región D, la EMD es el coste mínimo de convertir una pila en la otra, donde suponemos que el coste esta definido como la cantidad de tierra movida por la distancia en la que se mueve.

El cálculo de la EMD se puede formular y resolver como un problema de transporte. Supongamos que se tienen varios proveedores para abastecer varios consumidores. Asumimos P como los proveedores, el cuál posee m grupos, es decir $P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), \ldots, (p_m, w_{p_m})\}$, donde p_i es el clúster representativo y $w_{p_i} > 0$ es el peso del clúster p_i. Del mismo modo, se considera

ahora $Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), \ldots, (q_n, w_{q_n})\}$. Sea $D = [d_{i,j}]$ la distancia entre p_i y q_j. Queremos encontrar el flujo $F = [f_{i,j}]$, donde $f_{i,j}$ es el flujo entre p_i y q_j, que minimice el coste total. El flujo se obtiene resolviendo el siguiente problema de programación lineal:

$$
\begin{array}{ll}
\text{min} & \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j} \\
\text{s.a.} & \sum_{j=1}^{n} f_{i,j} \leq w_{pi} \quad 1 \leq i \leq m \\
& \sum_{i=1}^{m} f_{i,j} \leq w_{qj} \quad 1 \leq j \leq n \\
& \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} = min \left\{ \sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj} \right\} \\
& f_{ij} \geq 0
\end{array}
$$

La distancia del movimiento de tierra se define como el trabajo normalizado por el flujo total, es decir:

$$d_{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}}$$

2.3. Selección Bayesiana de distancias basado en el modelo de Plackett-Luce

Uno de los principales objetivos de la optimización heurística es la comparación del rendimiento de algortimos en caso de la resolución de un problema concreto. Podemos enfocar este objetivo desde el punto de vista de la clasificación de los algoritmos cuando se aplican a un problema.
Para ello se va a proponer un enfoque Bayesiano, basado en el modelo de Plackett-Luce, permitiéndonos comparar un conjunto de algoritmo simultáneamente. A partir de este modelo se pueden sacar conclusiones eficaces e interpretables sobre la compración del rendimiento de los distintos algoritmos.
Se van a considerar tres algoritmos distintos, considerando para cada uno de ellos cada una de las tres distancias propuestas anteriormente compararando su rendimiento y realizando un análisis bayesiano para seleccionar la mejor de ellas como se hace en [1, 6] para la selección de algoritmos de clustering.

Se implementará una estrategia Bayesiana basada en el modelo de Plackett-Luce sobre rankings, para evaluar qué distancias funcionan mejor, a la hora de comparar los pacientes con los patrones considerados y decidir cuál se parece más diagnósticando asi el paciente.
La principal ventaja del modelo PL es que los parámetros normalizados del modelo representan directamente la probabilidad marginal de un algoritmo de ser posicionado en primera posición.

Para aplicar la estrategia, primeramente se necesita crear una matriz de ranking R. Un ranking $\sigma = (\sigma_1, \ldots, \sigma_n)$ es una permutacion de las distancias selecionadas que nos indican el orden de cada registro. Se debe tener en cuenta que se pueden repetir los órdenes de cada registro y desencadenar empates en los rankings, por lo que se van a estudiar dos casos:

- La matriz R no admite empates
- La matriz R admite empates

2.3.1. Modelo Bayesiano de Plackett-Luce

El modelo de Plackett-Luce (PL) es un proceso secuencial en el que el nuevo algoritmo a clasificar se selecciona aleatoriamente entre los restantes.
Denotando por $\sigma = (\sigma_1, \ldots, \sigma_n)$ un ranking de tamaño n, donde $\sigma_i = j$ conlleva que el algoritmo j-ésimo es colocado en la posición i-ésima, y el vector $w = (w_1, \ldots, w_n)$ representa los pesos asociados a los n elementos, entonces la probabilidad del modelo de PL de seleccionar un raking viene dada por:

$$P_{PL}(\sigma) = \prod_{i=1}^{n} \frac{w_{\sigma_i}}{\sum_{j=i}^{n} w_{\sigma_j}} \tag{2.1}$$

Para simplificar, se va a suponer que la suma de los pesos, w, es 1.

El modelo bayesiano nos proporciona una idea de la incertidumbre sobre el modelo tras integrar los datos generados en la comparación.
Los pesos w, se van hallar mediante el cálculo de las probabilidad a posterior $p(w|R)$ de los pesos dado el ranking. Mediante la regla de Bayes se obtiene:

$$P(w|R) \propto P(w)P(R|w) \tag{2.2}$$

donde R representa los datos (rankings), es decir el conjunto de rankings y w los parámetros del modelo Plackett-Luce.

En cualquier procedimiento Bayesiano uno de los puntos críticos es la definición de la distribución a priori de los parámetros del modelo, en nuestro caso se va a considerar cada σ_i como una variable aleatoria multinomial, por lo que se puede considerar el vector de pesos como los parámetros de esa distribución multinomial.

La forma tradicional de representar la incertidumbre sobre los parámetros de una distribución multinomial es emplear la distribución de Dirichlet.
Utilizando el modelo PL como distribución de muestreo y modelando la incertidumbre sobre los pesos del modelo PL mediante una distribución de Dirichlet, se tiene que:

$$P(w/R) \propto Dir(w, \alpha) \prod_{\sigma \in R} P_{PL}(\sigma; w) \tag{2.3}$$

donde α son los hipérparametros de la distribución a priori. Los métodos de Markov Chain Monte Carlo (MCMC) se pueden utilizar para muestrear la distribución a posteriori y analizar la distribución de los pesos muestreados.

Se va tener

$$P(R/w) = \prod_{\sigma \in R} P_{PL}(\sigma; w)$$

y

$$P(w) = Dir(w, \alpha)$$

La funcion de distribucion de Dirichlet $Dir(w, \alpha)$ es una familia de distribuciones multivariantes parametrizadas por un vector α de números reales positivos. La distribición de Dirichlet de orden $N \geq 2$ con parámetros $\alpha = (\alpha_1, \ldots, \alpha_N)$ tiene la siguiente función de densidad

$$f(w_1, \ldots, w_N; \alpha_1, \ldots, \alpha_N) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{N} w_i^{\alpha_i - 1}, \quad \beta(\alpha) = \frac{\prod_{i=1}^{N} \Gamma(\alpha_i)}{\Gamma(\prod_{i=1}^{N} \alpha_i)},$$

donde $\sum_{i=1}^{N} w_i = 1,\ w_i \geq 0,\ \forall i \in [1, N]$ y la funcion Γ es la función gamma, que esta definida por:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

Dado que se carece de información sobre el rendimiento del algoritmo para un conjunto de datos generales, se va a utilizar una distribución uniforme para el hiperparámetro α, dada por, $\alpha_i = \alpha = 1,\ \forall i \in \{1, \ldots, N\}$
Cuando tenemos muchas muestras, seleccionamos la que tiene la máxima probabilidad. Esos son los pesos más probables, con la interpretación de que la probabilidad de que cada distancia sea la mejor para el tipo de datos considerados. Para cuantificar la incertidumbre, trazamos diagramas de caja con las muestras.

La explicación anterior del cálculo de probabilidades a través de 2.1 lo vamos a asociar a una matriz **R sin empates**, por lo que los empates se van a eliminar de manera aleatoria.

En el otro caso, en el que la matriz **R admita empates** se va desarrollar una extensión del modelo anterior. En este caso la matriz de rankings, R, si que va admitir empates, en este caso, se asigna la misma posición a cada algoritmo, dejando la siguiente posición libre.

Se va a explicar el método de Plackett Luce admitiendo empates, [7], de manera que la matriz R se va a representar de la forma

$$R = (C_1, \ldots, C_J)$$

donde los elementos C_1 se clasifican por encima que los elementos de C_2 y así respectivamente. Si se tienen varios objetos en el conjunto C_j empatados en la clasificación se calcula para un conjunto S

$$f(S) = \delta_{|S|} (\prod_{i \in S} \alpha_i)^{\frac{1}{|S|}}$$

donde $|S|$ es la cardinal del conjunto S, δ_n es un parámetros que representa la prevalencia de los empates de orden n ($\delta_1 = 1$) y α_i es el parámetro que representa el valor de i.
Por tanto, la probabilidad del ranking R, permitiendo empates de orden D, viene dado por:

$$\prod_{j=1}^{J} \frac{f(C_j)}{\sum_{k=1}^{min(D_j, D)} \sum_{S \in \binom{A_j}{k}} f(S)} \tag{2.4}$$

donde D_j es el cardinal de A_j, A_j es el conjunto de alternativas entre las que se elige C_j, $\binom{A_j}{k}$ son todas la posiblidades de elegir valores de k sobre A_j. El valor de D se puede interpretar como el número máximo de empates observados en los datos, es decis $\delta_n = 0 \ \ \forall \ n > D$.
Cuando los parámetros α_i, cumplen $\sum_{i=1}^{n} \alpha_i = 1$ representan la probabilidad de que el elemento corresponiente ocupe el primer lugar en la clasifcación de todos los elementos, cuando la primera posición no está empatada.
Los parámetros de prevalencia de empates δ son interpretables a través de los elementos empatados, por ejemplo, δ_2 es interpretable a través de la probabilidad de que dos elementos dados empaten en primer lugar. Especificamente, esa probabilidad es $\delta_2/(2 + \delta_2)$.
Cuando no se observan empates de ordenes intermedios, la estimación de máxima verosimilitud de los parámetros de prevalencia de empate correspondientes es cero, por lo que estos parámetros se excluyen del modelo.
Anteriormente se han hallado los parámetros α_i que nos indican la probabilidad de que el elemento i quede en primera posición en caso de que no haya empates. Para calcular la probabilidad de que un elemento i se encuentre en primera posición considerando empates, se va a utilizar la fórmula 2.4.
La idea general es calcular la probabilidad de cada una de las distintas posibilidades S_j, $P(S_j)$ a tráves de 2.4 y posteriormente calcular la probabilidad de que i esté en primera posición de manera que:

$$P_i = \sum_{j \ / \ S_j(i)=1}^{J} P(S_j)$$

es decir, para hallar la probabilidad de que i se encuentre en primera posición habrá que sumar todas las posibilidades en las que i se encuentre en primera posición, incluyendo los empates.

2.4. Aplicación

En este apartado buscamos diagnosticar a un paciente con un cuadro clinico de medidas numéricas, comparándolo con patrones numéricos, de la manera más precisa posible.
Para ello, vamos a determinar la capacidad de cada funcion distancia presentada anteriormente para el diagnóstico del paciente. Con este fin se va a seguir el siguiente esquema:

- Generación de Rankings
- Aplicación del modelo de Plackett-Luce para obtener la distancia con mejor rendimiento
- Diagnóstico del paciente a partir de la distancia obtenida

Se van considerar una tabla de pacientes para la cual sabemos el diagnóstico. Se van aplicar las distintas distancias en consideración y se ve si aciertan o no. La mejor distancia será la que dé mas aciertos, cuantificándolas con probabilidades. Con esta distancia se va a diagnosticar a los pacientes que no sabemos su evaluación.

2.4.1. Generación de Rankings

En esta sección se va a presentar el procedimiento para la generación de los rankings a partir de la matriz de distancias. Las distancias se calculan por las 3 métricas definidas anteriormente,

entre los patrones y los registros.
Para este apartado se van considerar una tabla de pacientes para la cual sabemos el diagnóstico. Se van aplicar las distintas distancias en consideración, Hamming, euclidea y EMD, y se ve si aciertan o no.
Para entender el ejemplo, se va a realizar un ejemplo de juguete. Vamos a partir de una tabla con la información de los pacientes y con el diagnóstico que le da cada una de las distancias consideradas.

Paciente	Diangóstico verdadero	Diagnóstico por d_H	Diagnóstico por d_E	Diagnóstico por d_{EMD}
1	1	1	1	1
2	2	2	1	2
3	5	5	5	2
4	3	3	2	3
5	4	5	4	1

A partir de la tabla anterior, se va crear la matriz de manera que:

$$D_{ij} = \begin{cases} 1 & \text{si el paciente i es diagnosticado correctamente por la distancia j} \\ 0 & \text{en otro caso} \end{cases}$$

De manera que obtenemos lo siguiente:

Paciente	Hamming	Euclídea	EMD
1	1	1	1
2	1	0	1
3	1	1	0
4	1	0	1
5	0	1	0
Suma	4	3	3

sumando por columnas, sabemos cuál ha acertado más sobre este grupo de pacientes, y asi obtenemos la primera fila de nuestra matriz $Perfomance$, a partir de la cual vamos a crear la matriz R de los rankings.

Repitiendo el proceso anterior el número de iteraciones que se quiera, 10 en nuestro caso, vamos a obtener la siguiente matriz, $Perfomance$

Iteración	Hamming	Euclídea	EMD
1	4	3	3
2	5	2	1
3	4	2	3
4	4	4	3
5	3	4	3
6	3	3	3
7	5	3	2
8	4	1	3
9	4	5	2
10	5	3	3

En la anterior tabla se exponen los aciertos, entre la correcta clasficación del paciente en cada iteración.
A partir de esta tabla se va asignar una posición más alta en el ranking cuanto mayor sea el número de aciertos, entendiendo como posición más alta queda en primer lugar y el mayor número de aciertos 5. Además, en caso de empate se asigna la misma posición a cada algoritmo, dejando la siguiente posición libre, con la idea de que una vez se desempate uno de los algoritmos empatados tome esa posición. Obteniendo la matriz R de rankings

Nº	Hamming	Euclidea	EMD
1	1	2	2
2	1	2	3
3	1	3	2
4	1	1	3
5	2	1	2
6	1	1	1
7	1	2	3
8	1	3	2
9	2	1	3
10	1	2	2

Tabla 2.1: Rankings generados a partir de los aciertos

Examinando la tabla anterior se puede observar el rendimiendo de la distancia de Hamming. Utilizamos en la siguiente sección el modelo de Plackett-Luce cómo método de evaluación estadística permitiéndonos determinar cuál es la probabilidad de cada algoritmo de ser el mejor para el problema de diagnóstico de pacientes.

2.4.2. Aplicación del modelo de Plackett-Luce

En esta sección se emplea el modelo de Plackett-Luce para comparar la capacidad de diagnosticar de las distintas distancias presentadas. Distinguiendo los casos anteriormente estudiados.

R no admite empates

Concretamente, se va tomar como entrada la tabla 2.1 generada a partir de la aplicacion de las distintas distancias y se van a eliminar los empates de manera aleatoria. Generando una nueva matriz R

Nº	Hamming	Euclidea	EMD
1	1	2	3
2	1	2	3
3	1	3	2
4	1	2	3
5	3	1	2
6	2	3	1
7	1	2	3
8	1	3	2
9	2	1	3
10	1	3	2

Tabla 2.2: Rankings generados a partir de 2.1 eliminando los empates de forma aleatoria

Ahora partimos de la matrices de la forma de 2.2, donde ya se han eliminado los empates. Debido a la aleatoriedad introducida para desempatar en las posiciones del raking, se va ejecutar en bucle 1000 veces eliminando los primeros 100 registros, el modelo de Plackett-Luce (2.3) para calcular cuales son las probabilidades marginales medias de cada algoritmo de quedar en primera posición, consiguiendo asi disolcer el efecto de la aleatoriedad y conseguir una probabilidad marginal más exacta.
Los resultados obtenidos son:

Hamming	Euclidea	EMD
0.4293	0.3375	0.2332

Los resultados de la tabla anterior nos indican de que la mejor distancia es la de Hamming, con una probabilidad del 42.93 %. A continuación la distancia euclidea con una probabilidad del 33.75 % y finalmente la distancia EMD con una probabilidad del 23.32 %. Para nuestros fines, la distancia Hamming parece parece tener un mejor rendimiento. Una vez que hemos obtenido una muestra grande de pesos dibujamos un *boxplot*, figura 2.1, usando todas las muestras para tener una idea de la incertidumbre.

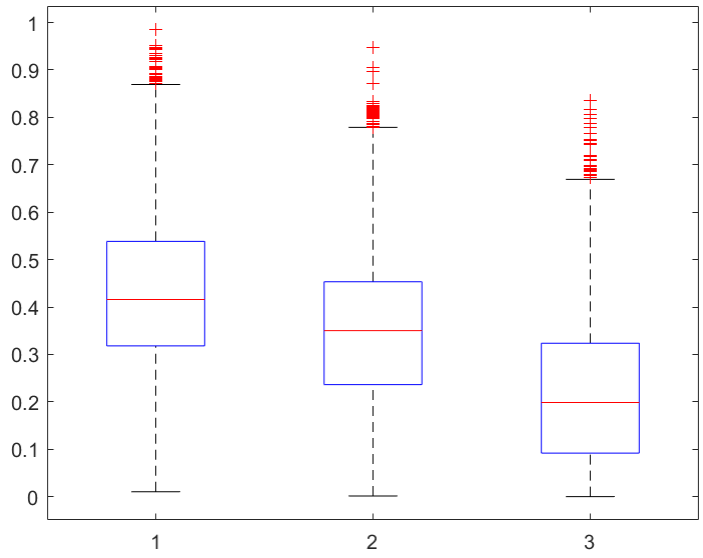


Figura 2.1: Representación de los diagramas de cajas de la interefencia Bayesiana que muestran la distribución posterior empírica de que un algoritmo sea el mejor clasificado.

R admite empates

En este caso se va a realizar una única iteracción, tomando como entrada la tabla 2.1, comenzamos calculando los α_i y δ_i a partir del modelo de Plackett-Luce (2.3) obteniendo:

$$\alpha_1 = 0.79739001$$

$$\alpha_2 = 0.13612835$$

$$\alpha_3 = 0.06648164$$

donde se puede comprobar que

$$\sum_{i_1}^{3} \alpha_i = 1$$

Los parámetros de prevalencia de empates δ toman los siguientes valores:

$$\delta_2 = 0.57440580, \qquad \delta_3 = 0.79125103$$

A partir de lo anterior se calculan las probabilidades ajustadas con 2.4 para todas las combinaciones de elección/alternativa en los datos. Obtenemos los siguientes casos:

- No hay empates en primer lugar

Descripción	Casos			Probabilidad
Probabilidad de que Hamming esté en primer lugar	1	2	3	0.52149419
	1	3	2	
	1	2	2	
Probabilidad de que Euclidea esté en primer lugar	2	1	3	0.08902813
	3	1	2	
	2	1	2	
Probabilidad de que EMD esté en primer lugar	2	3	1	0
	3	2	1	
	2	2	1	

- Empatan 2 en primer lugar

Descripción	Casos			Probabilidad
Hamming y Euclidea empatadas en primer lugar	1	1	3	0.12376766
Hamming y EMD empatadas en primer lugar	1	3	1	0
Euclidea y EMD empaten en primer lugar	3	1	1	0

- Empatan 3 en primer lugar

Descripción	Casos			Probabilidad
Hamming, Euclidea y EMD estén en primer lugar	1	1	1	0.1000000

Hecho lo anterior, la probabilidad de que el patron j de los J que se consideran sea el primero es la probabilidad de la unión de los eventos en que está primero, sea solo o empatado (o sea todos los rankings que has listado en los que salen primero solo o empatado). La probabilidad de una unión de eventos es la suma de las probabilidades de cada uno, menos sumas y mas sumas de probabilidades de intersecciones de 1,2,3,4,..., eventos. Si consideramos como evento base un ranking individual con todos sus grupos de empate y posiciones perfectamente definidos (sus elementos fijados), no va a haber intersecciones. Por tanto, las probabilidades que buscamos, parece que serían simplemente sumas de las probabilidades de los rankings de interés.

Es decir,

- Probabilidad de que distancia de Hamming se encuentre en primera posición

$$P_H = 0.52149419 + 0.12376766 + 0.10000000 = 0.7452619$$

- Probabilidad de que distancia Euclidea se encuentre en primera posición

$$P_E = 0.08902813 + 0.12376766 + 0.10000000 = 0.3127958$$

- Probabilidad de que distancia EMD se encuentre en primera posición

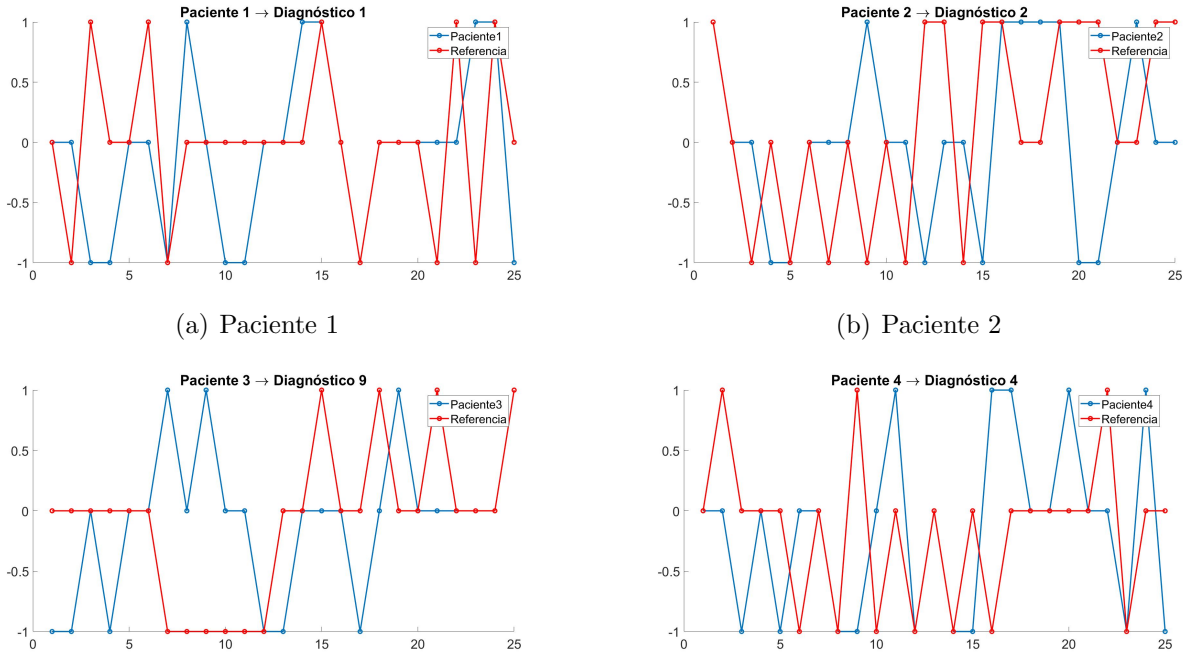
$$P_{EMD} = 0.1000000 = 0.10000000$$

De ambos estudios se puede sacar la misma conclusión, la distancia de *Hamming* es la mejor para nuestro conjunto de datos.

2.4.3. Diagnóstico de pacientes

Una vez que se ha obtenido que la mejor distancia para el diagnóstico de pacientes es la distancia de Hamming, se va aplicar dicha distancia para el diagnóstico de los pacientes de una base de datos sintética que se ha creado.
Debido a la confidencialidad de los datos sobre los que se ha trabajado, se van a generar los patrones y la base de datos de los pacientes.
Se van a generar M variables numéricas ya normalizadas, es decir vamos a tener una base de datos de los N pacientes, de $N \times M$, por lo que las columnas de dicha matriz $v_1, \ldots, v_M$ representan a las variables médicas simuladas, y $T_1, \ldots, T_N$ representa al paciente i.
Por otro lado, se va a generar la base de datos referente a los K patrones, de $K \times M$, las columnas de dicha matriz representan a las mismas variables médicas que las simuladas anteriormente, y $P_1, \ldots, P_K$ representa el patrón i.
En nuestro caso en concreto, se van a tener 100 variables médicas distintas, vamos a generar registros para 1000 pacientes y vamos a generar 10 patrones distintos.
Una muestra de las bases sintéticas creadas las podemos encontrar en el apéndice A.

Una vez generadas las bases de datos anteriores, se va a proceder al diagnóstico de pacientes por comparación utilizando la distancia de Hamming. Comparando cada paciente T_i con cada patrón P_k, diagnósticando con el que tenga una menor distancia.
Los resultados obtenidos se van a devolver en forma de gráfica, con el objetivo de hacerlo más visual, de manera que el valor de referencia es la enfermedad (patrón) con la que se ha diagnósticado al paciente y los valores azules no indican los valores de las medidas de cada paciente.
Una muestra de los diagnósticos obtenidos para los 4 primeros pacientes vienen representados en la figura 2.2.



(a) Paciente 1

(b) Paciente 2

(c) Paciente 3

(d) Paciente 4

Figura 2.2: Muestra del diagnóstico de cada paciente obtenido por distancia Hamming, mostrando en azul las medidas de las variables clínicas del pacientes y en rojo las variables médicas del patrón del referencia con el que se ha diagnósticado al paciente.

2.5. Patrones en días concretos: diagnóstico

Una vez que un paciente es diagnósticado, uno de los datos mas importantes de detectar es el día en el que comenzo el brote. Es decir, teniendo el historial del paciente, que contiene la evolución a lo largo del tiempo de las variables médicas que se hayan medido a dicho paciente. A priori no se tiene información acerca de día del brote, pero se van a usar herramientas gráficas para determinarlo.
Para la elección del día del brote, se va seguir el criterio de detección cómo el dia que tras el análisis de un mapa de calor, generado a partir de los datos históricos del paciente, se observe un desbande llamativo con respecto al resto.

Mediante la normalización presentada en la introducción del capítulo, visualizamos rápidamente qué variables suelen estar por encima o por debajo de su rango de normalidad. Además, cuando las variables salen de su rango de normalidad, puede indicar que el paciente empeore o responda al tratamiento. Sólo analizando la Fig 2.3, observamos que en el día 17 un gran número de variables toman valores anormales, lo que requiere una acción adicional para estabilizar al paciente.
Para el tratamiento de los valores faltantes, se va proponer imputarlos por el último dato no faltante, además eliminamos las filas o columnas que tengan más de un 50 % de datos *missing*.

Se va a proponer una función que automatice la detección del día anómalo.

Sea X_{ik} los datos de un paciente, de manera que $i \in \{1, \ldots, I\}$ son las variables médicas medidas en el paciente y $k \in \{1, \ldots, K\}$ nos indica el día, siendo K el número de días del histórico del paciente. Entonces se va a calcular el número de cambios en el dia k como:

$$nc_k = \sum_{i=1}^{N-1} |X_{ik} - X_{i+1,k}| \quad \forall k$$

Entonces, para la elección del día, nos vamos a quedar con el máximo del vector anterior, es decir,

$$dia = \text{máx}\,(nc_1, \ldots, nc_K)$$

De la manera anterior se obtiene de forma automática el día en el que se producen más cambios, pero se debe tener en cuenta que la medicina no es una ciencia exacta, y que muchas veces el médico trabaja de manera subjetiva. Este proceso nos ayuda a determinar el día, pero puede que haya veces que el críterio cambie, dependiendo de cada paciente.
Por ejemplo, se va considerar el *heatmap* de la figura 2.3, donde tenemos los datos normalizados, los recuadros negros nos indican los datos que faltan y las variables para las que no se tiene rango de normalidad.

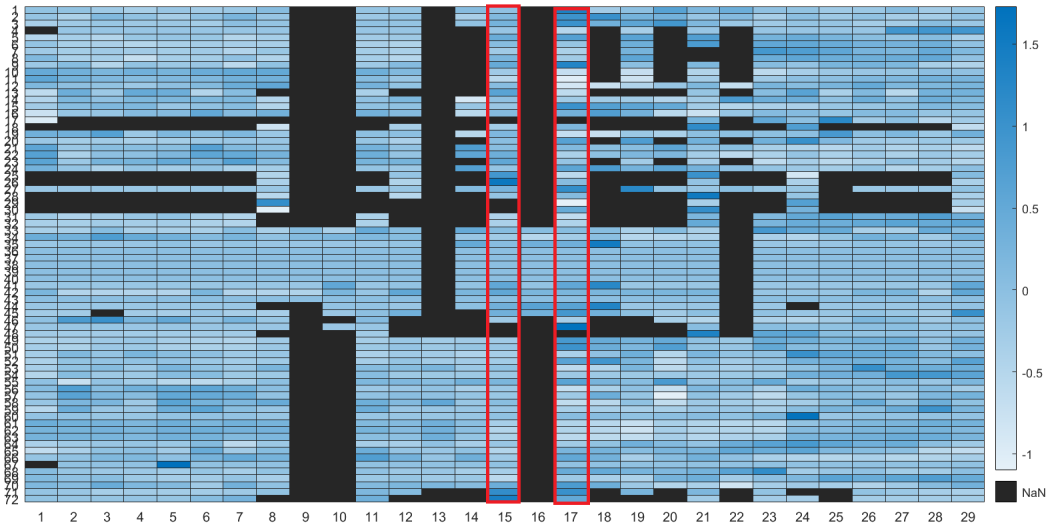


Figura 2.3: Mapa de calor de un paciente aleatorio

Utilizando el criterio automático anterior, se obtendría que el día anómalo es el 17, pero también se puede obervar que el día 15 es llamativo. En este caso, el dia 15 es anterior al 17 por lo que quizas en este caso nos beneficie elegir como día de brote el 15. De esta manera se puede observar como la elección de este dia es totalmente subjetivo y dependiente del criterio del profesional.
Se pueden definir listas de patrones -101, buscarlos todos los diferentes días, y mostrar qué patrones están presentes cada día. Para concluir qué es dominante y que podría estar empeorando el estado del paciente automáticamente, podríamos definir funciones de riesgo para los diferentes diagnósticos clave.

Capítulo 3

Patrones como magnitudes gráficas

El diagnóstico por imagen pretende obtener información sobre el estado de los órganos o partes del cuerpo de un paciente mediante imágenes, que luego son interpretadas por el personal médico con los conocimientos adecuados. El desarrollo de herramientas de detección automática puede ayudar a los médicos en esa tarea. En el caso más sencillo, las imágenes adoptan la forma de curvas que representan secuencias de datos registradas en momentos consecutivos. Pensemos, por ejemplo, en los electrocardiogramas (ECG).
Un electrocardiograma es un gráfico de la actividad eléctrica del corazón [8] que representa la tensión en función del tiempo, véase la Figura 3.1.
Se registran colocando electrodos en la piel, que detectan pequeños cambios eléctricos resultantes de la despolarización y repolarización del músculo cardíaco durante cada latido (ciclo cardíaco). Las alteraciones en el patrón del ECG indican anomalías cardíacas. Numerosos diagnósticos se basan en los patrones observados.

Nuestro objetivo en esta sección es introducir un procedimiento automático para identificar alteraciones en los patrones descritos por curvas unidimensionales. La idea es la siguiente, primero definimos un conjunto de curvas de referencia correspondientes a diferentes patologías conocidas. Dado otra curva, la comparamos con los patrones de referencia mediante una distancia adecuada. La distancia más pequeña seleccionaría un posible diagnóstico. Ilustraremos el proceso con ejemplos tomados de la electrocardiografía.

Para determinar el diagnósticos, se van a utilizar como función distancia, d, tres distancias distintas: la *time warping distance*, distancia EMD *(Eart Mover's distance)* y *Wassertein-1 distance*, una variante de la distancia EMD.

Desde el punto de vista medico, la posibilidad de disponer de herramientas que, a partir del ECG, puedan resaltar la actividad auricular y ventricular, son realmente interesantes y ademas hoy en día muy necesarias, ya que permiten el desarrollo de tecnicas de ayuda al diagnóstico clíncio como la detetección de otras patologías.

3.1. Estructura del electrocardiograma y alteraciones básicas

Un electrocardiograma (ECG) es un examen que registra la actividad eléctrica del corazón durante un periodo de tiempo. Esta actividad eléctrica se registra desde la superficie corporal del paciente, usando electrodos colocados sobre la piel en el tórax del paciente. Se dibuja en un papel mediante una representación gráfica o trazado, donde se observan diferentes ondas que representan los estímulos eléctricos de las aurículas y los ventrículos.

El ECG tiene una serie de deflexiones a las que se le asigna el nombre de las letras P,Q,R,S y T. A partir de las diferencias entre ellas, se describen las distintas enfermedades cardiovasculares, se distinguen el complejo PQRST (onda: P, Q, R, S y T) complejo QRS, intervalo P-Q y Q-T y por último el segmento ST.

Los electrocardiogramas repiten periódicamente la estructura representada en la figura 3.1.

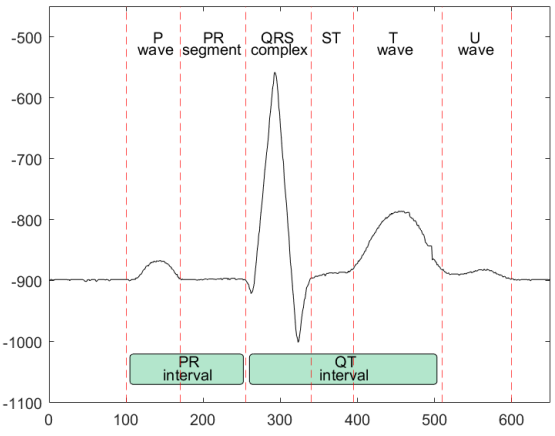


Figura 3.1: Representación esquemática de un ECG.

Las diferentes regiones están diseñadas por P, Q, R, S y T [9]:

- Onda P
  La onda P es la primera onda del ciclo cardiaco, representa la despolarización de las aurículas. La longitud habitual es inferior a 0.11 s en los adultos. La amplitud usual es inferior a 0.25 mV. Su forma es suave y redondeada.

- Complejo QRS
  Está formado por una secuencia de ondas que representan la despolarización de los ventrículos. La duración habitual es de unos $0,06$ s - $0,10$ s. Una pequeña onda negativa Q va seguida de una gran onda R positiva y una pequeña onda R negativa.

- Onda T
  La siguiente onda representa la repolarización ventricular. La longitud habitual es inferior

a 0, 20 s en los adultos. La amplitud habitual está entre 0, 2 y 0, 3 mV. Su forma es suave y redondeada.

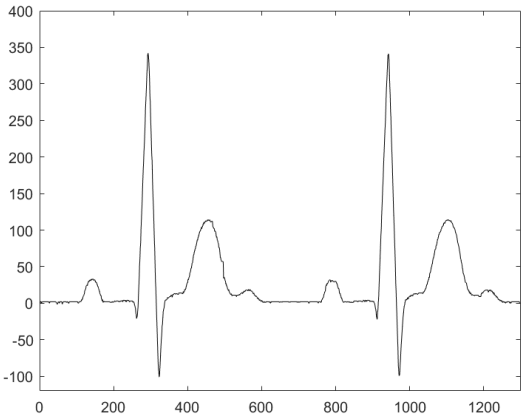
- Onda U
  Se cree que esta última onda diminuta representa la repolarización del músculo papilar. Puede no verse y a menudo se ignora.

La onda P y el complejo QRS están separados por el segmento PR, mientras que el segmento ST separa el complejo QRS y la onda T. La onda P con el segmento PR forman el intervalo PR. El complejo QRS, el segmento ST y la onda T forman el intervalo QT.
Las variaciones en la estructura del corazón y su entorno (incluida la composición de la sangre) alteran estas entidades. La presencia, la ausencia y el tamaño de las diferentes partes del ECG caracterizan diferentes anomalías cardíacas, que comprenden alteraciones del ritmo cardíaco (como la taquicardia ventricular y la fibrilación auricular), perturbaciones del flujo sanguíneo de las arterias coronarias (como el infarto de miocardio y el infarto de miocardio y la isquemia miocárdica), así como alteraciones electrolíticas (como la hiperpotasemia e hipopotasemia).

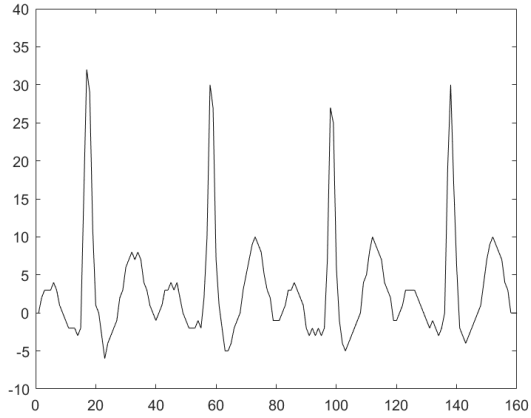
Seleccionamos algunos patrones representativos para ilustrar el método veáse en Figura 3.2.

- **ECG normal**, repite la estructura básica P-Q-R-S-T correspondiente a un ritmo sinusal de entre 60 y 100 latidos por minuto (bpm).
- **Taquicardia sinusal** cuando el ritmo sinusal es más de 100 latidos por minuto. Puede ser normal durante el ejercicio, pero anormal en otros casos.
- **Bradicardia sinusal**, alteración de la frecuencia del impulso eléctrico del corazón producido por el nodo sinoauricular, que lleva a que el ritmo de contracción de este se enlentezca por debajo de lo normal, es decir, por debajo de 60 latidos por minuto.
- **Arritmia sinusal**. La arritmia sinusal es un fenómeno fisiológico normal y se considera una variación del ritmo sinusal normal, caracterizada por las variaciones en los intervalos P-P superiores a 0.12 s con una morfología normal de onda P
- **Síndrome de QT largo**, es una afección caracterizada por una grave alteración en la repolarización ventricular, traducida electrocardiográficamente por una prolongación del intervalo QT. Esto significa que el músculo cardíaco tarda más de lo normal en recargarse entre latidos. Es un efecto secundario conocido de una una amplia gama de medicamentos. Una prolongación excesiva del QT puede desencadenar taquicardias y torsades de pointes (TdP). Puede hacer que el corazón lata rápido y de manera caótica. Estos latidos rápidos del corazón pueden provocar que te desmayes de forma repentina.
- **Torsado de Pointes** es una taquicardia ventricular, rápida y polimórfica, caracterizada por la fluctuación de los complejos QRS en torno a la línea de isoeléctrica, además suele estar asociado con síndrome del QT largo. Puede conducir a una fibrilación ventricular potencialmente mortal.

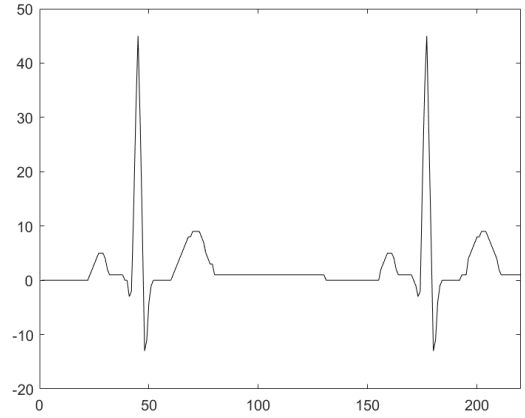
(a) ECG Normal



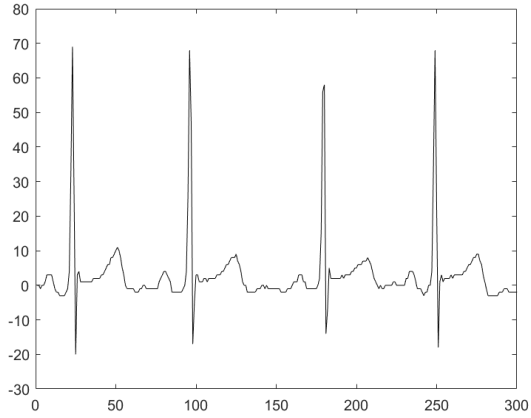
(b) Taquicardia sinusal



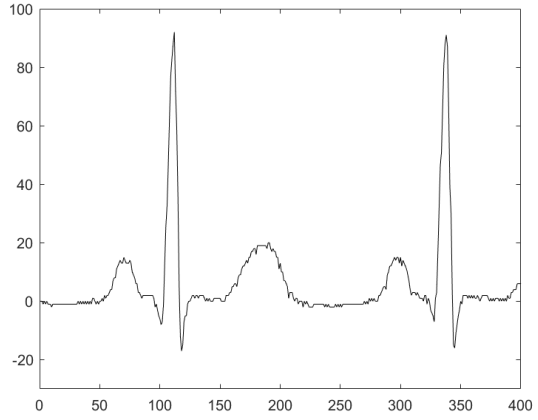
(c) Bradicardia sinusal



(d) Arritmia sinusal



(e) QT largo



(f) Torsades de Pointes

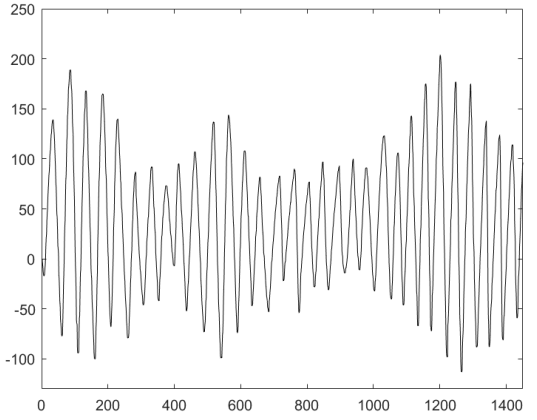
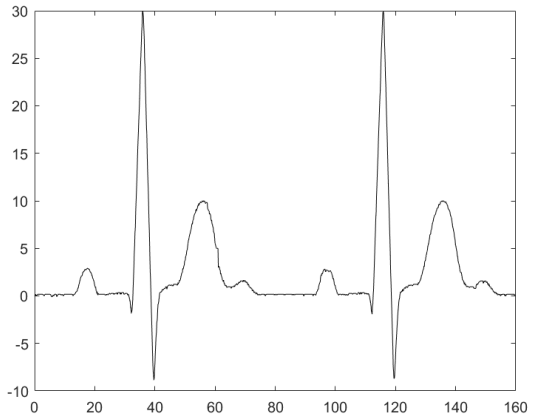


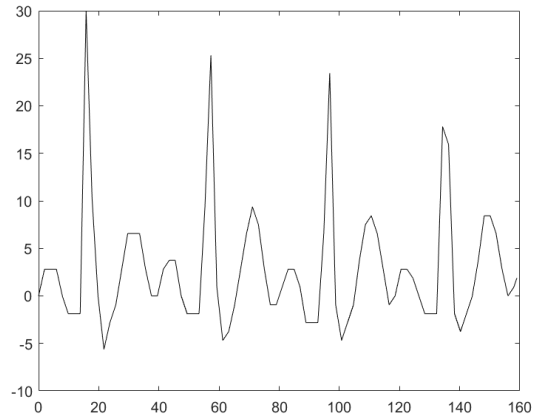
Figura 3.2: Patrones de referencia de ECG.

En la figura 3.2 se puede observar como los patrones considerados no están en la misma escala, por lo que se va a reescalar los patrones intentando que tengan la misma anchura y misma altura, obteniendo la Figura 3.3

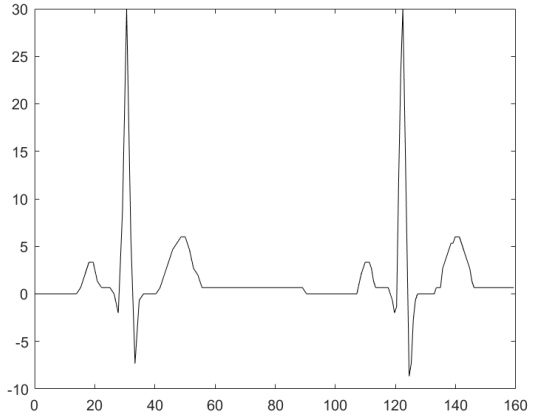
(a) ECG Normal



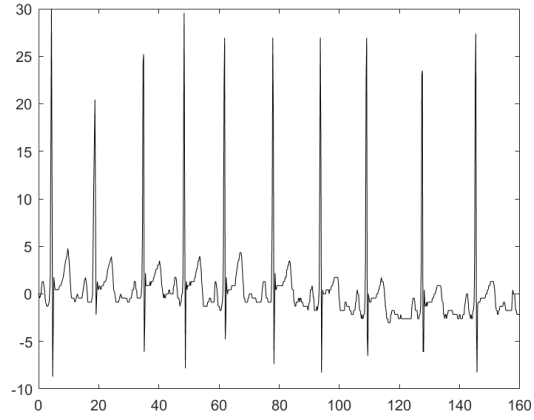
(b) Taquicardia sinusal



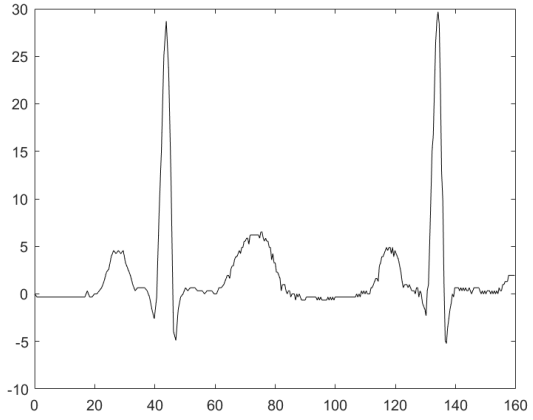
(c) Bradicardia sinusal



(d) Arritmia sinusal



(e) QT largo



(f) Torsades de Pointes

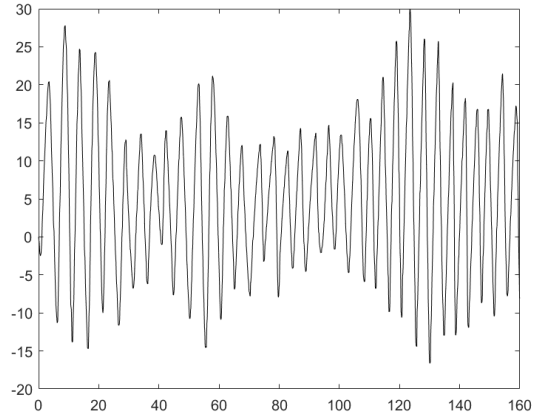


Figura 3.3: Patrones de referencia de ECG reescalados.

3.2. *Time warping distance*

Buscamos identificar patrones de ECG mediante distancias adecuadas. La *time warping distance* (TWD) es uno de los algoritmos para medir la similitud entre dos secuencias temporales, que pueden variar en velocidad. [10, 11]. En general, DTW es un método que calcula una coincidencia óptima entre dos secuencias dadas (por ejemplo, series de tiempo), con alguna condiciones:

- Cada índice de la primera secuencia coincide con uno o más índices de la segunda secuencia, y viceversa
- El primer índice de la primera secuencia se empareja con el primer índice de la segunda secuencia (tiene que ser su única coincidencia)
- El último índice de la primera secuencia se empareja con el último índice de la segunda secuencia (tiene que ser su única coincidencia)

Definimos un coste calculando la suma de las diferencias absolutas de los valores de la secuencia para cada par de índices emparejados. El emparejamiento óptimo minimiza el coste, siempre que se cumplan las condiciones anteriores.

Sean $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$ dos series con longitud n y m respectivamente.
La definición formal de la distancia *dynamic time warping* entre dos series, viene determinada por

$$DTW(\langle\,\rangle, \langle\,\rangle) = 0,$$

$$DTW(X, \langle\,\rangle) = DTW(\langle\,\rangle, Y) = \infty,$$

$$DTW(X, Y) = d(x_i, y_j) + min \begin{cases} DTW(X, Y[2:-]) \\ DTW(X[2:-], Y) \\ DTW(X[2:-], Y[2:-]) \end{cases}$$

donde $\langle\,\rangle$ indica la serie vacia, $[2:-]$ indica una submatriz cuyos elementos incluyen desde el segundo elemento hasta el último elemento de una matriz unidimensional, $d(x_i, y_j)$ indica la distancia entre dos puntos x_i e y_j que puede ser representada por las diferentes medidas de distancia, por ejemplo, la distancia euclidea $d(x, y) = |x - y|$.
La distancia DTW de dos series se puede calcular mediante el método de programación dinámica basado en la matriz de distancias acumuladas, cuyo algoritmo consiste principalmente en construir una matriz de distancias acumuladas:

$$DTW(i, j) = d(x_i, y_j) + min\,\{DTW(i-1, j), DTW(i, j-1), DTW(i-1, j-1)\}$$

donde cada elemento $DTW(i, j)$ indica la distancia DTW entre las secuencias $X_{1:i}$ e $Y_{1:j}$.

En la Figura 3.4 ilustra la diferencia entre la distancia Euclidea y la DTW.

(a) Distancia Euclidea (b) Distancia DTW

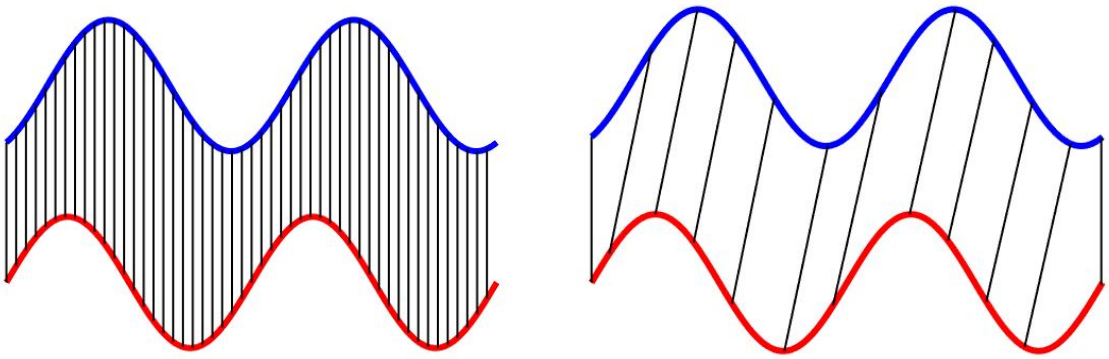


Figura 3.4: Distancia Euclidea vs *dynamic time warping*.

En la sección 3.4, se va aplicar esta distancia para la clasificación de ECG.

3.3. Distancia *Wasserstein-1*

La EMD que hemos considerado hasta ahora es muy general, puesto que admite valores negativos, teniendo en cuenta que en nuestro estudio se comparan imágenes que siempre dan números positivos igual tendría sentido introducir una variante de la distancia que se ajuste mejor a nuestro caso.
Recordemos que la distancia EMD se planteaba como un problema de Transporte Óptimo (OT, *Optimal Transport*). OT desempeña un papel crucial en muchas áreas como la dinámica de fluidos, procesnamiento de imágenes, el aprendizaje automático, ...
Se propone un algoritmo rápido para el cálculo de la distancia *Wasserstein-1* entre dos distribuciones definidas en una cuadrícula, particularizando así la distancia EMD.
En el artículo [12], se proponen dos algortimos que calculan la distancia *Wasserstein-1* entre dos imágenes, utilizando una de las tres normas propuestas: norma-1, norma-2 y norma-inf.
Dado un vector $\vec{x} = (x_1, x_2, ..., x_m)$ las normas anteriores se definen como:

$$||x||_1 = \sum_{k=1}^{m} |x_k|, \quad ||x||_2 = \left(\sum_{k=1}^{m} |x_k| \right)^{\frac{1}{2}}, \quad ||x||_\infty = \max_{1 \leq k \leq m} |x_k|$$

Dadas dos distribuciones de probabilidad 2D, o dos imágenes, ρ_0 y ρ_1, queremos encontrar un transporte de una a la otra con la mínima energía:

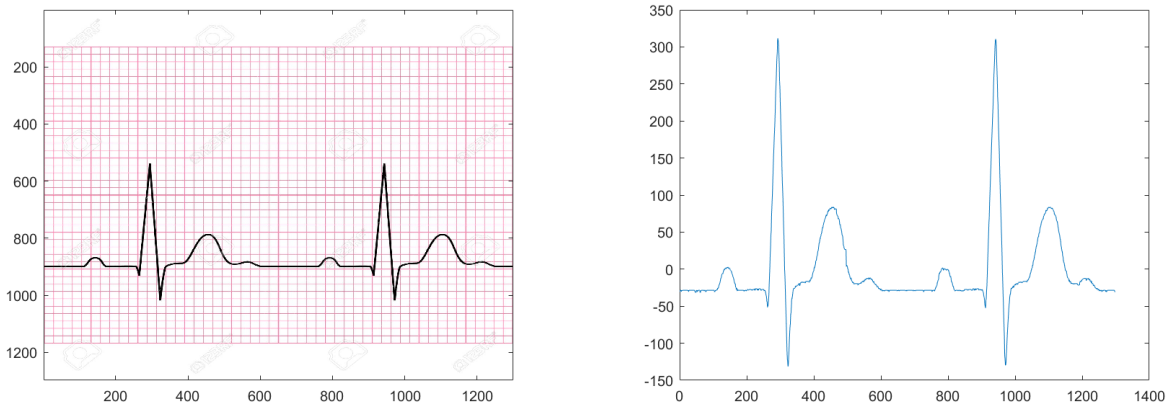
$$\begin{aligned} \min \quad & \int_x ||m(x)||_p \, dx \\ \text{s.a.} \quad & divergence_h(m) = \rho^0 - \rho^1, \\ & \textit{bajo la condición en la frontera de flujo nulo} \end{aligned}$$

donde p puede tomar los valores 1,2 o infinito, por lo que $||m(x)||_p$ representa la norma de orden p y h es el tamaño de paso espacial.

Se desarrollan dos algoritmos: *Algoritmo 1M* y *Algoritmo 2M* para la resolución anterior, propuestos e implementados en [12], además se pueden encontrar en el anexo B.
Se utilizan estos dos algoritmos para calcular la distancia *Wasserstein-1* entre los patrones de y los pacientes. La ventaja de estos algoritmos es que compara las dos imágenes directamente sin la necesidad de normalizarlas. Además cabe destacar la eficiencia del algoritmo 2M, que apenas tarda unos segundos en calcular la matriz distancias entre pacientes y patrones. Sin embargo, el algoritmo 1M no es tan eficiente en este aspecto, puesto que tarda en calcular la misma matriz unas horas, esto posiblemente sea debido a que se comparan imágenes reales que están en distintas escalas.

3.4. Aplicación y resultados

En esta sección, probamos el rendimiento de TWD y EMD para clasificar ECGs. En primer lugar normalizamos los ECG para poder compararlos, como se ilustra en la Figura 3.5.



(a) Imagen inicial

(b) Curva normalizada

Figura 3.5: Ejemplo de obtención de la curva normalizada a partir de la gráfica de un ECG normal.

La subfigura (a) muestra un ECG estándar y la (b) representa su normalización.
Se va a explicar brevemente como se lleva a cabo la normalización anterior, se utiliza el software `MATLAB`, el cual nos va leer la imagen y nos devuelve la matriz, $A_{N\times M}$, de colores asociados a dicha imagen. Como esta en formato `RGB` nos devuelve 3 matrices que dan para cada punto, (i,j), el porcentaje de cada color. Debemos identificar que valores de las matrices, $A(i,j,1)$, $A(i,j,2)$, $A(i,j,3)$ se correspondan con el color de la curva del ECG.
Para facilitar el paso anterior, se va crear una matriz estado, de manera que

$$Estado_{ij} = \begin{cases} 1 & \text{si la coordenada (i,j) es del color del ECG} \\ 0 & \text{en otro caso} \end{cases}$$

Una vez que tenemos la matriz anterior, realmente a nosotros únicamente nos interesan los pares $\{(i,j)/\ Estado_{ij} = 1\}$. Finalmente nos quedamos con una colección de puntos bidimensional

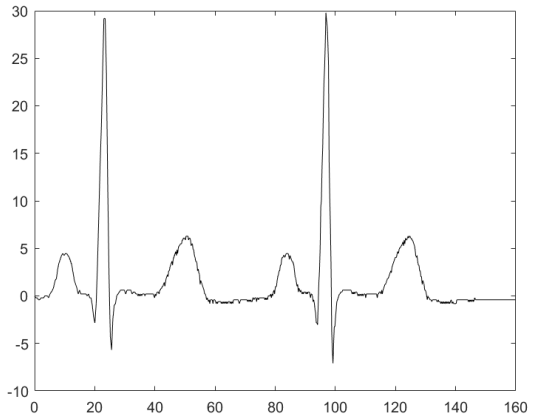
que corresponden a las filas y columnas de la matriz anterior.

Para evitar contratiempos, estos puntos los vamos a normalizar entorno a la media, es decir, cada curva es una secuencia (x_i, y_i), $i = 1, \ldots, n$. Calculamos la media $\mu = \frac{\sum_{i=1}^{n} y_i}{n}$ y consideramos la secuencia normalizada $(x_i, y_i - \mu)$, $i = 1, \ldots, n$.

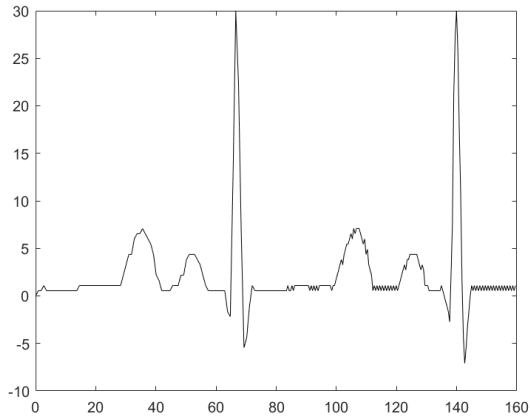
La figura 3.6 muestra ECGs normalizados que queremos clasificar que muestran un tipo de anormalidad u otra, es la base de datos que se va a considerar como pacientes. Los pacientes que considerados presentan el siguiente diágnostico:

- Paciente 1 $\rightarrow$ ECG normal
- Paciente 2 $\rightarrow$ ECG QT largo
- Paciente 3 $\rightarrow$ Torsado de pointes
- Paciente 4 $\rightarrow$ Bradicardia sinusal
- Paciente 5 $\rightarrow$ Taquicardia sinusal paciente

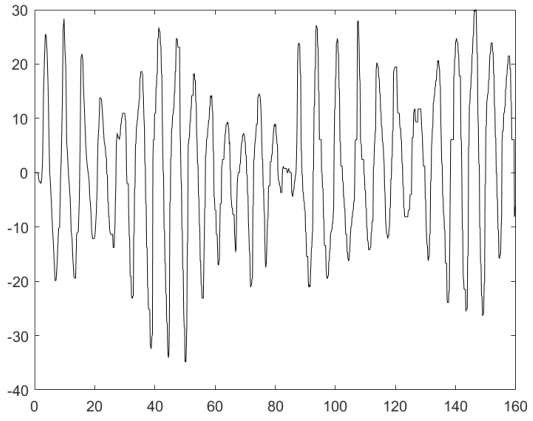
(a) Paciente 1



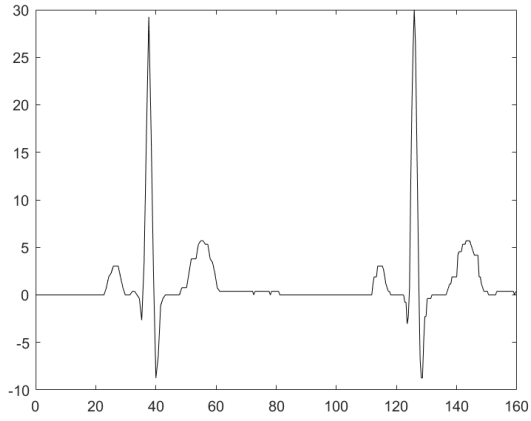
(b) Paciente 2



(c) Paciente 3



(d) Paciente 4



(a) Paciente 5

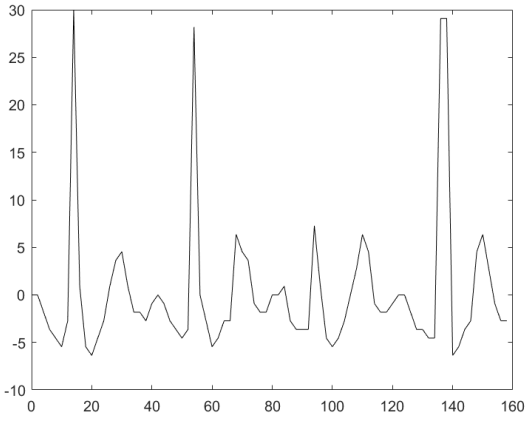


Figura 3.6: ECGs normalizados y reescalados para ser diagnosticados por comparación con los de referencia recogidos en la figura 3.3.

Para ello, calculamos la distancia de cada secuencia normalizada a los patrones. Las tablas 3.1 y 3.2 reproducen las matrices de distancia resultantes para la TWD y la EMD.

Patrón \Paciente	1	2	3	4	5
ECG Normal	772	414	9974	327	663
Taquicardia sinusal	1807	539	10046	413	197
Bradicardia sinusal	1086	259	10381	113	277
Arritmia sinusal	2099	1120	8550	1105	975
QT largo	1379	236	10174	313	378
Torsades de pointes	10170	9852	5939	9862	8947
Diagnóstico	Normal	QT largo	TdP	Bradicardia	Arritmia

Tabla 3.1: Distancias entre los patrones de ECG y los ECGs de los pacientes calculados, utilizando TWD y el diagnóstico propuesto, basado en la menor distancia.

Patrón \Paciente	1	2	3	4	5
ECG Normal	175.9249	95.9059	273.9574	108.5616	79.0234
Taquicardia sinusal	65.2876	79.0439	139.7507	74.7188	77.9492
Bradicardia sinusal	60.6457	76.1689	172.1750	72.2178	67.4229
Arritmia sinusal arrythmia	191.0266	121.2034	290.5680	108.8583	76.6855
QT largo	135.7548	116.9913	194.5916	104.3210	72.2278
Torsades de pointes	473.4429	309.0894	542.9015	278.8493	131.6373
Diagnosis	Bradicardia	Bradicardia	Taquicardia	Bradicardia	Bradicardia

Tabla 3.2: Distancias entre los patrones de ECG y los ECGs de los pacientes calculados utilizando el EMD y el diagnóstico propuesto, basado en la menor distancia.

Para visualizar el rendimiento de cada estrategia, creamos la tabla 3.3 cuyas entradas son

$$\begin{cases} 1 & \text{si el paciente i es diagnosticado correctamente por la distancia j} \\ 0 & \text{en otro caso} \end{cases}$$

Paciente	TWD	EMD
1	1	0
2	1	0
3	1	0
4	1	1
5	1	0
Éxitos	5	1

Tabla 3.3: Diagnóstico correcto con TWD frente a diagnóstico correcto con EMD.

Obsérvese que el TWD obtiene mejores resultados clasificando a todos los pacientes correctamente, sin embargo la distandia EMD no se puede afirmar que sea un buen clasificador, para

optimizar su funcionamiento se van a considerar variantes de la EMD 2D (*Wasserstein-1 distance*) explicado en la sección 3.3.

Las tablas 3.4 y 3.5 represetan las matrices de distancia obtenidas para la norma 1 para los algoritmos 1M y 2M respectivamente. El resto de matrices distancias se encuentran en el anexo B.

Patrón \Paciente	1	2	3	4	5
ECG Normal	0.2543	0.2703	0.1886	0.2635	0.4157
Taquicardia sinusal	0.4474	0.2907	0.3775	0.3028	0.0070
Bradicardia sinusal	0.4240	0.1817	0.3483	0.1982	0.2033
Arritmia sinusal	0.3225	0.2087	0.2032	0.2088	0.3347
QT largo	0.3959	0.0588	0.3167	0.0732	0.2271
Torsades de pointes	0.2796	0.1671	0.1317	0.1639	0.241
Diagnóstico	Normal	QT largo	TdP	QT largo	Taquicardia

Tabla 3.4: Distancias entre las imágenes de los patrones de ECG y los ECGs de los pacientes calculados, utilizando el algoritmo 1M, y la norma p=1, calculo del diagnóstico propuesto, basado en la menor distancia.

Patrón \Paciente	1	2	3	4	5
ECG Normal	0.2615	0.2761	0.1867	0.2678	0.4197
Taquicardia sinusal	0.4596	0.2885	0.3757	0.3033	0.0070
Bradicardia sinusal	0.4349	0.1824	0.3481	0.1974	0.2043
Arritmia sinusal	0.3313	0.2082	0.2024	0.2090	0.3351
QT largo	0.4069	0.0601	0.3153	0.0745	0.2252
Torsades de pointes	0.2791	0.1673	0.1313	0.1646	0.2410
Diagnóstico	Normal	QT largo	TdP	QT largo	Taquicardia

Tabla 3.5: Distancias entre las imágenes de los patrones de ECG y los ECGs de los pacientes calculados, utilizando el algoritmo 2M, y la norma p=1, calculo del diagnóstico propuesto, basado en la menor distancia.

Finalmente en tabla 3.6 se pueden ver los éxitos de la distancia *Wasserstein*, para analizar el rendimiento de los dos algoritmos propuestos y las distintas normas.

	Algoritmo 1M			Algoritmo 2M		
Paciente	$p=1$	$p=2$	$p=\infty$	$p=1$	$p=2$	$p=\infty$
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	1	1	1	1	1	1
4	0	0	0	0	0	0
5	1	1	1	1	1	1
Éxitos	4	4	4	4	4	4

Tabla 3.6: Diagnóstico correcto con la distancia *Wasserstein-1* comparando los algoritmos 1M y 2M, y las normas 1,2 e infinito.

Recopilando todos los resultados obtenidos, obtenemos los resultados de la tabla 3.7, donde podemos ver que la distancia que mejor funciona como clasificador es la distancia TWD, clasificando correctamente a los 5 pacientes.
La siguiente distancia mejor posicionada es la distancia *Wasserstein-1* la cuál ha clasificado bien 4 pacientes, confundiendo el paciente 4. El ECG del paciente mal clasificado muestra una bradicardia, y se ha clasificado con un QT largo. La confunsión puede ser debida a que la bradicardia presenta un ritmo cardiaco más lento de lo normal y el QT largo presenta una prolongación del intervalo QT, lo que se puede confundir con un ritmo cardiaco lento. Otra razón de la confusión puede ser debido a que estamos comparando imágenes reales que tienen escalas distintas lo que hace que haya una variabilidad en la comparativa de las imágenes y pueda provocar este tipo de confusiones.
Por último, con respecto a la distancia EMD considerada no se puede considerar como buen clasificador, esto puede ser debido a que la EMD que hemos considerado para el estudio es muy general, puesto que admite valores negativos, teniendo en cuenta que en nuestro caso se comparan imágenes que siempre dan números positivos, puede dar lugar a confusiones.

Paciente	TWD	EMD	Wasserstein-1
1	1	0	1
2	1	0	1
3	1	0	1
4	1	1	0
5	1	0	1
Éxitos	5	1	4

Tabla 3.7: Diagnóstico correcto con las distancias TWD,EMD y *Wasserstein-1*.

Capítulo 4

Conclusiones

La extracción de información de datos clínicos reales de forma automática se enfrenta a una serie de retos, como la falta de disponibilidad de cantidades de datos suficientemente grandes o el carácter incompleto de los registros. Para cada paciente sometido a la misma enfermedad, pueden haberse monitorizado variables ligeramente diferentes o los intervalos de tiempo entre las pruebas pueden variar en gran medida.
El diseño de herramientas automatizadas para la detección de anomalías en el cuadro clínico de un paciente ofrece una gran oportunidad para mejorar la gestión de los casos clínicos y la investigación clínica. Los datos médicos suelen almacenarse en forma de matrices curvas o imágenes. Las secuencias temporales de los resultados de las pruebas de laboratorio, por ejemplo, adoptan la forma de matrices numéricas.

Se ha hablado de un tipo de normalización, relacionado con los rangos de normalidad, que nos permiten extraer información con significado médico.
A continuación, ilustramos cómo proponer posibles diagnósticos buscando patrones numéricos específicos en cada periodo explotando la normalización de los datos. Realizando un análisis bayesiano basado en el modelo de Plackett-Luce, comparamos el rendimiento de las tres distancias consideradas, para identificar los mejores métodos con una incertidumbre cuantificada, de tres distancias: distancia de Hamming, Euclídea y EMD.
El modelo de Plackett-Luce destaca a la distancia de Hamming cómo la mejor para clasificar pacientes.
Utilizando la distancia de Hamming para comparar el resultado de los análisis de laboratorio en diferentes días, o para diferentes pacientes, uno podría clasificar automáticamente los perfiles de los pacientes.
Además se añade un estudio de patrones en días concretos, a través de herramientas gráficas como los mapas de calor, detectando días anómalos donde se observe un desbande llamativo con respecto al resto de días.

En la última parte del trabajo se desarrolla varios métodos automáticos para el diagnóstico por imagen, concretamente para el diagnóstico de electrocardiogramas. Se ha probado el potencial de las distancias *Time warping*, EMD y *Wassertein-1* para clasificar correctamente los patrones básicos de anormalidad. Dos de ellas clasificaron casi correctamente a todos los pacientes. Además, la distancia *Time Warping* parece proporcionar la herramienta más robusta para este propósito.

El desarrollo de estas herramientas, que a partir del ECG, puedan ser capaces de resaltar la actividad auricular y ventricular, suponen una gran oportunidad para mejorar la gestión estratégica de las unidades, el manejo de casos clínicos concretos y la investigación clínica.

A modo de conclusión final, las técnicas de desarrolladas aquí pueden sentar una base para el cribado automático de información médica basada en la comparación de patrones. En el proyecto se justifica la importancia de dichas herramientas automáticas para el diagnóstico de pacientes, incluyendo los recursos humanos, (personal especializado con conocimientos clínicos) para la extracción de información clínica relevante. Estas herramientas pueden ser de gran ayuda para identificar y anticiparse a las necesidades de cada paciente.

Bibliografía

[1] Ana Carpio, Alejandro Simón, Luis F de Villa, Bayesian inference for clustering analysis and hyperparameter selection, preprint 2020, arxiv.org/abs/2009.11531

[2] Encyclopedia of Mathematics. http://encyclopediaofmath.org/index.php?title=Hamming_distance&oldid=39148

[3] Waggener, Bill (1995). Pulse Code Modulation Techniques (https://books.google.com/books?i d=8l_o6kI3760C&pg=PA206). Springer. p. 206. ISBN 9780442014360.

[4] Wikipedia contributors. (2021, 26 febrero). Earth mover's distance. Wikipedia. https://en.wikipedia.org/wiki/Earth_mover%27s_distance

[5] Yossi Rubner; Carlo Tomasi; Leonidas J. Guibas (1998). .ª Metric forDistributions with Applications to Image Databases". ProceedingsICCV 1998: 59–66. doi:10.1109/ICCV.1998.710701, ISBN 81-7319-221-9

[6] Ana Carpio, Alejandro Simón, Luis F de Villa, Ranking analysis for hyperparameter and clustering algorithm selection, preprint 2021

[7] H. L. Turner et al, "Modelling rankings in R: the PlackettLuce package", Published online: 12 February 2020

[8] Lilly, Leonard S, ed. (2016). Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty (sixth ed.). Lippincott Williams & Wilkins. p. 74. ISBN 978-1451192759.

[9] Klabunde, R.E. (2005). Electrical activity of the heart. Cardiovascular physiology concepts. Lippincott Williams & Wilkins. ISBN 0-7817-5030-X.

[10] Yingmin Li, Huiguo Chen, Zheqian Wu, "Dynamic Time Warping Distance Method for Similarity Test of Multipoint Ground Motion Field", Mathematical Problems in Engineering, vol. 2010, Article ID 749517, 12 pages, 2010. https://doi.org/10.1155/2010/749517

[11] Gold, Omer; Sharir, Micha (2018). "Dynamic Time Warping and Geometric Edit Distance: Breaking the Quadratic Barrier". Association for Computing Machinery. 14 (4).

[12] Jialin Liu and Wotao Yin and Wuchen Li and Yat Tin Chow, "Multilevel Optimal Transport: a Fast Approximation of Wasserstein-1 distances"(2019).

[13] Ana Carpio, Alejandro Simón, Alicia Torres, Luis F. de Villa, Pattern recognition in data as a diagnosis tool for immune disorders, preprint 2021.

Apéndice A

Anexo I: Base de datos sintética de datos numéricos

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14
1	0	0	-1	-1	0	0	-1	1	0	-1	-1	0	0	1
2	1	0	0	-1	-1	0	0	0	1	0	0	-1	0	0
3	-1	-1	0	-1	0	0	1	0	1	0	0	-1	-1	0
4	0	0	-1	0	-1	0	0	-1	-1	0	1	-1	0	-1
5	1	1	1	-1	-1	1	-1	0	0	0	1	1	0	-1
6	-1	-1	1	0	0	0	1	0	1	0	1	0	-1	1
7	0	1	0	-1	1	0	1	1	-1	1	1	0	-1	-1
8	-1	0	1	-1	0	0	-1	1	0	1	-1	1	-1	0
9	0	1	0	0	-1	0	0	1	0	0	0	0	0	1
10	0	0	0	1	0	0	0	1	0	0	1	-1	1	1
11	0	-1	0	0	0	1	1	0	1	1	-1	1	-1	-1
12	-1	-1	-1	0	0	1	-1	-1	0	-1	-1	0	-1	0
13	1	-1	1	0	0	-1	-1	1	0	1	0	1	1	-1
14	1	0	0	0	0	-1	1	0	0	-1	0	-1	0	0
15	-1	0	1	-1	1	0	0	0	1	1	-1	0	-1	0
16	0	0	1	0	1	0	0	0	0	0	-1	0	1	-1
17	-1	-1	-1	0	0	0	0	1	1	0	0	1	0	-1
18	0	0	0	0	-1	-1	0	0	-1	0	-1	0	0	0
19	0	1	-1	0	0	0	0	-1	0	0	1	0	1	0
20	0	0	0	-1	0	-1	1	0	1	0	0	0	-1	1
21	1	0	1	0	1	0	-1	0	-1	1	0	0	1	-1
22	0	1	-1	-1	0	0	0	1	0	-1	0	1	1	0
23	0	0	-1	0	1	-1	0	0	0	-1	1	-1	0	1
24	0	1	0	0	0	0	0	0	-1	0	0	-1	0	1
25	0	0	1	1	0	-1	0	-1	0	-1	-1	-1	1	0
26	0	0	1	0	0	0	0	0	1	-1	0	1	1	1
27	-1	0	-1	1	0	0	0	1	-1	-1	1	0	-1	0

Figura A.1: Muestra de la base de datos sintética de los pacientes considerados.

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14
1	0	-1	1	0	0	1	-1	0	0	0	0	0	0	0
2	1	0	-1	0	-1	0	-1	0	-1	0	-1	1	1	-1
3	1	1	-1	0	0	-1	1	1	0	0	-1	1	0	0
4	0	1	0	0	0	-1	0	-1	1	-1	0	-1	0	-1
5	-1	0	0	0	0	0	0	-1	-1	0	0	-1	1	0
6	-1	-1	1	1	0	0	0	1	1	0	-1	-1	0	0
7	0	0	0	0	0	0	0	0	1	1	0	0	-1	0
8	1	1	0	0	-1	0	-1	0	1	1	0	0	0	0
9	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	0	0
10	-1	0	0	-1	0	0	1	1	0	-1	1	0	0	-1

Figura A.2: Base de datos sintética para considerar distintos patrones para tener de referencia.

Apéndice B

Wasserstein-1 distance

B.0.1. Algoritmos 1M y 2M

Algoritmo 1 Algoritmo 1M

Input: ρ^0, ρ^1, h, L, *a sequence of tolerances* $\{\epsilon_l\}_{l=1}^{L}$
Output: $m_{h_L}^{K_L}$, $\phi_{h_L}^{K_L}$
Initialization: Let $m_{h_0}^{K_0}$=0, $\phi_{h_0}^{K_0}$=0
for $l = 1, 2, \ldots, L$ **do**
Initialize the current level
$m_{h_l}^0$=Interpolate($m_{h_{l-1}}^{K_{l-1}}$) , $\phi_{h_l}^0$=Interpolate($\phi_{h_{l-1}}^{K_{l-1}}$)
Call Algorithm 1:
$(m_{h_L}^{K_L}, \phi_{h_L}^{K_L})$=Algorithm 1 $(\rho^0, \rho^1, h_l, m_{h_l}^0, \phi_{h_l}^0, \epsilon_l)$
end for

Algoritmo 2 Algoritmo 2M

Input: ρ^0, ρ^1, h, L, *a sequence of tolerances* $\{\epsilon_l\}_{l=1}^{L}$
Output: $m_{h_L}^{K_L}$, $\phi_{h_L}^{K_L}$ (Obtain $\phi_{h_L}^{K_L}$ from $\varphi_{h_L}^{K_L}$)
Initialization: Let $m_{h_0}^{K_0}$=0, $\phi_{h_0}^{K_0}$=0
for $l = 1, 2, \ldots, L$ **do**
Initialize the current level
$m_{h_l}^0$=Interpolate($m_{h_{l-1}}^{K_{l-1}}$) , $\varphi_{h_l}^0$=Interpolate($\varphi_{h_{l-1}}^{K_{l-1}}$)
Call Algorithm 2:
$(m_{h_L}^{K_L}, \varphi_{h_L}^{K_L})$=Algorithm 2 $(\rho^0, \rho^1, h_l, m_{h_l}^0, \varphi_{h_l}^0, \epsilon_l)$
end for

B.0.2. Matrices distancia

Patrón \Paciente	1	2	3	4	5
ECG Normal	0.2473	0.2625	0.1833	0.2540	0.3994
Taquicardia sinusal	0.4436	0.2781	0.3587	0.2914	0.0066
Bradicardia sinusal	0.4189	0.1783	0.3355	0.1907	0.1965
Arritmia sinusal	0.3181	0.2044	0.1959	0.2034	0.3282
QT largo	0.3903	0.0579	0.3051	0.0723	0.2174
Torsades de pointes	0.2705	0.1661	0.1266	0.1646	0.2334
Diagnóstico	Normal	QT largo	TdP	QT largo	Taquicardia

Tabla B.1: Distancias entre las imágenes de los patrones de ECG y los ECGs de los pacientes calculados, utilizando el algoritmo 1M, y la norma p=2, calculo del diagnóstico propuesto, basado en la menor distancia.

Patrón \Paciente	1	2	3	4	5
ECG Normal	0.2546	0.2718	0.1989	0.2646	0.4133
Taquicardia sinusal	0.4733	0.2893	0.3844	0.3029	0.0071
Bradicardia sinusal	0.4409	0.1874	0.3586	0.1996	0.20585
Arritmia sinusal	0.3222	0.2247	0.2107	0.2204	0.3634
QT largo	0.4026	0.0631	0.3273	0.0761	0.2261
Torsades de pointes	0.2765	0.1845	0.1368	0.1813	0.2546
Diagnóstico	Normal	QT largo	TdP	QT largo	Taquicardia

Tabla B.2: Distancias entre las imágenes de los patrones de ECG y los ECGs de los pacientes calculados, utilizando el algoritmo 1M, y la norma p=∞, calculo del diagnóstico propuesto, basado en la menor distancia.

Patrón \Paciente	1	2	3	4	5
ECG Normal	0.2493	0.2629	0.1838	0.2552	0.3995
Taquicardia sinusal	0.4437	0.2776	0.3578	0.2915	0.0068
Bradicardia sinusal	0.4190	0.1777	0.3335	0.1911	0.1975
Arritmia sinusal	0.3176	0.2061	0.1953	0.2050	0.3291
QT largo	0.3909	0.0592	0.3044	0.0728	0.2171
Torsades de pointes	0.2718	0.1658	0.1276	0.1641	0.2310
Diagnóstico	Normal	QT largo	TdP	QT largo	Taquicardia

Tabla B.3: Distancias entre las imágenes de los patrones de ECG y los ECGs de los pacientes calculados, utilizando el algoritmo 2M, y la norma p=2, calculo del diagnóstico propuesto, basado en la menor distancia.

Patrón \Paciente	1	2	3	4	5
ECG Normal	0.2520	0.2662	0.1919	0.2583	0.4043
Taquicardia sinusal	0.4479	0.2807	0.3650	0.2952	0.0075
Bradicardia sinusal	0.4231	0.1803	0.3422	0.1940	0.2020
Arritmia sinusal	0.3204	0.2165	0.2029	0.2148	0.3396
QT largo	0.3948	0.0612	0.3134	0.0746	0.2201
Torsades de pointes	0.2752	0.1751	0.1323	0.1739	0.2380
Diagnóstico	Normal	QT largo	TdP	QT largo	Taquicardia

Tabla B.4: Distancias entre las imágenes de los patrones de ECG y los ECGs de los pacientes calculados, utilizando el algoritmo 2M, y la norma p=∞, calculo del diagnóstico propuesto, basado en la menor distancia.