MEDIDAS DE DISTINGUIBILIDAD ENTRE DISTRIBUCIONES DE PROBABILIDAD

Aspectos teóricos y aplicaciones al estudio de las series temporales.

Tesis Doctoral

Leonardo Esteban Riveaud

Director: P. W. Lamberti

Facultad de Matemática Astronomía, Física y Computación.
Universidad Nacional de Córdoba
Marzo 2020

Tesis de Doctorado

L. E. Riveaud[1],
1 L. E. Riveaud FaMAF, Córdoba, Argentina
∗ **E-mail: leoriveaud@gmail.com**

Contenidos

1 Resumen

Esta tesis se encuadra en el marco de la Teoría de la Información. Uno de los propósitos de esta área de la ciencia es encontrar y estudiar las propiedades de los cuantificadores de información, conocidos como entropías. Como aporte original en esta temática, introducimos una nueva entropía que generaliza la entropía de Shannon. Realizamos un exhaustivo estudio de sus propiedades y exploramos posibles ámbitos de aplicación. Por otro lado, y como objetivo principal de este trabajo, definimos y estudiamos cuantificadores de distinguibilidad entre distribuciones de probabilidad. Estos cuantificadores de distinguibilidad, también llamados divergencias, permiten discriminar dos distribuciones de probabilidad, que por su propio carácter estadístico, es difícil o imposible discriminar. Se investiga el mapeo de señales a distribuciones de probabilidad y por el uso de las divergencias se logra estudiar las propiedades estadísticas de las señales. Definimos aquí dos divergencias distintas: una proveniente de una novedosa interpretación de la divergencia de Kullback-Leibler; la otra generaliza a la distancia Euclidiana y permite una nueva interpretación de la divergencia de Jensen-Shannon. En los dos casos se hicieron aplicaciones tanto a señales de origen natural como simuladas.

2 Introducción

A finales de la segunda guerra mundial Claude E. Shannon estudiaba de forma teórica los sistemas de comunicación. Su objetivo era dar un modelo matemático de las comunicaciones. Sus estudios fueron publicados en 1948 bajo el título "A Mathematical Theory of Comunications" [1]. Con él daba inicio a la teoría de comunicaciones, o como la conocemos hoy en día, a la Teoría de la Información. En este marco teórico, se presentaba a la información trasmitida por los canales desde una perspectiva probabilística. Con el objetivo de encontrar una cantidad que represente, en un sentido probabilístico, a la información, Shannon introdujo el concepto de Entropía de Información. Esta cantidad representa la incertidumbre referente a un canal de información. Por su similitud con la entropía de Gibbs, desde la física se intentó relacionar esta cantidad representativa de la información con los sistemas físicos. Fue en 1957 que E. Jaynes [2] publicó su trabajo pionero, en el cual mediante un método variacional lograba resultados de relevancia en física usando la entropía de Shannon. A partir de ese momento se formalizó la idea de interpretar a los sistemas físicos desde la perspectiva de la teoría de la información.

Luego de la publicación del trabajo de Shannon se buscó un concepto de información más general. Así surgieron entropías como la de Renyi [3], Havrda-Charvat-Tsallis [4] y Salicrú [5] por dar algunos ejemplos. A estas entropías se las conoce como entropías generalizadas ya que mediante un parámetro o una función, generalizaban a la entropía de Shannon. Esto llevó a formalizar el concepto de entropía y enunciar las propiedades que tenía que cumplir un funcional para poder ser llamado de esa manera. En esta tesis se hará uso de estas entropías.

En 1951, con el objetivo de encontrar una función que represente la cantidad de información extra (en bits) que se necesita para escribir un mensaje construido con símbolos de un alfabeto dado con una distribución de probabilidad Q (probabilidad de ocurrencia de los símbolos) cuando en realidad se esperaba uno basado en una distribución de probabilidad P, S. Kulback y R. A. Leibler [6] introdujeron el concepto de Entropía Relativa. Si bien en 1927 R. Fisher había trabajado con esta idea, se puede decir que por primera vez se definía una función que midiera la disimilitud entre dos distribuciones de probabilidad. A esta medida se la conoce actualmente como divergencia de Kulback-Leibler (K-L). Esto dió lugar a una nueva rama dentro de la Teoría de Información. Esta línea de investigación se encarga de encontrar medidas de disimilitud entre distribuciones de probabilidad. Una particularmente relevante para este trabajo de tesis, es la divergencia de Jensen-Shannon [7] y depende exclusivamente de la entropía de Shannon. Rápidamente surgieron otras medidas de disimilitud entre distribuciones de probabilidad como son los casos de la divergencia de Renyi [8], la f-divergence [9], la divergencia de Bregman [10] o la Non-logaritmic Jensen-Shannon [11], las cuales se estudiarán en extenso en el presente trabajo. Esto obligó a los estudiosos del área a dar un marco formal al concepto de divergencia concensuando las propiedades que debía cumplir un funcional de este tipo y ser utilizado como medida de disimilitud entre distribuciones de probabilidad.

Las aplicaciones de estas medidas se han dado en diferentes áreas de la ciencia, tanto en la física [12], en las ciencias sociales [13], como en la estadística en general ([14], [15]).

En el presente trabajo se hará uso de las entropías y divergencias antes mencionadas, pero tendrá como eje principal el desarrollo teórico de nuevas medidas, tanto de entropías como de divergencias generalizadas.

Otra medida importante dentro de la ciencia de la computación y de la Teoría de Información es la *Complejidad*. Kolmogorov en su trabajo de 1963 [16] definió una complejidad algorítmica. Con esta medida pretendía expresar la mínima descripción que podía tener un cierto código. En otros campos del conocimiento, tales como la física y la biología se han buscado definiciones de medidas de complejidad requiriendo propiedades mínimas que una tal medida debiera tener. El éxito en esta búsqueda ha sido relativo, pues no siempre las medidas introducidas han reflejado las características que se buscan describir. En la presente tesis avanzaremos en la introducción de una medida de complejidad siguiendo los criterios de razonabilidad que una medida como esta debe tener.

Al estudiar fenómenos naturales se pueden tomar dos enfoques. El primero de ellos es dar un modelo matemático que describa a dicho fenómeno y por su intermedio poder predecir la evolución temporal de dicho fenómeno. El otro camino consiste en hacer mediciones de ciertas propiedades del sistema y por medio de los resultados, inferir la dinámica subyacente en la evolución del mismo. Para sistemas complejos las ecuaciones que los representan pueden tener comportamientos caóticos. Es por eso que parte de nuestro estudio será analizar sistemas deterministas con un comportamiento caótico y compararlos con sistemas representados por un proceso estocástico.

Para sistemas muy complejos, como suelen ser los sistemas biológicos, dar un marco teórico o modelo matemático exacto puede ser casi imposible. Por eso se opta por el segundo camino. En éste, lo único que se tiene, es una secuencia de mediciones sucesivas en el tiempo, es decir, una serie temporal.

El objetivo principal de la tesis será utilizar las medidas de disimilitud entre distribuciones de probabilidad desarrolladas en este trabajo y mediante diferentes métodos de segmentación analizar series temporales tanto de origen natural como simuladas.

2.1 Organización de la Tesis

El presente trabajo de tesis está dividido en ocho capítulos.

El capítulo 2 titulado *"Marco Teórico"*, provee de las principales herramientas teóricas necesarias para el desarrollo de los siguientes capítulos. En él se brindarán nociones básicas de teoría de probabilidad y procesos estocásticos, y se detallarán de forma resumida las principales propiedades de las funciones convexas. También, en una sección aparte, se describirán los principales conceptos de la teoría de la información.

El capítulo 3 titulado *"Introducción a la problemática de las distancias entre distribuciones de probabilidad"*, consta de una introducción a la problemática principal de este trabajo. En él argumentaremos sobre la importancia de medir distancias entre distribuciones de probabilidad y daremos el marco formal que respalda a esta rama de la teoría de la información. También mostraremos las consecuencias de utilizar una sola cadena de muestra y el porqué recurrir a la teoría de la información para el análisis de series temporales.

El capítulo 4 titulado *"Aspectos Metodológicos"*, consta de la descripción de los métodos de mapeo y segmentación (método del puntero y método de la ventana móvil). Allí mostramos los argumentos de porqué es necesario hacer un mapeo de una serie temporal a una secuencia simbólica y porqué se recurre a métodos de segmentación para el análisis de una serie temporal.

El capítulo 5 titulado *"Divergencia tipo Bregman"*, consta del desarrollo teórico y definición de una nueva divergencia que toma como argumento a entropías generalizadas. Se define una nueva divergencia a partir de una interpretación de la divergencia de Kulback-Leibler como un operador de entropías y de la simetrización de la divergencia de Bregman. Se mostrarán las condiciones necesarias que debe cumplir una entropía generalizada para que esta nueva divergencia cumpla con las propiedades deseadas. Luego se aplicará la divergencia desarrollada en el capítulo a series temporales simuladas y de origen natural. En los dos casos se usará el mapeo de Bandt y Pompe (introducido en el capítulo anterior) y el método de segmentación del puntero. Para el caso de series temporales de origen natural se usarán señales tomadas de un electrocardiograma referidas a situaciones de una fibrilación auricular y en una ritmia normal del corazón [17].

El capítulo 6 titulado *"Divergencia Gamma"*, consta del desarrollo teórico de una nueva divergencia generalizada a partir de propiedades de funciones convexas. Allí mostramos que tanto la divergencia de Jensen-Shannon como el cuadrado de la métrica euclídea pertenecen a una familia de divergencias que dependen de una cierta función $g(x)$. Se mostrarán las condiciones que tienen que cumplir las funciones $g(x)$ para que esta familia de divergencias cumplan con las condiciones deseadas. En otra sección del capítulo se definirá una nueva entropía generalizada que está íntimamente relacionada con la divergencia Gamma. Se demostrará que esta nueva entropía cumple con las propiedades que debe cumplir una entropía generalizada. En otra sección se demostrará que para un cierto conjunto de funciones $g(x)$ la raiz cuadrada de la divergencia Gamma cumple la desigualdad triangular, es decir, es una métrica. En la última sección del capítulo se mostrarán las aplicaciones a series temporales simuladas y de origen natural. En los dos casos se utilizará el mapeo de Bandt y Pompe y el método de segmentación de la ventana móvil. En el caso de las series temporales de origen natural se usaron señales de un electroencefalograma referidas a los distintos estadíos de sueño.

El capítulo 7 titulado *"Complejidad Estadística"*, consta del desarrollo y definición de una complejidad estadística. Para esto usamos la divergencia y entropía generalizadas introducidas en el capítulo anterior, y siguiendo el modelo de la complejidad de Lamberti, Rosso y Pastino definimos una complejidad generalizada que depende de una cierta función $g(x)$.

El capítulo 8 titulado *"Caos y Ruido"*, consta de la diferenciación y distinguibilidad de series temporales producidas por sistemas caóticos y ruidos coloreados mediante el uso de la divergencia Jensen-Shannon. Este capítulo tiene una línea de investigación distinta a los capítulos anteriores ya que está enfocado a la aplicación de herramientas de teoría de información y no al desarrollo teórico de nuevas herramientas. Con el propósito de estudiar sistemas caóticos y ruidos coloreados desarrollamos un mapeo diferente al usado en los demás capítulos. Éste consta de dos etapas, la primera es generar una cadena binaria y luego, a partir de esta generar una cadena alfabetizada. Para todos los análisis se usará el metodo de segmentación de la ventana móvil [18].

Aunque en cada capítulo se hará una conclusión técnica y parcial referida al mismo, dedicaremos las últimas páginas de la tesis para hacer una conclusión general.

3 Capítulo II: Marco Teórico

A los fines de presentar la notación e introducir los principales conceptos utilizados a lo largo de la tesis, acercamos al lector una breve introducción a la teoría de probabilidades y a los procesos estocásticos. Se comentarán también las propiedades básicas de las funciones convexas. Por último se darán algunas nociones básicas de teoría de la información.

3.1 Breviario de Teoría de probabilidad

3.1.1 Medida de Probabilidad

Para entender el concepto de distribución de probabilidad es bueno introducir primero la visión frecuentista de la misma. La axiomatización del concepto de probabilidad surge mucho después y es, en algún sentido, mucho más abstracta. Es por eso que para delucidar este concepto es bueno realizar un ejemplo representativo.

Supongamos que tenemos una bolsa llena de bolas de distintos colores, y queremos saber qué probabilidad hay de agarrar una de color rojo. Lo primero que debemos hacer es contar la cantidad total de bolas que hay en la bolsa, y luego contar la cantidad de bolas rojas. El cociente de estas dos cantidades (si el número de bolas totales es lo suficientemente grande) nos dará la probabilidad deseada:

$$probabilidad_R = \frac{cantidad \quad de \quad bolas \quad rojas}{cantidad \quad de \quad bolas} \tag{1}$$

La cantidad de bolas totales reprensenta nuestro "ensamble" y el hecho de elegir (al azar) una bola roja se lo llama evento. Para dar una fundamentación axiomática de la probabilidad es necesario brindar conceptos de teoría de la medida. Es por eso que empezaremos nuestra descripción de distribución de probabilidad con la fundamentación de un espacio de probabilidad.

Si hacemos un análisis de la descripción frecuentista de la probabilidad nos percataremos de que si algún suceso es un evento su complemento también lo es, y si dos sucesos son eventos la unión de los mismos también lo es. La siguiente definición no es más que la formalización de lo dicho anteriormente ([19]).

$\sigma-$álgebra: Sea Ω nuestro conjunto de eventos, se dice que una familia $\mathcal{F}$ de subconjuntos de Ω es un $\sigma-$álgebra (familia de eventos) sobre Ω si cumplen:

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$ (donde A^c es el complemento de A)
3. $A_1, A_2, \cdots \in \mathcal{F} \Longrightarrow \bigcup_{i\geq 1} A_i \in \mathcal{F}$

Llamamos Espacio Medible al par $(\Omega, \mathcal{F})$.

A veces nos es más cómodo trabajar sobre un tipo particular de espacio medible. Es por eso que necesitamos de una función que nos lleve del espacio medible original al espacio medible donde nosotros queremos trabajar. A esta función se la llama función medible y en teoría de probabilidad, Variable Aleatoria (VA). En nuestro caso el espacio medible en el que trabajaremos será $\mathbb{R}^n$.

Variable aleatoria: Sean $(\Omega, \mathcal{F})$ y $(\Omega', \mathcal{F}')$ dos espacios medibles y sea $X : \Omega \longrightarrow \Omega'$ un mapa entre ambos espacios. Diremos que X es una variable aleatoria si

$$X^{-1}(A') \in \mathcal{F}, \quad \forall A' \in \mathcal{F}' \tag{2}$$

Lo que hemos hecho hasta ahora es construir un espacio al que se le pueda aplicar una medida, en nuestro caso será una medida de probabilidad. En Teoría de la Información, en la mayoría de los casos, se trabaja con la medida de probabilidad y no con las funciones medibles.

Probabilidad: Si es $\mathcal{F}$ es un $\sigma-$álgebra sobre Ω se dice que $P : \mathcal{F} \longrightarrow \mathbb{R}$ es una probabilidad si se cumple

1. $0 \leq P(A_i) \leq 1$ para $A_i \in \mathcal{F}$
2. $P(\Omega) = 1$
3. $A_1, A_2, ... \in \mathcal{F}$ con $A_i \bigcap A_j = \emptyset$ si $i \neq j \Longrightarrow P\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} P(A_i)$

Llamaremos a la terna $(\Omega, \mathcal{F}, P)$ espacio de probabilidad.

Un concepto relacionado a la probabilidad es la distribución de probabilidad. Supongamos que nuestra VA puede tomar todos los valores de la recta real. La distribución de probabilidad $f(x)$ es una función tal que $f(x)dx$ es la probabilidad de que la VA tome los valores entre x y $x + dx$. Formalmente definimos a la distribución de probabilidad de la siguiente manera.

Distribución de Probabilidad Continua: Sea $f : \Omega \to \mathbb{R}$ una función no negativa. Se dice que f es una densidad de probabilidad en $\mathbb{R}$ si

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{3}$$

A esta condición se la llama normalización de la distribución de probabilidad.

Supongamos ahora que nuestra VA puede tomar un conjunto numerable de valores, es decir, nuestro espacio muestral es numerable. En consecuencia la distribución de probabilidad va a ser un conjunto de valores positivos que sumados cumplan con la condición de normalización. Formalmente se la define de la siguiente manera.

Distribución de Probabilidad Discreta: Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad con Ω a lo sumo numerable. En este caso podemos tomar a $\mathcal{F}$ como un conjunto de partes de Ω. Definimos función de distribución de probabilidad discreta p asociada a P por

$$p : \Omega \longrightarrow [0, 1] \tag{4}$$

de la siguiente manera

$$p(\omega) = P(\{\omega\}) \tag{5}$$

con la siguiente propiedad: para cada $A \subset \Omega$ podemos determinar la probabilidad usando la función distribución

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} p(\omega) \tag{6}$$

y la propiedad de normalización

$$P(\Omega) = \sum_{\omega \in \Omega} P(\{\omega\}) = \sum_{\omega \in \Omega} p(\omega) = 1 \tag{7}$$

El concepto de distribución de probabilidad discreta nos acompañará en toda la tesis, ya que las señales con las que trabajamos son mapeadas a una secuencia de símbolos pertenecientes a un alfabeto finito.

Supongamos que deseamos saber la probabilidad de que una VA tome un valor determinado y a la vez otra VA tome otro valor determinado. En otras palabras lo que estamos buscando es la intersección de dos eventos particulares. La probabilidad de la intersección de dos eventos tiene un nombre particular y se la llama probabilidad conjunta y se define de la siguiente manera:

Probabilidad Conjunta: Formalmente hablando, si $A_i, A_j \in \mathcal{F}$ definimos a la probabilidad conjunta como

$$P(A_i \cap A_j) := P(A_i, A_j) \tag{8}$$

Se dice que dos eventos son independientes si son independientes en probabilidad:

Independencia en Probabilidad: Definimos independencia en probabilidad cuando sucede que

$$P(A_i, A_j) = P(A_i)P(A_j) \tag{9}$$

Una cantidad importante es la Esperanza o primer momento de una VA discreta definida por

$$E[X] = \sum_{i=1}^{N} x_i p(x_i) \tag{10}$$

y se la puede interpretar como la generalización de la media aritmética.

Momentos: El k-ésimo momento de una variable aleatoria X se define como

$$E[X^k] := \sum_i x_i^k p(x_i) \tag{11}$$

Varianza: Sea X una VA discreta se define varianza como

$$\sigma^2 = E[(X - E[X])^2] \tag{12}$$

Como veremos más adelante la varianza es un caso particular de la autocovarianza cuando las dos variables aleatorias son la misma. Particularmente en este caso se quiere cuantificar el promedio de la distancia de los valores respecto de la media. Aunque para medir estas desviaciones en distribuciones de probabilidad de las VA que tienen colas pesadas se necesitan medidas de dispersión más robustas. En nuestra caso será de gran importancia para definir el concepto de significancia que se detallará más adelante.

3.2 Teoría de los Procesos Estocásticos

Una serie temporal es un conjunto de valores observados (experimentalmente o generados por algún algoritmo) de una variable dada, x_t, representando t al tiempo en que ese valor se observa. Diremos que la serie temporal es discreta si el conjunto de valores de tiempos t, es discreto. En los últimos años ha crecido notablemente el número de métodos de estudio de las propiedades estadísticas de las series temporales. Ellos se originan en conceptos clásicos de la estadística, en la teoría de la información, y en otras áreas de la matemática. A su vez la importancia de las series temporales es que una gran variedad de fenómenos naturales pueden representarse por medio de ellas. Secuencias genéticas, registros electrofisiológicos (EEG y ECG), dinámica de motores moleculares, modelos climáticos, etc., son claros ejemplos de esta presencia ubicua y de la importancia de las series temporales, en las ciencias naturales. Desde un punto de vista formal, una serie temporal se puede representar como un proceso estocástico. Es por ello que a continuación repasaremos las principales propiedades de este tipo de procesos (basado en [21]).

Proceso Estocástico Al ser los valores x_t de carácter impredecible es natural que los tratemos como si fueran una variable aleatoria X_t. Se define como Proceso Estocástico a la familia de variables aleatorias $\{X_t; \ \forall\, t \in T\}$; definidas en un espacio de probabilidad $(\Omega, \psi, \mathbb{P})$. El parámetro t no pertenece al espacio muestral, es decir, no es un evento. Esto significa que para cada t tendremos una variable aleatoria $X_t : \Omega \to \mathbb{R}$; que vista como una aplicación es $X : T \times \Omega \to \mathbb{R}$.

Se puede ver a la serie temporal $\{x_1, x_2, ..., x_n\}$ como una realización "particular" finita del proceso estocástico $\{X_t; t \in T\}$. Viéndolo así sería :

$$X_{-\infty}, ..., X_{-1}, \underbrace{X_0, ..., X_n}_{x_0,...,x_n}, X_{n+1}... \tag{13}$$

La sucesión de arriba es un proceso estocástico, y la de abajo una serie temporal. El hecho que tenga las letras en minúsculas es porque son un valor específico (particular), dentro del rango de la variable aleatoria, que ha tomado esta.

Autocovarianza: Sea $\{X_t, t \in T\}$ un proceso estocástico tal que $Var(X_t) < \infty$ para cada $t \in T$, luego la función autocovarianza $\gamma_X(h)$ de $\{X_t\}$ se define como

$$\gamma_X(r, s) := Cov(X_r, X_s) = E[(X_r - E[X_r])(X_s - E[X_s])] \tag{14}$$

Podemos definir la autocovarianza también de la siguiente manera

$$\gamma_X(h) := Cov(X_{t+h}, X_t) = E[(X_{t+h} - E[X_{t+h}])(X_t - E[X_t])] \tag{15}$$

y la función autocorrelación para el mismo proceso se define como

$$\rho(h) := \frac{\gamma(h)}{\gamma(0)} \tag{16}$$

La autocorrelación es muy útil en la estadística cuando se tiene una sola serie temporal y deseamos inferir el valor futuro de la serie, ya que nos brinda un conocimiento de como se relacionan los valores de la serie entre sí.

Existen procesos en la naturaleza que mapeados a una serie temporal presentan cierto tipo de patrones, uno de ellos es el de la estacionareidad. En este tipo de situaciones la señal se modela con un proceso estacionario, es decir, como un proceso en el cual sus variables aleatorias que lo componen tienen la misma esperanza y también en el cual la autocovarianza se repite para una cierta distancia de tiempo. Formalmente se lo define de la siguiente manera

Proceso Estacionario: La serie de tiempo $X_t, t \in \mathbb{Z}$ se dice *estacionaria* si

1. $E|X_t|^2 < \infty$ para todo $t \in \mathbb{Z}$
2. $EX_t = m$ para todo $t \in \mathbb{Z}$
3. $\gamma_X(r, s) = \gamma_X(r + t, s + t)$ para todo $t, r, s \in \mathbb{Z}$

Los procesos estacionarios son de gran utilidad teórica especialmente para entender la relación entre el determinismo y la aleatoreidad, como se ve reflejado en el teorema de descomposición de Wold:

Teorema: *Descomposición de Wold*

Si $\sigma^2 > 0$ entonces podemos expresar a cualquier proceso estocástico estacionario X_t como

$$X_t = V_t + \sum_{j=0}^{\infty} \psi_j Z_{t-j} \tag{17}$$

donde

1. $\psi_0 = 1$ y $\sum_{j=0}^{\infty} \psi_j^2$
2. $\{Z_t\} \sim WN\,(0, \sigma^2)$
3. $Z_t \in \mathcal{M}_t$ para cada $t \in \mathbb{Z}$
4. $E\,(Z_t V_s) = 0$ para todo $s, t \in \mathbb{Z}$
5. $V_t \in \left\{\bigcap_{n=-\infty}^{\infty} \mathcal{M}_n\right\}$
6. El proceso estocástico $\{V_t\}$ es Determinista

Los procesos estocásticos $\{X_j\}$ y $\{V_j\}$ y los coeficientes $\{\psi_j\}$ están unívocamente determinados por la ecuación (17) y las condiciones (1–6). Dejaremos de lado la demostración.

Por más que no utilizaremos este teorema nos surgió la necesidad de expresar este resultado ya que nos muestra las concecuencias de tener un proceso estacionario. *El hecho de poder descomponer un proceso estocástico en la suma de dos procesos completamente antagónicos nos afirma nuevamente lo importante que es estudiar los procesos deterministas y aleatorios.* Se estudiarán este tipo de procesos en el capítulo 8.

Teniendo ya el concepto de autocovarianza es posible definir el Ruido Blanco. Más allá que este proceso se utilizará de forma explícita en el capítulo de "Caos y Ruido", es de gran importancia en toda la investigación aunque sea de forma implícita. Esto se debe a que en muchos momentos de la tesis trabajamos con secuencias simbólicas que tienen distribuciones que podrían ser generadas por este tipo de proceso.

Definición: Ruido Blanco: Sea $\{Z_t\}$ un proceso estocástico con varianza σ^2. Se lo llamará ruido blanco si cumple con las siguientes condiciones:

1. $E(Z_t) = 0$, para todo t entero.
2. $\gamma(h) = \sigma^2 \cdot \delta(h)$.

De manera equivalente, esta última condición se la puede escribir como:

$$\gamma(h) = \begin{cases} \sigma^2 & ;\text{si } h = 0 \\ 0 & ;\text{si } h \neq 0 \end{cases} \tag{18}$$

Se suele resumir los parámetros del proceso poniendo

$$\{Z_t\} \sim WN(0, \sigma^2) \tag{19}$$

3.3 Propiedades de las Funciones Convexas

En esta sección describiremos las propiedades de los conjuntos y funciones convexas a utilizar en las siguientes secciones (para más detalles ver [22]). La definición de función convexa es fundamental para entender gran parte de nuestra investigación y será usada en reiteradas ocaciones.

Definición: (Segmento Línea) Sean x e y dos puntos de $\mathbb{R}^n$. Definimos como segementos línea $\overline{xy}$ a todos los puntos de la forma $\alpha x + \beta y$ donde $\alpha \geq 0$ y $\beta \geq 0$ con $\alpha + \beta = 1$.

Definición: (Conjunto Convexo) Decimos que $\mathcal{S}$ es un conjunto convexo si para cada par de puntos x e y se cumple que $\overline{xy} \subset \mathcal{S}$.

Definición: (Función Convexa) Sea f una función a valores reales definida sobre un dominio convexo $\mathcal{D} \subset \mathbb{R}^n$. Luego f es convexa si

$$f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y) \tag{20}$$

para todo x e y en $\mathcal{D}$ y para todo $\alpha \geq 0$, $\beta \geq 0$ con $\alpha + \beta = 1$. Decimos que f es cóncava si la desigualdad se cumple de manera invertida.

Uno de los teoremas fundamentales para la tesis, y que lo usaremos en uno de los trabajos de investigación es el siguiente. Este afirma que una función es convexa si y solo si su hessiano es positivo.

Teorema: Sea f una función a valores reales definida en el intervalo abierto $(a, b) \in \mathbb{R}$ y si suponemos que f'' existe en (a, b). Entonces f es convexa si y solo si $f''(t) \geq 0$ para todo $t \in (a, b)$.

PRUEBA: Suponemos que f es convexa con x e y en el intervalo (a, b) con $x < y$. Dada $t_1 \equiv y, t_2, t_3, \cdots$ una sucesión decreciente convergente a x, y tomamos $f(t_1) \equiv f(y), f(t_2), f(t_3), \cdots$ las correpondientes imágenes, luego

$$f'(x) = \lim_{k\to\infty} \frac{f(x) - f(t_k)}{x - t_k} = \lim_{k\to\infty} m_k \tag{21}$$

donde m_k es la pendiente del segmento línea que une los puntos $(x, f(x))$ y $(t_k, f(t_k))$. Pero m_k es una sucesión decreciente, por lo tanto, $f'(x) \leq m_k$ para cada k y en particular

$$f'(x) \leq m_1 = pendiente\overline{(x, f(x))(y, f(y))} \tag{22}$$

Entonces $f'(x) \leq f'(y)$. Se sigue que $f'(x)$ es nodecreciente y por lo tanto $f''(x) \geq 0$ en (a, b).

Ahora suponemos que $f''(x) \geq 0$ en (a, b) por lo tanto $f'(x)$ es nodecreciente. Sean x e y dos puntos de (a, b) con $x < y$ y $z = \alpha x + \beta y$ una combinación convexa de esos dos puntos. Del teorema fundamental del cálculo tenemos que

$$f(z) - f(x) = \int_x^z f'(t)dt \leq f'(z)(z - x) \tag{23}$$

ya que $f'(t)$ toma un máximo en z. Por otra parte tenemos también

$$f(y) - f(z) = \int_z f'(t)dt \geq f'(z)(y - z) \tag{24}$$

ya que $f'(t)$ toma un mínimo en z. Esto nos da

$$f(z) \leq f(x) + f'(z)(z - x) \tag{25}$$

$$f(z) \leq f(y) - f'(z)(y - z) \tag{26}$$

Se sigue que tenemos

$$\begin{aligned} f(z) =& \alpha f(z) + \beta f(z) \leq \\ & \leq \alpha[f(x) + f'(z)(z - x)] + \beta[f(y) - f'(z)(y - z)] = \\ & = \alpha f(x) + \beta f(y) \end{aligned} \tag{27}$$

mostrando que $f(x)$ es convexa.

El corolario siguiente será necesario para las próximas aplicaciones de las propiedades de las funciones convexas, específicamente para el entendimiento de la divergencia de Bregman que se detallará en capítulos posteriores, ya que garantiza su positividad.

Corolario: Sea $f(\vec{x})$ una función convexa, definimos como $g(\vec{x})$ al hiperplano tangente en $\vec{z}$ dado por

$$g(\vec{x}) = f(\vec{z}) + \sum_{i=1}^{N}(x_i - z_i)\frac{\partial f}{\partial x_i}\bigg|_{\vec{x}=\vec{z}} = (\vec{x} - \vec{z}) \cdot \nabla f + f(\vec{z}) \tag{28}$$

Como $f(\vec{x})$ es convexa y doblemente diferenciable tenemos que $\frac{\partial^2 f}{\partial x_i \partial y_i} \geq 0$. Por lo tanto el plano tangente no toca ningún punto de $f(\vec{x})$ para cualquier punto $\vec{z}$. Entonces se obtiene que para cualquier $\vec{x}$

$$f(\vec{x}) - g(\vec{x}) \geq 0 \qquad \forall \vec{z} \tag{29}$$

Donde la igualdad se cumple para $\vec{x} = \vec{z}$.

Teorema: (Desigualdad de Jensen) Sea f una función a valores reales definida sobre el subconjunto convexo $\mathcal{D}$ de $\mathbb{R}^n$, y sea $\alpha_1 x_1 + \cdots + \alpha_m x_m$ una combinación convexa de los puntos $x_1, \cdots, x_m$ de $\mathcal{D}$. Con $\alpha_i \geq 0$ para $i : 1, \cdots, m$ y $\alpha_1 + \cdots + \alpha_m = 1$, entonces si f es convexa tenemos que

$$f(\alpha_1 x_1 + \cdots + \alpha_m x_m) \leq \alpha_1 f(x_1) + \cdots + \alpha_m f(x_m) \tag{30}$$

Prueba: Demostraremos para $m = 3$, ya que la extensión se hace por inducción. Sea $\beta = \alpha_2 + \alpha_3$ luego $\frac{\alpha_2}{\beta} x_2 + \frac{\alpha_3}{\beta} x_3$ es una combinación convexa para los puntos x_2 y x_3. Pero se sigue cumpliendo que $\alpha_1 + \beta = 1$ con $\beta \geq 0$, concluimos entonces que

$$\begin{aligned} f(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3) &= f\left(\alpha_1 x_1 + \beta\left(\frac{\alpha_1}{\beta} x_2 + \frac{\alpha_3}{\beta} x_3\right)\right) \\ &\leq \alpha_1 f(x_1) + \beta f\left(\frac{\alpha_1}{\beta} x_2 + \frac{\alpha_3}{\beta} x_3\right) \\ &\leq \alpha_1 f(x_1) + \alpha_2 f(x_2) + \alpha_3 f(x_3) \end{aligned} \tag{31}$$

Por inducción se puede mostrar para $m \geq 3$.

Si a los pesos $\{\alpha_i\}$ los tomamos como probabilidades de los valores $\{x_i\}$, y a los valores $\{x_i\}$ los tomamos como valores de una cierta variable aleatoria, la desigualdad de Jensen toma la siguiente forma

$$E(f(X)) \geq f(E(X)) \tag{32}$$

Teorema: Sea f una función a valores reales definida sobre el subconjunto convexo $\mathcal{D}$ de $\mathbb{R}^n$ y sea g una función convexa definida sobre $I \subset \mathbb{R}$. Suponemos que $f(\mathcal{D}) \subset I$ y que g es una función no decreciente, luego la composición $g \circ f$ es convexa en $\mathcal{D}$.

PRUEBA: Siendo f convexa y g convexa y no decreciente tenemos

$$\begin{aligned} g(f(\alpha x + \beta y)) &\leq g(\alpha f(x) + \beta f(y)) \\ &\leq \alpha g(f(x)) + \beta g(f(y)) \end{aligned} \tag{33}$$

Mayorización: En este apartado daremos algunas de las definiciones de la teoría de mayorización. Esta teoría es muy extensa y tiene muchas aplicaciones.

Diremos que el vector $P = \{p_1, \cdots, p_n\}$ mayoriza al vector $Q = \{q_1, \cdots, q_n\}$ si

$$\sum_{i=1}^{n-1} p_i \geq \sum_{i=1}^{n-1} q_i \tag{34}$$

y

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i \tag{35}$$

Esto trae como consecuencia que para cualquier función convexa ϕ se tenga que

$$\sum \phi(p_i) \geq \sum \phi(q_i) \tag{36}$$

Esta propiedad será de mucha importancia para una parte de la investigación, específicamente en la definición de una entropía generalizada.

3.4 Elementos de Teoría de la Información

3.4.1 Entropía

La entropía es un concepto fundamental en la física. Aunque sea su interpretación estadística la que más nos interesa, es importante explicar el significado macroscópico de esta, ya que este fue el primero que se conoció. Es por eso que daremos una breve reseña del significado que tiene esta para la termodinámica. Entre los años 1700 y 1900 en plena era industrial la máquina de vapor era la principal protagonista. Se observaba que durante una parte del ciclo la máquina extraía calor de una fuente caliente y entregaba calor a una fuente fría. Es por eso que se decía que un motor funcionaba entre dos fuentes. Los científicos de la época enunciaron el segundo principio de la termodinámica de la siguiente manera: *Es imposible construir un motor que, funcionando según un ciclo, no produzca otro efecto que extraer calor de una fuente y realizar una cantidad equivalente de trabajo.* Fue Clausius [23] quien demostró que para un proceso reversible que va de un estado A hasta un estado B existe una función de estado, que llamó entropía y que cumple con

$$\int_a^b \frac{\delta Q}{T} = S(b) - S(a) \tag{37}$$

que en su forma diferencial tiene la siguiente expresión

$$\frac{\delta Q}{T} = dS \tag{38}$$

En equilibrio las variables termodinámicas toman los valores que maximizan a la entropía.

Fue el físico austríaco Ludwing Boltzmann [24] en la década de los setenta del siglo XIX quien pudo darle una interpretación estadística a la entropía. Sus ideas quedaron plasmadas en la famosa expresión

$$S = k \ln(W) \tag{39}$$

donde k es una constante y W es la cantidad de microestados del sistema. Una formulación alternativa a esta fue realizada por Josiah Willard Gibbs [25] quien expresó a la entropía de un sistema en la forma:

$$S = -k \sum_i p_i \ln p_i \tag{40}$$

donde p_i es la probabilidad del microestado i. Es esta formulación la que tendremos más en cuenta ya que, como se verá, guarda una cercana similitud con la entropía de información de Shannon.

3.4.2 Entropía de información y Teorema de Kinchin

Supongamos que tenemos un canal de comunicación que nos manda información mediante símbolos con cierta probabilidad. Esto nos permite repesentar nuestro canal mediante una sucesión de eventos $\{A_1, \cdots, A_n\}$ con sus respectivas probabilidades de ocurrencia $\{p_1, \cdots, p_n\}$, con $\sum_{i=1}^{n} p_i = 1$ y $p_i \geq 0$. Llamaremos a $H[p_1, \cdots, p_n] = -\sum_i p_i \log p_i$ *entropía de información* con la condición de que $p_i \log p_i = 0$ si $p_i = 0$. Se la puede interpretar como la incerteza en la información que envía el canal. Si sólo un símbolo es trasmitido, digamos el símbolo j, $p_j = 1$ y $p_i = 0$ para todo $i \neq j$. Entonces $H = 0$. A su vez si todos los símbolos tienen la misma probabilidad de ocurrencia, la entropía es máxima. Esta afirmación se puede demostrar usando la desigualdad de Jensen. En efecto, usando la convexidad de la función $\phi(x) = x \log(x)$,

$$\phi\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) \leq \frac{1}{n}\sum_i \phi(p_i) \tag{41}$$

tenemos que

$$\phi\left(\frac{1}{n}\right) = \frac{1}{n}\log\frac{1}{n} \leq \frac{1}{n}\sum_{i=1}^{n} p_i \log p_i = -\frac{1}{n}H[p_1, \cdots, p_n] \tag{42}$$

lo que nos da que

$$H[p_1, \cdots, p_n] \leq \log n = H\left[\frac{1}{n}, \cdots, \frac{1}{n}\right] \tag{43}$$

Otra importante propiedad resulta de tener otro canal de información con eventos $\{B_1, \cdots, B_n\}$ con sus respectivas probabilidades $\{q_1, \cdots, q_n\}$. Si estos canales son independientes entonces la información conjunta será igual a la suma de las dos informaciones por separado. Si los canales son independientes entonces la propabilidad conjunta también lo es, esto es, $\pi_{ik} = p_i q_k$

$$H(AB) = -\sum_{i,k} \pi_{ik} \log \pi_{ik} \sum_i \sum_k p_i q_k (\log p_i + \log q_k) = H(A) + H(B) \tag{44}$$

Ahora veremos que si en nuestro canal existe un evento con probabilidad cero, la entropía de información no cambia. Esto se ajusta a lo que uno en la cotidianeidad entiende de información. Se puede ver por definición que

$$H[p_1, \cdots, p_n, 0] = H[p_1, \cdots, p_n] \tag{45}$$

Relación entre la Entropía y la Información Mutua Se puede definir a la entropía condicional de la variable aleatoria Y dado X como la incertidumbre producida solamente por Y ya que se observó X. Su formulación matemática es la siguiente

$$H(Y|X) = -\sum_i p(X = x_i) \sum_j p(Y = y_j|X = x_i) \ln p(Y = y_j|X = x_i) \tag{46}$$

Esta definición nos permite ver que

$$H(Y|X) = H(X,Y) - H(X) \tag{47}$$

donde $H(X)$ es la entropía referida a la variable aleatoria X y $H(X,Y)$ se la puede interpretar como la incertidumbre producida por las dos variables en conjunto. Podemos escribir nuevamente esta ecuación de la siguiente manera

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y); \tag{48}$$

sumando y restando $H(Y)$ tenemos que

$$H(X,Y) = H(X) + H(Y) - H(Y) + H(Y|X); \tag{49}$$

Teniendo esto en cuenta podemos definir a la Información Mutua como

$$I(X:Y) \doteq H(Y) - H(Y|X) = H(X) - H(X|Y); \tag{50}$$

La cantidad $I(X:Y)$ representa la incertidumbre compartida entre las dos variables.

Se puede ver que si tomamos la misma variable aleatoria, es decir, calcular la auto-información mutua, tenemos

$$I(X:X) = H(X) - H(X|X) = H(X); \tag{51}$$

ya que $H(X|X) = 0$. Esto se debe a que la incertidumbre de la variable aleatoria X dado que se observó X tiene que ser 0.

3.4.3 Entropías Generalizadas

Por diversas motivaciones, se han introducido otras expresiones de incerteza. Entre estas, hay dos que particularmente nos resultarán de interés en el presente trabajo:

Entropía de Rényi: La entropía de Rényi [3] generaliza a la entropía de Shannon mediante un parámetro α. Es utilizada en el contexto de ecología como índice de diversidad y en mecánica cuántica como medida de entrelazamiento. Nosotros la tomaremos en el contexto de la Teoría de Información como una generalización de la entropía de Shannon para la obtención de divergencias generalizadas.

Sea $P = \{p_1, \cdots, p_n\}$ una distribución de probabilidad discreta de una variable aleatoria X, sea $\alpha > 0$ y $\alpha \neq 1$ definimos a la entropía de Renyi como

$$H_\alpha^R[P] = \frac{1}{1-\alpha} \cdot \log\left(\sum_{i=1}^{n} p_i^\alpha\right) \tag{52}$$

Se puede ver que $\sum_i p_i^\alpha > 1$ para $0 < \alpha < 1$ y es $\sum_i p_i^\alpha < 1$ para $\alpha > 1$, usando la definición (52) vemos que es no negativa para $\alpha > 0$. También se puede ver que si tomamos una distribución uniforme, es decir, $\{p_i\} = 1/n \ \forall i$ tenemos

$$H_\alpha^R[P] = \frac{1}{1-\alpha} \log\left(\frac{1}{N^\alpha} \sum_i\right) = \frac{1}{1-\alpha} \log\left(N^{1-\alpha}\right) = \log n \tag{53}$$

esto vale para cualquier α y es, además, el valor máximo de la entropía.

Supongamos que tenemos dos canales de comunicación independientes representados por una distribución de probabilidad $p_{ij} = p_i p_j$ se puede ver que

$$H_\alpha^R[P] = \frac{1}{1-\alpha} \log\left(\sum_{ij} p_i^\alpha p_j^\alpha\right) = \frac{1}{1-\alpha} \left(\log\left(\sum_i p_i^\alpha\right) + \left(\sum_i p_j^\alpha\right)\right) \tag{54}$$

Esto nos dice que es aditiva cuando las distribuciones de probabilidad son independientes.

Si derivamos respecto de α tenemos

$$\begin{aligned}
\frac{dH_\alpha^R[P]}{d\alpha} &= \frac{1}{(1-\alpha)^2} \log\left(\sum_j p_j^\alpha\right) + \frac{(1-\alpha)}{(1-\alpha)^2} \frac{\sum_i \log(p_i) p_i^\alpha}{\sum_j p_j^\alpha} = \\
&= \frac{1}{(1-\alpha)^2} \sum_i z_i \log\left(\sum_j p_j^\alpha\right) + \frac{(1-\alpha)}{(1-\alpha)^2} \sum_i z_i \log p_i = \\
&= \frac{1}{(1-\alpha)^2} \sum_i z_i \log\left(\frac{p_i^\alpha}{z_i}\right) + \frac{(1-\alpha)}{(1-\alpha)^2} \sum_i z_i \log p_i = \\
&= -\frac{1}{(1-\alpha)^2} \sum_i z_i \log\left(\frac{z_i}{p_i}\right)
\end{aligned} \tag{55}$$

donde $z_i = \frac{p_i^\alpha}{\sum_j p_j^\alpha}$. La última sumatoria es positiva ya que es la divergencia de Kullback-Leibler (en el próximo capítulo lo demostraremos), es decir, que la entropía de Renyi es no creciente respecto α.

Hay tres valores destacables del parámetro, $\alpha = 0$, $\alpha = 1$ y $\alpha = \infty$.

Se puede ver fácilmente que para $\alpha = 0$ se obtiene el valor máximo de entropía

$$H_0^R[P] = \frac{1}{1-0} \cdot \log\left(\sum_{i=1}^n p_i^0\right) = \log n \tag{56}$$

Si hacemos el límite para α tendiendo a 1 se obtiene la entropía de Shannon

$$\lim_{\alpha \longrightarrow 1} H_\alpha^R[P] = -\sum_i p_i \log p_i \tag{57}$$

En cambio, si realizamos el límite para $\alpha = \infty$ se obtiene el mínimo valor de entropía

$$H_\infty^R[P] = -\log(\max_i p_i) = min_i(-\log p_i) \tag{58}$$

Usando estos resultados podemos escribir la siguiente desigualdad

$$H_0^R \geq H_1^R \geq H_\infty^R \tag{59}$$

Otra propiedad importante es que si tenemos una distribución determinista, es decir, $P = (1, 0, \cdots, 0)$ la entropía de Renyi vale cero.

$$H_\alpha^R[P] = \frac{1}{1-\alpha}\log(1) = 0 \tag{60}$$

Un resultado importante relaciona la entropía de Renyi con la energía libre $F = -T\ln Z$ [27], donde Z es la función partición. Supongamos que tenemos un sistema físico en equilibrio a temperatura T_0 y la distribución de probabilidad de los estados está dada por

$$p_i = e^{-E_i/T_0} \tag{61}$$

Supongamos ahora que tenemos el sistema en otra temperatura T, nuestra distribución de probabilidad va a estar dada por

$$\frac{e^{-E_i/T}}{Z} \tag{62}$$

la función partición está dada por

$$Z = \sum_i e^{-E_i/T} \tag{63}$$

Si ponemos $\alpha = T_0/T$ tenemos que

$$H_{T_0/T}^R[P] = \frac{1}{1-T_0/T}\ln\left(\sum_i p_i^{T_0/T}\right) = \frac{T}{T-T_0}\ln\left(\sum_i e^{-E_i/T}\right) = -\frac{F}{T-T_0} \tag{64}$$

Entropía de Havrda-Charvat-Tsallis (HCT): La entropía de Havrda, Charvat y Tsallis ([28], [4]) también generaliza la entropía de Shannon mediante un parámetro. Se la utiliza en el ámbito de la Teoría de la Información para estudiar intercambios de información de sistema que están fuera del equilibrio.

Sea $P = \{p_1, \cdots, p_n\}$ una distribución de probabilidad discreta de una variable aleatoria X, sea $\alpha > 0$ y $\alpha \neq 1$ definimos a la entropía de Havrda, Charvat y Tsallis como

$$H_\alpha^T = \frac{1}{\alpha-1}\cdot\left(1 - \sum_k p_k^\alpha\right) \tag{65}$$

La entropía HCT no es aditiva para distribuciones independientes, es decir, si tenemos

una distribución $\pi = PQ$

$$\begin{aligned} H_\alpha^T(\pi) &= \frac{1}{\alpha-1}\left(1-\sum_{ij}^{\alpha}\pi_{ij}\right) = \frac{1}{\alpha-1}\left(1-\sum_i p_i^\alpha \sum_j q_j^\alpha\right) = \\ &= \frac{2}{\alpha-1} - \frac{1}{\alpha-1}\sum_i p_\alpha - \frac{1}{\alpha-1}\sum_j q_\alpha - \frac{1}{\alpha-1}\left(1-\sum_i p_i^\alpha - \sum_j q_j^\alpha - \sum_i p_i^\alpha \sum_j q_j^\alpha\right) = \\ &= \left(\frac{1}{\alpha-1}\left(1-\sum_i p_i^\alpha\right)\right) + \left(\frac{1}{\alpha-1}\left(1-\sum_j p_j^\alpha\right)\right) + (1-\alpha)H_\alpha^T(p)H_\alpha^T(Q) = \\ &= H_\alpha^T(P) + H_\alpha^T(Q) + (1-\alpha)H_\alpha^T(P)H_\alpha^T(Q) \end{aligned} \tag{66}$$

Como dijimos anteriormente la entropía HCT tiende a la entropía de Shannon, en el límite para $\alpha \longrightarrow 1$

$$\lim_{\alpha\longrightarrow 1}\frac{1}{\alpha-1}\left(1-\sum_i p_i^\alpha\right) = -\sum_i p_i \log p_i \tag{67}$$

Otra propiedad importante es que si uno tiene una distribución determinista, es decir, $P = (1, 0, \cdots, 0)$ la entropía HCT vale cero.

$$H_\alpha^T[P] = \frac{1}{\alpha-1}(1-1) = 0 \tag{68}$$

3.4.4 Divergencias

Una divergencia es una medida de disimilitud entre dos distribuciones de probabilidad. Es decir, es un funcional que como argumento tiene dos (o más) distribuciones de probabilidad, pudiendo ser simétrica o no respecto a sus argumentos. Nosotros trabajamos con divergencias entre distribuciones de probabilidad discreta. A este fin utilizaremos estas herramientas para el análisis de series temporales. En el contexto del análisis de señales las divergencias tienen muchos usos; uno de ellos es cuantificar cambios de la señal que a simple vista no se observan. Una divergencia tiene las siguientes propiedades

- $D(P||Q) \geq 0$
- $D(P||Q) = 0$ si y solo si $P = Q$
- $D(P||Q) \neq D(Q||P)$, si es no simétrica
- $D(P||Q) = D(Q||P)$, si es simétrica

donde P y Q son distribuciones de probabilidad discreta. Una de las divergencias más conocidas y usadas es la entropía relativa o Divergencia de Kullback-Leibler, definida como

$$D_{KL}(P||Q) = \sum_i p_i \log_2\left(\frac{p_i}{q_i}\right) \tag{69}$$

donde $P = \{p_i\}_{i=1}^N$ y $Q = \{q_i\}_{i=1}^N$ representan la distribución de probabilidad de N estados posibles de una variable aleatoria X. Esta divergencia es no simétrica y definida positiva. A esta divergencia se la puede interpretar como la cantidad de información extra (en bits) que uno necesita si observa un código basado en una distribución de probabilidad Q cuando en realidad esperaba uno basado en una distribución de probabilidad P. En los siguientes párrafos demostraremos la positividad de esta divergencia usando la desigualdad de Jensen.

$$\begin{aligned} -D_{KL}(P||Q) = -\sum_i p(x_i) \log\left(\frac{p(x_i)}{q(x_i)}\right) = \\ = \sum_i p_{x_i} \log\left(\frac{q(x_i)}{p(x_i)}\right) \leq \\ \leq \log \sum_i p(x_i) \frac{q(x_i)}{p(x_i)} = \\ = \log \sum_i q(x_i) = \\ = 0 \end{aligned} \tag{70}$$

es decir $D_{KL}(P||Q) \geq 0$ como queríamos demostrar (para más detalles ver [30]). Obviamente $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Podemos escribir a la D_{KL} en términos de la entropía de la siguiente manera

$$D_{KL}(P||Q) = -H[P] - \sum_i p_i \log q_i \tag{71}$$

En estos términos se entiende más porqué se la considera como un extra de bits respecto del modelo referencial P. También se puede ver de la definición que $D_{KL}(P||Q) = 0$ si y solo si $P = Q$. La mejor forma de verlo es mirando la desigualdad del segundo y tercer renglón de la ecuación (70), ya que la igualdad se cumple si y solo si $p = q$. La divergencia de K-L está definida si y solo si $q(x)$ es distinto de cero cuando $p(x)$ es igual a cero.

Ahora mostraremos la relación que hay entre la información mutua y la entropía relativa. Para ello necesitamos escribir a la información mutua de una manera diferente.

$$\begin{aligned} I(X:Y) = H(Y) - H(Y|X) = \\ = -\sum_i p_{y_i} \log p(y_i) + \sum_{i,j} p(x_j, y_i) \log(p(y_i|x_j)) = \\ = -\sum_{i,j} p(x_j, y_i) \log(p(y_i)) + \sum_{i,j} p(x_j, y_i) \log\left(\frac{p(x_j, y_i)}{p(x_j)}\right) = \\ = \sum_{i,j} p(x_j, y_i) \log\left(\frac{p(x_j, y_i)}{p(x_j)p(y_i)}\right) \end{aligned} \tag{72}$$

Esto nos da que

$$I(X:Y) = \sum_{i,j} p(x_j, y_i) \log\left(\frac{p(x_j, y_i)}{p(x_j)p(y_i)}\right) \tag{73}$$

Teniendo esto podemos establecer la relación con la entropía relativa

$$I(X:Y) = D_{KL}(P(X,Y)||P(X)P(Y)) \tag{74}$$

Visto de otro modo es la distancia entre la probabilidad conjunta y la probabilidad conjunta si los eventos fuesen independientes. La divergencia de K-L también se puede expresar como el valor de expectación:

$$E_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} \tag{75}$$

Asociada con la entropía de Renyi, se puede introducir la divergencia de Renyi

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left(\sum_i \frac{p_i^\alpha}{q_i^{\alpha-1}} \right) \tag{76}$$

para valores de alfa $0 < \alpha < \infty$ y $\alpha \neq 1$. Cuando realizamos el límite de alfa tendiendo a 1 se obtiene la divergencia de Kullcack-Leibler.

La divergencia de Kullback-Leibler pertenece a una familia de divergencias conocidas como Csiszár [31] o divergencias-f, definidas como

$$D_f(P||Q) = \sum_i p_i f\left(\frac{q_i}{p_i}\right) \tag{77}$$

donde f es una función convexa, y además se tiene que satisfacer $f(1) = 0$ y $f''(1) = 1$. Se puede demostrar usando la desigualdad de Jensen que esta divergencia es positiva

$$D_f(P||Q) = E_P(f(X)) \geq f(E(X)) = f(1) = 0 \tag{78}$$

También se puede ver que es igual a cero si y solo si $P = Q$. Si $P = Q$ tenemos

$$D_f(P||P) = \sum_i p_i f(1) = 0 \tag{79}$$

ya que por definición $f(1) = 0$.

La reversa la dejaremos para el apéndice. Una expansión de Taylor puede mostrar el comportamiento cuadrático de las divergencias Csiszar cuando estudiamos distribuciones de probabilidad cercanas

$$D_f(P||P + \delta P) = \frac{1}{2} \sum_i \frac{\delta p_i^2}{p_i} + o\left(\sum_i \delta p_i^2 \right) \tag{80}$$

Una divergencia muy utilizada en el contexto del análisis de señales [32] y en otros campos de la física [33] es la divergencia de Jensen Shannon (D_{JS}). Esta se puede relacionar con D_{KL} de la siguiente manera

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right) \tag{81}$$

es decir, es una simetrización de la divergencia D_{KL}. Como se puede ver corrige el problema que tenía la D_{KL} de que q_i no podía valer cero. A la distribución $M := \frac{P+Q}{2}$ se la entiende como distribución mezcla. Otra manera de escribir la D_{JS} es mediante la entropía de Shannon dada por $H[P] = -\sum_i p_i \log_2 p_i$

$$D_{JS}(P||Q) = H\left[\frac{P+Q}{2}\right] - \frac{1}{2}H[P] - \frac{1}{2}H[Q] \tag{82}$$

La demostración de que la divergencia de la Jensen-Shannon cumple con las propiedades correctas referidas a una divergencia simétrica lo dejaremos para el capítulo de la divergencia Gamma que es parte de nuestra investigación. Esto se debe a que la divergencia de Jensen-Shannon es un caso particular de la divergencia Gamma.

Uno puede generalizar el concepto de distribución mezcla para pesos diferentes a $1/2$ y más de dos distribuciones, es decir, $M := \sum_{i=1}^{n} \pi_i P_i$ con $\pi_1, \pi_2, \cdots \pi_n \geq 0$ y $\sum_i \pi_i = 1$. En este sentido la divergencia de Jensen-Shannon construida para una mezcla generalizada es

$$D_{JS}^{\pi_1,\cdots,\pi_n}(P_1, \cdots, P_n) = H\left(\sum_{i=1}^{n} \pi_i P_i\right) - \sum_{i=1}^{n} \pi_i H(P_i) \tag{83}$$

La divergencia de Jensen-Shannon tiene una propiedad muy importante en el contexto geométrico, y es que la raiz cuadrada de D_{JS}

$$\sqrt{D_{JS}(P||Q)} \tag{84}$$

cumple con la desigualdad triangular, es decir, que es una métrica.

4 Capítulo III: Problemática de las distancias entre distribuciones de probabilidad

4.1 Consecuencias estadísticas de tener una sola secuencia de muestra

En algunas ocasiones es posible realizar experimentos controlados (repetibles varias veces bajo idénticas condiciones iniciales) que permiten obtener varias realizaciones particulares del mismo proceso estocástico. Sin embargo, esta no es la situación más común. Casi siempre la serie temporal se refiere a un período muestral que tan solo es una parte de la historia del proceso estocástico que subyase en dicha serie. En lo que sigue haremos un resumen de la estrategia utilizada en los casos de tener solo una secuencia pero desde la perspectiva de la Inferencia Estadística y no desde la Teoría de la Información. Esto nos ayudará a entender porqué elegimos el enfoque previsto por la Teoría de la Información.

En el Análisis de Series Temporales es de interés poder inferir el valor que tomará la variable aleatoria X_N dado los datos observados $(x_1, ..., x_{N-1})$. Cuando tenemos una sola cadena nos vemos obligados a hacer ciertas suposiciones tales como la estacionareidad y ergodicidad del proceso subyacente a la serie temporal.

Cuando el proceso es estacionario los parámetros μ, γ_0,γ_1,γ_2,..., es decir, la media y la autocovarianza, pueden estimarse a partir de una sola cadena de observaciones. Para una muestra $x_0, ..., x_n$ asociada al proceso estocástico $\{X_t\}$ los estimadores son

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=0}^{n} x_i \tag{85}$$

$$\hat{\gamma}_k = \frac{1}{n}\sum_{i=0}^{n-k}(x_i - \hat{\mu}_X)(x_{i+k} - \hat{\mu}_X) \tag{86}$$

Una pregunta a hacerse es si los estimadores son consistentes con los valores reales de la media (μ) y de la autocovarianza (γ_k, para cualquier k). Esta pregunta se responde si imponemos la condición de ergodicidad en el proceso. Pero esto no es suficiente ya que necesitamos saber el proceso subyacente a la muestra. Este problema se puede solucionar imponiendo un modelo.

No resolveremos aquí de manera exhaustiva este problema; solo daremos un bosquejo de como se resolvería desde la perspectiva de la Inferencia Estadística. Como se verá en la secciones siguientes, la *metodología* usada, en conjunto con las herramientas de teoría de la información, nos permitirán estudiar la distinguibilidad entre dos series temporales teniendo una sola secuencia de muestra.

4.2 Distancias entre distribuciones de probabilidad

Desde la ciencia se ha intentado siempre caracterizar objetos y procesos. Para caracterizarlos o diferenciarlos es siempre necesario un cuantificador, es decir, algo que dado ciertas cantidades nos de un número de referencia. Desde la teoría de la información se ha buscado cuantificadores de distiguibilidad entre distribuciones de probabilidad con el objetivo de distinguir fragmentos de una señal, estados cuánticos o procesos estadísticos, por ejemplo. Para definir a un cuantificador de distinguibilidad recurrimos a la definición de métrica de un espacio métrico. Una métrica nos da una referencia de cuán distintos son dos elementos de un espacio métrico y es por eso que nos basamos en esta definición. En esta tesis se presentan dos casos en los que las divergencias son métrica, uno como la raiz cuadrada de la divergencia de Jensen-Shannon, y el otro corresponde a ciertos ejemplos de divergencias Gamma.

Como se vió en el capítulo anterior las herramientas presentadas se aplicarán a distribuciones de probabilidad discreta. Esto, a su manera, nos da un marco de referencia de que tipos de problemas podemos abordar en el contexto del análisis de señales. Durante la tesis se aplicarán las herramientas desarrolladas a señales obtenidas por procesos reales o señales simuladas por un programa, sea cual sea el caso, es importante aclarar ciertas cuestiones que van más allá del origen de la señal. Como dijimos anteriormente el objetivo primario es obtener una distribución de probabilidad discreta y para eso es necesario que nuestra señal contenga un alfabeto finito. A esta señal la llamaremos cadena simbólica. Sea $\mathcal{C} = \{c_1, \cdots, c_N\}$ una cadema de símbolos donde N es un número finito, y con $\{c_i \in \mathcal{A}\}_{i \leq N}$, donde $\mathcal{A}$ es un alfabeto finito. Ya teniendo a nuestra disposición la cadena simbólica es posible calcular la distribución de probabilidad discreta en base a un algoritmo frecuentista. Esto es principalmente una consecuencia de que estamos trabajando con alfabetos finitos. Este sería el paso básico en los estudios que haremos, las herramientas que utilizaremos dependerán de forma exclusiva de esta función.

Obviamente no siempre tenemos en nuestras manos una secuencia simbólica. En muchos casos, tanto en señales reales como simuladas, los datos obtenidos son números reales. Si lo expresamos matemáticamente, definimos a $\mathcal{S} = \{s_1, \cdots, s_N\}$ como una secuencia de números reales. En este caso la situación es bastante diferente ya que no tenemos un alfabeto conocido. El análisis de secuencias de números reales se hará de una manera indirecta. Es decir, la idea sería poder transformar esta secuencia de números reales en una cadena o secuencia de símbolos. A este tipo de transformación se la llama mapeo. Es aquí donde surgen varios interrogantes: ¿perdemos información al realizar el mapeo?, si es así ¿qué ventajas traería realizar el mapeo? ¿Qué perdemos y qué ganamos al hacer el mapeo? Como es de imaginar perdemos información cuantitativa. Esto se debe a que no existe una transfomación o función unívoca entre los reales y los enteros. Hacemos la relación con los enteros porque cualquier alfabeto finito tiene un mapa unívoco con los enteros. Pero si perdemos información cuantitativa ¿por qué buscar un mapeo a una secuencia simbólica? Podríamos decir, erróneamente, que es porque usamos distribuciones

de probabilidad discreta, pero esto en realidad es la consecuencia. La verdadera razón está en lo que buscamos de la secuencia de números reales, y esto es, su estructura, su dinámica, sus fluctuaciones, sus cambios. No estamos interesados, en primera instancia, en los valores que toma la señal de números reales sino en analizarla desde otra perspectiva. Es por eso que si el mapeo es el adecuado podrá plasmar todas estas características. En consecuencia, obtenemos un secuencia de símbolos que representa en alguna medida a la señal original pero a la cual se le es posible calcular o estimar su distribución de probabilidad discreta. Las herramientas que se presentaron en el marco teórico y las que se prensentarán en los futuros capítulos son básicamente medidas de distinguibilidad entre dos distribuciones. Pero el trasfondo de esto es ver que los cuantificadores van a depender de estas propiedades (estructura, dinámica, etc.). Si mediante el mapeo nuestra distribución de probabilidad representa de algún modo cierta estructura y dinámica, la medida de distinguibilidad entre dos distribuciones cuantificará diferencias de estructura o dínamica de las series. Como se verá en las aplicaciones se podrá distinguir cambios cualitativos importantes en señales de sueño, o en ECG, o en señales simuladas de mapas caóticos y ruidos.

5 Capítulo IV: Aspectos Metodológicos

Daremos aquí algunos detalles de cómo implementaremos las herramientas teóricas desarrolladas en la tesis. Sea $\mathcal{S} = \{s_1, \cdots, s_N\}$ una secuencia de números reales. Como aclaramos en la sección anterior estamos interesados en distinguir propiedades tales como la estructura y la dinámica de una serie temporal. Es por eso que se recurre a una transformación o mapeo de esta secuencia de números reales a un secuencia simbólica. Pero aún teniendo la cadena simbólica si el mapeo no es el adecuado la obtención de información de estructura se puede perder. Supongamos que tenemos dos secuencias $\mathcal{C}_1 = \{0, 0, 1, 1\}$ y $\mathcal{C}_2 = \{0, 1, 0, 1\}$. Como se puede ver estructuralmente son distintas pero tienen la misma distribución de probabilidad de ocurrencia de los símbolos 0 y 1. Esto es un problema ya que la entropía será la misma y las medidas de distinguibilidad las tomarán como idénticas. Es por eso que a veces se recurre a otro tipo de mapeo como veremos en el capítulo 8. Esto nos deja una enseñanza importante: la distribución de probabilidad no distingue estructura de por sí, es necesario un mapeo adecuado y como se verá más adelante un proceso de segmentación. Pero para aclarar o resumir podemos decir que hay tres pasos básicos: el mapeo a un secuencia simbólica, la segmentación y la posterior obtención de la distribución de probabilidad. En los párrafos siguientes detallaremos uno de los mapeos que más usaremos por su simpleza y eficacia: el mapeo de Bandt y Pompe [34].

5.1 Mapeo de Bandt y Pompe

El mapeo de Bandt y Pompe consiste en tomar una serie temporal $\mathcal{X} = \{x_1, \cdots, x_N\}$ y transformarla a una secuencia simbólica $\mathcal{C}$ de la siguiente manera. Sea una serie de tiempo discreta de valor real $\{x(t)\}_{t\geq 0}$ y sean dos enteros $d \geq 2$ y $\tau \geq 1$, siendo d la *dimensión de inmersión* y τ el *tiempo de retardo* respectivamente. Lo que haremos será armar vectores llamados vectores de permutación, que tienen una dimensión d. El parámetro τ representa la distancia al vecino que queremos tomar en cuenta. Formalmente se escribe de la siguiente manera: A partir de la serie de tiempo original se introduce el vector d-dimensional $Y^t_{d,\tau}$

$$Y^t_{d,\tau} \to (x_{t-(d-1)\tau}, \cdots, x_{t-\tau}, x_t), \qquad t \geq (d-1)\tau \tag{87}$$

Nos moveremos a lo largo de cada posición dentro de la serie de tiempo y así formaremos $N - (d-1)\tau$ vectores d-dimensionales $Y^t_{d,\tau}$. Existen condiciones tanto para d como para τ con el fin de que el vector $Y^t_{d,\tau}$ conserve las propiedades dinámicas del sistema completo (teorema de Takens). A continuación, las componentes de la trayectoria del espacio de la fase $Y^t_{d,\tau}$ se ordenan de forma ascendente. Entonces se puede definir un vector permutación $\Pi^t_{d,\tau}$, cuyas componentes son las posiciones de los valores ordenados de $Y^t_{d,\tau}$. Como ejemplo, se toma la serie temporal $x_t = (1.7; 2.1; 1.5; 1.4; 2)$ y se aplica a ella el mapeo de Bandt y Pompe para $d = 3$ y $\tau = 1$. De este procedimiento se obtienen los vectores $Y^t_{d,\tau}$ correspondientes a la serie X_t que son $Y^1_{3,1} = (1.7; 2.1; 1.5)$, $Y^2_{3,1} = (2.1; 1.5; 1.4)$

y $Y^3_{3,1} = (1.5; 1.4; 2)$, y los vectores de permutación correspondientes son $\Pi_1 = (1; 2; 0)$, $\Pi_2 = (2; 1; 0)$ y $\Pi_3 = (1; 0; 2)$. Cada uno de estos vectores representa un patrón (o forma). Existen d! posibles patrones. Luego se calcula de forma frecuentista las probabilidades de aparición de cada uno de los patrones, es decir, la distribución de probabilidad. En la figura 1 se ejemplifica el método.

Podemos explicar el mapeo como una receta. Supongamos que tenemos una secuencia de números reales $\mathcal{X} = \{x_1, \cdots, x_N\}$, partimos del casillero 1, es decir, de x_1, y armamos un vector de dimensión d con los siguientes $(d-1)$ vecinos. Esto se debe a que τ es igual a 1. Pero si esto no fuese así, si por ejemplo tenemos $\tau = 2$, tomaríamos d números eligiendo uno por medio, es decir, estaríamos en la secuencia a $(d-1)2$ casilleros más adelante. Esto se puede generalizar para cualquier entero $\tau \geq 1$, es decir, que si nos paramos en el casillero t y armamos el vector estaríamos en el casillero $t + (d-1)\tau$. Pero volvamos al caso más común, y el que vamos a utilizar, en el que $\tau = 1$. En este caso ya armado el vector $Y^1_{(d,\tau)}$, armamos otro vector llamado vector de permutación en el cual cada casillero tiene la posición de una lista de orden creciente de magnitud (empezando en 0). Existirán $d!$ diferentes formas de patrones del vector de permutación, a cada una de esas formas le asignamos una letra o un número de un alfabeto finito $\mathcal{A}$. A esta letra o número lo denominaremos a_i con $i \leq d!$. Teniendo ya este vector, que lo denominamos $\Pi^1_{(d,\tau)}$, corremos el cursor para el casillero 2. Así obtenemos el vector $Y^2_{(d,\tau)}$, y por lo tanto, el vector $\Pi^2_{(d,\tau)}$. El algoritmo sigue hasta que el cursor toca el casillero $t = (N - (d-1)\tau)$. Ya habiendo recorrido toda la serie vamos formando una secuencia símbolica de largo $N - (d-1)\tau$ en la cual en cada casillero habrá una letra a_j perteneciente al alfabeto finito $\mathcal{A}$. Teniendo ya esta secuencia simbólica podemos estimar de forma frecuentista la distribución de probabilidad de los elementos de $\mathcal{A}$. Teniendo la distribución de probabilidad discreta podemos aplicar las herramientas de la teoría de la información.

Hay una deficiencia en el algoritmo de Bandt y Pompe, y tiene que ver sobre qué sucede si hay dos números iguales en el vector $Y^t_{d,\tau}$ (para algún tiempo t), pues en la lógica utilizada por el algoritmo se necesita que todos los números del vector $Y^t_{d,\tau}$ sean ditintos. Lo que se hace en el algoritmo es hacer una excepción y dar un orden creciente aunque no exista. Por ejemplo, si tenemos el caso $Y^t_{d,\tau}$ de forma $(2.1; 5; 3.2; 5)$, el vector permutación es $(0; 2; 1; 3)$. Es decir, el primero de los números iguales es el menor en el vector permutación.

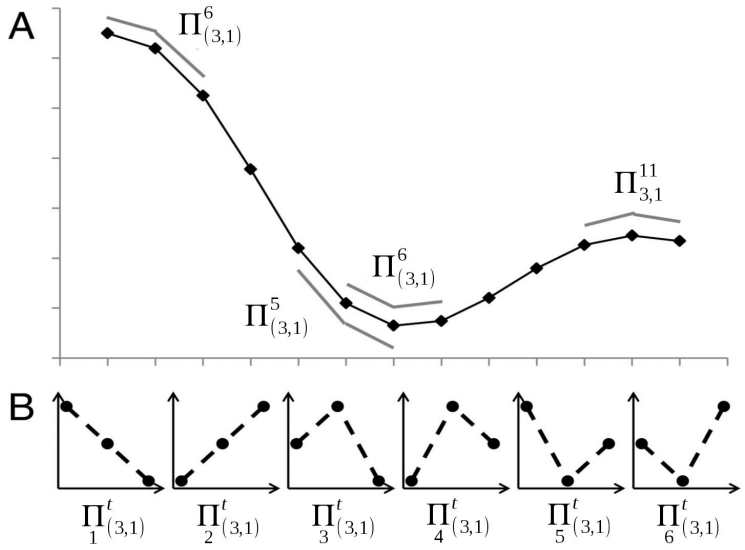

Figura 1. Ejemplo gráfico del método del mapeo de Bandt y Pompe. En la parte (A) se muestra como se generan los vectores de permutación $\Pi^t_{(\tau,d)}$ a lo largo de la señal a analizar. En este caso particular para cada tiempo t de la señal tomamos $d = 3$ y $\tau = 1$. Se ve como van apareciendo los patrones a medida que la señal evoluciona. En la parte (B) para una dimensión de inmersión $d = 3$ se tienen $d! = 6$ diferentes posibles patrones.

5.2 Diferentes tipos de segmentación: Ventana y Puntero

Es importante primero entender a qué nos referimos con segmentación. Cuando analizamos una señal y queremos detectar cambios en su estructura o dinánica no alcanza con realizar el mapeo de Bandt y Pompe ya que luego haríamos estadística sobre toda la secuencia entera. Es decir, tendríamos una distribución de probabilidad de los patrones de toda la secuencia y no por partes. Si queremos realizar un análisis de cambios de estructura y dinámica en ciertas regiones de la señal es necesario segmentar. Esto nos permitirá encontrar cambios locales dentro de la secuencia. Estos cambios locales o por regiones de la secuencia son invisibles si hacemos estadística sobre toda la señal. Para este tipo de análisis entonces es que se decide analizar la secuencia por regiones como es el caso del método de la ventana. También se puede hacer un análisis dinámico dividiendo la secuencia en dos por un cursor y mover el cursor desde el inicio de la secuencia hasta el final como es el caso del método del puntero. Empecemos analizando el método de la ventana.

Supongamos que ya hemos mapeado nuestra señal con algún método, por ejemplo el de Bandt y Pompe. Llamemos $\mathcal{C} = \{c_1, \cdots, c_N\}$ a esta secuencia de símbolos. Tomemos una sección de esta secuencia de largo $L = L_1 + L_2 + 1 = 2n + 1$, con $L_j = n$. A esta sección la llamaremos ventana. El hecho que tenga largo $2n + 1$ es porque el centro de la ventana no se tiene en cuenta y necesitamos que cada lado de la ventana tenga el mismo

largo, es decir, de largo n. El algoritmo sería el siguiente. En un principio la ventana estará centrada en el casillero $n+1$ de la secuencia, es decir, c_{n+1} y tendrá de cada lado n casilleros a la izquierda y a la derecha. Esto significa que el borde izquierdo de la ventana estará en el principio de la secuencia, o sea, en c_1. Luego lo que hacemos es calcular de forma frecuentista la distribución de probabilidad de los símbolos del lado izquierdo y derecho de la ventana. Llamararemos DPI_{n+1} y DPD_{n+1} distribución de probabilidad del lado izquierdo y distribución de probabilidad del lado derecho respectivamente, referidas al centro de la ventana c_{n+1}. Teniendo esto estamos en condiciones de utilizar las herramientas de la teoría de la información que nos permiten distinguir entre dos distribuciones, por ejemplo la divergencia de Jensen-Shannon. Guardamos el valor obtenido por la divergencia y procedemos a correr el centro de la ventana un casillero, esto es, a c_{n+2} y a hacer el mismo procedimiento. Es decir, calcular las distribuciones de probabilidad en cada lado de la ventana, DPI_{n+2} y DPD_{n+2}. El hecho de mover el centro de la ventana implicará que las distribuciones probabilidad de ocurrencia de los símbolos de cada lado de la ventana cambien, y en consecuencia el valor de la divergencia. Realizamos el mismo algoritmo hasta que el centro de la ventana llegue a $c_{N-(n+1)}$, es decir, hasta que el borde derecho de la ventana esté en el final de la secuencia. Con este método de segmentación podemos ver como es el cambio de la estructura de la secuencia por regiones y distinguir en que parte tiene cambios más bruscos o repentinos y comparar con otras regiones de la secuencia.

Otro método de segmentación con el que trabajaremos, se conoce como método del puntero. El procedimiento es el siguiente. Primero colocamos el cursor o puntero en algún casillero que esté después del primer elemento c_1, supongamos que está en algún casillero cualquiera c_i con $i \geq 1$. En consecuencia nos quedan dos subsecuencias I y D a la izquierda y a la derecha del puntero respectivamente. Como el puntero puede estar en cualquier casillero de la secuencia el largo de las subsecuencias va a ser distinto. Entonces si quisiéramos calcular la distribución de probabilidad de cada subsecuencia hay que tener en cuenta esto. La solución a este problema es asignarle pesos estadísticos a cada subsecuencia. El peso estadístico de la subsecuencia I es L_I/N, donde L_I es el largo de la subsecuencia de la izquierda y N el largo total de la secuencia. A continuación se calculan las distribuciones de probabilidad de ocurrencia de símbolos P_I y P_D referidas a las subsecuencias de izquierda y derecha respectivamente. Para cuantificar la diferencia entre las dos distribuciones podemos utilizar, por ejemplo, la siguiente divergencia

$$D_{JS}(P_I||P_D) = H\left(\pi_I P_I + \pi_D P_D\right) - \pi_I H(P_I) + \pi_D H(P_D) \tag{88}$$

es decir, la divergencia de Jensen-Shannon *generalizada* para dos distribuciones. El algoritmo a seguir será empezar en el casillero 2 e ir moviendo el puntero y calcular la divergencia para todas las posiciones del puntero hasta llegar al casillero $N-1$. La idea es almacenar los valores que la divergencia va tomando y comparar la diferencia de valores para diferentes posiciones del puntero.

6 Capítulo V: Divergencia tipo Bregman

Presentaremos aquí una divergencia que depende de entropías generalizadas. El capítulo se presentará de forma cronológica en referencia a los resultados obtenidos ya que de esta manera podremos ver las motivaciones y las causas de porqué tomamos ciertas decisiones en la investigación. Como se verá esta divergencia surge de una interpretación de la divergencia Kullback-Leiber y concluye relacionándose con la divergencia de Bregman.

Este capítulo consta principalmente de un aporte teórico, pero para investigar la eficacia de estas divergencias, concluiremos el capítulo aplicándolas a distintas series temporales tanto simuladas como de origen natural. En el caso de series temporales de origen natural elegimos señales correspondientes a un electrocardiograma (ECD), una de ellas pertenece a una fibrilación auricular y la otra a un estadío normal del corazón. En la sección de aplicaciones de este capítulo detallaremos el origen de las señales para un mejor entendimiento de los resultados obtenidos.

6.1 Visión alternativa de la D_{KL}

Podemos pensar a la Divergencia de Kullback-Leibler (D_{KL}) como un operador en función de la entropía de Shannon. Para esto partimos de su definición:

$$D_{KL}\Big(P||Q\Big) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) = \sum_i p_i \left(\log(p_i) - \log(q_i)\right) \tag{89}$$

Teniendo en cuenta la definición de la Entropía de Shannon

$$H\Big(P\Big) = -\sum_j p_j \log(p_j) \tag{90}$$

La divergencia de K-L toma la forma

$$D_{KL}\Big(P||Q\Big) = -H\Big(P\Big) - \sum_i p_i \log(q_i) \tag{91}$$

y si calculamos la derivada parcial de esta entropía tenemos que

$$\frac{\partial}{\partial p_i}\Big(H\Big(P\Big)\Big) = -\left(\log(p_i) + 1\right) \tag{92}$$

Por lo tanto

$$D_k\Big(P||Q\Big) = \sum_i p_i \left(\log(p_i) + 1 - 1 - \log(q_i)\right) = \sum_i p_i \left(-\frac{\partial H\Big(P\Big)}{\partial p_i} + \frac{\partial H\Big(Q\Big)}{\partial q_i}\right) \tag{93}$$

Finalmente obtenemos la siguiente expresión

$$D_{KL}\big(P||Q\big) = \sum_i p_i \left(\frac{\partial H\big(Q\big)}{\partial q_i} - \frac{\partial H\big(P\big)}{\partial p_i} \right) \tag{94}$$

que es la divergencia de K-L expresada como un operador en función de la entropía de Shannon. Esta es otra manera de ver e interpretar a la D_{KL}. Una pregunta interesante es si podemos tomarlo como un operador que es función de diferentes entropías.

$$D_{H_\alpha}(P||Q) = \sum_i p_i \left(\frac{\partial}{\partial p_i}\bigg|_{p_{i=q_i}} - \frac{\partial}{\partial p_i}\bigg|_{p_{i=p_i}} \right) [H_\alpha] \tag{95}$$

¿Seguirá siendo una divergencia? Para empezar y como ejemplo de prueba tomaremos el operador anterior y le aplicaremos la entropía de Renyi, esto nos ayudará a dilucidar la respuesta.

El operador resultante es la siguiente función

$$D_{H_\alpha}\left(P||Q\right) \doteq \sum_{i=1}^{N} p_i \left(\frac{\partial H_\alpha}{\partial q_i} - \frac{\partial H_\alpha}{\partial p_i} \right) \tag{96}$$

donde H_α es la Entropía de Renyi, o sea

$$H_\alpha\left(P\right) = \frac{1}{1-\alpha} \log \left(\sum_{j=1}^{N} p_j^\alpha \right) \tag{97}$$

Es necesario preguntarse entonces si la función (96) tenderá a la divergencia de K-L cuando el parámetro α tienda a 1. Vale recalcar que

$$\lim_{\alpha\to 1} \left[\frac{\partial H_\alpha}{\partial p_i} \right] \neq \frac{\partial}{\partial p_i} \left(\lim_{\alpha\to 1} H_\alpha \right) \tag{98}$$

Teniendo en cuenta esto calculemos el límite de α tendiendo a 1 para D_{H_α}. Primero veremos qué pasa si introducimos el límite dentro de la sumatoria respecto de i de la función (96).

$$\lim_{\alpha\to 1} D_{H_\alpha}\left(P||Q\right) = \lim_{\alpha\to 1} \sum_{i=1}^{N} p_i \left(\frac{\partial H_\alpha}{\partial q_i} - \frac{\partial H_\alpha}{\partial p_i} \right) = \sum_{i=1}^{N} p_i \left(\lim_{\alpha\to 1} \left[\frac{\partial H_\alpha}{\partial q_i} \right] - \lim_{\alpha\to 1} \left[\frac{\partial H_\alpha}{\partial p_i} \right] \right) \tag{99}$$

Ahora calculemos el límite

$$\lim_{\alpha\to 1} \left[\frac{\partial H_\alpha}{\partial p_i} \right] = \lim_{\alpha\to 1} \left[\frac{\partial}{\partial p_i} \left(\frac{1}{1-\alpha} \log \left(\sum_{j=1}^{N} p_j^\alpha \right) \right) \right] = \lim_{\alpha\to 1} \left[\frac{1}{1-\alpha} \frac{\partial}{\partial p_i} \left(\log \left(\sum_{j=1}^{N} p_j^\alpha \right) \right) \right] \tag{100}$$

sabiendo que

$$\frac{\partial}{\partial p_i}\left(\log\left(\sum_{j=1}^{N} p_j^{\alpha}\right)\right) = \frac{\alpha \cdot p_i^{\alpha-1}}{\sum_{j=1}^{N} p_j^{\alpha}} \tag{101}$$

Entonces el límite toma la forma

$$\lim_{\alpha\to 1}\left[\frac{\partial H_\alpha}{\partial p_i}\right] = \lim_{\alpha\to 1}\left[\frac{1}{1-\alpha}\cdot\frac{\alpha \cdot p_i^{\alpha-1}}{\sum_{j=1}^{N} p_j^{\alpha}}\right] \tag{102}$$

Pero si $p_i \neq 0$ para todo $i : 1, \ldots, N$; entonces

$$\lim_{\alpha\to 1} p_i^{\alpha-1} = 1 \tag{103}$$

y también

$$\lim_{\alpha\to 1}\sum_{j=1}^{N} p_j^{\alpha} = \sum_{j=1}^{N} p_j = 1 \tag{104}$$

Luego tenemos que

$$\lim_{\alpha\to 1}\left[\frac{\partial H_\alpha}{\partial p_i}\right] = \infty \tag{105}$$

Como dijimos anteriormente si cambiamos el límite por la derivada tenemos

$$\frac{\partial}{\partial p_i}\left(\lim_{\alpha\to 1}\left[H_\alpha\right]\right) = \frac{\partial\left(H\right)}{\partial p_i} \tag{106}$$

donde H es la Entropía de Shannon. Entonces tenemos que

$$\frac{\partial}{\partial p_i}\left(\lim_{\alpha\to 1}\left[H_\alpha\right]\right) = \frac{\partial\left(H\right)}{\partial p_i} = \frac{\partial}{\partial p_i}\left(-\sum_{j=1}^{N}\log(p_j)p_j\right) = -\Big(\log(p_i)+1\Big) \tag{107}$$

Esto es un problema porque no solo no conmutan sino que el límite que nosotros necesitamos diverge.

Ahora queremos hacer el mismo análisis pero sin meter los límites dentro de la sumatoria en i de la ecuación (96), es decir,

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = \lim_{\alpha\to 1}\sum_{i=1}^{N} p_i\left(\frac{\partial H_\alpha}{\partial q_i} - \frac{\partial H_\alpha}{\partial p_i}\right) \tag{108}$$

Si calculamos las derivadas parciales tenemos que

$$\frac{\partial H_\alpha}{\partial q_i} = \frac{1}{1-\alpha}\cdot\frac{\alpha \cdot q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^{\alpha}} \tag{109}$$

y para p_i tenemos

$$\frac{\partial H_\alpha}{\partial p_i} = \frac{1}{1-\alpha} \cdot \frac{\alpha \cdot p_i^{\alpha-1}}{\sum_{j=1}^{N} p_j^\alpha} \tag{110}$$

Reemplazando en la ec. (108) obtenemos la siguiente expresión

$$\lim_{\alpha \to 1} D_\alpha \left(P||Q\right) = \lim_{\alpha \to 1} \left[\frac{\alpha}{1-\alpha} \cdot \sum_i p_i \left(\frac{q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^{\alpha-1}}{\sum_{j=1}^{N} p_j^\alpha} \right) \right] \tag{111}$$

Colocando p_i adentro del paréntesis tenemos

$$\lim_{\alpha \to 1} D_\alpha \left(P||Q\right) = \lim_{\alpha \to 1} \left[\frac{\alpha}{1-\alpha} \cdot \sum_i \left(\frac{p_i q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \right) \right] \tag{112}$$

Como se puede ver es un límite indeterminado de la forma

$$\lim_{\alpha \to 1} D_\alpha \left(P||Q\right) = \frac{0}{0} \tag{113}$$

Pero si aplicamos la regla de L'Hopital tenemos que

$$\lim_{\alpha \to 1} D_\alpha \left(P||Q\right) = \lim_{\alpha \to 1} \left[\frac{\frac{d}{d\alpha} \left(\alpha \cdot \sum_i \left[\frac{p_i q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \right] \right)}{\frac{d(1-\alpha)}{d\alpha}} \right] \tag{114}$$

Por simplicidad de cálculos es conveniente renombrar a la sumatoria que está arriba de la división como

$$A \doteq \sum_i \left[\frac{p_i q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \right] \tag{115}$$

Es visible ahora que lo que estamos haciendo es

$$\lim_{\alpha \to 1} D_\alpha \left(P||Q\right) = \lim_{\alpha \to 1} \left[\frac{\frac{d}{d\alpha}\Big(\alpha A\Big)}{\frac{d}{d\alpha}(1-\alpha)} \right] \tag{116}$$

Los detalles de este límite se encuentran en el Apendice A.

Tras algunos cálculos simples se llega a

$$\begin{aligned} \lim_{\alpha \to 1} D_\alpha \left(P||Q\right) = \lim_{\alpha \to 1} \Bigg\{ & \frac{\sum_i p_i q_i^{\alpha-1}}{\sum_e q_e^\alpha} - 1 - \frac{\alpha \cdot \sum_i p_i q_i^{\alpha-1} \cdot ln(q_i)}{\sum_e q_e^\alpha} + \\ & + \frac{\alpha \cdot \sum_e q_e^\alpha \cdot ln(q_e) \sum_i p_i q_i^{\alpha-1}}{\sum_e q_e^\alpha} + \frac{\alpha \cdot \sum_i p_i^\alpha \cdot ln(p_i)}{\sum_j p_j^\alpha} - \\ & - \frac{\alpha \cdot \sum_j p_j^\alpha \cdot ln(p_j) \sum_i p_i^\alpha}{\sum_j p_j^\alpha} \Bigg\} \end{aligned} \tag{117}$$

Se utilizó logaritmo natural en vez de logaritmo en base 10, pues esto no influye en los cálculos y tampoco en los resultados. Tomando el límite tenemos que

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = 1 - 1 - \sum_i p_i \cdot ln(q_i) + \sum_e q_e \cdot ln(q_e) + \sum_i p_i \cdot ln(p_i) - \sum_j p_j \cdot ln(p_j) \quad (118)$$

Como resultado final tenemos

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = -H\Big(Q\Big) - \sum_i p_i \cdot ln(q_i) \quad (119)$$

Comparando con la ecuación (91) vemos que difieren. La conclusión sería que el límite de la función (96) no es el deseado, es decir, no converge a la divergencia de K-L. Es por eso que en los siguientes párrafos nos dedicaremos a solucionar este problema simetrizando el operador (96).

6.1.1 Solución del problema: Simetrización de la distancia, definición de $\mathcal{D}$

Como podemos ver de (119) intercambiamos las distribuciones resulta

$$\lim_{\alpha\to 1} D_\alpha\left(Q||P\right) = -H\Big(P\Big) - \sum_i q_i \cdot ln(p_i) \quad (120)$$

Esto nos muestra que si simetrizamos la función (96) para la entropía de Renyi

$$\mathcal{D}_{\mathcal{H}_\alpha}[P,Q] = D_{H_\alpha}(P||Q) + D_{H_\alpha}(Q||P) \quad (121)$$

y aplicamos el límite de α tendiendo a 1 nos da

$$\begin{aligned}\lim_{\alpha\to 1} \mathcal{D}_{\mathcal{H}_\alpha}[P,Q] = \left(-H(P) - \sum_i p_i \log(q_i)\right) + \left(-H(Q) - \sum_i q_i \log(p_i)\right) = \\ = D_{KL}(P||Q) + D_{KL}(Q||P)\end{aligned} \quad (122)$$

es decir converge a lo esperado, a la simetrización de la divergencia K-L.

Esto nos permite analizar el operador (95) desde otra perspectiva. Es decir, extender los resultados obtenidos (el operador aplicado a la entropía de Renyi) a otras entropías. Esto nos lleva a la definición de una nueva divergencia simétrica

$$\mathcal{D}_{\mathcal{H}'}[P,Q] = \frac{1}{2}\sum_i p_i \left(\frac{\partial H'}{\partial q_i} - \frac{\partial H'}{\partial p_i}\right) + \frac{1}{2}\sum_i q_i \left(\frac{\partial H'}{\partial p_i} - \frac{\partial H'}{\partial q_i}\right) \quad (123)$$

que también se puede escribir de la siguiente manera

$$\mathcal{D}_{\mathcal{H}'}[P,Q] = \frac{1}{2}\sum_i (p_i - q_i) \left(\frac{\partial H'}{\partial q_i} - \frac{\partial H'}{\partial p_i}\right) \quad (124)$$

donde H' representa una entropía cualquiera. Más adelante detallaremos que tipo de entropía utilizaremos, pero a priori la entenderemos como una entropía generalizada cualquiera. En los siguientes párrafos nos dedicaremos a introducir la divergencia de Bregman y mostrar la relación con la divergencia dada en (124).

Definición:(Divergencia Bregman) Sea $\psi : \mathcal{S} \longrightarrow \mathbb{R}$ una función estrictamente convexa definida en el conjunto $\mathcal{S} \subseteq \mathbb{R}^N$ tal que sea diferenciable en el interior de $\mathcal{S}$. La divergencia de Bregman queda definida por la siguiente expresión

$$d_\psi(P,Q) = \psi(P) - \psi(Q) - \sum_{i=1}^{N}(p_i - q_i)\frac{\partial\psi(Q)}{\partial q_i} \tag{125}$$

y cumple con las siguientes propiedades

1. **No Negativa:**
$$d_\psi(P,Q) \geq 0 \tag{126}$$
con la igualdad si y solo si $P \equiv Q$. Esta propiedad es consecuencia directa de la convexidad de ψ

2. **Extensiva:**
$$d_{\lambda\psi}(P,Q) = \lambda d_\psi(P,Q) \qquad (\lambda \geq 0) \tag{127}$$

3. **Convexa:** Es convexa en el primer argumento, pero no necesariamente en el segundo. Esto también es consecuencia directa de la convexidad de ψ

A la divergencia $d_\psi(P,Q)$ se la puede interpretar como la resta entre la función ψ valuada en el punto P menos el valor de la función del plano tangente en el punto Q valuado en P. Al ser la función ψ estrictamente convexa esta resta siempre es positiva. No solo eso sino que también el plano tangente no va a pasar otra vez por la función ψ. Esto nos asegura que $d_\psi(P,Q)$ es cero si y solo si $P = Q$, es decir, cuando la resta valga cero. El hecho que sea estrictamente convexa ψ nos dice que la propiedad de no negatividad se da si y solo si se cumple

$$\frac{\partial^2\psi}{\partial p_i \partial p_j} > 0 \tag{128}$$

Más adelante aplicaremos esta condición para definir el conjunto de funciones ψ que utilizaremos. Vale recalcar que la divergencia de Bregman está íntimamente relacionada con la divergencia de K-L. Si definimos a $\psi = -H$ donde H es la entropía de Shannon obtenemos la divergencia de K-L.

Si simetrizamos la divergencia de Bregman, es decir

$$\mathcal{D}_\psi(P||Q) = \frac{1}{2}d_\psi(P,Q) + \frac{1}{2}d_\psi(Q,P) \tag{129}$$

surge la relación con la divergencia (124) propuesta anteriormente. Para eso es necesario definir a $\psi := -H^G$, donde H^G es una entropía generalizada. Si reemplazamos ψ en la definición anterior tenemos que

$$\begin{aligned}\mathcal{D}_{H^G}(P||Q) = \frac{1}{2}\left(-H^G(P) + H^G(Q) + \sum_{i=1}^{N}(p_i - q_i)\frac{\partial H^G(Q)}{\partial q_i}\right) + \\ + \frac{1}{2}\left(-H^G(Q) + H^G(P) + \sum_{i=1}^{N}(q_i - p_i)\frac{\partial H^G(P)}{\partial p_i}\right)\end{aligned} \tag{130}$$

Si realizamos unos simples pasos algebraicos obtenemos

$$\mathcal{D}_{H^G}(P||Q) = \frac{1}{2}\sum_{i=1}^{N}(p_i - q_i)\left(\frac{\partial H^G(Q)}{\partial q_i} - \frac{\partial H^G(P)}{\partial p_i}\right) \tag{131}$$

es decir, que es igual a la divergencia (124). Estas son dos maneras de llegar a la misma divergencia, una es simetrizando el operador en función de entropías (95) y otra es simetrizando la divergencia de Bregman y definiendo a la función $\psi = -H^G$.

Es importante analizar qué sucede cuando tomamos dos distribuciones similares. En este tipo de análisis buscamos que el desarrollo sea de orden cuadrático. Esto se debe a que queremos relacionarlo con el concepto de métrica Riemanniana

$$G = \sum_{i,j=1}^{m} g_{ij} dx_i \otimes dx_j \tag{132}$$

donde g_{ij} son los elementos de una matriz y G un tensor métrico. En una primera instancia el operador de entropías definido en la ecuación (95) tenía la problemática que su expansión no era de orden cuadrático, pero esto se soluciona simetrizando el operador, es decir, definiendo la nueva divergencia mostrada en la ecuación (124) y (131). Esta es otra de las razones por las que se decide simetrizar y utilizar esa divergencia.

Para esto definiremos a la distribución Q de la siguiente manera. Para cada elemento de la distribución Q tenemos que $q_i = p_i + \delta p_i$ donde $\sum_i \delta p_i = 0$. Entonces si hacemos un desarrolo de Taylor a primer orden de la derivada parcial de la entropía generalizada tenemos que

$$\left.\frac{\partial H^G}{\partial p_i}\right|_{q_i = p_i + \delta p_i} \simeq \left.\frac{\partial H^G}{\partial p_i}\right|_{p_i} + \frac{\partial^2 H^G}{\partial p_i \partial p_j}\delta p_j \tag{133}$$

Si reemplazamos en la divergencia generalizada tenemos que

$$D^{H^G}(P||P+\delta P) = \frac{1}{2}\sum_{i,j}(p_i - p_i - \delta p_i)\left(\frac{\partial H^G}{\partial p_i} + \frac{\partial^2 H^G}{\partial p_i \partial p_j}\delta p_j - \frac{\partial H^G}{\partial p_i}\right) \tag{134}$$

Esto nos da el siguiente resultado

$$D^{H^G}(P||P+\delta P) \simeq -\frac{1}{2}\sum_{i,j}\frac{\partial^2 H^G}{\partial p_i \partial p_j}\delta p_i \delta p_j \tag{135}$$

Si definimos los coeficientes

$$g_{ij} = -\frac{1}{2}\frac{\partial^2 H_G}{\partial p_i \partial p_j} \tag{136}$$

vemos que hay una coincidencia, a orden más bajo, entre la expresión (135) y (132). En general, esta igualdad a orden más alto falla. En el caso de que la entropía que usemos sea la de Shannon la métrica resultante es la métrica de Fisher (la cual efectivamente es una

métrica Riemanniana) y la correspondiente distancia geodésica entre dos distribuciones de probabilidad P y Q está dada por

$$W(P,Q) = 2 \cdot arcos\left(\sum_i \sqrt{p_i q_i}\right) \tag{137}$$

6.2 Entropías Generalizadas

Las entropías que utilizaremos pertenecen a un grupo llamado entropías generalizadas y nacen de la idea de generalizar la entropía de información de Shannon. Haciendo un análisis global de las propiedades que tienen estas, podemos decir que en general el concepto de entropía generalizada está asociado con las siguientes propiedades

1. Que sea continua en $\{p_i\}$ para todo $i : 1, \cdots, N$.
2. Que sea no negativa.
3. Que sea cero en el caso determinista, es decir, cuando para un cierto $p_i = 1$ y para los demás la probabilidad es cero.
4. Que sea máxima en el caso de una distribución uniforme, es decir, que para todo i se tiene que $p_i = \frac{1}{N}$.
5. Que sea una función cóncava $H^G\left(\sum_i \lambda_i p_i\right) \geq \sum_i \lambda_i H^G(P)$.

Entropía de Salicrú La entropía generalizada que utilizaremos fue definida por primera vez por M. Salicrú, M.L. Menendez, D. Morales, L. Pardo [35]. Ellos generalizaron el concepto de entropía creando una entropía que también incluye a las entropías de Renyi y de HCT, y es de la forma

$$H_{(h,\phi)}[P] = h\left(\sum_i \phi(p_i)\right) \tag{138}$$

con las siguientes condiciones

- h creciente y ϕ cóncava
- h decreciente y ϕ convexa

y con $\phi(0) = 0$ y $h(\phi(1)) = 0$.

Podemos dar como ejemplo las tres entropías antes mencionadas. Si $h(y) = y$ y $\phi(y) = -y \ln y$ obtenemos la entropía de Shannon, si utilizamos $h(y) = \frac{\ln y}{1-\alpha}$ y $\phi(y) = y^\alpha$ obtenemos la entropía de Renyi y si usamos $h(y) = \frac{y-1}{1-\alpha}$ y $\phi(y) = y^\alpha$ obtenemos la entropía de HCT.

Haciendo un estudio de las propiedades de la entropía $H_{(h,\phi)}[P]$ podemos decir que es invariante ante permutaciones de las componentes de P. También se puede ver que posee la propiedad de ser expansible.

$$H_{(h,\phi)}(p_1, \cdots, p_N, 0) = H_{(h,\phi)}(p_1, \cdots, p_N) \tag{139}$$

Es importante ver que si uno tiene una distribución determinista, es decir, $P = (1, 0, \cdots, 0)$ la entropía de Salicrú vale cero.

Si usamos propiedades de mayorización tenemos que si

$$\sum_{i=1}^{N-1} p_i \leq \sum_{i=1}^{N-1} q_i \tag{140}$$

y

$$\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} q_i \tag{141}$$

implica que

$$H_{(h,\phi)}[P] \geq H_{(h,\phi)}[Q] \tag{142}$$

6.2.1 Conjunto de entropías para la nueva divergencia

Ya descriptas las propiedades de las entropías de Salicrú, podemos ahora imponer las condiciones que tienen que cumplir para que la divergencia definida en este capítulo cumpla con las condiciones deseadas. Como dijimos podemos partir de la simetrización de la divergencia de Bregman, esto resulta en

$$\mathcal{D}^{\psi} = \frac{1}{2} d_{\psi}(P, Q) + \frac{1}{2} d_{\psi}(Q, P) \tag{143}$$

Si reemplazamos $\psi = -H_G$, donde H_G ahora es una entropía de Salicrú, podemos escribir la divergencia de la siguiente manera

$$\mathcal{D}_{H^G}(P||Q) = \frac{1}{2} \sum_{i=1}^{N} (p_i - q_i) \left(\frac{\partial H^G(Q)}{\partial q_i} - \frac{\partial H^G(P)}{\partial p_i} \right) \tag{144}$$

Como vimos de la definición de la divergencia de Bregman podemos obtener una condición para que esta función cumpla las condiciones de positividad (y que sea 0 si y solo si $P = Q$). Esa condición se deriva requiriendo

$$-\frac{\partial^2 H_{(h,\phi)}[P]}{\partial p_i \partial p_j} > 0 \tag{145}$$

Sabiendo que

$$\frac{\partial H_G}{\partial p_i} = \frac{dh}{dy} \frac{d\phi_i}{dp_i} \tag{146}$$

donde $y = \sum_i \phi(p_i)$. Luego aplicando el signo menos y la derivada parcial en j tenemos

$$-\frac{\partial^2 H_{(h,\phi)}[P]}{\partial p_i \partial p_j} = -\frac{\partial}{\partial p_j}\left(\frac{\partial H_G}{\partial p_i}\right) = -\frac{\partial}{\partial p_j}\left(\frac{dh}{dy}\frac{d\phi_i}{dp_i}\right) \tag{147}$$

si derivamos el producto tenemos que

$$-\frac{\partial^2 H_{(h,\phi)}[P]}{\partial p_i \partial p_j} = -\frac{\partial}{\partial p_j}\left(\frac{dh}{dy}\right)\frac{d\phi_i}{dp_i} - \delta_{ij}\frac{dh}{dy}\frac{d^2\phi_i}{dp_i^2} \tag{148}$$

aplicando la regla de la cadena resulta

$$-\frac{\partial^2 H_{(h,\phi)}[P]}{\partial p_i \partial p_j} = -\frac{d^2h}{dy^2}\frac{d\phi_j}{dp_j}\frac{d\phi_i}{dp_i} - \delta_{ij}\frac{dh}{dy}\frac{d^2\phi_i}{dp_i^2} \tag{149}$$

Pasando en limpio nos da que la condición para que las entropías puedan ser utilizadas en la divergencia es

$$-h''(y)\phi'(p_i)\phi'(p_j) - h'(y)\phi''(p_i)\delta_{ij} > 0 \tag{150}$$

con $y = \sum_i \phi(p_i)$. Dado esto tenemos las condición que deben cumplir las entropías de Salicrú para que la divergencia esté definida correctamente.

6.3 Asignación de pesos estadísticos a la divergencia $\mathcal{D}_{H_G}$

Hemos visto que el disponer de una divergencia como la de Jensen-Shannon, nos permite desarrollar técnicas de análisis de series temporales, de manera muy eficiente y con resultados altamente significativos. Es por ello que en esta sección nos proponemos extender esos métodos al uso de las divergencias tipo Bregman. Un primer paso es definir divergencias tipo Bregman pero con la posibilidad de asignar pesos a las diferentes distribuciones de probabilidad. Sean π_P y π_Q dos números no negativos tal que $\pi_P + \pi_Q = 1$. Estos números se los puede interpretar como pesos estadísticos de las distribuciones P y Q respectivamente. Proponemos definir la divergencia tipo Bregman con pesos de la siguiente forma:

$$\mathcal{D}_{H_G}^{\pi_P \pi_Q}\Big(P||Q\Big) = \pi_P d_\psi(P||M) + \pi_Q d_\psi(Q||M) \tag{151}$$

donde $\psi = -H_G$ y $M = \pi_P P + \pi_Q Q$, que se lo puede interpretar como la distribución mezcla. Como se puede ver si optamos por utilizar la entropía de Shannon el resultado es la divergencia de la Jensen-Shannon generalizada por pesos mostrada en la ecuación (83) para el caso de dos distribuciones solamente, esto es

$$D_{JS}^{\pi_P \pi_Q}(P||Q) = H\left(M\right) - \pi_P H(P) - \pi_Q H(Q) \tag{152}$$

En los siguientes apartados utilizaremos la forma generalizada de esta nueva divergencia propuesta mostrada en la ecuación (151). Esta generalización por pesos de nuestra divergencia la utilizaremos y aplicaremos principalmente en dos entropías generalizadas bien establecidas en la literatura, estas son la entropía de Renyi y la de HCT.

6.4 El caso de la entropía de Renyi

En esta sección mostraremos los resultados obtenidos cuando utilizamos la entropía de Renyi en esta nueva divergencia. Por sustitución directa resulta

$$\mathcal{D}_{H_\alpha^R}(P||Q) = \frac{1}{2}\sum_{i=1}^{N}(p_i - q_i)\left(\frac{\partial H_\alpha^R(Q)}{\partial q_i} - \frac{\partial H_\alpha^R(P)}{\partial p_i}\right) \tag{153}$$

La forma explícita de la divergencia es

$$\mathcal{D}_{H_\alpha^R}\Big(P||Q\Big) = \frac{\alpha}{1-\alpha}\left(\frac{\sum_i p_i q_i^{\alpha-1}}{Q_\alpha} + \frac{\sum_i q_i p_i^{\alpha-1}}{P_\alpha} - 2\right) \tag{154}$$

donde $Q_\alpha = \sum_j q_j^\alpha$ y $P_\alpha = \sum_j p_j^\alpha$. Es necesario imponer las condiciones sobre la entropía para garantizar las propiedades deseadas de la divergencia, es decir, su no negatividad y que sea igual a cero si y solo si las dos distribuciones son iguales. Para esto necesitamos utilizar la ecuación (150) y ver para que rango del parámetro α se satiface. Explícitamente tenemos

$$-h''(y)\phi'(p_i)\phi'(p_j) - h'(y)\phi''(p_i)\delta_{ij} = \alpha\frac{p_i^{\alpha-2}}{P_\alpha} - \frac{\alpha}{\alpha-1}\frac{p_i^{\alpha-1}p_j^{\alpha-1}}{P_\alpha^2} > 0 \tag{155}$$

La desigualdad se cumple para los $\alpha \in (0,1)$. Para este rango del parámetro tenemos garantizadas las propiedades de la divergencia.

La divergencia generalizada mediante pesos estadísticos utilizando la entropía de Renyi es

$$\begin{aligned}\mathcal{D}_{H_\alpha^R}^{\pi_p\pi_q}(P||Q) = \pi_p H_\alpha[P] - \pi_p H_\alpha[M] + \frac{\alpha}{1-\alpha}\frac{\pi_p}{P_\alpha}\sum_i (m_i - p_i)\, p_i^{\alpha-1} + \\ + \pi_q H_\alpha[Q] - \pi_q H_\alpha[M] + \frac{\alpha}{1-\alpha}\frac{\pi_q}{Q_\alpha}\sum_i (m_i - q_i)\, q_i^{\alpha-1}\end{aligned} \tag{156}$$

6.5 El caso de la entropía de HCT

En esta sección aplicaremos la nueva divergencia a la entropía de HCT. Es directo verificar que para la entropía de HCT el límite sobre la divergencia cuando el parámetro α tiende a 1 es

$$\lim_{\alpha\to 1}\Big[\mathcal{D}_{H_\alpha^T}\Big(P||Q\Big)\Big] = \mathcal{D}_H\Big(P||Q\Big) \tag{157}$$

donde H es la entropía de Shannon. Los detalles de este límite se encuentran en el Apéndice A.

La forma explícita de la nueva divergencia construida con la entropía de HCT es

$$\mathcal{D}_{H_\alpha^T}\Big(P||Q\Big) = \frac{\alpha}{1-\alpha}\left(\sum_i q_i p_i^{\alpha-1} + \sum_i p_i q_i^{\alpha-1} - Q_\alpha - P_\alpha\right) \tag{158}$$

donde $Q_\alpha = \sum_j q_j^\alpha$ y $P_\alpha = \sum_j p_j^\alpha$. Ahora lo que necesitamos es ver para que rango del parámetro α se verifica (150). Haciendo los cálculos obtenemos que

$$-h''(y)\phi'(p_i)\phi'(p_j) - h'(y)\phi''(p_i)\delta_{ij} = \delta_{ij}\alpha p_i^{\alpha-2} > 0 \tag{159}$$

esto nos da que para cualquier $\alpha > 0$ tenemos garantizadas las propiedades de la divergencia.

Para las aplicaciones que haremos necesitamos tener la forma generalizada mediante pesos estadísticos de la divergencia siendo su forma explícita la siguiente expresión

$$\begin{aligned} \mathcal{D}_{H_\alpha^T}^{\pi_p\pi_q}(P|Q) = \pi_p H_\alpha[P] - \pi_p H_\alpha[M] + + \frac{\alpha}{1-\alpha}\pi_p \sum_i (m_i - p_i)\, p_i^{\alpha-1} + \\ + \pi_q H_\alpha[Q] - \pi_q H_\alpha[M] + \frac{\alpha}{1-\alpha}\pi_q \sum_i (m_i - q_i)\, q_i^{\alpha-1} \end{aligned} \tag{160}$$

Al igual que en el caso de Renyi las aplicaciones a señales reales y simuladas se harán al final de este capítulo mostrando también las respectivas conclusiones.

6.6 Aplicaciones

En esta sección mostraremos las aplicaciones de los resultados teóricos obtenidos en señales reales y simuladas. En cualquiera de los dos casos utilizaremos el método del puntero detallado en el capítulo "Aspectos Metodológicos".

6.6.1 Series Simuladas

En esta sección se hará un análisis sobre secuencias binarias de símbolos creadas por un algoritmo que detallaremos en los párrafos siguientes. Tomemos una secuencia $\mathcal{S}$ de largo $L_\mathcal{S}$ compuesta por dos secuencias $\mathcal{S}_1$ y $\mathcal{S}_2$ de largo $L_2 = L - L_1$ del mismo tamaño. Unimos las dos secuencias formando la secuencia $\mathcal{S}$. Las secuencias $\mathcal{S}_1$ y $\mathcal{S}_2$ serán binarias con una distribución de probabilidad $P_k = \{s_k, 1 - s_k\}$ con $k : 1, 2$. Realizaremos una cantidad N_e de muestras de estas secuencias (en consecuencia de la secuencia total también). Lo que haremos es mover el puntero x a lo largo de toda la secuencia total, es decir, $1 < x < L_\mathcal{S}$. Entonces calcularemos la distribución de probabilidad de aparición de 0 y 1 de cada lado del puntero esto nos da dos distribuciones $\mathcal{P}_I(x)$ y $\mathcal{P}_D(x)$ a la izquierda y derecha respectivamente. En la práctica, como vimos en la exposición de este método, es necesario pesar estadísticamente las distribuciones y las subsecuencias de cada lado del cursor ya que no van a tener el mismo largo. La asignación de pesos es de la siguiente manera; sea $\pi_I = \frac{x}{L_\mathcal{S}}$ y $\pi_D = 1 - \pi_I$. Para este tipo de aplicación utilizaremos la versión de la divergencia generalizada por pesos para las dos entropías Renyi y HCT, esto es $\mathcal{D}_{H_\alpha^R}^{\pi_I\pi_D}(\mathcal{P}_I(x)||\mathcal{P}_D(x))$ y $\mathcal{D}_{H_\alpha^T}^{\pi_I\pi_D}(\mathcal{P}_I(x)||\mathcal{P}_D(x))$ respectivamente y para cada posición del puntero x.

En la práctica utilizamos un ensamble de $N_e = 1000$ realizaciones, un largo total de secuencia de $L_{\mathcal{S}} = 40000$, y un largo para cada secuencia $\mathcal{S}_1$ y $\mathcal{S}_2$ de $L_{\mathcal{S}_1} = L_{\mathcal{S}_2} = 20000$. Realizamos un promedio de los valores obtenidos de las divergencias sobre el ensamble de realizaciones. Las distribuciones de las secuencias $\mathcal{S}_1$ y $\mathcal{S}_2$ están dadas por los parámetros $s_1 = \frac{2}{3}$ y $s_2 = \frac{1}{3}$ respectivamente.

En la figura 2 se puede ver la divergencia $\mathcal{D}_{H_\alpha^R}^{\pi_I \pi_D}(\mathcal{P}_I(x)||\mathcal{P}_D(x))$ en función de x para la entropía de Renyi para distintos valores de α, $\alpha = 0.2$, $\alpha = 0.4$, $\alpha = 0.6$ y $\alpha = 0.8$. Como se observa el valor máximo (que es un promedio) se da en la unión de las secuencias $\mathcal{S}_1$ y $\mathcal{S}_2$, es decir, cuando el puntero se encuentra en $x = L_{\mathcal{S}_1}$. Se puede ver que el valor máximo de la divergencia aumenta a medida que aumenta el valor de α. Es bueno aclarar que el hecho de que sea una curva suave es porque estamos promediando sobre un ensamble de realizaciones, si utilizaríamos una sola secuencia se verían fluctuaciones.

En la figura 3 se puede ver la divergencia $\mathcal{D}_{H_\alpha^T}^{\pi_I \pi_D}(\mathcal{P}_I(x)||\mathcal{P}_D(x))$ en función de x para la entropía de HCT para distintos valores de α, $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$ y $\alpha = 1.3$. Como se observa el valor máximo (que es un promedio) se da en la unión de las secuencias $\mathcal{S}_1$ y $\mathcal{S}_2$, es decir, cuando el puntero se encuenta en $x = L_{\mathcal{S}_1}$. Al igual que en el caso de la entropía de Renyi el valor máximo de la divergencia aumenta a medida que aumenta el valor del parámetro α.

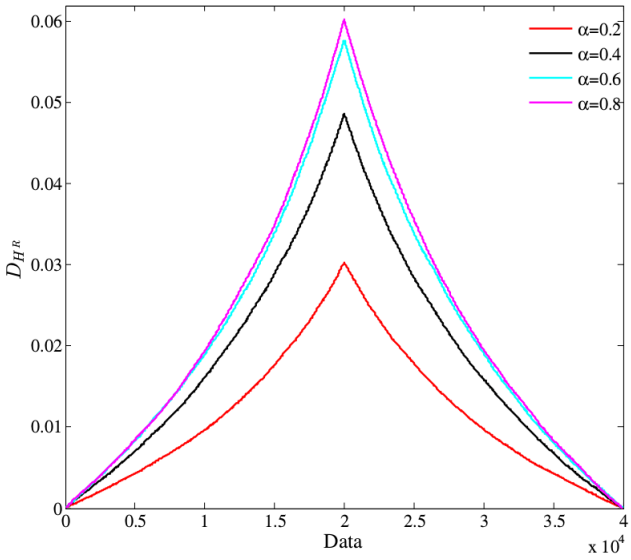

Figura 2. Promedio sobre un ensamble de la divergencia generalizada por pesos estadísticos para la entropía de Renyi, para distintos valores del parámetro α.

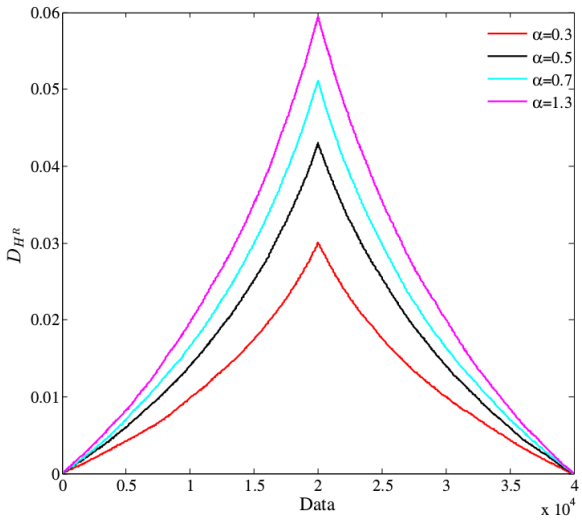

Figura 3. Promedio sobre un ensamble de la divergencia generalizada por pesos estadísticos para la entropía de HCT, para distintos valores del parámetro α.

6.6.2 Series Reales

Para el estudio de series temporales reales utilizaremos señales obtenidas por un electrocardiograma, y estudiaremos camparativamente las señales de un electrocardiograma en un ritmo normal del corazón con señales obtenidas en situación de una fibrilación auricular.

La fibrilación auricular (FA) es una arritmia cardíaca sostenida muy común, que ocurre en una parte importante de la población general. Una forma simple de describir esta enfermedad es la siguiente; la fibrilación auricular es un desarreglo en la contracción de las fibras de la aurícula. Esto hace que la aurícula bombee de forma defectuosa la sangre. Se asocia con una mortalidad y morbilidad significativas a través de la asociación del riesgo de muerte causada por accidentes cerebrovasculares, insuficiencia cardíaca y enfermedad coronaria. A pesar de la magnitud de este problema, la detección de FA sigue siendo problemática, ya que puede ser episódica. Por estas razones, es importante desarrollar métodos que puedan detectar la diferencia entre la FA y los ritmos sinusales normales (NSR), situación en la que el corazón está en una condición normal y sana, utilizando trazas de electrocardiograma (ECG).

Probamos nuestro método de análisis en la problemática de detectar las diferencias entre un registro de FA y uno de NSR. Con este objetivo, colocamos en un solo registro dos señales de ECG, una con FA seguida de otra con NSR. El ECG es de un solo electrodo y nos muestra en la señal valores de voltaje. Estos registros fueron tomados por el banco de datos Physionet [36]. La señal se normalizó dividiéndola por el máximo valor tal que los valores queden entre 0 y 1. Además se la cortó con la misma longitud ($L_{FA} = L_{NSR} = 20000$

puntos de datos) y se unió para formar la señal "completa". Las señales se mapearon previamente en una secuencia de alfabeto finito usando el método de mapeo de vector de permutación de Bandt y Pompe. Este mapeo requiere dos parámetros, generalmente denominados d (dimensión de la inmersión o del vector) y τ (retraso). En nuestro ejemplo, utilizamos los valores $d = 4$ y $\tau = 1$. Después de este procesamiento, aplicamos el método de segmentación descripto anteriormente (el del puntero) mediante la entropía de Renyi con diferentes valores de α. La figura 4 muestra el comportamiento de la divergencia $\mathcal{D}_{H_\alpha^R}$ en función del cursor que se mueve a lo largo de la serie fusionada. Esta cantidad alcanza su valor máximo en el punto exacto donde cambia la dinámica de la señal de ECG. Esto sucede porque la distribución de probabilidad empírica de los vectores de permutación es diferente cuando la señal está en FA que en NSR. Esta diferencia se detecta claramente por la divergencia para todos los valores de α, pero las detecciones más claras ocurren para α bajo.

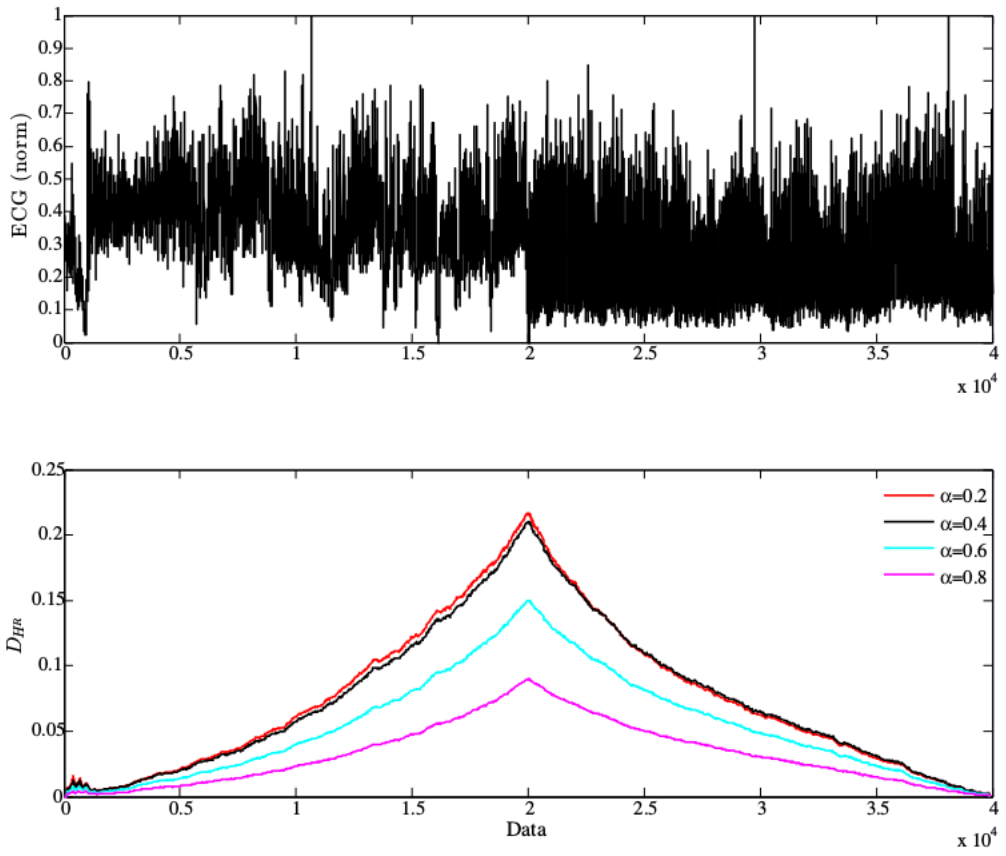

Figura 4. Divergencia generalizada por pesos para un análisis de una señal normalizada de ECG. La señal de ECG es una combinación de dos secuencias $\mathcal{S} = \mathcal{S}_1\mathcal{S}_2$ con longitud $L_{\mathcal{S}_1} = L_{\mathcal{S}_2} = 20000$ respectivamente. La primera parte de la señal pertenece a la traza de fibrilación auricular y la segunda es Rítmica sinusal normal. La señal se discretizó usando el enfoque de vector de permutación con el parámetro $d = 4$ y $\tau = 1$. La divergencia se tomó usando diferentes valores de α indicados en los gráficos.

6.7 Conclusiones

En primera instancia pudimos ver una nueva forma de interpretación de la divergencia de Kulback-Leibler. Esto llevó a pensarla como un operador que depende de una entropía. Esto nos permitió generalizar este operador para diferentes entropías generalizadas. Por cuestiones técnicas y de convergencia hacia la K-L original que depende de la entropía de Shannon fue necesario simetrizar el operador, lo que nos permitió mostrar que la simetrización de dicho operador era igual a la simetrización de la divergencia de Bregman. Esto nos garantizó las propiedades necesarias para que esta nueva divergencia cumpla con las condiciones que debe tener una divergencia como se la entiende en la literatura.

Otra conclusión importante fue dar las condiciones que tienen que cumplir las entropías derivadas de la entropía de Salicrú para que esta nueva divergencia (la simetrización de la divergencia de Bregman) cumpla con las condiciones correctas de una divergencia. Es decir, que sea mayor que cero e igual a cero si y solo si las distribuciones son iguales. Sabemos que por construcción se cumple la simetría respecto de sus argumentos.

Por último, se aplicó esta nueva divergencia en series simuladas y de origen natural. Para esto fue necesario generalizar la divergencia por pesos estadísticos y así aplicar los métodos de segmentación correctos. Respecto a la eficiencia de este nuevo cuantificador de distinguibilidad entre distribuciones pudimos ver en el apartado de aplicaciones que tanto para series simuladas como para las de origen natural, la divergencia cumplió el objetivo de diferenciar y distinguir las series comparadas. Esto nos garantiza que es una buena herramienta para el análisis de series temporales.

7 Capítulo VI: Divergencia Gamma

Como hemos visto en los capítulos anteriores las funciones convexas guardan una estrecha relación con los conceptos de divergencia y entropía. En este capítulo utilizaremos, como en el capítulo anterior, a las funciones convexas y sus propiedades para definir un cuantificador de distinguibilidad entre distribuciones de probabilidad. Este capítulo se lo puede dividir en varias partes interrelacionadas entre sí. En primer lugar definiremos una divergencia generalizada que depende de las propiedades de ciertas funciones convexas, que llamaremos divergencia γ y como mostraremos, cumple con todos los requisitos que esperamos de un cuantificador de distinguibilidad entre distribuciones de probabilidad. Luego del estudio teórico de las propiedades de esta divergencia introducimos una entropía generalizada con una estrecha relación con las funciones convexas usadas al definir la divergencia. Como se verá esta entropía cumple con todos los requisitos que tiene que cumplir una entropía generalizada mostrados en el capítulo anterior. En tercer lugar, y gracias a la relación que tienen la divergencia γ y la entropía generalizada, podemos encontrar que para ciertas funciones convexas la raiz cuadrada de la divergencia γ cumple con la desigualdad triangular permitiendo así definir una nueva métrica. Por último, para verificar la eficiencia de esta divergencia generalizada se mostrarán las aplicaciones a series temporales reales y simuladas.

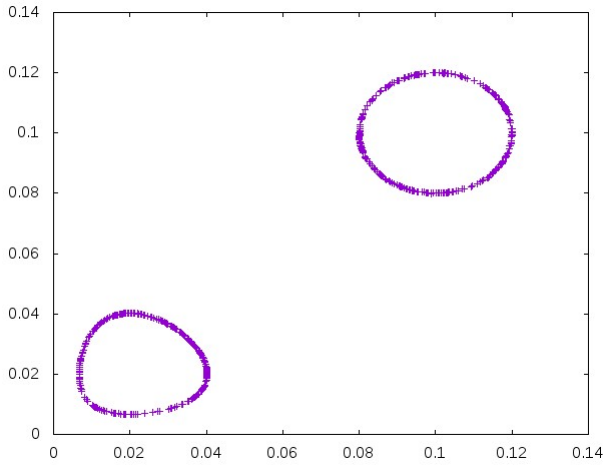

Figura 5. Conjuntos convexos formados por la divergencia de Jensen Shannon (izquierda abajo) y por el cuadrado de la métrica euclídea (derecha arriba).

Las motivaciones para esta sección de la tesis surgen de la observación de la figura 5. Ésta muestra los conjuntos C_1 (para la D_{JS}) y C_2 (para el cuadrado de la métrica euclídea) cuando se le asigna un valor determinado (radio) a las respectivas funciones. Es decir, son los valores o puntos compatibles a un cierto "radio" de los respectivos funcionales. En este caso la $D_{JS}(P, K) = 0.05$ para una distribución fija K y para el cuadrado de

la métrica euclídea se eligió el radio 0.02 también para la distribución K fija. Como se puede observar los dos conjuntos compatibles con sus respectivos "radios" son conjuntos convexos. Esto nos hace pensar que estas dos divergencias tienen algo en común, y como veremos más adelante estas pertenecen a una familia particular de funcionales.

Como primer paso veremos la relación que existe entre el cuadrado de la métrica euclídea y la divergencia de Jensen-Shannon. Sean $P = \{p_i\}_{i=1}^N$ y $Q = \{q_i\}_{i=1}^N$ distribuciones de probabilidad de N estados de una variable aleatoria discreta. Podemos escribir el cuadrado de la métrica euclídea como

$$E(X, P) = \sum_i (x_i - p_i)^2 = 2x_i^2 + 2p_i^2 - (x_i + p_i)^2 \tag{161}$$

Haciendo unos simples pasos algebraicos podemos escribirla de la siguiente manera

$$E(Q||P) = \sum_i 2\left(q_i I(q_i) + p_i I(p_i) - (p_i + q_i) I\left(\frac{p_i + q_i}{2}\right)\right) \tag{162}$$

donde $I(\cdot)$ es la función identidad. De la misma manera podemos escribir a D_{JS} como

$$D_{JS}(Q||P) = \sum_i \frac{1}{2} p_i \cdot \ln(p_i) + \frac{1}{2} x_i \cdot \ln(q_i) - \frac{1}{2}(p_i + q_i) \cdot \ln\left(\frac{q_i + p_i}{2}\right) \tag{163}$$

Como se puede ver las dos divergencias tienen la misma estructura, es decir, son ejemplos de una función más general de la forma

$$\mathcal{D}(Q||P) = \sum_i q_i g(q_i) + p_i g(p_i) - (q_i + p_i) g\left(\frac{q_i + p_i}{2}\right) \tag{164}$$

En el caso del cuadrado de la métrica euclídea la función $g(y) = 2I(y)$ y en el caso de la D_{JS} es igual a $g(y) = \frac{1}{2}\ln y$. En los siguientes párrafos mostraremos las propiedades que tiene que tener la función $\mathcal{D}(Q||P)$ para que sea una divergencia. Para eso enunciaremos el siguiente teorema:

Teorema: Sea $Q \in (\mathbb{R}^+)^N$ y $P \in (\mathbb{R}^+)^N$ y $g : \mathbb{R}^+ \longrightarrow \mathbb{R}$. Si la función $g(y)$ es tal que $f(y) := y \cdot g(y)$ es una función convexa, luego la función definida por

$$\mathcal{D}_{\gamma_g}(Q||P) = \sum_i \gamma_g(q_i, p_i) \tag{165}$$

con

$$\gamma_g(q_i, p_i) = q_i g(q_i) + p_i g(p_i) - (q_i + p_i) g\left(\frac{q_i + p_i}{2}\right) \tag{166}$$

satisface las siguientes condiciones

1. $\mathcal{D}_{\gamma_g}(Q||P) = D_{\gamma_g}(P||Q)$ $\quad$ (*Simetría*)
2. $\mathcal{D}_{\gamma_g}(Q||P) > 0 \qquad$ para $\quad Q \not\equiv P \quad$ (*Positividad*)
3. $\mathcal{D}_{\gamma_g}(Q||P) = 0$ si y solo si $Q \equiv P$

Prueba: Siendo D_{γ_g} una suma de las funciones $\gamma_g(q_i, p_i)$ si esta función es simétrica, positiva y se anula si y solo si $Q \equiv P$ entonces D_{γ_g} también lo será. Usaremos también la siguiente notación

$$m_i := \frac{x_i + p_i}{2} \tag{167}$$

Simetría:

Surge de observar que

$$\gamma_g(q, p) = \gamma_g(p, q) \tag{168}$$

y por lo tanto $\mathcal{D}_{\gamma_g}(Q||P)$ también lo es. $\diamondsuit$

Positividad:

Para probar la desigualdad $\mathcal{D}_{\gamma_g}(Q||P) \geq 0$ primero demostraremos que

$$\gamma_g(x_i, p_i) \geq 0, \quad \forall i \tag{169}$$

Por la hipótesis del teorema tenemos que para todo $t \in [0, 1]$ y $q_i, p_i \in \mathbb{R}^+$ se tiene que

$$tf(q) + (1-t)f(p) \geq f(tq + (1-t)p) \tag{170}$$

Si reemplazamos $f(y)$ tenemos

$$t(q \cdot g(q)) + (1-t)(p \cdot g(p)) - (tq + (1-t)p)\, g\,(tq + (1-t)p) \geq 0 \tag{171}$$

y si elegimos $t = 1/2$ obtenemos

$$\frac{q}{2}g(q) + \frac{p}{2}g(p) - \frac{(q+p)}{2}g\left(\frac{q+p}{2}\right) \geq 0 \tag{172}$$

Usando la definición (166) tenemos

$$\frac{\gamma_g(q, p)}{2} \geq 0 \Longrightarrow \gamma_g(q, p) \geq 0 \tag{173}$$

por lo tanto la suma de cantidades positivas es positiva, y así

$$\mathcal{D}_{\gamma_g}(Q||P) = \sum_i \gamma_g(q_i, p_i) \geq 0 \tag{174}$$

$\diamondsuit$

$\mathcal{D}_{\gamma_g}(Q||P) = 0$ si y solo si $Q \equiv P$:

$\Rightarrow$: Si reemplazamos $Q \equiv P$ en la definición de $\gamma_g(q_i, p_i)$ dada en (166) tenemos

$$\gamma_g(q, q) = q \cdot g(q) + q \cdot g(q) - (q + q)g\left(\frac{q + q}{2}\right) = 0 \tag{175}$$

Como se cumple para todo i entonces $\mathcal{D}_{\gamma_g}(Q||P) = 0$

$\Leftarrow$:

Comenzamos afirmando que

$$\mathcal{D}_{\gamma_g}(Q||P) = 0 \tag{176}$$

pero como para cada i $\gamma_g(q_i, p_i)$ es positiva la igualdad anterior es cierta si y solo si

$$\gamma_g(q_i, p_i) = 0, \qquad \forall i \tag{177}$$

haciendo unos simples pasos algebraicos tenemos que en términos de $f(y)$

$$f(m_i) = \frac{f(q_i) + f(p_i)}{2} \tag{178}$$

Al ser $f(y)$ convexa y distinta de la identidad entonces la igualdad anterior es cierta si y solo si $q_i = m_i$ y $p_i = m_i$, lo que es equivalente a decir $x_i = p_i$ $\qquad \diamondsuit \bullet$

Enunciado el teorema estamos en condiciones de encontrar el subconjunto de funciones $g(y)$ que satisfacen la hipótesis del teorema. Sabemos que si $f(y)$ es dos veces diferenciable, es estrictamente convexa si y solo si

$$\frac{d^2 f}{dy^2} = \frac{d^2(yg(y))}{dy^2} > 0 \tag{179}$$

Si reemplazamos $f(y) := y \cdot g(y)$ obtenemos la siguiente condición para las funciones $g(y)$

$$y\frac{d^2 g}{dy^2} + 2\frac{dg}{dy} > 0 \qquad \forall y \geq 0 \tag{180}$$

7.1 Linealidad

A veces en la práctica se nos presenta la posibilidad de tener una función que se escriba como la sumatoria de varias funciones. Mediante algunos cálculos podemos verificar que si esta descomposición cumple con ciertas condiciones al resultado de la sumatoria también se le puede aplicar la divergencia. Sea $g(y) = \sum_{k=1}^{m} \alpha_k g_k(y)$, con $\alpha_k > 0$ para todo $k : 1, \cdots, m$. Las funciones $g_k(y)$ satisfacen las hipótesis del teorema. Luego remplazando en la ecuación (166)

$$\gamma_g(q_i, p_i) = \sum_k \alpha_k \gamma_{g_k}(q_i, p_i) \tag{181}$$

donde

$$\gamma_{g_k}(q_i, p_i) := q_i \cdot g_k(q_i) + p_i \cdot g_k(p_i) - (q_i + p_i) \cdot g_k\left(\frac{q_i + p_i}{2}\right) \tag{182}$$

La divergencia toma la siguiente forma

$$\mathcal{D}_{\gamma_g}(Q||P) = \sum_k \alpha_k \mathcal{D}_{\gamma_{g_k}}(Q||P) \tag{183}$$

con

$$\mathcal{D}_{\gamma_{g_k}}(Q||P) := \sum_i \gamma_{g_k}(q_i, p_i) \tag{184}$$

Ahora mostraremos que γ_g cumple con las propiedades del teorema.

Podemos ver que para todo k, $\mathcal{D}_{\gamma_{g_k}} \geq 0$, ya que cada $g_k(y)$ satisface las hipótesis del teorema. Por lo tanto si $\alpha_k > 0$ tenemos

$$\mathcal{D}_{\gamma_g}(Q||P) \geq 0 \tag{185}$$

Siendo $\mathcal{D}_{\gamma_{g_k}}(Q||P) = \mathcal{D}_{\gamma_{g_k}}(P||Q)$ para todo k,

$$\mathcal{D}_{\gamma_g}(Q||P) = \mathcal{D}_{\gamma_g}(P||Q) \tag{186}$$

Como $\mathcal{D}_{\gamma_{g_k}}(Q||P) = 0$ si y solo si $X \equiv P$ para todo k por hipótesis, y $\alpha_k > 0$ para todo k, obtenemos

$$Q \equiv P \Longrightarrow \mathcal{D}_{\gamma_g}(Q||P) = \sum_k \alpha_k \mathcal{D}_{\gamma_{g_k}}(Q||P) = 0 \tag{187}$$

Si $\mathcal{D}_{\gamma_g}(Q||P) = 0$ implica que cada término debe ser igual a cero. Como $\alpha_k > 0$ para todo k luego $\mathcal{D}_{\gamma_{g_k}}(Q||P) = 0$. Pero por hipótesis cada $\mathcal{D}_{\gamma_{g_k}}$ cumple con el teorema. Entonces

$$\mathcal{D}_{\gamma_{g_k}}(Q||P) = 0 \Longrightarrow Q \equiv P \qquad \forall k \tag{188}$$

y por lo tanto $\mathcal{D}_{\gamma_g}(Q||P) = 0 \Longrightarrow Q \equiv P$

7.2 Distribuciones Similares

Sea $q_i = p_i + \delta p_i$ para cada i, con $\sum_{i=1}^{N} \delta p_i = 0$, y sea $g(y)$ una función diferenciable. Luego la divergencia $\mathcal{D}_{\gamma_g}$ es

$$\mathcal{D}_{\gamma_g}(P+\delta P||P) = \sum_i \gamma_g(p_i+\delta p_i, p_i) = \sum_i (p_i+\delta p_i)\left(g\left(p_i+\delta p_i\right) - g\left(p_i+\frac{\delta p_i}{2}\right)\right) + \\ + p_i \cdot \left(g(p_i) - g\left(p_i+\frac{\delta p_i}{2}\right)\right) \tag{189}$$

Si tomamos el término lineal del desarrollo de Taylor de $g(y)$ valuado en $y = p_i + \delta p_i$ tenemos que

$$g(p_i+\delta p_i) \simeq g(p_i) + \dot{g}(p_i)\delta p_i \tag{190}$$

donde $\dot{g}(y) := \frac{dg}{dy}$. Luego la aproximación de $\mathcal{D}_{\gamma_g}$ es

$$\mathcal{D}_{\gamma_g}(P||P+\delta P) \simeq \sum_i (p_i+\delta p_i)\left(\dot{g}(p_i)\delta p_i + g(p_i) - \dot{g}(p_i)\frac{\delta p_i}{2} - g(p_i)\right) + p_i\left(g(p_i) - \dot{g}(p_i)\frac{\delta p_i}{2} - g(p_i)\right) \tag{191}$$

Haciendo unos pasos algebraicos tenemos

$$\mathcal{D}_{\gamma_g}(P||P+\delta P) \simeq (p_i+\delta p_i)\dot{g}\frac{\delta p_i}{2} - p_i\dot{g}(p_i)\frac{\delta p_i}{2} \tag{192}$$

Cancelando los términos iguales obtenemos

$$\mathcal{D}_{\gamma_g}(P+\delta P||P) \simeq \sum_i \dot{g}(p_i)\frac{(\delta p_i)^2}{2} \tag{193}$$

El coeficiente que acompaña a $(\delta p_i)^2$ no tiene por que ser una métrica riemanianna

7.3 Generalización de la divergencia con pesos

Como vimos en el capítulo Aspectos Metodológicos la utilización de punteros es muy práctica y efectiva. Para poder utilizarlo de manera correcta debemos desarrollar de forma teórica una versión generalizada de la divergencia que utilice pesos estadísticos. En los párrafos siguientes veremos como esta generalización está arraigada a la definición de la divergencia y a las propiedades de convexidad de las funciones. En la ecuación (165) asumimos que las distribuciones de probabilidad P y Q tenían el mismo peso ($\frac{1}{2}$). Una directa generalización es asignarle diferentes pesos a las distribuciones de probabilidad.

Sea $\pi_P, \pi_Q \geq 0$, $\pi_P + \pi_Q = 1$ pesos arbitrarios para las distribuciones de probabilidad P y Q. Podemos ver en la ecuación (171) una natural asignación tomando $t = \pi_X$ y $(1-t) = \pi_P$.

$$\begin{aligned}\mathcal{D}_{\gamma_g}^{\pi_q \pi_p}(Q||P) &= \sum_i \pi_q q_i g(q_i) + \pi_p p_i g(p_i) - m_i g(m_i) = \\ &= \sum_i \pi_q q_i \Big(g(q_i) - g(m_i)\Big) + \pi_p p_i \Big(g(p_i) - g(m_i)\Big)\end{aligned} \tag{194}$$

donde $m_i = \pi_p p_i + \pi_q q_i$. Esta asignación nos asegura que $\mathcal{D}_{\gamma_g}^{\pi_x \pi_p}(X||P) \geq 0$, ya que

$$\gamma_g^{\pi_q \pi_p}(q_i, p_i) = \pi_q q_i g(q_i) + \pi_p p_i g(p_i) - m_i g(m_i) \geq 0 \qquad \forall\, i : 1, \ldots, N \tag{195}$$

Esta generalización en función de pesos estadísticos nos permitirá trabajar correctamente con la metodología de punteros.

7.4 Generalización para más de dos distribuciones

En la práctica se nos puede presentar la situación de trabajar con varias señales a la vez. Esto presenta un dilema a la hora de poder encontrar cantidades que cuantifiquen diferencias entre las señales. Es posible encontrar en la literatura un cuantificador de distancias entre varias distribuciones de probabilidad como es el caso de la versión generalizada de la divergencia Jensen-Shannon.

Es posible generalizar la $\mathcal{D}_{\gamma_g}$ para más de dos distribuciones de probabilidad. Sea $f(y)$ una función convexa y sean $\{y^1, \ldots, y^W\}$ valores del dominio de $f(y)$. Usando la desigualdad de Jensen tenemos

$$f\left(\sum_{k=1}^{W} \pi^k y^k\right) \leq \sum_{k=1}^{W} \pi^k f(y^k) \tag{196}$$

donde $\pi^k \in [0, 1]$, usando la definición de $f(y) := yg(y)$, tenemos

$$\sum_{k=1}^{W} \pi^k y^k g(y^k) - \left(\sum_{k=1}^{W} \pi^k y^k\right) \cdot g\left(\sum_{k=1}^{W} \pi^k y^k\right) \geq 0 \tag{197}$$

El lado izquierdo es justamente la extensión de $\gamma_g^{\pi_x \pi_p}(x_i, p_i)$ para W distribuciones de probabilidad, asignándoles $\{\pi^1, \ldots, \pi^W\}$ como los pesos de las distribuciones, y la llamamos $\gamma_g^{\pi^1, \ldots, \pi^W}(p^1, \ldots, p^W)$. Luego

$$\mathcal{D}_{\gamma_g}^{\pi^1, \ldots, \pi^W}(P^1, \ldots, P^W) := \sum_{i=1}^{N} \gamma_g^{\pi^1, \ldots, \pi^W}(p_i^1, \ldots, p_i^W) = \sum_{i}^{N} \left[\left(\sum_{k=1}^{W} \pi^k p_i^k g(p_i^k)\right) - m_i g(m_i)\right] \geq 0 \tag{198}$$

donde $m_i := \sum_{k=1}^{W} \pi^k p^k$.

7.5 Una propuesta de entropía

Con los conceptos mostrados en el capítulo Marco Teórico pudimos entender a la entropía como una representación de la información. Es el teorema de Kinchin el que expresa esta idea de forma muy concreta. Pero es posible generalizar el concepto de entropía propuesto por Shannon dando las principales características que debe tener una entropía en general. Hay muchas maneras de definir una entropía generalizada dependiendo del objetivo que tengamos. En los casos de las entropías HCT y Renyi se buscó generalizar al concepto de entropía mediante un parámetro. En el caso de la entropía de Salicrú la intención era englobar un conjunto de entropías mediante la definición de las dos funciones h y ϕ. En nuestro caso buscamos una entropía que se relacione conceptualmente con la divergencia γ. En los siguientes párrafos repasaremos las propiedades que un funcional tiene que cumplir para que se interprete como una entropía. Sea $H_G[P]$ un funcional de P se puede decir que es una entropía generalizada si cumple con las siguientes propiedades:

- ser continua para cada p_i
- ser no negativa
- $H_G[P, 0] = H_G[P]$
- ser igual a cero en el caso determinístico, es decir, cuando para un $p_i = 1$ y para los otros $p_i = 0$
- ser máxima cuando se tiene una distribución uniforme, es decir, cuando $p_i = \frac{1}{N}\ \forall i$
- ser una función cóncava respecto de P

Si tenemos en cuenta la última propiedad podemos definir una divergencia llamada coloquialmente "tipo Jensen" de la siguiente forma

$$D^G(P, Q) = H_G\left(\frac{P+Q}{2}\right) - \frac{1}{2}H_G[P] - \frac{1}{2}H_G[Q] \tag{199}$$

En esta tesis presentaremos una entropía generalizada que se relaciona con $D_{\gamma_g}(P, Q)$ a través de (199). Sea $f(x) = xg(x)$ una función convexa. Sea $R = \{1, 0, \cdots, 0\}$ una distribución de probabilidad discreta y sea $P = \{p_1, \cdots, p_n\}$ una distribución arbitraria. Con argumentos de mayorización tenemos

$$\sum_{i=1}^{n-1} r_i \geq \sum_{i=1}^{n-1} p_i \tag{200}$$

con

$$\sum_{i=1}^{n} r_i = \sum_{i}^{n} p_i \tag{201}$$

Entonces para cualquier función convexa f tenemos que

$$\sum_i f(r_i) \geq \sum_i f(p_i) \tag{202}$$

por lo tanto para cualquier distribución P y dada la definición de f y R tenemos que

$$g(1) \geq \sum_i p_i g(p_i) \tag{203}$$

Sea $Q = \{1/n, \cdots, 1/n\}$ la distribución uniforme. Sea $\phi(y)$ una función convexa. Si usamos la desigualdad de Jensen

$$\phi\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) \leq \frac{1}{n}\sum_{i=1}^{n}\phi(p_i) \tag{204}$$

y usando que $\phi(x) = f(x) = xg(x)$ tenemos que

$$\frac{1}{n}g(1/n) \leq \frac{1}{n}\sum_{i=1}^{n} p_i g(p_i) \tag{205}$$

$$n\frac{1}{n}g(1/n) \leq \sum_{i=1}^{n} p_i g(p_i) \tag{206}$$

$$\sum_{i=1}^{n}\frac{1}{n}g(1/n) \leq \sum_{i=1}^{n} p_i g(p_i) \tag{207}$$

Por lo tanto la función $\sum_{i=1}^{n} p_i g(p_i)$ es mínima cuando la distribución es uniforme. Estos resultados nos muestran que la función

$$H_g[P] = g(1) - \sum_{i=1}^{n} p_i g(p_i) \tag{208}$$

tiene un máximo cuando la distribución es uniforme ($P = \{1/n, \cdots, 1/n\}$) y es igual a cero cuando es determinística ($P = \{1, 0, \cdots, 0\}$). También podemos ver que la función $H_g[P]$ cumple con

$$H_g[P, 0] = H_g[P] \tag{209}$$

Otra importante propiedad es que es cóncava. En efecto, esto se deduce a partir de que $f(x)$ es convexa por definición. Sea $H_{h,\phi}[P] = h\left(\sum_i \phi(p_i)\right)$ la entropía definida por Salicrú. Si definimos a h como la identidad y ϕ como

$$\phi(p_i) = \frac{g(1)}{n} - p_i g(p_i) \tag{210}$$

obtenemos que $H_g[P]$ es un ejemplo de una entropía de Salicrú. Si usamos la definición (199) y remplazamos H_G con H_g tenemos

$$D^g(P,Q) = g(1) - \sum_i \frac{(p_i+q_i)}{2} g(m_i) - \frac{1}{2}g(1) + \sum_i \frac{p_i}{2} g(p_i) - \frac{1}{2}g(1) + \sum_i \frac{q_i}{2} g(q_i) \quad (211)$$

Haciendo unos pasos algebraicos obtenemos

$$D^g(P,Q) = \frac{1}{2}\left(\sum_i p_i g(p_i) + q_i g(q_i) - (p_i+q_i)g(m_i)\right) \quad (212)$$

donde $m_i = \frac{p_i+q_i}{2}$. Esto es igual a

$$D^g(P,Q) = \frac{1}{2}\mathcal{D}_{\gamma_g}(P,Q) \quad (213)$$

Esto nos dice que si usamos la divergencia "tipo Jensen" (199) y le aplicamos la entropía generalizada definida anteriormente H_g obtenemos la divergencia γ multiplicada por un factor. Es decir, que la entropía H_g guarda una estrecha relación con la divergencia γ definida al principio del capítulo. Además, esta forma de escribir la divergencia γ nos servirá para la siguiente sección en la búsqueda de una métrica.

7.6 Propiedad de métrica de las divergencias

A lo largo de esta tesis, no hemos con familias de divergencias entre distribuciones de probabilidad. Aún cuando ellas cumplen propiedades de una distancia entre elementos del simplex de las distribuciones de probabilidad discreta, no todas son una verdadera métrica. Por ejemplo, la divergencia de Kulback-Leibler no verifica la desicualdad triangular, en cambio se ha podido demostrar que la raiz cuadrada de la divergencia de Jensen-Shannon la verifica. Si estamos trabajando con un funcional como es en este caso la divergencia γ sería importante verificar si ésta la cumple o no, ya que nos permitiría definir una métrica. El funcional

$$E(X,P) = \sum_i (x_i - p_i)^2 \quad (214)$$

y la divergencia de Jensen Shannon

$$D_{JS}(Q||P) = \sum_i \frac{1}{2}p_i \cdot \ln(p_i) + \frac{1}{2}x_i \cdot \ln(q_i) - \frac{1}{2}(p_i+q_i) \cdot \ln\left(\frac{q_i+p_i}{2}\right) \quad (215)$$

tienen la particularidad que la raiz cuadrada de los dos cumplen la desigualdad triangular y por consiguiente son métrica. Hay varias formas de demostrar la desigualad triangular de $\sqrt{D_{JS}}$, la que tomaremos como referencia en este caso es la que se encuentra en el trabajo de Briet y Harremoes [37]. El objetivo es encontrar alguna condición para $g(y)$ tal que $\sqrt{D_{\gamma_g}}$ cumpla la desigualdad triangular. Ellos utilizan el siguiente teorema

Teorema: Sea (X, d) un espacio de distancia, luego $(X, d^{1/2})$ es un espacio métrico si y solo si (X, d) es un espacio definido negativo.

Ahora daremos la definición de espacio negativo. Sea (X, d) un espacio de distancia. Luego se dice que d es definida negativa si y solo si para todo conjunto finito de $(c_i)_{i\leq n}$ de números reales tales que $\sum_i c_i = 0$, y para todo conjunto finito de $(P_i)_{i\leq n}$ puntos en X se cumple que

$$\sum_{i,j} c_i c_j d(P_i, P_j) \leq 0 \tag{216}$$

En este caso se dice que (X, d) es un espacio de distancia de tipo negativo.

Entonces tenemos que demostrar que

$$\sum_{i,j} c_i c_j d(P_i, P_j) \tag{217}$$

es menor o igual que cero para todo conjunto finito de puntos de X y para todo conjunto de números reales $(c_i)_{i\leq n}$ y con la condición que su suma sea igual a cero. En nuestro caso la distancia en juego va a ser la divergencia γ multiplicada por un factor $1/2$.

$$d(P_i, P_j) = H_g\left(\frac{P_i + P_j}{2}\right) - \frac{1}{2}H_g(P_i) - \frac{1}{2}H_g(P_j) \tag{218}$$

El objetivo será encontrar funciones $g(y)$ que permitan definir un conjunto de distancia de tipo negativo. Luego si reemplazamos en la ecuación (217) y si tenemos en cuenta que $\sum_i c_i \sum_j c_j H_g(P_j) = \sum_j c_j \sum_i c_i H_g(P_i)$ ya que la sumatoria es sobre todo el conjunto de números c_i y sobre todo el conjunto de puntos de X, obtenemos la siguiente igualdad

$$\begin{aligned}\sum_{i,j} c_i c_j d(P_i, P_j) &= \sum_{i,j} c_i c_j \left[H_g\left(\frac{P_i + P_j}{2}\right) - \frac{1}{2}H_g(P_i) - \frac{1}{2}H_g(P_j)\right] = \\ &= -\sum_i c_i \sum_j c_j H_g(P_j) + \sum_{i,j} c_i c_j H_g\left(\frac{P_i + P_j}{2}\right)\end{aligned} \tag{219}$$

viendo que el factor $\sum_i c_i = 0$ en el primer sumando y luego reemplazando la definición de H_g teniendo en cuenta que $g(1)$ es una constante numérica tenemos

$$\begin{aligned}\sum_{i,j} c_i c_j d(P_i, P_j) &= 0 + \sum_{i,j} c_i c_j \left[g(1) - \sum_k \left(\frac{p_{ik} + p_{jk}}{2}\right) g\left(\frac{p_{ik} + p_{jk}}{2}\right)\right] = \\ &= -\sum_{i,j}\sum_k \left(\frac{p_{ik} + p_{jk}}{2}\right) g\left(\frac{p_{ik} + p_{jk}}{2}\right)\end{aligned} \tag{220}$$

El objetivo ahora es encontrar una función $g(y)$ que nos permita factorizar la sumatoria. Una buena propuesta es una función $g(y)$ que cumpla la siguiente condición

$$g(x+y) = \frac{h(x)h(y)}{x+y} \tag{221}$$

Como ejemplo podemos dar

$$g(x+y) = \frac{e^{x+y}}{x+y} \tag{222}$$

Para este caso resulta

$$\begin{aligned}\sum_{i,j} c_i c_j d(P_i, P_j) &= -\sum_{i,j} c_i c_j \sum_k \left(\frac{p_{ik}+p_{jk}}{2}\right) \frac{1}{\left(\frac{p_{ik}+p_{jk}}{2}\right)} h\left(\frac{p_{ik}}{2}\right) h\left(\frac{p_{jk}}{2}\right) = \\ &= -\sum_k \left[\sum_i c_i h\left(\frac{p_{ik}}{2}\right)\right]\left[\sum_j c_j h\left(\frac{p_{jk}}{2}\right)\right] = \\ &= -\sum_k \left[\sum_i c_i h\left(\frac{p_{ik}}{2}\right)\right]^2 < 0\end{aligned} \tag{223}$$

Esto significa

$$\sum_{i,j} c_i c_j d(P_i, P_j) < 0 \tag{224}$$

lo que implica que

$$\sqrt{d(P_i, P_j)} = \sqrt{H_g\left(\frac{P_i+P_j}{2}\right) - \frac{1}{2}H_g(P_i) - \frac{1}{2}H_g(P_j)} \tag{225}$$

con la $g(y)$ definida como (221) cumple con la desigualdad triangular lo que nos permite definirla como una métrica.

7.7 Aplicaciones

En esta sección mostraremos aplicaciones en señales reales y simuladas de la divergencia γ. Primero daremos la noción de significancia y su importancia en el análisis de series temporales y luego mostraremos los casos de análisis de series simuladas y reales, las dos con el método de la ventana.

7.7.1 Significancia

El concepto de la significancia surge de la necesidad de saber si dos distribuciones de probabilidad son realmente diferentes o estamos midiendo una fluctuación estadística. Este problema fue planteado inicialmente por Grosse y colaboradores [32] aplicado particularmente a la divergencia de Jensen-Shannon.

La idea es calcular la divergencia γ para un grupo de secuencias con una misma distribución de probabilidad. En un análisis teórico el valor de la divergencia sería cero, pero esto no ocurre por las fluctuaciones estadísticas. Generamos M secuencias de largo $L_i = N$ con un alfabeto de largo $L_\alpha = \alpha$ con una distribución teórica P. La distribución de ocurrencia de simbolos asociada a cada secuencia es entonces $\mathcal{P}_i$ para $i : 1, \cdots, M$. Para todos los posibles pares de secuencias calculamos la divergencia γ y luego medimos el valor medio y la desviación estándar como $\mu^{\mathcal{P}} = \langle \mathcal{D}_{\gamma_g}(\mathcal{P}_i||\mathcal{P}_j)\rangle$ y $\sigma^{\mathcal{P}} = \langle (\mathcal{D}_{\gamma_g}(\mathcal{P}_i||\mathcal{P}_j) - \mu^{\mathcal{P}})^2\rangle^{1/2}$ respectivamente, finalmente definimos la significancia como

$$\mathcal{S} = \mu^{\mathcal{P}} + \sigma^{\mathcal{P}} \tag{226}$$

Podemos decir entonces que dos distribuciones P y Q son estadísticamente distinguiblessi se cumple la condición

$$\mathcal{D}_{\gamma_g}(P||Q) > \mathcal{S} \tag{227}$$

Como se puede ver $\mathcal{S}$ depende del largo de la secuencia L_i y del largo del alfabeto L_α. Por lo tanto, podemos hacer una tabla de $\mathcal{S}$ para diferentes funciones $g(y)$ y para distintos valores de L_i y L_α. Mostramos un ejemplo para secuencias binarias en la tabla 1.

Tabla 1. Valores de significancia ($\mathcal{S}$) para diferentes funciones $g(y)$ aplicado a secuencias binarias $L_\alpha = 2$ con diferentes longitudes de cadena L_N. Todos los valores son del orden 10^{-3}.

L_N	e^x	$log(x)$	$\sqrt{x}$	$sinh(x)$
100	5.9	4.4	5.5	5.8
500	1.1	8.6	1.0	1.1
2000	0.57	4.1	5.4	5.8
5000	0.11	8.8	1.1	1.1
10000	0.06	4.2	5.5	5.9

En la práctica, cuando comparamos dos secuencias simuladas, cada una con una distribución de probabilidad, vamos a obtener dos significancias distintas una para cada secuencia simulada. Esto es consecuencia del método que elegimos para calcular la significancia. En estos casos lo que se hace es elegir la significancia mayor. Esto se debe a que al elegir la más grande estamos siendo más específicos y rigurosos al momento de decidir que valores de la divergencia γ son los que se deben tomar como cuantificadores de distinguibilidad y no como fluctuaciones estadísticas de las series.

7.7.2 Series Simuladas

Utilizamos la simulación con el método de la ventana para estudiar el comportamiento de la divergencia γ para diferentes funciones $g(x)$ como detector de cambios en la distribución

de probabilidad de una secuencia. En este ejemplo, generamos una secuencia compuesta de dos subsecuencias de longitud $L_{S_1} = L_{S_2} = L/2$. La subsecuencia de longitud $L_{S_k}(k = 1, 2)$ se genera a partir de una distribución de probabilidad $P_{s_k} = [s_k, 1 - s_k]$. Sobre cada realización de secuencias formamos una ventana de largo L_V centrada en el cursor , donde cada lado de la ventana (izquierda y derecha) tiene la misma longitud L_{V_I} y L_{V_D} respectivamente.

Movemos el cursor y estimamos una distribución $P_I(i)$ basada en las frecuencias relativas de 0 y 1 en el lado izquierdo de la ventana (V_I) y hacemos lo mismo para el lado derecho de la ventana (V_D), es decir, $P_D(i)$. Calculamos la divergencia γ para cada posición i, $D_{\gamma_g}(i) = D_{\gamma_g}(P_I(i)||P_D(i)))$ donde el cursor i recorre toda la secuencia (desde L_{V_I} hasta $L - L_{V_I}$). El máximo de esta cantidad $D_{\gamma_g}(i) = D_\gamma max$, en función de la posición del cursor, se interpreta como la detección de una posición donde cambia la distribución de símbolos. La Figura 6 muestra la divergencia γ para cuatro $g(x)$ diferentes aplicadas sobre una secuencia binaria combinada. Los primeros 2000 puntos se generaron con una distribución de probabilidad $P = [0.5, 0.5]$ y los siguientes 2000 puntos con $Q = [0.4, 0.6]$. El análisis se realizó utilizando cuatro funciones diferentes $(e^x, \log(x), \sqrt{x}, sinh(x))$ que satisfacen la condición formal dada en el teorema principal de este capítulo. La longitud de las ventanas que se utilizaron son de $L_{V_I} = L_{V_D} = 1000$ puntos de datos. La línea de guión horizontal es el valor de significancia $\mathcal{S}(g)$ tomado de la tabla 1 y la línea de guión vertical determina el punto donde la divergencia alcanza su máximo. Se puede ver claramente que se detecta los cambios en el punto exacto donde la secuencia cambia de distribución de probabilidad.

Con la idea de estudiar el límite de detección de la divergencia, generamos de la misma manera que antes, cuatro secuencias combinadas con distribuciones de probabilidad cada vez más parecidas o cercanas. Las secuencias se analizaron con la función $g(x) = e^x$ y se tomó la significancia $\mathcal{S}(g)$ de la tabla 1. La longitud de las ventanas utilizadas fue $L_{V_I} = L_{V_D} = 1000$ puntos de datos. La Figura 7 muestra que a medida que las distribuciones de probabilidad se vuelven cada vez más similares el valor máximo de divergencia $(D_\gamma max)$ se acerca al umbral de significancia, sin poder alcanzarlo cuando $P = [0.5, 0.5]$ y $Q = [0.51, 0.49]$ (Figura 7 D). Para este caso, la divergencia γ no puede distinguir el cambio entre las dos secuencias. Como pudimos ver, el umbral de detección depende de la función $g(x)$ utilizada, así como de la longitud de las ventanas tomadas para el análisis. Como se explicó en el capítulo anterior, la suavidad en las curvas de los gráficos se debe a que hacemos un promedio de los valores de la divergencia γ para cada punto del cursor de la ventana. El promedio se hizo sobre 1000 señales simuladas.

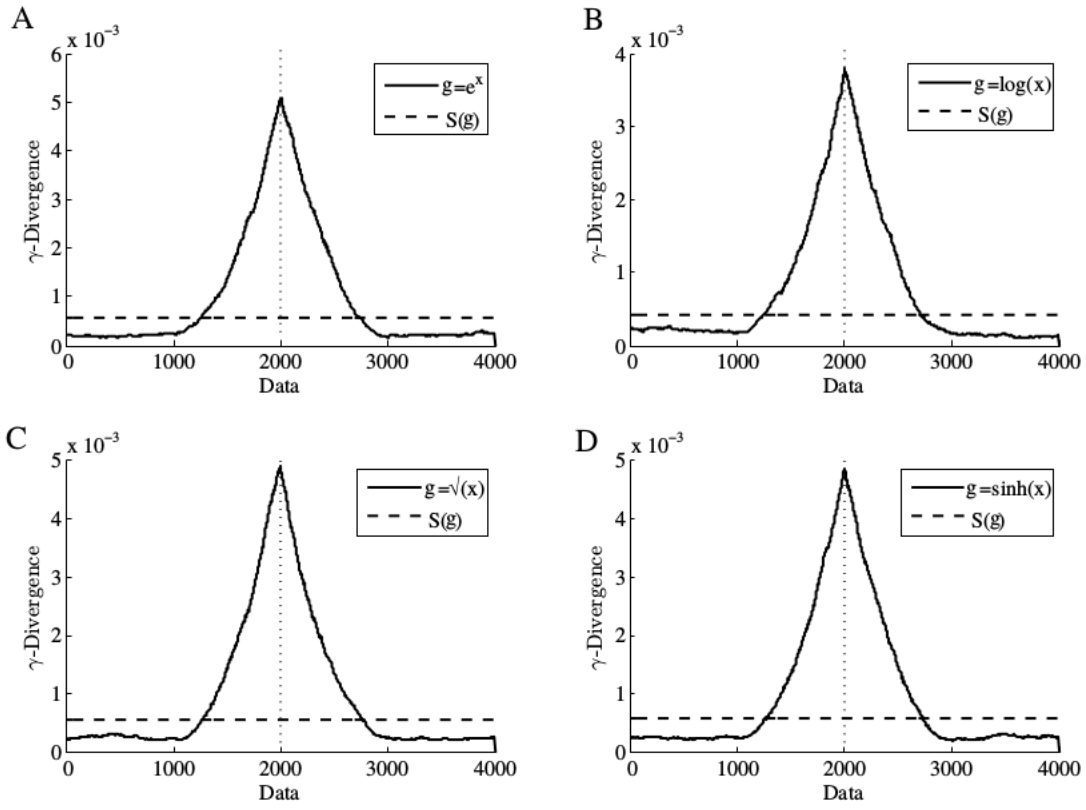

Figura 6. Análisis sobre secuencias binarias combinadas, los primeros 2000 puntos tienen una distribución $P = [0.5, 0.5]$ y los segundos 2000 puntos una distribución $Q = [0.4, 0.6]$. La longitud de las ventanas son de $L_{V_I} = L_{V_D} = 1000$. La línea horizontal representa la significancia para cada una de las funciones $g(x)$ usadas (sacadas de la tabla 1). Se puede ver que para todas las funciones $g(x)$ usadas la divergencia detecta el cambio de distribución en la secuencia total.

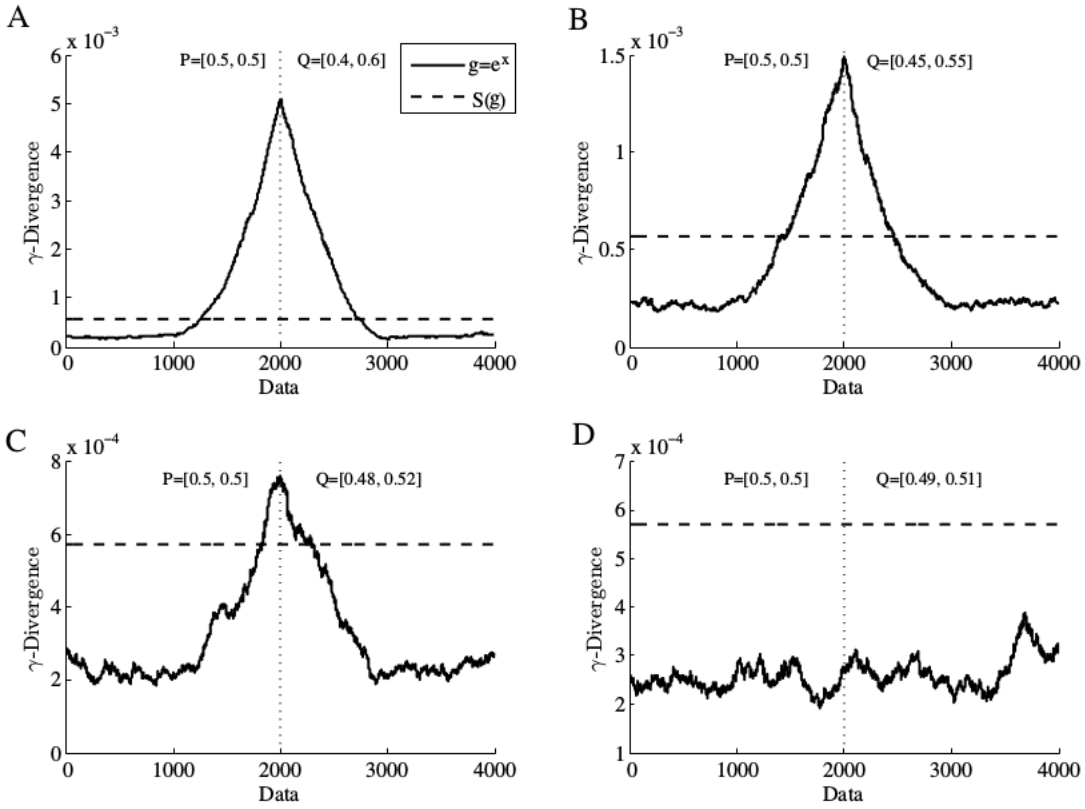

Figura 7. Análisis de una secuencia binaria combinada usando la divergencia γ para dos distribuciones de probabilidad P y Q cada vez más similares. La función usada es $g(x) = e^x$ y la longitud de la ventana es $L_{V_I} = L_{V_D} = 1000$. La línea horizontal punteada es la significancia $\mathcal{S}(g)$, y la línea vertical punteada representa el máximo de la divergencia para los casos A, B y C. Como P y Q son cada vez más similares el máximo de la divergencia se acerca al valor de la significancia. En el caso D los valores de la divergencia no pueden superar al valor de la significancia mostrando que las dos subsecuencias son indistinguibles para la divergencia.

7.7.3 Series Reales

En el segundo ejemplo, utilizamos la divergencia γ para detectar la transición de estados de un EEG (electroencefalograma) en un paciente en estado de sueño. El sueño es una actividad dinámica, durante la cual tienen lugar muchos procesos vitales para la salud y el bienestar. Es esencial para ayudar a mantener el estado de ánimo, la memoria y el rendimiento cognitivo . Hay cinco etapas de sueño. La etapa sws 1 y sws 2 representan un sueño ligero, en el que se puede entrar y salir del sueño y puede despertarse fácilmente. Las etapas sws 3 y sws 4 tienen ondas cerebrales extremadamente lentas y estas son las etapas de sueño más profundo. Durante el REM (movimiento ocular rápido), las ondas

cerebrales imitan la actividad durante el estado de vigilia. Estas cinco etapas progresan cíclicamente desde sws 1 hasta REM y luego comienzan de nuevo en sws 1. Es muy importante que estos ciclos se mantengan para la salud. Para esto, se miden varios parámetros electrofisiológicos como el EEG. El desarrollo de herramientas que puedan detectar los cambios en las etapas del sueño a través de la señal EEG es esencial para el estudio de pacientes con trastornos del sueño. Los datos fueron tomados del banco de datos de Physionet The Sleep-EDF Database Expandido [38]. La grabación del canal EEG fue (Fpz-Cz) con un solo electrodo puesto en la parte frontal de la cabeza, y la frecuencia de muestreo fue de 100 Hz. Las señales tienen 6 etapas diferentes: despertar, ojos cerrados, REM, sws 1, sws 2, sws 3 y sws 4. Las grabaciones para cada etapa son 6000 puntos de datos (60 segundos) formando una señal de 36000 puntos como se muestra en Figura 8A. Las señales se procesaron previamente con un filtro de paso de banda entre $0.5-60$ Hz (es decir, las frecuencias que nos importan ya que son las frecuencias fisilógicas del cerebro). Mapeamos la señal utilizando el método de vector de permutación, con los parámetros $d = 4$ y $\tau = 1$. Usando una ventana de largo $L_{V_I} = L_{V_D} = 1000$, la divergencia γ se aplicó usando cuatro funciones diferentes $(g(x) = e^x, \log(x), \sqrt{x}, sinh(x))$. Como se muestra en la Figura 8B, la divergencia pudo revelar el paso entre todas las etapas del sueño, para las cuatro funciones $g(x)$, siendo $g(x) = \log(x)$ la que mejor detecta. El progreso entre sws 2 y sws 3 tiene los valores más altos, esto se debe a que cuando un paciente entra en un sueño profundo, el cuerpo se vuelve menos sensible a los estímulos externos, por lo que la información que el cerebro debe manejar es más pequeña, haciendo que las ondas sean más lentas. Las ondas cerebrales son muy similares a las que aparecen en estado de coma. La transición entre los estados sws 1 y sws 2 no está absolutamente clara, esto podría basarse en que estas dos etapas son de sueño ligero.

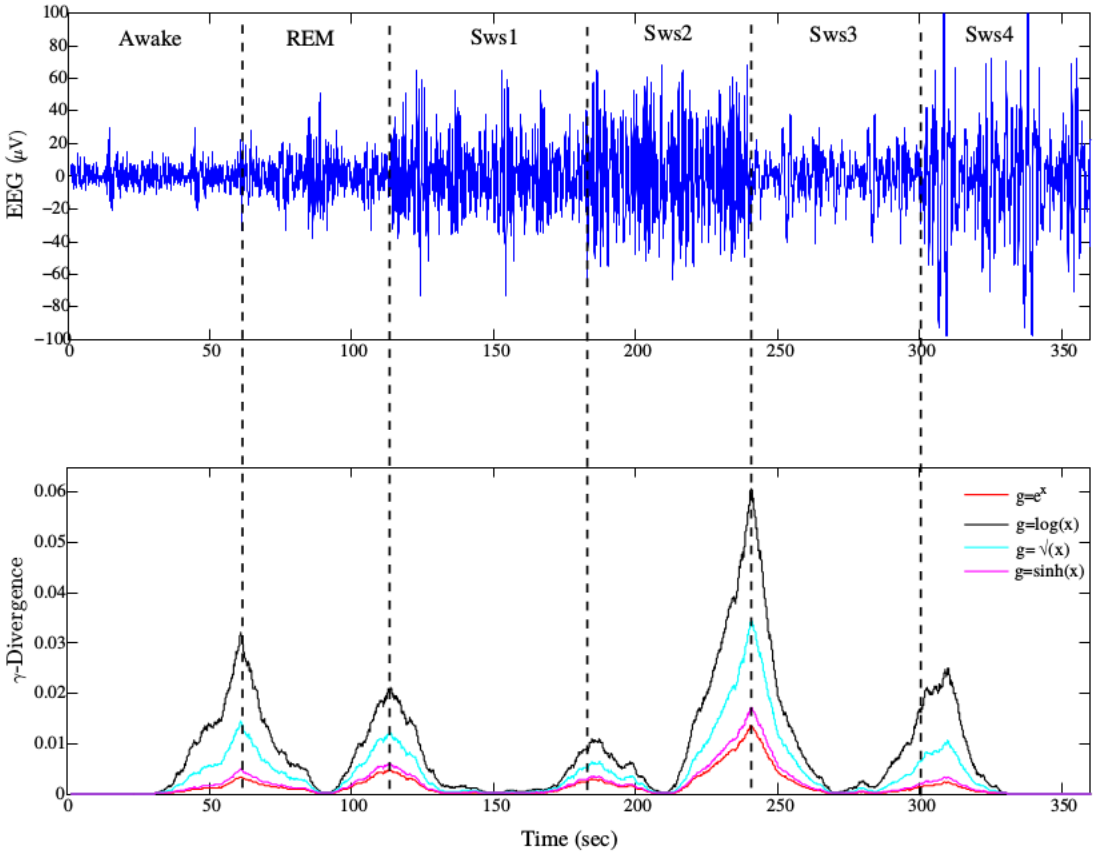

Figura 8. Análisis de la divergencia γ sobre una señal de EEG referida al sueño usando el método de la ventana. Usamos cuatro funciones $g(x)$ distintas para el análisis. La transición entre los estados de sueño es sobre un tiempo de 60 segundos. La señal fue mapeada mediante el método de vector de permutación con parámetros $d = 4$ y $\tau = 1$. Los valores máximos de la divergencia aparecen en la transición entre los distintos estados de sueño.

7.8 Conclusiones

En primer lugar pudimos ver que existe una estrecha relación entre el cuadrado de la métrica euclídea y la divergencia de Jensen-Shannon estableciendo que pertenecen a una misma familia de funcionales. Esto nos permitió introducir una familia de funcionales que dependen de funciones convexas y así ver que tanto la divergencia de Jensen-Shannon como la métrica euclídea son ejemplos particulares de esta familia de funcionales. El hecho que la definición de la divergencia γ sea a partir de funciones convexas nos permitió demostrar que este funcional generalizado cumple con todos los requisitos de una divergencia, en este caso generalizada para un grupo de funciones convexas. Esto muestra, a primera vista,

la cercana relación que tienen las funciones convexas y las divergencias, por lo menos en este caso. En segundo lugar pudimos generalizar esta divergencia para más de dos distribuciones de probabilidad, permitiendo así, una posible aplicación en un análisis de varias señales a la vez.

Seguidamente, se logró definir una entropía generalizada también a partir de las propiedades de las funciones convexas. Esta entropía incluye a la entropía de Shannon si elegimos a $g(x) = \log x$, y es un caso particular de una familia de entropías más grande llamadas Salicrú. Como pudimos ver la entropía generalizada definida por nosotros cumple con todos los requisistos que se le puede pedir a una entropía en el contexto de la teoría de la información. Al final de esa sección tambíen pudimos ver que esta entropía generalizada guarda una estrecha relación con la divergencia γ vía la fórmula de la divergencia "tipo Jensen" que depende de entropías generalizadas. También se pudo encontrar, gracias a esta forma de escribir la divergencia γ, que existe un conjunto de funciones $g(x)$ para las cuales la raiz de la divergencia γ es una métrica, dándole un valor agregado a esta divergencia generalizada.

Por último, se verificó la eficiencia de la divergencia utilizando el método de segmentación por ventana en aplicaciones a series simuladas y reales. Para una consistente aplicación se tuvo que recurrir a la definición del concepto de significancia para tener en claro cuando dos distribuciones eran en verdad distinguibles. Tanto en series simuladas como en las aplicaciones a EEG en etapas de sueño se pudo observar que la divergencia detectaba los cambios de la señal para distintas funciones $g(x)$, mostrando así la eficiencia de la misma.

8 Capítulo VII: Complejidad

El concepto de complejidad atraviesa varias áreas de la ciencia, y por ende no tiene un solo significado. Existe la noción de lo que puede ser complejo en el colectivo de las personas, pero en cada rama de la ciencia existe una definición distinta y acorde a lo que se busca describir. Por eso dedicaremos unos párrafos para aclarar qué es lo que buscamos representar cuando decimos que algo es complejo. Estudiaremos la complejidad desde la perspectiva de la teoría de la información brindando una nueva definición basada en resultados mencionados anteriormente en la tesis. Daremos una definición de la complejidad desde la perspectiva de la estadística. Una de estas consideraciones es definir que sistemas tienen una baja complejidad, es decir, que consideramos como sistemas no complejos.

Para esto es bueno hacer una comparación con la complejidad de Kolmogorov para tener una guía en el camino hacia nuestra definición.

Máquina de Turing: Haremos una descripción informal de la máquina de Turing ya que solo necesitamos entender el concepto. Pensaremos a una máquina de Turing como una máquina que posee un algoritmo, un cabezal de lectura y escritura y una cinta o cadena de caracteres (que puede ser infinita). El procedimiento se basa en leer el caracter que está escrito en la cinta, borrarlo y escribir un caracter nuevo para luego moverse a la derecha o izquierda dependiendo de lo que exprese la función transición. Esta función la podemos definir de la siguiente manera. Sea P el conjunto de algoritmos (escritos en binario) y G el conjunto de cadenas simbólicas. La máquina de Turing queda definida por la siguiente función transición

$$\phi : P \longrightarrow G \tag{228}$$

Se dice que $p \in P$ es una descripción de $s \in G$ si

$$\phi(p) = s \tag{229}$$

Teniendo esta definición estamos en condiciones de describir la

Complejidad de Kolmogorov: Formalmente hablando se define a la complejidad de Kolmogorov como

$$K(s)_\phi = \min\{|p| : \phi(p) = s\} \tag{230}$$

Donde $|P|$ es la cantidad de bits del algoritmo que describe a s. Esto nos dice que la complejidad de Kolmogorov no es más que la cantidad de bits de la mínima descripción de s. A simple vista se puede observar que la complejidad de Kolmogorov es incomputable, es decir, no existe una máquina de Turing tal que dada una cierta cadena de símbolos s nos de la cantidad de bits de la mínima descripción.

Supongamos que tenemos una secuencia binaria y queremos cuantificar su complejidad. Si quisiéramos hacerlo desde la perspectiva de la complejidad de Kolmogorov, es decir, con

una complejidad algorítmica, nos encontraremos que para las secuencias que tengan algún tipo de patrón (y en un caso extremo que sea una cadena en que aparezca un solo símbolo en todos los casilleros) la complejidad es baja. En cambio, para secuencias en las cuales no existe ningún patrón aparente, es decir, secuencias aleatorias la complejidad tiende al largo de la cadena, esto es, a su máximo. Desde esta perspectiva, el algoritmo que puede describir la cadena es la cadena misma. Por esta razón tiene una complejidad alta pues no podemos describir o generar la cadena con un algoritmo que tenga un largo menor ya que no hay un patrón aparente. Pero si miramos el problema desde la perspectiva de la estadística, la situación es muy distinta. Empecemos mostrando una deficiencia de usar estadística para analizar la estructura. Supongamos que tenemos dos cadenas binarias como las siguientes

$$01010101010101010101010101010101 \tag{231}$$

$$00011011100110100110011100011 \tag{232}$$

La primer cadena, como podemos observar, es periódica, en cambio, la segunda se asemeja a una secuencia aleatoria ya que no hay un patrón aparente. Estructuralmente son muy distintas. Si las analizamos con una complejidad algorítmica la primera tendría una complejidad baja (ya que podemos describirla, informalmente, como "15 veces 01") y la segunda una complejidad alta. En este caso la complejidad algorítmica representaría muy bien a la complejidad estructural de la cadena. Pero desde la perspectiva de la estadística las dos cadenas tienen la misma distribución de probabilidad $P = \{1/2, 1/2\}$. Pero la situación cambia si tomamos como referencia lo que se entiende como un sistema físico complejo. Por ejemplo un sistema periódico como un cristal y un sistema aleatorio como un gas son sistemas que desde la física estadística son poco complejos. Si volvemos al caso de la secuencia aleatoria, estadísticamente es poco compleja, ya que tiene un distribución uniforme, como la periódica. Es decir, necesitamos poca imformación para describir estadísticamente a los dos sistemas. Es por eso que para una complejidad estadística una secuencia periódica y una aleatoria tienen baja complejidad. Nosotros definiremos una complejidad desde esta perspectiva, es decir, una complejidad que interprete lo que es estadísticamente complejo. Nuestro modelo a seguir de complejidad será la propuesta por López-Ruiz, Mancini y Calbet (LMC) [39], que definieron una complejidad desde la mecánica estadística.

Complejidad de LMC: Supongamos que tenemos un sistema físico el cual cumple con la estadística de Boltzmann-Gibbs. Esto nos dice que la entropía estadística del sistema esta definida como

$$H(p_1, \cdots, p_N) = -k \sum_{i=1}^{N} p_i \log p_i \tag{233}$$

Si hacemos una expansión de Taylor alrededor de $H_{max} = k \log N$

$$H(p_1, \cdots, p_N) = k \log(N) - \frac{Nk}{2} \sum_{i=1}^{N} \left(p_i - \frac{1}{N} \right)^2 + \cdots \tag{234}$$

Llamando a $D = \sum_{i=1}^{N} \left(p_i - \frac{1}{N} \right)^2$ desequilibrio y multiplicando por H a ambos lados tenemos

$$H^2 = H \cdot H_{max} - \frac{Nk}{2}(H \cdot D) + k^2 f(N, P) \tag{235}$$

Se define a la complejidad como $C_{LMC} = H \cdot D$. Entonces

$$C_{LMC} = cte \cdot H(H_{max} - H) + \frac{2k}{N} f(N, P) \tag{236}$$

Es aquí donde se puede ver mejor el concepto de desequilibrio. Esta definición de complejidad cumple con propiedades que buscamos en este tipo de medidas. Por ejemplo para un sistema perfectamente ordenado se tiene que $H = 0$ y por lo tanto $C_{LMC} = 0$. Para un sistema completamente desordenado se tiene que $D = 0$ y por lo tanto $C_{LMC} = 0$. La C_{LMC} para otros sistemas tendrá valores intermedios (siempre positivos).

Desde la teoría de la información se ha propuesto un idea que conserva las propiedades fundamentales de la complejidad C_{LMC} reemplazando a D por la divergencia de la Jensen-Shannon. Esta propuesta fue mostrada por primera vez por Lamberti, Rosso y Plastino (LRP) [40].

Complejidad LRP: Definimos a la complejidad LRP de la siguiente manera

$$C_{LRP}[P] = H[P] \cdot D_{JS}(P||1/N) \tag{237}$$

Por las propiedades de la divergencia y de la entropía se puede ver que para un sistema completamente desordenado tenemos que $C_{LRP} = 0$ y para un sistema perfectamente ordenado se tiene que $C_{LRP} = 0$. La complejidad, de la misma manera que C_{LMC}, tendrá valores intermedios (siempre positivos). En este trabajo generalizaremos esta complejidad.

Como se puede ver cualquiera de estas medidas de complejidad están dentro un contexto referido a la complejidad estadística, en la cual un sistema determinista (con distribución $P = \{1, 0, \cdots, 0\}$) y un sistema completamente aleatorio (con distribución uniforme) tienen una complejidad igual a cero. Es importante para interpretar correctamente los valores de una complejidad estadística del estilo de las que vimos anteriormente, que tanto la entropía como lo que se utilice como desequilibrio estén normalizados.

Nosotros propondremos en esta tesis una complejidad que generaliza la complejidad propuesta por Lamberti, Rosso y Plastino ($\mathcal{C}_{LRP}$), por lo tanto es una complejidad estadística que es función de una distribución de probabilidad. Nuestra propuesta es la siguiente

$$\mathcal{C}_g(P) = H_g(P) \cdot \mathcal{Q}_0 \mathcal{D}_{\gamma_g}(P||P_e) = H_g(P)\mathcal{Q} \tag{238}$$

donde $H_g(P) = g(1) - \sum_i p_i g(p_i)$ es la entropía generalizada expuesta en la ecuación (208) y $\mathcal{D}_{\gamma_g}$ es la divergencia γ propuesta en el capítulo anterior. La constante Q_0 es una constante de normalización. Como dijimos anteriormente es necesario normalizar tanto la entropía como al desequilibrio. Para esto hay que encontrar la constante $\mathcal{Q}_0$. Vale recordar que la función $g(x)$ es tal que $x \cdot g(x)$ sea convexa. Para normalizar la entropía H_g tenemos que dividirla por su valor máximo, esto es

$$\mathcal{H}_g(P) = \frac{H_g(P)}{H_g(P_e)} = \frac{g(1) - \sum_{i=1}^{N} p_i g(p_i)}{g(1) - g(1/N)} \tag{239}$$

donde $P_e = \{1/N, \cdots, 1/N\}$. Ahora nos falta normalizar el desequilibrio generalizado, para eso tenemos que encontrar la constante $\mathcal{Q}_0$. Esta constante se calcula mediante la divergencia γ. Supongamos que tenemos la divergencia γ de dos distribuciones P y Q

$$\mathcal{D}_{\gamma_g}(P||Q) = \sum_{i=1}^{N} p_i g(p_i) + q_i g(q_i) - (p_i + q_i) g\left(\frac{p_i + q_i}{2}\right) \tag{240}$$

donde $P = \{1, \cdots, 0\}$ es una distribución determinista y $Q = P_e$ es una distribución uniforme, resultando

$$\frac{1}{\mathcal{Q}_0} = g(1) + g\left(\frac{1}{N}\right) - \left(\frac{N+1}{N}\right) g\left(\frac{N+1}{2N}\right) - \left(\frac{N-1}{N}\right) g\left(\frac{1}{2N}\right) \tag{241}$$

Entonces podemos generalizar el concepto de desequilibrio de la siguiente manera

$$\mathcal{Q}_g[P] = \mathcal{Q}_0 \cdot \mathcal{D}_{\gamma_g}(P||P_e) = \mathcal{Q}_0 \left[g\left(\frac{1}{N}\right) + \sum_{i=1}^{N} p_i g(p_i) - \left(p_i + \frac{1}{N}\right) g\left(\frac{p_i}{2} + \frac{1}{2N}\right) \right] \tag{242}$$

Por lo tanto la complejidad estadística generalizada es

$$\mathcal{C}_g[P] = \mathcal{H}_g[P] \cdot \mathcal{Q}_g[P] \tag{243}$$

Para entender estas ideas de información, desequilibrio y complejidad estadística es importante ver cómo se comportan en función de diferentes distribuciones de probabilidad. Para eso tomaremos un conjunto de distribuciones de probabilidad, que cumplen con la condición de normalización $\sum_i p_i = 1$, que van desde la distribución determinista $P = \{1, 0, \cdots, 0\}$ hasta la distribución uniforme $P_e = \{1/N, \cdots, 1/N\}$. Tomaremos para este cálculo $N = 4$. En la figura 9A se pueden ver los cálculos, para diferentes distribuciones de probabilidad, del desequilibrio, de la entropía y de la complejidad estadística para $g(x) = e^x$. Como se puede observar en este gráfico el desequilibrio es máximo cuando la distribución es determinista y mínimo cuando la distribución es uniforme. Contrariamente la entropía tiene su mínimo para la distribución determinista y su máximo para la

distribución uniforme. Por su lado la complejidad estadística tiene un mínimo en estos dos extremos y un máximo en un valor intermedio.

Por otro lado en la figura 9B mostramos el cálculo de la complejidad en función de la distribución de probabilidad para diferentes funciones $g(x)$ que cumplan con el teorema expuesto en el capítulo anterior. Las funciones $g(x)$ utilizadas son $e^x, \log(x), \sqrt{x}$ y $sinh(x)$.

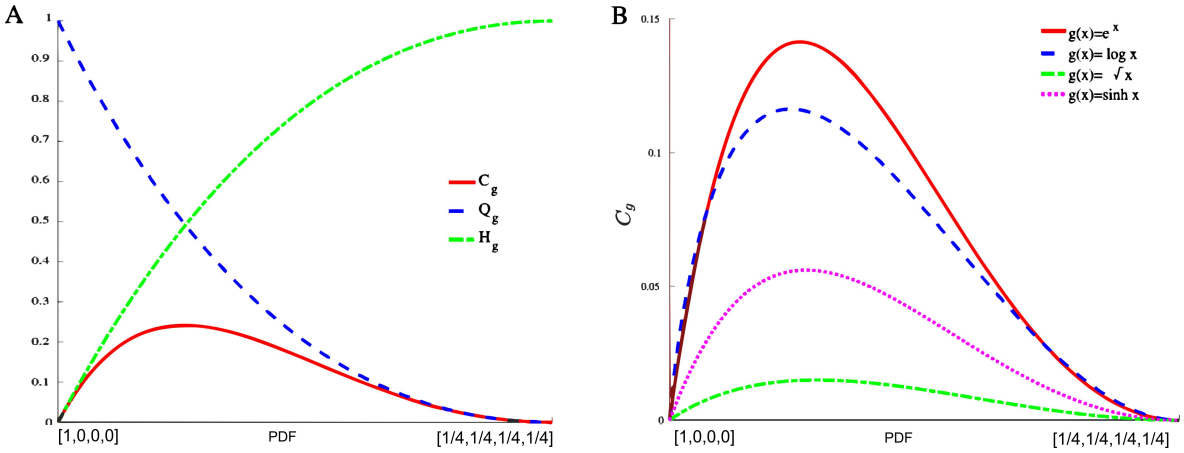

Figura 9. A)Valores de la entropía generalizada $\mathcal{H}_g$, del desequilibrio generalizado $\mathcal{Q}_g$ y de la complejidad estadística $\mathcal{C}_g$ para todas las posibles distribuciones de probabilidad $P = \{p_1, p_2, p_3, p_4\}$ que van del perfecto orden (estadísticamente hablando) hasta la distribución uniforme usando $g(x) = e^x$. B) Muestra la complejidad estadística $\mathcal{C}_g$ para el mismo grupo de distribuciones de probabilidad de A) para $g(x) = \{e^x, \log(x), \sqrt{x}, sinh(x)\}$. Los mismos resultados se obtienen si tomamos distribuciones de probabilidad con diferente N.

.

8.1 Conclusiones

Lo que se buscó en este capítulo es una medida que cuantifique cuán complejas son, estadísticamente hablando, las series temporales. Para esto tuvimos que definir qué es para nosotros complejo y que no, y adecuar nuestra definición de complejidad a estos requerimientos. Siguiendo el esquema de Lamberti, Rosso y Plastino fue necesario definir el desequilibrio y cuál entropía debíamos usar. Por una cuestión de consistencia teórica al elegir a la divergencia γ como desequilibrio, fue necesario utilizar a la entropía H_g definida en el capítulo anterior ya que se relaciona de forma directa con esta divergencia. Ya definida la complejidad se pudo verificar que cumple con los requisitos que le habíamos impuesto.

9 Capítulo VIII: Caos y Ruido

A diferencia de los capítulos anteriores donde el objetivo era hacer un aporte teórico en la definición de cuantificadores de distinguibilidad o de complejidad, en este nos enfocamos en el análisis de series temporales dadas ciertas herramientas teóricas introducidas en el marco teórico. Como se sabe, los procesos caóticos y los ruidos tienen un origen teórico antagónico (unos son deterministas y los otros estadísticamente aleatorios). Aún así comparten un grado de impredecibilidad muy importante. Es por eso que las señales producidas por estos pueden ser mal interpretadas. Nos enfocaremos, mediante el uso de herramientas de teoría de información conocidas en la literatura y ciertos métodos de segmentación, en realizar un análisis de series temporales producidas por estos procesos con la intención de distinguirlos o diferenciarlos.

Si bien la medida para cuantificar la disimilitud entre la series temporales será la divergencia de Jensen-Shannon, aquí usaremos un mapeo distinto a los usados en los capítulos anteriores. Como detalleremos luego haremos dos procesos de mapeo: el primero es pasar la serie temporal a un secuencia simbólica binaria, y el segundo consiste en transformar la cadena simbólica binaria en una cadena alfabetizada. En primera instancia daremos conceptos teóricos de sistemas dinámicos y de ruidos coloreados dando algunos ejemplos para esclarecer. En segundo lugar presentaremos el modelo de mapeo y segmentación utilizado. Luego definiremos el concepto de significancia usado. Y por último mostraremos los resultados para un conjunto de sistemas caóticos y de ruidos coloreados.

9.1 Teoría de los Sistemas Dinámicos

Un buen comienzo para el estudio de sistemas caóticos es definir con claridad lo que es un sistema dinámico, ya que los procesos caóticos son un caso particular de ellos. Entendemos a un Sistema Dinámico como un sistema que evoluciona en el tiempo. Un sistema dinámico puede estar descripto por un conjunto de ecuaciones diferenciales, ecuaciones integrales, ecuaciones en diferencias finitas o una mezcla de todas. Llamaremos *estados* del sistema a los valores que irá tomando mientras el tiempo transcurre y se le dirá *Espacio de Estados* $\mathcal{E}_{\mathcal{X}}$ al conjuntos de esos valores. Utilizaremos el término *camino trazado* para referirnos al conjunto de valores que toma el sistema (o sea los estados) mientras evoluciona dada una condición inicial. Para terminar, se le dirá *Diagrama de Fase* al conjunto de todas las trayectorias posibles. Este nos ayudará a entender el comportamiento del sistema dinámico dada cualquier condición inicial.

Nos vemos tentados a relacionar un sistema dinámico con el concepto de serie temporal. Como pudimos ver una serie temporal no es más que una serie de valores ordenados por un parámetro. En este contexto diremos entonces que un camino (una serie de estados) es una serie temporal, con la particularidad que los valores para un tiempo $t = n$, o sea, x_n, dependen explícitamente del valor inicial x_0. Esto se debe por supuesto a que son sistemas deterministas. Una diferencia entre el análisis de series temporales y el estudio

de sistemas dinámicos es que a estos últimos se los estudian en situaciones límite, es decir, trataremos de entender el comportamiento para valores grandes de tiempo.

Definición: Sistemas Dinámicos

Llamamos Sistema Dinámico a la terna $(\mathcal{E}_\mathcal{X}, \mathcal{T}, \mathbf{f})$, donde $\mathcal{E}_\mathcal{X}$ es el espacio de fase o de estados; $\mathcal{T}$ es el conjunto de tiempo y $\mathbf{f}$ es el flujo (en el caso discreto se lo llama mapa) del sistema y cumple con la aplicación $f : \mathcal{T} \times \mathcal{E}_\mathcal{X} \longrightarrow \mathcal{E}_\mathcal{X}$. y con las propiedades:

1. $\mathbf{f}(\cdot, \cdot)$ es continua
2. $\mathbf{f}(0, x) = x$ para toda $x \in \mathcal{E}_\mathcal{X}$
3. $\mathbf{f}(t, \mathbf{f}(s, x)) = f(t + s, x)$ para todo $t, s \in \mathcal{T}$ y $x \in \mathcal{E}_\mathcal{X}$

Diremos que el sistema dinámico es discreto si el conjunto $\mathcal{T}$ es un subconjunto de los números enteros, y fluye de modo creciente y de un paso, o sea, $t : 1, 2, \ldots, n$. Si $\mathcal{T}$ es $\mathbb{R}^+$ o $\mathbb{R}$ diremos que es continuo.

Como dijimos anteriormente un sistema dinámico va a estar dado por una ecuación en diferencias o una ecuación diferencial depende del caso. Si estamos trabajando con un sistema discreto la función flujo $\mathbf{f}$ será una relación recursiva del tipo

$$x_{k+1} = \mathbf{f}(x_k) \tag{244}$$

Como se puede ver consideramos sistemas *autónomos*, es decir, sistemas que no dependan explícitamente del tiempo. Esto nos permite considerar arbitrariamente el instante de tiempo inicial en $t = 0$ ya que el estado del sistema solo depende del estado anterior y no del instante de tiempo.

Si el sistema fuese continuo podría presentarse de la forma de una ecuación diferencial, por ejemplo

$$\frac{dx}{dt} = \mathbf{f}(x(t)) \tag{245}$$

Nosostros nos centraremos en los sistemas discretos ya que trabajamos con secuencias simbólicas. A continuación daremos un conjunto de definiciones que nos ayudarán en el futuro para poder interpretar correctamente los ejemplos que utilizaremos.

Definición: Trayectoria

Ya teniendo definido nuestro sistema dinámico podemos definir, para cada $x \in \mathcal{X}$, una *sucesión* de puntos $\mathcal{E}_\mathcal{X}$

$$\gamma(x, \mathbf{f}) = \Big(x, \mathbf{f}(x), \mathbf{f}^2(x), \ldots, \mathbf{f}^n(x), \ldots\Big) \tag{246}$$

y se define como Órbita al *conjunto* de los puntos de la *trayectoria*, es decir,

$$O(x, \mathbf{f}) = \Big\{x, \mathbf{f}(x), \mathbf{f}^2(x), \ldots, \mathbf{f}^n(x), \ldots\Big\} \tag{247}$$

donde $\mathbf{f}^k(x) = \overbrace{\mathbf{f} \circ \mathbf{f} \circ \cdots \circ \mathbf{f}}^{k}$, es decir, la composición de la función $\mathbf{f}$, k veces. A la sucesión (246) se la conoce como *trayectoria de x bajo f*. La interpretación que uno le puede dar a esta sucesión $\gamma(x, \mathbf{f})$ es la siguiente: supongamos que tenemos un objeto que se mueve bajo ciertas leyes deterministas y para el tiempo $t = 0$ se encuentra en la posición x; entonces en el tiempo $t = 1$, el mismo objeto se econtrará $\mathbf{f}(x)$, y para el tiempo $t = 2$, estará $\mathbf{f}\Big(\mathbf{f}(x)\Big)$ y así sucesivamente.

Para el estudio adecuado de los sistemas dinámicos tendremos que analizar todas las órbitas posibles, es decir el mapa $\psi_t : \mathcal{E}_\mathcal{X} \longrightarrow \mathbb{R}$. También es útil estudiar el estado asintótico de estas órbitas, esto es, estudiar cuando n tiende a infinito. De hecho un comportamiento asintótico que nos importa es el estado estacionario del sistema. A algunos de estos estados se los conoce como puntos de equilibrio (puntos fijos), órbitas periódicas, eventualmente periódicas, estables y conjuntos límite. Muchos de estos comportamientos asintóticos son casos particulares de un grupo general llamado *Atractor*. Un Atractor es un *Conjunto Invariante*, es por esto que es adecuado a la cronología del conocimiento definir primero lo que es un conjunto invariante.

Definición: Conjunto Invariante

Se dice que $\mathcal{V} \subseteq \mathcal{E}_\mathcal{X}$, siendo $\mathcal{E}_\mathcal{X}$ el espacio de los estados, es un conjunto invariante del sistema dinámico definido por $\mathbf{f}$, si $\mathcal{V} = \mathbf{f}\Big(\mathcal{V}\Big)$. Esto es, si el resultado de la iteración del conjunto $\mathcal{V}$ es el propio conjunto.

Definicón: Atractor

Entendemos intuitivamente al atractor como la región del espacio de los estados al cual convergen las órbitas dadas ciertas condiciones iniciales. Podemos definir a los atractores de un sistema dinámico $\mathbf{f}(\mathcal{X}, t)$, como un subconjunto invariante $\mathcal{A}$ del espacio de los estados $\mathcal{E}_\mathcal{X}$ tal que

- Sea D un subconjunto del espacio de los estados, que contiene a $\mathcal{A}$, entonces este convergerá al atractor, esto es:

$$\lim_{t \longrightarrow \infty} |\mathbf{f}(D, t) - \mathcal{A}| = 0 \tag{248}$$

  donde $\mathcal{A} \subset D$.

- $D \supset f^t(t, D)$, para $t > 0$.

Vale aclarar que se utiliza como equivalente las notaciones $\mathbf{f}^n(\cdot)$ y $\mathbf{f}(n, \cdot)$, para remarcar la iteración n de nuestro sistema dinámico. El hecho de que puedan coexistir varios atractores da lugar al concepto de *Cuenca de Atracción*, que no es más que el conjunto de todas las condiciones iniciales que conducen a un determinado atractor.

Daremos ahora algunos casos particulares de los atractores.

Definición: Punto de equilibrio

Un punto x_e se denomina punto de equilibrio si

$$\mathbf{f}(x_e) = x_e \tag{249}$$

Se puede ver que el punto de equilirio es un estado estacionario del sistema. Diremos que x_e es estable si para todo $\epsilon > 0$, existe un $\delta > 0$ tal que

$$|x_0 - x_e| \leq \delta \Longrightarrow |x_n - x_e| \leq \epsilon;\ \forall n \geq 1 \tag{250}$$

Si el punto no es estable, lo llamaremos simplemente inestable. Un ejemplo simple podría ser la función $\mathbf{f}(x_t) = \frac{1}{x_t}$, aquí el estado $x_t = 1$ es un punto de equilibrio.

Otro concepto importante que debemos tener en cuenta es la *periodicidad* de las órbitas. Como ya vimos la órbita relacionada al estado x_0 de un sistema dinámico discreto puede escribirse como

$$O(x_0, \mathbf{f}) = \left\{x_0, \mathbf{f}(x_0), \mathbf{f}^2(x_0), \ldots, \mathbf{f}^n(x_0), \ldots\right\} \tag{251}$$

Se dirá que $O(x_0, \mathbf{f})$ es periódica de período $p \in \mathbb{N}$ si $\mathbf{f}^p(X_0) = x_0$, es decir, que contendrá a los puntos $\{x_0, \mathbf{f}(x_0), \ldots, \mathbf{f}^p(x_0)\}$. Notemos que si $p = 1$ entonces la órbita periódica se convierte en un punto de equilibrio. Se llamará *órbita eventualmente periódica* si para $\mathbf{f}^{p+n}(x_n) = x_n$ para algún $n > 1$ y $p > 2$.

Definición: Sistemas Caóticos

Diremos que un sistema dinámico $(\mathcal{E}_\mathcal{X}, \mathcal{T}, \mathbf{f})$ es caótico si cumple las siguientes propiedades

- Es sensible respecto a las condiciones iniciales
- Es topológicamente transitivo
- Sus puntos periódicos son densos en $\mathcal{E}_\mathcal{X}$

Ahora nos ocuparemos en entender cada una de estas propiedades.

Definición: Topológicamente transitivo

La función $f : \mathcal{E}_\mathcal{X} \longrightarrow \mathcal{E}_\mathcal{X}$, se llama topológicamente transitiva si para cualquier par de conjuntos abiertos $U, V \subset \mathcal{E}_\mathcal{X}$ existe $k > 0$ tal que $\mathbf{f}^k(U) \bigcap V \neq \emptyset$. Es decir, que si nuestro sistema es topológicamente transitivo, entonces a $\mathcal{E}_\mathcal{X}$ no se lo podrá descomponer en dos subconjuntos disjuntos invariantes. Una observación importante es que si nuestro sistema tiene una órbita densa es topológicamente transitivo. Esto es si para todo punto $q \in \mathcal{E}_\mathcal{X}$ existe una sucesión $n \longrightarrow \infty$ tal que $\mathbf{f}^n(p) = q$, para algún $p \in \mathcal{E}_\mathcal{X}$.

Definición: Sensible a las condiciones iniciales

La función $f : \mathcal{E}_{\mathcal{X}} \longrightarrow \mathcal{E}_{\mathcal{X}}$ tiene dependencia sensible a las condiones iniciales si existe un $\delta > 0$, tal que para cualquier $x \in \mathcal{E}_{\mathcal{X}}$ y para cualquier vecindad N de x existe $y \in N$ y $n \geq 0$ tal que $|\mathbf{f}^n(x) - \mathbf{f}^n(y)| > \delta$.

Definición: Puntos periódicos densos

Diremos que los puntos periódicos de un sistema dinámico $\left(\mathcal{E}_{\mathcal{X}}, \mathbf{f}\right)$ son *densos* si para cualquier subconjunto U de $\mathcal{E}_{\mathcal{X}}$, siempre existe un punto periódico en U.

Teniendo estas definiciones en cuenta podríamos interpretar a los sistemas dinámicos caóticos como impredecibles, irreducibles pero también con un grado de *regularidad*, y esto último es consecuencia de que tienen puntos periódicos densos.

En las últimas décadas hubo grandes avances en la teoría del caos. En 1992 se demostró que la sensibilidad en las condiciones iniciales es consecuencia de las otras dos condiciones. En 1994 se demostró que un sistema dinámico definido en un intervalo por una función $\mathbf{f}$ topológicamente transitiva, es caótico. Y finalmente en 1997 se demostró que un sistema dinámico $\mathbf{f}$ es caótico si y solo si para cualquier par de subconjuntos abiertos U y V, existe una órbita periódica que visita a ambos.

9.1.1 Algunos Ejemplos

Tent-map: En la presente sección reflejaremos la teoría estudiada recientemente en un ejemplo clásico de la literatura, "*La Tienda de Campaña*" (también se la conoce como Tent-Map [41]). Haremos primero un análisis general del mapa discreto para luego concentrarnos en las regiones caóticas.

Se define al mapa "Tent-Map" como una aplicación $T : [0, 1] \longrightarrow \left[0, \frac{\beta}{2}\right]$, con $0 < \beta \leq 2$, de la forma,

$$T(x) = \begin{cases} \beta x; & 0 \leq x < 0,5 \\ \beta(1-x); & 0,5 \leq x \leq 1 \end{cases} \tag{252}$$

Como se puede ver tiene un solo máximo en $x = 0,5$, y al tener dos intervalos, $E_1 = [0;\ 0,5)$ y $E_2 = [0,5; 1]$, se lo puede escribir de la siguiente manera

$$T(x) = x\beta\mathcal{Y}_{E_1} + \beta(1-x)\mathcal{Y}_{E_2} \tag{253}$$

Y si utilizamos la función *signo* también se lo puede expresar como

$$T(x) = x\beta\frac{1 - S(x-0,5)}{2} + \beta(1-x)\frac{1 + S(x-0,5)}{2} \tag{254}$$

con

$$S(x) = \begin{cases} 1; & x \geq 0 \\ -1; & x < 0 \end{cases} \tag{255}$$

Teniendo esto podemos reescribir la ecuación del mapa de la forma

$$T(x) = \frac{\beta}{2}\Big(1 - 2x \cdot S(x - 0,5) + S(x - 0,5)\Big) \tag{256}$$

Y recordando que $|x-0,5| = (x-0,5)\cdot S(x-0,5)$, la expresión que se usa habitualmente para el "Tent-Map" es

$$T(x) = \frac{\beta}{2}\Big(1 - 2|x - 0,5|\Big) \tag{257}$$

Como se puede ver el comportamiento de $T(x)$ está definido por el parámetro β. En los mapas unidimensionales a este parámetro se lo suele denominar también como *parámetro de bifurcación* ya que determina el comportamiento asintótico. En función de β se pueden definir distintos comportamientos asintóticos del Tent-Map

1. $(0 < \beta < 1)$: El atractor está compuesto por un único punto fijo atractivo en el origen $(x = 0)$, de modo que cualquier condición inicial, al cabo de algunas iteraciones, converge hacia él.

2. $(\beta = 1)$: Todos los puntos de la región $[0; 0,5]$ son puntos fijos, mientras que los de la región $(0,5; 1]$ se mapean en la región $[0; 0,5]$ tras una única iteración, siendo por lo tanto puntos eventualmente fijos.

3. $(1 < \beta < \sqrt{2})$: Órbitas cuasi-periódicas. Aunque para el espacio de las fases $[0, \beta/2]$, el atractor no ocupa el espacio completo, permanece confinado en dos estrechas bandas dentro de la región $[\beta(2-\beta)/2; \beta/2]$, entre las que va saltando alternativamente.

4. $(\sqrt{2} < \beta < 2)$: Región caótica. Nuevamente el comportamiento asintótico de las secuencias generadas es limitado a la región $[\beta(2 - \beta)/2; \beta/2]$, aunque en este caso la cubre por completo y el comportamiento de la señales dentro de la misma es caótico.

5. $(\beta = 2)$: Nos encontramos con caos completamente desarrollado, es decir, la secuencia simbólica generada cubre el espacio de las fases por completo.

La demostración de porqué para $\beta = 2$ es un sistema dinámico caótico excede esta tesis, nos confomaremos con decir que para este valor de parámetro se cumplen las tres condiciones dichas al principio.

Mapa Logístico: El Mapa Logístico [42] es sin duda el mapa polinómico más conocido y simple. Este mapa es fruto de la discretización de la ecuación logística propuesta por el biólogo belga Verhulst. Esta se utiliza para modelar el crecimiento/decrecimiento de una población con recursos limitados. Formalmente hablando podemos decir que el Mapa Logístico es una aplicación $f : [0; 1] \longrightarrow [0; 1]$ con

$$f(x) = \lambda x(1 - x) \tag{258}$$

siendo λ un parámetro que varía de 0 a 4. En biología este mapa proporciona un valor normalizado de la población comprendido entre 0 (extinción) y 1 (máxima población posible). Este mapa puede presentar un comportamiento dinámico muy complejo que incluye puntos fijos, ciclo límite de todos los períodos y caos, dependiendo únicamente del valor del parámetro λ. Se pueden distinguir varios comportamientos variando este parámetro:

- $(0 \leq \lambda < 1)$: El atractor es un punto fijo en el origen.

- $(1 \leq \lambda < 3)$: El atractor es un punto fijo en $x = 1 - \frac{1}{\lambda}$.

- $(3 \leq \lambda \leq 3,57)$: Aparecen atractores periódicos de orden 2^n $(n = 1, 2, 3, \ldots)$ crecientes a medida que aumenta λ, y cada vez con menor separación entre ellos (es decir, la distancia entre el valor de λ correspondiente al período 2^n y al período 2^{n+1} es cada vez menor).

- $(3,57 \leq \lambda \leq 4)$: Regiones caóticas intercaladas de regularidad (ciclo límite de período de no potencia de 2). Por ejemplo, en $\lambda = 3,6786\ldots$ aparece el primer ciclo límite de período impar, y en $\lambda = 3,8284\ldots$ aparece el ciclo límite de período 3, que por el teorema de Sarkovskii implica la existencia de ciclos límite de todos los períodos para algún parámetro del mapa, y garantiza que el mapa sea capaz de generar secuencias caóticas para un cierto rango de su parámetro.

- $(\lambda = 4)$: Caos completamente desarrollado, y la secuencia simbólica (caótica) generada cubre todo el espacio de fases por completo.

Nosotros, como se especificará en la siguiente sección utilizaremos $\lambda = 4$. Dejaremos a un lado en este trabajo la demostración de porqué el Mapa Logístico es caótico para $\lambda = 4$, ya que excede nuestras necesidades.

Ahora describiremos de forma breve los mapas que utilizamos en este capítulo.

El generador de congruencia lineal: [43]

$$X_{n+1} = (aX_n + c) \quad mod(m) \tag{259}$$

donde

m, $0 < m$ se lo llama módulo
a, $0 < a < m$ se lo llama multiplicador
c, $0 \leq c < m$ se lo llama incremento
X_0, $0 < X_0 < m$ es la semilla o valor inicial

Y la función *mod* es la función módulo usada en informática, o sea el resto de la división euclídea entre dos números. Valores de parámetros utilizados: $a = 7141$, $c = 54773$ y $m = 259200$.

Mapa de Gauss: [44]

$$x_{n+1} = \frac{1}{x_n} \quad mod(1) \tag{260}$$

Valor inicial $x_0 = 0.1$.

Mapa de Pinchers: [45]

$$x_{n+1} = |tanh(s(x_n - c))| \tag{261}$$

donde los valores de los parámetros utilizados son $s = 2$ y $c = 0.5$; y la condición inicial es $x_0 = 0$.

Modelo de población de Ricker: [46] Es un modelo de población discreto que proporciona el número esperado de individuos a un cierto tiempo, dada una población a un tiempo anterior

$$x_{n+1} = Ax_n e^{-x_n} \tag{262}$$

donde el valor del parámetro utilizado es $A = 20$ y la condición inicial es $x_0 = 0.1$.

Mapa seno circular: [47]

$$x_{n+1} = x_n + \Omega - \frac{K}{2\pi}\sin(2\pi x_n) \tag{263}$$

Los valores de los parámetros utilizados son $\Omega = 0.5$ y $K = 2$. La condición inicial utilizada es $x_0 = 0.1$.

Mapa seno: [48]

$$x_{n+1} = \mu \sin(\pi x_n) \tag{264}$$

Los valores de los parámetros utilizados son $\mu = 1$, y la condición inicial usada es $x_0 = 0.1$.

Mapa de Spence: [49]

$$x_{n+1} = |\log(x_n)| \tag{265}$$

La condición inicial utilizada es $x_0 = 0.5$.

Es importante aclarar que para todas las series temporales usadas referidas a sistemas caóticos se descartan los primeros 1000 valores y se toman los siguientes. Esto se hace para que la secuencia no dependa tanto de la condición inicial, ya que después de los 1000 valores se puede ver el comportamiento intrínseco de la señal.

9.2 Ruidos

Para este trabajo utilizaremos ruidos coloreados, o también llamados ruidos de Hurst, y son caracterizados por su espectro de potencia, es decir, por su densidad espectral. Esta depende de la frecuencia de la siguiente manera

$$f(\nu) = \frac{1}{\nu^{\alpha}} \tag{266}$$

Pero para entender de forma constructiva esta definición primero debemos entender que es el espectro de potencias y la densidad espectral. Por eso daremos algunas nociones y definiciones en los siguientes párrafos.

Supongamos que tenemos un proceso estocástico y lo escribimos de la siguiente manera

$$X_t = \sum_{j=1}^{n} A(\nu_j) e^{it\nu_j} \tag{267}$$

donde $-\pi < \nu_1 < \nu_2 < \cdots < \nu_n = \pi$ y $A(\nu_1), \cdots, A(\nu_n)$ son coeficientes no-correlacionados $\Big(E(A(\nu_j)A(\nu_i)) = 0\Big)$ de valor complejo, tal que

$$E(A(\nu_j)) = 0, j = 1, ..., n \tag{268}$$

y

$$E(A_{\nu_j}, \overline{A(\nu_j)}) = \sigma_j^2 \tag{269}$$

Se puede ver que la ecuación (267) es la transformada de Fourier discreta. De las ecuaciones (268 y 269) podemos ver que para un proceso estacionario se tiene que

$$E(X_t) = 0 \tag{270}$$

y

$$E\Big(X_{t+h} X_t\Big) = \sum_{j=1}^{n} \sigma_j^2 e^{ih\nu_j} \tag{271}$$

Podemos ver que el proceso es estacionario ya que el momento y la autocovarianza no dependen de t. Si escribimos la expresión anterior como una integral podemos ver que para el proceso estacionario $\{X_t\}$ la autocovarianza es

$$\gamma(h) = \int_{(-\pi,\pi]} e^{ih\nu} dF(\nu) = \int_{(-\pi,\pi]} e^{ih\nu} f(\nu) d\nu \tag{272}$$

A la función F se la conoce como distribución espectral del proceso $\{X_t\}$ y le asigna un peso a cada frecuencia del intervalo $(-\pi, \pi]$ y a $f(\nu)$ se la conoce como Densidad Espectral. Podemos ver que si hacemos la transformada de la autocovarianza tenemos

$$f(\nu) = \int_{-\infty}^{\infty} \gamma(\tau) e^{i\tau\omega} d\tau \tag{273}$$

que es la Densidad Espectral. Como se ve la densidad espectral depende de la frecuencia y la autocovarianza del tiempo. Las utilidades de esta perspectiva pueden verse en la caracterización de ruidos.

9.2.1 Algunos Ejemplos

Ruido Blanco En este capítulo mostraremos al ruido blanco desde la perspectiva de la densidad espectral. Sea $\{Z_t\}$ un proceso estocástico con varianza σ^2 se lo llamará ruido blanco si cumple con las siguientes condiciones:

1. $E(Z_t) = 0$, para todo t entero.
2. $\gamma(h) = \sigma^2 \cdot \delta(h)$.

De esta definición se pueden sacar varias conclusiones importantes. La primera, es que es estacionario ya que el momento es finito y la autocovarianza no depende del parámetro t. Por hipótesis el espacio es L^2, lo que implicaría $E|Z_t|^2 < \infty$. La segunda, por como está definida la autocovarianza, el ruido blanco es un proceso estocástico incorrelacionado, es decir, hay una independencia (por lo menos en el sentido lineal) entre cada proceso para cada tiempo t. Esto nos dice que el valor que pueda tomar la variable aleatoria Z_i es independiente del valor que haya tomado la variable aleatoria Z_{i-1}. Como habíamos dicho un proceso completamente aleatorio es aquel que su valor no depende en absoluto del valor que haya tomado el sistema en el momento anterior. También se puede ver que si tomamos la definición de densidad espectral, tenemos que

$$f(\nu) = \int \gamma(\tau)_{WN} e^{i\nu\tau} d\tau = \int \sigma^2 \delta(\tau - 0) e^{i\nu\tau} d\tau = \sigma^2 \tag{274}$$

Vemos que es constante, he aquí el porqué del nombre blanco. ya que tiene una distribución uniforme en las frecuencias.

Ruido Rosa: Podemos definir entonces al Ruido Rosa haciendo uso del análisis esprectral,

$$f(\nu) \propto \frac{1}{\nu} \tag{275}$$

donde ν representa la frecuencia, y $f(\nu)$ la densidad espectral. Como se puede ver la principal diferencia con el ruido blanco es que su densidad espectral depende de la frecuencia. De la misma manera que con el ruido blanco, uno define al ruido rosa sin recurrir a la distribución de probabilidades y mucho menos al proceso estocástico subyacente. En varias publicaciones (como es el caso de Halley y colaboradores) se intenta llegar a una descripción del ruido rosa mediante un *Ruido Auto-Regresivo*

$$X_t \approx \mu + \sum_{j=1}^{K} A_t^{(j)} \tag{276}$$

La ecuación (276) se puede interpretar como la combinación de K procesos auto-regresivos más un "residuo" de bajas frecuencias. No utilizaremos la descripción (276) del proceso estocástico subyacente al ruido rosa, pero sirve para ver la estructura complicada que tiene. Nosotros tomaremos secuencias generadas por un algoritmo estándar citado al final del trabajo.

Ahora resumiremos brevemente los demás ruidos que utilizaremos.

Ruido Azul: El ruido azul tiene un espectro de potencia proporcional ν

$$f(\nu) \propto \nu \tag{277}$$

es decir, para $\alpha = -1$.

Ruido Rojo o Ruido marrón: El ruido rojo es similiar al rosa en referencia a su correlación. Este tiene una densidad espectral inversamente proporcional a ν^2

$$f(\nu) \propto \frac{1}{\nu^2} \tag{278}$$

es decir, para $\alpha = 2$.

Ruido Violeta: El ruido violeta es similar al azul en referencia a su correlación. Su densidad espectral es proporcional a ν^2

$$f(\nu) \propto \nu^2 \tag{279}$$

es decir, para $\alpha = -2$.

Para generar todos los ruidos usamos los algoritmos extraidos de las referencias ([50], [51])

9.3 Mapeo

En esta sección daremos los detalles del mapeo y de la segmentación. El algoritmo de mapeo es diferente al que hemos usado en capítulos anteriores (Bandt y Pompe). Es por eso que se prefirió detallarlo en este capítulo directamente y no incluirlo en el capítulo de Aspectos Metodológicos. El algoritmo de segmentación es el de la ventana móvil mostrado en el capítulo de Aspectos Metodológicos.

9.3.1 Correspondencia entre Serie Temporal y Cadena Simbólica Binaria

Para poder analizar series temporales dentro del marco de la teoría de la información debemos primero transformarlas a cadenas simbólicas. En este primer paso del mapeo lo que haremos será transformar una serie temporal a una secuencia simbólica binaria. El método que elegimos es el presentado por Yang, Yein y Hseu en su trabajo del 2003 [52]. Es importante ver que sea cual sea el método que elijamos, tendrá que ser capaz de que la cadena simbólica resultante exprese las propiedades que tenía la serie temporal. Esto plantea un problema de subjetividad, en el cual, lo que pesará serán nuestros objetivos. Para nuestra investigación será importante que la cadena simbólica binaria $\mathcal{C}$ represente de la mejor manera posible la estructura de la serie temporal.

Sea $\mathcal{S} = \{x_1, \cdots, x_N\}$ nuestra serie temporal, definimos al elemento "n-ésimo" de la cadena binaria $\mathcal{C} = \{a_1, \cdots, a_{N-1}\}$, como

$$c_n = \begin{cases} 0 & ; if\ x_n \leq x_{n+1} \\ 1 & ; if\ x_n > x_{n+1} \end{cases} \tag{280}$$

La cadena $\mathcal{C}$ tiene un elemento menos, es decir, llegará hasta a_{N-1}. Se ve claramente que lo que se busca con esta asignación es darle importancia al crecimiento y decrecimiento de la serie temporal, en consecuencia lo que se logra es mapear la estructura general de la serie.

9.3.2 Algoritmo de mapeo y segmentación

Este algoritmo consta principalmente de tres partes. La primera consiste en construir una cadena binaria formada por la unión de dos de las series temporales mencionadas más arriba. En la segunda parte, se construye una cadena simbólica distinta a partir de la cadena binaria anteriormente obtenida. A esta cadena se la llamará cadena alfabetizada, ya que la construiremos en base a valores determinados por *Palabras* de un largo "L" fijado con anterioridad. La estadística se hará en base a esta última cadena simbólica. Por último, se aplicará el método de segmentación de la ventana móvil.

Creación de la Cadena simbólica (binaria) doble: Elijamos dos series temporales que estén a nuestra disposición, y llamémoslas $\mathcal{S}_1$ y $\mathcal{S}_2$. Ahora "peguemos" el casillero final de la $\mathcal{S}_1$ con el primer casillero de la $\mathcal{S}_2$. Si suponemos que cada una de ellas tiene largo

N, entonces el largo de la serie temporal doble será 2N. Llamaremos "Serie Total" ($\mathcal{S}_T$), a la unión de las dos Series (1 y 2). Este método transformará nuestra serie temporal total a una cadena binaria de largo $(2N-1)$, y la llamaremos Cadena Binaria Total, $\mathcal{C}_T$. El hecho que tenga un elemento menos es consecuencia directa del método elegido. Para poder esclarecer el proceso descripto anteriormente, haremos un ejemplo con pocos casilleros.

Tomemos los últimos casilleros de una serie temporal producida por el mapa logístico de largo N y los primeros casilleros de una serie temporal producida por un proceso de ruido blanco, entonces tenemos

$$Mapa\ Log = \begin{cases} \dots \\ 0.904065 \\ 0.346926 \\ 0.906274 \\ 0.339767 \\ 0.897301 \\ 0.368606 \\ 0.930943 \\ 0.257154 \\ 0.764103 \\ 0.720999 \\ 0.804638 \\ 0.628782 \\ 0.933661 \end{cases}$$

$$Ruido\ Blanco = \begin{cases} 0.003740 \\ 0.146042 \\ 0.247883 \\ -0.040866 \\ 0.229108 \\ 0.044863 \\ 0.004388 \\ 0.207709 \\ -0.127349 \\ -0.162176 \\ -0.109847 \\ -0.097632 \\ 0.128830 \\ \dots \end{cases}$$

El tramo de la Cadena Binaria Total correspondiente a los 2×13 valores anteriores, es

$$\cdots 10101010101010011011000\cdots$$

vemos claramente que en este tramo tenemos 25 valores en vez de 26 como en el tramo de la $\mathcal{S}_T$ detallado anteriormente.

Construcción de la cadena alfabetizada: Supongamos que tenemos la cadena binaria total, $\mathcal{C}_T$, de $(2 \cdot N - 1)$ casilleros. Ahora tomemos un segmento de largo L, que llamaremos *Palabra*, y leemos los símbolos que tiene en su interior. Este segmento se irá moviendo de a un casillero a la vez, hasta detenerse en el último casillero. Nuestros resultados fueron obtenidos con $L = 8$. En los siguientes párrafos se detallarán cada uno de los pasos en la construcción de la cadena alfabetizada:

1. Tomemos los primeros L casilleros y leamos sus símbolos. La idea será corresponder a este conjunto ordenado de L símbolos un solo número natural perteneciente al intevalo $I = [0, 2^{L+1} - 1]$.

2. La elección del número que asignaremos a la *Palabra* de L elementos será de la siguiente manera:

$$a_i = \sum_{j=0}^{L-1} c_{i+j} \cdot 2^j \tag{281}$$

   Donde a_i es el elemento i-ésimo de la cadena alfabetizada, $\mathcal{C}_A$, y c_{i+j} es el elemento (i+j)-ésimo de la cadena total binaria, $\mathcal{C}_T$. Vemos que la sumatoria es hasta $(L-1)$, esto es porque empieza en cero y la palabra tiene largo L. Nuestro primer elemento de la cadena alfabetizada será para $i = 0$. Entonces tendremos que

$$a_0 = c_0 2^0 + c_1 2^1 + \cdots + c_{L-1} 2^{L-1} \tag{282}$$

3. El siguiente paso es mover la *Palabra* de largo L un casillero a la derecha, entonces nuestro siguiente elemento de la Cadena Alfabetizada es

$$a_{i=1} = c_1 2^0 + c_2 2^1 + \cdots + c_L 2^{L-1} \tag{283}$$

   Un punto a tener en cuenta es que hay una relación unívoca entre la palabra contenida en la ventanita y el número correspondiente a la Cadena Alfabetizada.

4. Seguiremos repitiendo este proceso hasta el casillero $\Big[(2N-1)-(L-1)\Big]$, es decir, hasta que la *Palabra* toque el último casillero. Esto significa que el primer casillero de la *Palabra* estará en el casillero $\Big[(2N-1)-(L-1)\Big]$ de la Cadena Total Binaria, $\mathcal{C}_T$.

5. Luego la Cadena Alfabetizada, $\mathcal{C}_A$ tendrá $(L-1)$ casilleros menos que la Cadena Total Binaria, o sea contendrá $\Big[(2N-1)-(L-1)\Big]$ casilleros.

Uno podría preguntarse por qué construir otra cadena si ya tenemos una secuencia de alfabeto finito en la cual se puede implementar fácilmente la teoría de la información. Más allá de que la elección del valor L es algo subjetivo del investigador, una de las razones más importantes, es poder tener en cada elemento de la Cadena Alfabetizada una noción o representación de los cambios de la serie temporal original. Como dijimos anteriormente los caracteres 0 y 1 de la cadena $\mathcal{C}_T$ representan si creció o decreció con respecto al valor anterior de la serie. Entonces en la Cadena Alfabetizada, $\mathcal{C}_A$, se tiene en cuenta los cambios en un cierto período de tiempo, definido por nosotros. Por lo tanto, $\mathcal{C}_A$ representará cambios locales de la serie temporal original (locales de largo L) en vez de cambios puntuales como en la cadena $\mathcal{C}_T$. Hacer estadística sobre la cadena $\mathcal{C}_A$ significará ver que cambios locales son más probables que otros.

Análisis de la Cadena Alfabetizada: Teniendo ya la cadena $\mathcal{C}_A$ estamos en condiciones de implementar la Teoría de la Información. Para esto utilizaremos el algoritmo de la ventana móvil descripto anteriormente. Teniendo todos los valores correspondientes de la divergencia se construirá un gráfico dando el valor de la D_{JS} para cada valor del cursor. En los gráficos se muestra en realidad un valor promedio de la D_{JS}, es decir, se realiza varias veces el mismo algoritmo (1000 veces) y se promedia. Es por eso que los gráficos se verán suaves y sin fluctuaciones. Lo que buscamos son los valores del cursor para los cuales la D_{JS} toma los valores más altos. Esos casilleros (o sea los valores del cursor) corresponderán a las posiciones de la cadena alfabetizada donde la diferencia entre las distribuciones de probabilidad de la ventana son más grandes. Esto se corresponderá con los mayores cambios "estructurales" de la serie temporal.

9.4 Medida de Disimilitud

Como dijimos en la introducción a este capítulo el objetivo principal del mismo es mostrar las diferencias o similitudes estadísticas que hay entre cadenas producidas por sistemas dinámicos y ruidos. También se comparan cadenas de sistemas dinámicos entre sí. Para esto es necesario una medida de disimilitud entre distribuciones de probabilidad ya que nuestro análisis es desde la perspectiva de la estadística. En este capítulo elegiremos la divergencia de Jensen-Shannon presentada en el marco teórico.

9.4.1 Significancia

Como vimos en capítulos anteriores es bueno preguntarse si las dos distribuciones son diferentes o si los valores de la divergencia son fruto de una fluctuación estadística. La respuesta a esta inquietud nos la da la definición de significancia. Para cada análisis compararemos el valor de la divergencia con el valor de la significancia, si este no supera

al segundo se entenderá que las dos secuencias son estadísticamente iguales, es decir, la divergencia no es capaz de diferenciarlas. En los siguientes párrafos detallaremos los pasos a seguir para construir la significancia.

Supongamos que tenemos las cadenas alfabetizadas de un sistema caótico (por ejemplo el mapa logístico) y la cadena alfabetizada de un ruido coloreado (por ejemplo ruido rosa). El primer paso a seguir es generar varias cadenas del mapa logístico y varias cadenas del ruido rosa y formar dos grupos grandes de cadenas diferentes. Lo que haremos es calcular la "auto-D_{JS}", es decir, compararemos (con la divergencia) todos los pares de cadenas del mapa logístico, y haremos lo mismo con todos los pares de cadenas del grupo del ruido rosa. Esto es

$$D_{JS}^{W} = D_{JS}(P^{W_i}||P^{W_j}), \qquad i : 1, \cdots, N \qquad i \neq j \tag{284}$$

donde W puede ser el mapa logístico o el ruido rosa. Es importante aclarar que se calcula la divergencia para las cadenas pertenecientes a un mismo grupo, es decir, no comparamos cadenas del mapa logístico con cadenas del ruido rosa, es por eso que se le dice "auto-D_{JS}". Con estos valores de la divergencia calculamos el promedio de la auto divergencia, esto es,

$$\mu^{W} = \langle D_{JS}^{W} \rangle \tag{285}$$

Es decir tendremos dos valores de promedio μ^{W} y $\mu^{\widetilde{W}}$. Luego calculamos la varianza para cada grupo de cadenas, esto es

$$\sigma^{W} = \langle (D_{JS}(P^{W_i}||P^{W_j}) - \mu^{W})^2 \rangle^{1/2} \tag{286}$$

Es decir, tendremos dos valores de varianza σ^{W} y $\sigma^{\widetilde{W}}$. Teniendo esto podemos definir la significancia para cada grupo de cadenas

$$S^{W} = \mu^{W} + \sigma^{W}, \qquad S^{\widetilde{W}} = \mu^{\widetilde{W}} + \sigma^{\widetilde{W}} \tag{287}$$

Entonces si comparamos una cadena de un sistema dinámico con otra de otro sistema dinámico o con una de un ruido coloreado podemos definir la significancia para esa cadena "pegada" de la siguiente manera

$$S^{W\widetilde{W}} = max\left[S^{W}, S^{\widetilde{W}}\right] \tag{288}$$

Entonces diremos que dos cadenas de diferentes orígenes son distintas o distinguibles si para algún momento la divergencia de Jensen-Shannon cumple con

$$D_{JS}(P||Q) \geq S^{W\widetilde{W}} \tag{289}$$

Esto nos dice que para cada comparación de cadenas referentes a diferentes orígenes (puede ser un sistema dinámico o un ruido coloreado) tendremos una significancia distinta a calcular.

9.5 Resultados

Esta sección se divide en dos partes, la primera está dedicada a la detección de cambios en un secuencia de símbolos mediante el método de la ventana móvil y en la otra mostramos las matrices distancia (luego detallaremos que son). Para empezar tenemos que aclarar el origen de la series temporales que usamos. Los sistemas caóticos usados son los siguientes:

- El generador de congruencia lineal.
- El mapa gaussiano.
- El mapa logístico.
- El mapa de Pinchers.
- El modelos de población de Ricker.
- El círculo seno.
- El mapa seno.
- El mapa de Spencer.
- El mapa Tienda de Campaña.

Y los ruidos coloreados usados son

- El ruido blanco.
- El ruido azul.
- El ruido rosa.
- El ruido marrón.
- El ruido violeta.

Aquí, usamos el esquema de ventana móvil de largo $L_I = L_D = 20000$, propuesto para detectar cambios en una señal. Para este propósito, utilizamos dos señales diferentes x_1 y x_2 de igual longitud L_{x_1},$L_{x_2} = 50000$ símbolos, que se fusionan en una sola secuencia. Puede darse la situación de que la señal x_1 sea caótica y x_2 sea un ruido, o dos secuencias caóticas diferentes. Los ejemplos de dos secuencias normalizadas combinadas se trazan en las Figuras 10(a) y 10(b). En ambas figuras, graficamos los resultados de aplicar el procedimiento de segmentación para diferentes combinaciones de señales. La D_{JS} alcanza su valor máximo exactamente en el punto de fusión de las dos secuencias, que está marcado con una línea vertical punteada. Se observaron resultados similares en el caso de que ambas secuencias sean generadas por procesos caóticos. En todos los casos el valor de

D_{JS} alcanza varios órdenes de magnitud más altos que los correspondientes a una sola secuencia estacionaria referidas a la "auto-divergencia", es decir, a la significancia.

Otro análisis interesante es ver la robustez de la divergencia de Jensen-Shannon para la detección de cambios dinámicos, bajo diferentes contenidos de ruido en las señales de procesos caóticos. Con este objetivo, utilizamos dos mapas caóticos (el Tent map y el mapa de Riker) con diferentes niveles de ruido (en este caso se le suma un porcentaje de ruido blanco). Los porcentajes de ruido utilizados fueron $NSR = 0\%$, 1%, 2%, 5% y 10%. La Figura 11 muestra el comportamiento de la D_{JS} para diferentes contenidos de ruido en dos mapas caóticos. Los parámetros de la discretización son $d = 6$ (largo de la *palabra*) y $\tau = 1$ (cada cuantos casilleros se mueve la ventanita de las palabras); el ancho de las ventanas es $L_I = L_D = 20000$. De estos resultados podemos concluir que el método utilizado distingue los dos procesos dinámicos independientemente del nivel de ruido, pero cuando el nivel de ruido aumenta, los valores de la D_{JS} disminuyen.

9.5.1 Matriz Distancia

Para cada tipo de proceso mostrado anteriormente generamos $N_s = 10^6$ series temporales de $L_s = 10^6$ puntos de datos con parámetros idénticos y una inicialización aleatoria. Calculamos una matriz de distancia entre ruidos de color y sistemas caóticos, utilizando el criterio de significancia para el D_{JS} ya explicado. Lo que se hizo fue calcular el valor de la D_{JS} referida a toda la señal completa, es decir, la distribución de toda la señal de caos y la distribución de toda la señal de ruido y ponerlo en una matriz. La Figura 12 muestra la matriz para los parámetros correspondientes $\tau = 1$ y $d = 8$. Para diferentes dimensiones de inmersión o largo de la palabra, se obtuvo resultados similares. En el caso de la matriz de distancias de D_{JS} correspondiente al ruido-caos (Fig. 12), observamos que la mayoría de los mapas caóticos son distinguibles de los diferentes tipos de ruidos coloreados. Cuanto más bajos son estos valores de la D_{JS}, más similares son las secuencias. Solo para el caso particular del mapa generador de congruencia lineal (LCG) y el ruido blanco (WN), el valor de D_{JS} no pasa el criterio de significancia. Este hecho significa que el mapa LCG es un ejemplo de un generador de números aleatorios que pasa la prueba de Miller-Rabin. Por lo tanto, la distribución de las palabras correspondiente a LCG y WN son muy similares. La misma evaluación se ha realizado entre mapas caóticos. La figura 13 muestra la matriz de distancia correspondiente. Como se puede ver, todos los valores de la D_{JS}, excepto uno, están por encima del criterio de significancia. Lo que demuestra que nuestro método es adecuado para distinguir diferentes tipos de caos. Un análisis más detallado muestra una fuerte relación entre los valores de la D_{JS} y el diagrama de fase de los mapas caóticos. Por ejemplo, el mapa logístico y el mapa tienda tienen diagramas de fase similares, y el valor de la distancia correspondiente es nulo (no supera a la significancia). Un comportamiento paracido se puede encontrar entre el mapa de Pincher y el mapa de Spencer.

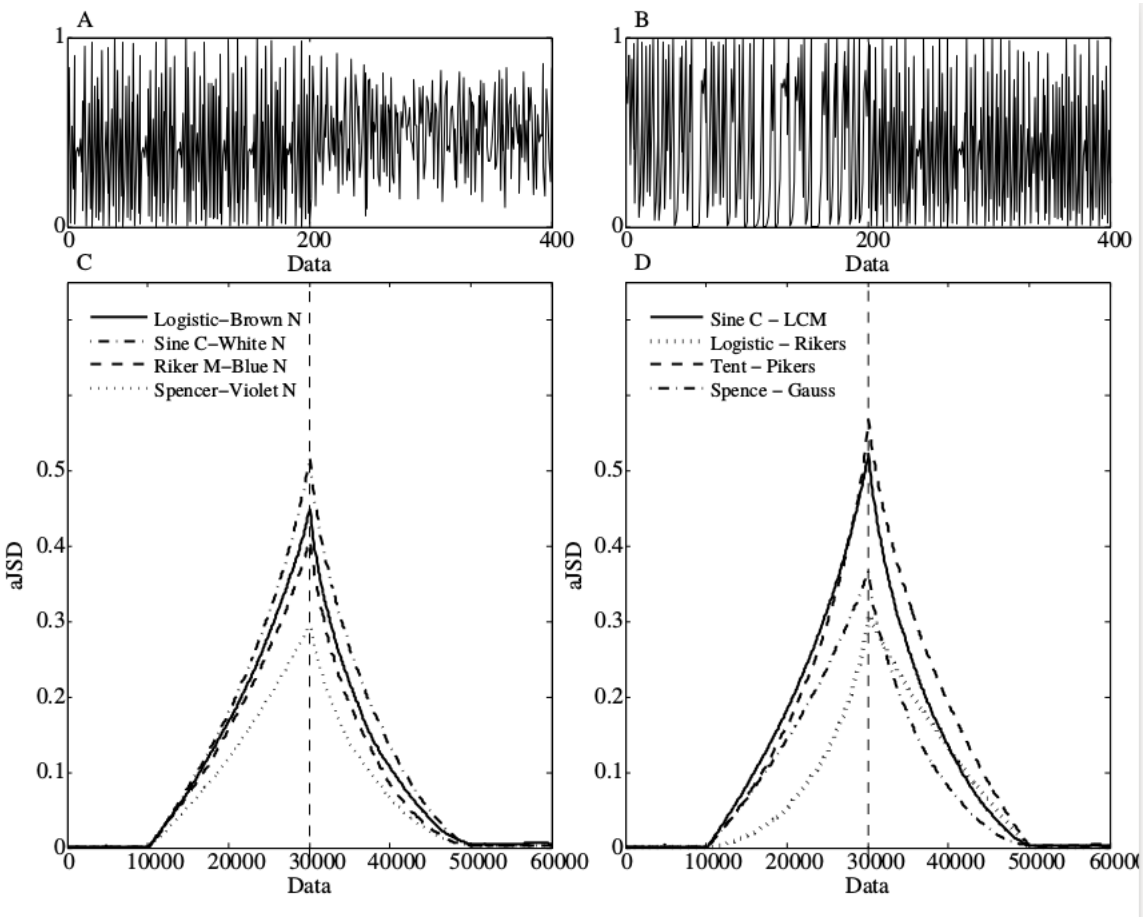

Figura 10. (A) Ejemplo de la combinación de una señal caótica y un ruido (Riker's maps y Ruido Azul), las señales se fusionaron en la posición intermedia (200 puntos). (B) Al igual que (A) se fusionaron dos señales pero ahora las dos de origen caótico (Sine Circle y Lineal Congruential Maps). (C) Los valores de la D_{JS} (en la figura aparace con el nombre de "aJSD" que significa alphabetic Jensen-Shannon Divergence) usando el método de la ventana móvil aplicado a cuatro composiciones de señales de caos-ruido. Los valores de parámetros elegidos fueron $d = 8$, $\tau = 1$ (cada cuantos casilleros se mueve la ventanita de las palabras) y un largo de ventana de 20000. (D) Al igual que (C) aplicado para cuatro composiciones de señales caóticas, usando los mismos parámetros. En ambos gráficos se puede ver que el valor máximo de la D_{JS} se alcanza en la mitad, es decir, donde se fusionan las señales.

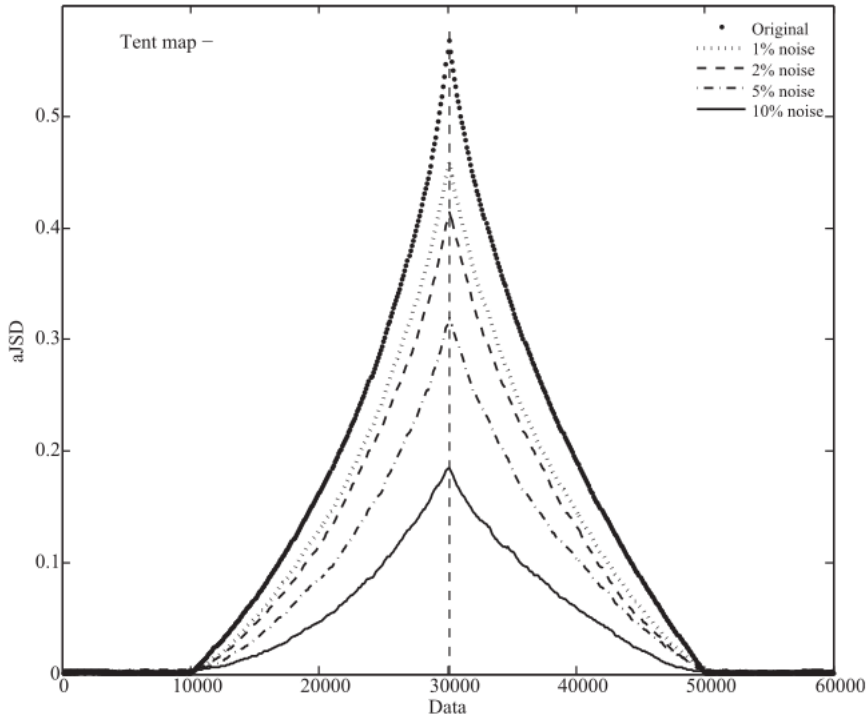

Figura 11. Se muestran los valores de la D_{JS} usando el método de la ventana móvil aplicado a dos señales caóticas (Tent maps y Riker's maps), a las dos señales se les va agregando un porcentaje de ruido blanco. Podemos ver que cuando crece el porcentaje de ruido decrece el valor máximo de la D_{JS}, sin embargo la D_{JS} es capaz de detectar el cambio en la señal fusionada. Los parámetros usados son $d = 6$, $\tau = 1$ y un largo de ventana de $L = 20000$.

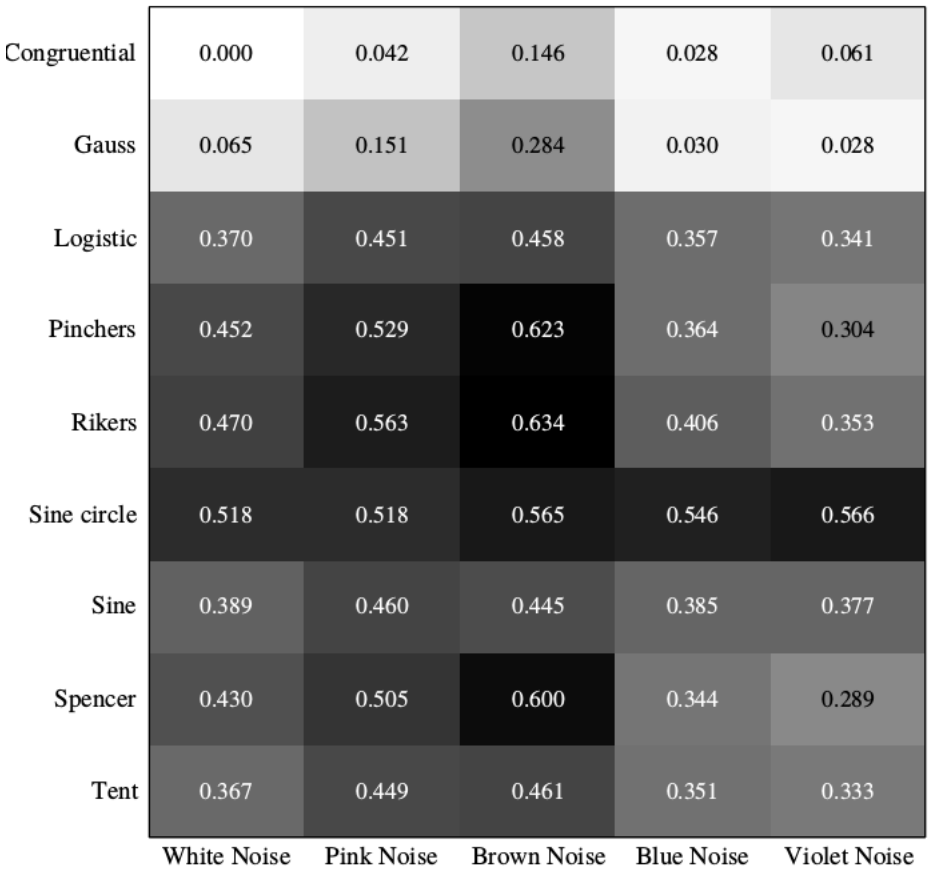

Figura 12. Matriz distancia referida a la D_{JS} entre señales caóticas y ruidos coloreados. Los parámetros usados son $d = 8$ y $\tau = 1$. Se puede observar que hay una buena discriminación entre las señales. Cuando el valor de la D_{JS} es 0.00 es porque no superó el valor de la significancia.

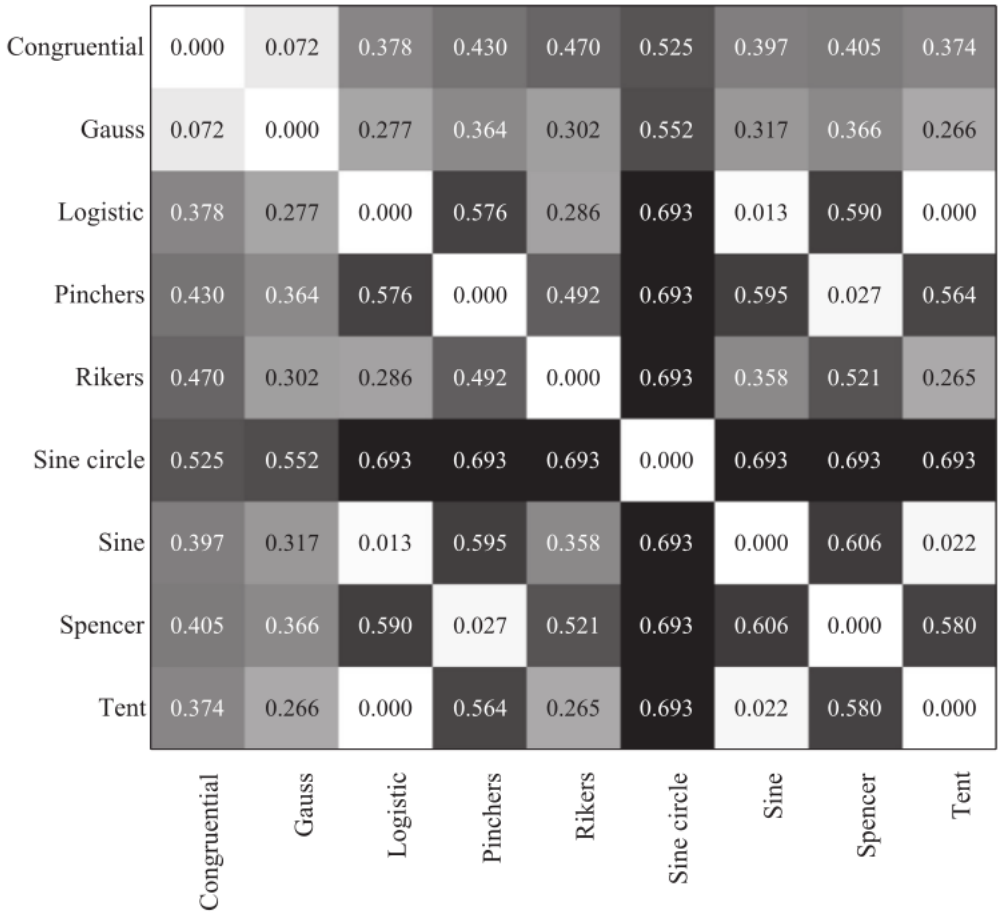

Figura 13. Matriz distancia referida a la D_{JS} entre diferentes señales caóticas. Los parámetros usados son $d = 8$ y $\tau = 1$. Cuando el valor de la D_{JS} es 0.00 es porque no superó el valor de la significancia.

9.6 Conclusiones del Capítulo

El objetivo principal del capítulo era generar un algoritmo que mediante la estadística se pueda analizar la estructura de una serie temporal y así distinguir o diferenciar dos señales. Como hemos dicho anteriormente en la tesis es necesario pasar de la señal original a una secuencia de símbolos con un alfabeto finito. En nuestro caso era un secuencia binaria que representaba el crecimiento y decrecimiento de la señal. Pero como nuestro objetivo es hacer estadística sobre la estructura local de la serie temporal tuvimos que realizar dos etapas más. La primera fue generar otra cadena simbólica a partir de la cadena binaria, esta cadena representaba la estuctura local de la serie. Es decir, nos daba un número que representaba unívocamente un cierto tipo de estructura de crecimiento y decrecimiento de la señal. Teniendo esta secuencia alfabetizada recién empezamos a hacer estadística. La siguiente etapa, ya que nuestro objetivo final era distinguir dos señales, fue realizar un proceso de segmentación, que en este caso fue el algoritmo de la ventana móvil. Cumpliendo estas tres etapas recién se aplicó la divergencia de Jensen-Shannon. Este proceso (las tres etapas en conjunto) fue fundamental y es la base por la que nuestro método pudo diferenciar señales de diferentes orígenes. Si nos enfocamos en el proceso de análisis los resultados fueron óptimos, y esto se debe a que el procedimiento en el mapeo de una señal a valores reales a una secuencia símbolica de alfabeto finito (en nuestro caso la cadena alfabetizada) fue representativo de la dinámica de la misma. Si no hubiese sido así la distinción entre las señales hubiese sido muy defectuosa.

Referido a la naturaleza de las señales y a los resultados obtenidos las conclusiones son varias. En primer lugar la distinguibilidad entre sistemas caóticos y ruidos coloreados fue óptima como se puede ver en el gráfico 10. Mostrando que aunque los dos procesos son de naturaleza impredecibles (siendo de orígenes antagónicos), la divergencia de Jensen-Shannon mediante un adecuado proceso de segmentación permite distinguir estos procesos. Los mismos resultados se obtuvieron en la distinguibilidad de patrones de estructura de dos señales caóticas pegadas.

Respecto al estudio estadístico de las señales completas (sin hacer el proceso de segmentación) mostradas en las dos matrices distancia, la distinción entre sistemas caóticos y ruidos coloreados fue óptima. Hay que hacer una salvedad con la situación del Ruido Blanco y el mapa generador de congruencia lineal (LCG), ya que el máximo de la divergencia en este caso no supera el umbral impuesto por la significancia. Esto se debe a que el mapa LCG se usa como un generador de números aleatorios. Esto marca la importancia de la significancia, ya que nos brinda una cantidad representativa de cuán lejos estamos del error estadístico. También pudimos ver que el uso de la divergencia Jensen-Shannon sirve para la distinguibilidad entre diferentes mapas caóticos, mostrando que cuando teníamos en la matriz distancia valores bajos de la divergencia estamos en presencia de dos mapas con diagramas de fase parecidos. Hay que hacer una salvedad en este caso también, ya que para el caso del Tent Map y el mapa logístico la divergencia tampoco superó el umbral de la significancia.

Por último pudimos ver la robustez del método en la distinguibilidad de mapas caóticos con la presencia de un porcentaje de ruido. Esto es muy útil en señales reales las cuales se las modela por un proceso dinámico y caótico y por el aparato de medición se filtra un porcentaje de ruido.

10 Conclusiones Generales

Aunque se hayan hecho conclusiones parciales en cada capítulo de la tesis, nos pareció oportuno finalizar nuestro trabajo con algunas consideraciones generales a modo de conclusión.

En primera instancia, podemos hacer algunas reflexiones sobre los avances teóricos obtenidos en este trabajo. A nuestro parecer, la importancia de desarrollar nuevas divergencias generalizadas no solo radica en tener una familia de divergencias sino en *interpretar o entender desde otra perspeciva a las divergencias ya establecidas*. Como se pudo ver en el capítulo 5 ("Divergencia tipo Bregman") se obtuvo una familia de divergencias a partir de entropías generalizadas. Aunque aportar a la literatura un conjunto de nuevas medidas de disimilitud es un hecho importante de por sí, es la interpretación de la divergencia de Kulback-Leibler como un operador de entropías generalizadas lo que le da un valor extra al capítulo. Esto nos permitió relacionarla con la divergencia de Bregman, y así, obtener una divergencia simétrica generalizada a partir de entropías generalizadas. En el caso del capítulo 6 ("Divergencia Gamma") el aporte teórico fue dar una familia de divergencias a partir de un conjunto de funciones $g(x)$ con vista a futuras aplicaciones. En primera instancia, lo interesante fue ver la relación que había entre el cuadrado de la métrica euclídea y la divergencia de Jensen-Shannon. El hecho que los dos funcionales formaran conjuntos convexos nos permitió observar que los dos pertenecían a una misma familia de divergencias. Esta familia fue generada a partir de funciones convexas y de sus propiedades. Esta nueva interpretación de la divergencia de Jensen-Shannon nos hace notar la fuerte relación que hay entre este tipo de divergencias y las funciones convexas. Tal es así, que la generalización por pesos estadísticos y la generalización para más de dos distribuciones surgen de aplicar las propiedades de las funciones convexas a la divergencia gamma.

Por otra parte, en el capítulo 6 introdujimos una nueva entropía generalizada H_g. Esta surge de la interpretación de la propiedades de concavidad de la entropía de Shannon y de verla como un caso particular de una familia de medidas de incerteza con la misma estructura. Lo que hace particular a esta entropía generalizada es su estrecha relación con la divergencia gamma, presentada en el mismo capítulo. El hecho de obtener la divergencia gamma a partir de una divergencia "tipo Jensen-Shannon" utilizando la entropía generalizada H_g le da un valor agregado tanto a la divergencia gamma como a la entropía H_g.

De estas dos ideas introducidas en el capítulo 6 pudimos definir una complejidad estadística acorde al modelo presentado por Lamberti, Rosso y Plastino. Esto recalca una vez más que la divergencia gamma y la entropía generalizada H_g no son dos conceptos que van por diferentes caminos sino todo lo contrario. En referencia a la complejidad generalizada los resultados fueron los deseados ya que cumplía con la propiedad de "U" invertida al igual que la complejidad de LRP.

Otra observasión importante a hacer es la importancia del método de mapeo y seg-

mentación utilizados al momento de aplicar las herramientas teóricas desarrolladas. Sea el mapeo de Bandt y Pompe o el utilizado en el capítulo 8 los dos lograron representar en una cadena simbólica las propiedades deseadas referidas a la señal original. Esto nos permitió, mediante el uso de métodos de segmentación, distinguir señales de diferentes orígenes. Estos dos procesos son, en nuestro caso, el nexo entre las divergencias desarrolladas y el análisis de series temporales. Como pudimos ver en el capítulo 5 se pudo detectar los cambios en una señal compuesta por una fibrilación auricular y una de un ritmo sinusal normal. Y en el capítulo 6 se pudieron distinguir los diferentes estadíos de sueño. En los dos casos se demostró que las herramientas teóricas desarrolladas eran útiles en el análisis de series temporales.

Como conclusión final podemos decir que el desarrollo de nuevas medidas de distinguibilidad no solo es importante en el contexto de la tería de la información sino que también en el análisis de series temporales.

11 Apéndice A: Límites del operador de entropías

Límite para la entropía de Renyi:

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = \lim_{\alpha\to 1}\sum_{i=1}^{N} p_i\left(\frac{\partial H_\alpha}{\partial q_i} - \frac{\partial H_\alpha}{\partial p_i}\right) \tag{290}$$

Si calculamos las derivadas parciales tenemos que

$$\frac{\partial H_\alpha}{\partial q_i} = \frac{1}{1-\alpha}\cdot\frac{\alpha\cdot q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} \tag{291}$$

y para p_i tenemos

$$\frac{\partial H_\alpha}{\partial p_i} = \frac{1}{1-\alpha}\cdot\frac{\alpha\cdot p_i^{\alpha-1}}{\sum_{j=1}^{N} p_j^\alpha} \tag{292}$$

Reemplazando en la ec. 108 obtenemos la siguiente expresión

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = \lim_{\alpha\to 1}\left[\frac{\alpha}{1-\alpha}\cdot\sum_i p_i\left(\frac{q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^{\alpha-1}}{\sum_{j=1}^{N} p_j^\alpha}\right)\right] \tag{293}$$

Colocando p_i adentro del paréntesis tenemos

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = \lim_{\alpha\to 1}\left[\frac{\alpha}{1-\alpha}\cdot\sum_i \left(\frac{p_i q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^{\alpha}}{\sum_{j=1}^{N} p_j^\alpha}\right)\right] \tag{294}$$

Como se puede ver es un límite indeterminado de la forma

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = \frac{0}{0} \tag{295}$$

Pero si aplicamos la regla de L'Hopital tenemos que

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = \lim_{\alpha\to 1}\left[\frac{\frac{d}{d\alpha}\left(\alpha\cdot\sum_i\left[\frac{p_i q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^{\alpha}}{\sum_{j=1}^{N} p_j^\alpha}\right]\right)}{\frac{d(1-\alpha)}{d\alpha}}\right] \tag{296}$$

Por simplicidad de cálculos es conveniente renombrar a la sumatoria que está arriba de la división como

$$A \doteq \sum_i\left[\frac{p_i q_i^{\alpha-1}}{\sum_{e=1}^{N} q_e^\alpha} - \frac{p_i^{\alpha}}{\sum_{j=1}^{N} p_j^\alpha}\right] \tag{297}$$

Es visible ahora que lo que estamos haciendo es

$$\lim_{\alpha\to 1} D_\alpha\left(P||Q\right) = \lim_{\alpha\to 1}\left[\frac{\frac{d}{d\alpha}\Big(\alpha A\Big)}{\frac{d}{d\alpha}(1-\alpha)}\right] \tag{298}$$

Podemos empezar ahora a calcular cada una de las derivadas respecto del parámetro α. Empecemos por la más corta;

$$\frac{d}{d\alpha}(1-\alpha) = -1 \tag{299}$$

La segunda derivada la haremos en varios pasos así no se ve tan engorroso. Para empezar la derivada de la parte superior de la ec.298 es

$$\frac{d\Big(\alpha A\Big)}{d\alpha} = A + \alpha\cdot\frac{d\Big(A\Big)}{d\alpha} \tag{300}$$

Ahora calculemos la derivada de A respecto del parámetro α, recordando la definición de A de la ec. 115, tenemos que

$$\frac{d\Big(A\Big)}{d\alpha} = \frac{d}{d\alpha}\left(\sum_i\left[\frac{p_i q_i^{\alpha-1}}{\sum_{e=1}^N q_e^\alpha} - \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha}\right]\right) = \sum_i p_i\left(\frac{d}{d\alpha}\left[\frac{q_i^{\alpha-1}}{\sum_{e=1}^N q_e^\alpha}\right] - \frac{d}{d\alpha}\left[\frac{p_i^{\alpha-1}}{\sum_{j=1}^N p_j^\alpha}\right]\right) \tag{301}$$

Derivando explícitamente tenemos que

$$\frac{d}{d\alpha}\left[\frac{q_i^{\alpha-1}}{\sum_{e=1}^N q_e^\alpha}\right] = \frac{d}{d\alpha}\left[q_i^{\alpha-1}\right]\left(\sum_e q_e^\alpha\right)^{-1} - q_i^{\alpha-1}\left(\sum_e q_e^\alpha\right)^{-2}\left(\frac{d}{d\alpha}\left[\sum_e q_e^\alpha\right]\right) \tag{302}$$

Donde

$$\frac{d}{d\alpha}\left[q_i^{\alpha-1}\right] = \frac{d}{d\alpha}\left[e^{ln\left(q_i^{-1}\right)}\cdot e^{\alpha\cdot ln(q_i)}\right] = q_i^{\alpha-1}\cdot ln(q_i) \tag{303}$$

y de la misma forma obtenemos

$$\frac{d}{d\alpha}\left[\sum_e q_e^\alpha\right] = \sum_e q_e^\alpha\cdot ln(q_e) \tag{304}$$

Reemplazando tenemos la siguiente expresión

$$\frac{d}{d\alpha}\left[\frac{q_i^{\alpha-1}}{\sum_{e=1}^N q_e^\alpha}\right] = \frac{q_i^{\alpha-1}\cdot ln(q_i)}{\sum_e q_e^\alpha} - \frac{q_i^\alpha\Big(\sum_e q_e^\alpha\cdot ln(q_e)\Big)}{\sum_e q_e^\alpha} \tag{305}$$

De igual manera podemos escribir

$$\frac{d}{d\alpha}\left[\frac{p_i^{\alpha-1}}{\sum_{j=1}^N p_j^\alpha}\right] = \frac{p_i^{\alpha-1}\cdot ln(p_i)}{\sum_j p_j^\alpha} - \frac{p_i^\alpha\left(\sum_j p_j^\alpha\cdot ln(p_j)\right)}{\sum_j p_j^\alpha} \tag{306}$$

Ahora estamos en condiciones de escribir la derivada respecto del parámetro α de A

$$\frac{dA}{d\alpha} = \sum_i p_i \left(\left(\frac{q_i^{\alpha-1} \cdot ln(q_i)}{\sum_e q_e^\alpha} - \frac{q_i^\alpha \Big(\sum_e q_e^\alpha \cdot ln(q_e) \Big)}{\sum_e q_e^\alpha} \right) - \left(\frac{p_i^{\alpha-1} \cdot ln(p_i)}{\sum_j p_j^\alpha} - \frac{p_i^\alpha \left(\sum_j p_j^\alpha \cdot ln(p_j) \right)}{\sum_j p_j^\alpha} \right) \right) \tag{307}$$

Si distribuimos el signo queda

$$\frac{dA}{d\alpha} = \sum_i p_i \left(\frac{q_i^{\alpha-1} \cdot ln(q_i)}{\sum_e q_e^\alpha} - \frac{q_i^\alpha \Big(\sum_e q_e^\alpha \cdot ln(q_e) \Big)}{\sum_e q_e^\alpha} - \frac{p_i^{\alpha-1} \cdot ln(p_i)}{\sum_j p_j^\alpha} + \frac{p_i^\alpha \left(\sum_j p_j^\alpha \cdot ln(p_j) \right)}{\sum_j p_j^\alpha} \right) \tag{308}$$

Si repartimos la sumatoria tenemos

$$\frac{dA}{d\alpha} = \frac{\sum_i p_i q_i^{\alpha-1} \cdot ln(p_i)}{\sum_e q_e^\alpha} - \frac{\sum_e q_e^\alpha \cdot ln(q_e) \sum_i p_i q_i^{\alpha-1}}{\sum_e q_e^\alpha} - \frac{\sum_i p_i^\alpha \cdot ln(p_i)}{\sum_j p_j^\alpha} + \frac{\sum_j p_j^\alpha \cdot ln(p_j) \sum_i p_i^\alpha}{\sum_j p_j^\alpha} \tag{309}$$

Usando la ec.299 podemos expresar al límite como

$$\lim_{\alpha \to 1} D_\alpha \left(P || Q \right) = \lim_{\alpha \to 1} (-1) \frac{d(\alpha A)}{d\alpha} = \lim_{\alpha \to 1} \left[-A - \alpha \cdot \frac{dA}{d\alpha} \right] \tag{310}$$

Reemplazando la ec. 297 y 309 obtenemos la siguiente expresión del límite que buscamos.

$$\begin{aligned} \lim_{\alpha \to 1} D_\alpha \left(P || Q \right) = \lim_{\alpha \to 1} \Bigg\{ & \frac{\sum_i p_i q_i^{\alpha-1}}{\sum_e q_e^\alpha} - 1 - \frac{\alpha \cdot \sum_i p_i q_i^{\alpha-1} \cdot ln(q_i)}{\sum_e q_e^\alpha} + \\ & + \frac{\alpha \cdot \sum_e q_e^\alpha \cdot ln(q_e) \sum_i p_i q_i^{\alpha-1}}{\sum_e q_e^\alpha} + \frac{\alpha \cdot \sum_i p_i^\alpha \cdot ln(p_i)}{\sum_j p_j^\alpha} - \\ & - \frac{\alpha \cdot \sum_j p_j^\alpha \cdot ln(p_j) \sum_i p_i^\alpha}{\sum_j p_j^\alpha} \Bigg\} \end{aligned} \tag{311}$$

Tomando el límite tenemos que

$$\lim_{\alpha \to 1} D_\alpha \left(P || Q \right) = 1 - 1 - \sum_i p_i \cdot ln(q_i) + \sum_e q_e \cdot ln(q_e) + \sum_i p_i \cdot ln(p_i) - \sum_j p_j \cdot ln(p_j) \tag{312}$$

Como resultado final tenemos

$$\lim_{\alpha \to 1} D_\alpha \left(P || Q \right) = -H\Big(Q\Big) - \sum_i p_i \cdot ln(q_i) \tag{313}$$

Límite para la entropía HCT: El procedimiento es similar al utilizado para la entropía de Renyi. Partiremos del límite sobre el operador de entropías 95 y verificaremos que resulta la divergencia de K-L. Tenemos entonces que

$$\frac{\partial H_\alpha^T}{\partial q_i} = \frac{-\alpha}{\alpha - 1} q_i^{\alpha-1} \tag{314}$$

entonces el operador aplicado a la entropía de HCT toma la forma

$$D_{H_\alpha^T}(P||Q) = \frac{\alpha}{\alpha - 1} \sum_i p_i^\alpha - p_i q_i^{\alpha-1} \tag{315}$$

Llamando A a la sumatoria de la divergencia y aplicando el límite de α tendiendo a 1 tenemos

$$\lim_{\alpha\to 1} D_{H_\alpha^T}(P||Q) = \lim_{\alpha\to 1} \frac{\frac{d(\alpha A)}{d\alpha}}{\frac{d(\alpha-1)}{d\alpha}} \tag{316}$$

donde

$$\frac{d(\alpha A)}{d\alpha} = A + \alpha \frac{dA}{d\alpha} \tag{317}$$

y

$$\frac{d(\alpha - 1)}{d\alpha} = 1 \tag{318}$$

tenemos entonces que la derivada de A respecto α es

$$\frac{dA}{d\alpha} = \sum_i p_i^\alpha \ln p_i + \sum_i p_i q_i^{\alpha-1} \ln q_i \tag{319}$$

de todo lo anterior tenemos que

$$\lim_{\alpha\to 1} D_{H_\alpha^T}(P||Q) = \lim_{\alpha\to 1} \left[\sum_i \left(p_i^\alpha - p_i q_i^{\alpha-1} \right) + \alpha \sum_i \left(p_i^\alpha \ln p_i \right) - \alpha \sum_i \left(p_i q_i^{\alpha-1} \ln q_i \right) \right] \tag{320}$$

lo que nos da

$$\lim_{\alpha\to 1} D_{H_\alpha^T}(P||Q) = \sum_i p_i \ln p_i - \sum_i p_i \ln q_i = -H(P) - \sum_i p_i \ln q_i \tag{321}$$

que es igual a la divergencia de K-L, es decir, lo que buscábamos. Esto nos permite simetrizar el operador, aplicar el límite y verificar que

$$\lim_{\alpha\to 1} \left[\mathcal{D}_{H_\alpha^T}\Big(P||Q\Big) \right] = \mathcal{D}_H\Big(P||Q\Big) \tag{322}$$

Referencias

1. C. E. Shanonn, A mathematical theory of comunications, Bell Syst. Tech, (1948). 379-423

2. E. T. Jaynes. Information theory and statistical mechanics. Physical review, 106(4):620, (1957).

3. A. Rényi, On measures of information and entropy, in: Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability, (1961), pp. 547–561.

4. J. Havrda, F. Charvat, Kybernetika 3 (1967) 30–35.

5. M. Salicrú, M.L. Menéndez, D. Morales, L. Pardo, Comm. Statist. Theory Methods 22 (1993) 2015–2031.

6. Kullback, S.; Leibler, R.A. . On Information and Sufficiency. Annals of Mathematical Statistics 22 (1): 79–86. (1951)

7. Endres, D. M.; J. E. Schindelin . A new metric for probability distributions. (2003)

8. Van Erven, Tim; Harremoës, Peter . Renyi Divergence and Kullback–Leibler Divergence. IEEE Transactions on Information Theory 60 (7): 3797-3820.(2014)

9. Csiszár, I.. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. Magyar. Tud. Akad. Mat. Kutato Int. Kozl. 8: 85–108.(1963)

10. Bregman, L. M.. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics. 7 (3): 200–217. (1967)

11. Lamberti, Pedro W. , Majtey, Ana P., 2003. Non-logarithmic Jensen–Shannon divergence, Physica A: Statistical Mechanics and its Applications, Elsevier, vol. 329(1), pages 81-90.

12. Majtey, A.; Lamberti, P.; Prato, D.. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. Physical Review A. 72 (5): 052310. (2005)

13. Jinghui Lu, Maeve Henchion, and Brian Mac Namee, Extending Jensen Shannon Divergence toCompare Multiple Corpa

14. Gyuhyeong Goh, Applications of Bregman DivergenceMeasures in Bayesian Modeling, (2015)

15. Extrinsic Jensen-Shannon divergence and noisy Bayesian active learning Mohammad Naghshvar, Tara Javidi, Kamalika Chaudhuri, 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton).(2013)

16. Kolmogorov, Andrey (1963). On Tables of Random Numbers. Sankhyā Ser. A. 25: 369–375.

17. Leonardo E. Riveaud, Diego Mateos, Steeve Zozor, Pedro W. Lamberti, Generalized divergences from generalized entropies, Physica A 510 (2018) 68–76

18. D. M. Mateos, L. E. Riveaud, and P. W. Lamberti, Detecting dynamical changes in time series by using the Jensen Shannon divergence, CHAOS 27, 083118, (2017)

19. Richard Durrett, Probability, Theory and Examples, Duxbury Pres,(1996)

20. W. Feller, Introducción a la Teoría de Probabilidades y sus Aplicaciones, Limusa

21. P. Brockwell, Time Series: Theory and Methods, Springer, (2006)

22. Steven R. Lay, Convex sets and their applications, Wiley, (1982)

23. Clausius R., The Mechanical Theory of Heat – with its Applications to the Steam Engine and to Physical Properties of Bodies, London: John van Voorst, (1867)

24. L. Boltzmann, Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen, (1872)

25. Gibbs J. W., Elementary Principles of Statistical Mechanics, Scribner's sons, (1902)

26. Kinchin A. I., Mathematical fountadions of information theory, Dover, (1957)

27. Baez, J. C., Renyi Entropy and Free Energy, arXiv:1102.2098v3, (2011)

28. C. Tsallis, Phys. Rev. E 58 (1998) 1442.

29. J. Havrda, F. Charvát, Kybernetika 3 (1967) 30–35.

30. T.M. Cover, J.A. Thomas, Elements of Information Theory, Wilet, USA, 2005.

31. I. Csiszàr, Studia Sci. Math. Hungar. 2 (1967) 299–318

32. I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. E. Stanley, Analysis of symbolic sequences using the Jensen-Shannon divergence, Phys. Rev. E 65(4), 041905 (2002).

33. A. Majtey, D. Prato, P.W. Lamberti, Phys. Rev. A 72 (2005) 052310.

34. C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. Physical Review Letters, 88:174102, (2002)

35. M. Salicru, M.L. Menéndez, D. Morales, L. Pardo, Comm. Statist. Theory Methods 22 (1993) 2015–2031.

36. Las series temporales están libremente disponible en PhysioNet (http://www.physionet.org/challenge/).

37. Jop Briet, Peter Harremoes, Flemming Topsoe, Properties of Classical and Quantum Jensen-Shannon Divergence, arXiv:0806.4472, (2009)

38. The Sleep-EDF Database [Expanded],Goldberger, A L and Others, Components of a new research resource for complex physiologic signals, physiobank, physiotoolkit, and physionet, american heart association journals, (2000)

39. R. Lopez-Ruiz, H. Mancini, X. Calvet, A Statistical Measure of Complexity, Physics Letters A, 2002

40. P.W. Lamberti, M.T. Martin, A. Plastino, O.A. Rosso, Intensive entropic non-triviality measure, Physica A: Statistical Mechanics and its Applications, 2004

41. M. W. Hirsch, S. Smale, and R. L. Devaney, Differential Equations, Dynamical Systems, and an Introduction to Chaos (Academic Press, 2004), Vol. 60.

42. R. M. May, "Simple mathematical models with very complicated dynamics," Nature 261(5560), 459–467 (1976).

43. D. E. Knuth, The Art of Computer Programming, Vol. 3, 2nd ed. (Adison-Wesley, 1998)

44. W. H. Steeb and M. A. Van Wy, Chaos and Fractals: Algorithms and Computations (Wissenschaftsverlag, 1992).

45. W. H. Steeb and M. A. Van Wy, Chaos and Fractals: Algorithms and Computations (Wissenschaftsverlag, 1992).

46. W. E. Ricker, "Stock and recruitment," J. Fish. Board Can. 11(5), 559–623 (1954)

47. V. I. Arnol'd, "Small denominators. I. Mapping the circle onto itself," Izv. Ross. Akad. Nauk. Ser. Mat. 25(1), 21–86 (1961)

48. S. H. Strogatz, Nonlinear Dynamics and Chaos: With Applications to Physics, Biology and Chemistry (Perseus Publishing, 2001).

49. R. Shaw, Strange Attractors, "Chaotic Behavior and Information Flow," Z. Naturforsch. 38a, 80–112 (1981).

50. H. A. Larrondo, http://www.mathworks.com/matlabcentral/fileexchange/3538
for program: noisef k.m (2012).

51. O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, "Distinguishing noise from chaos," Phys. Rev. Lett. 99(15), 154102 (2007).

52. A. C. Yang, S. Hseu, H. Yien, A. L. Goldberger, and C. Peng, "Linguistic analysis of the human heartbeat using frequency and rank order statistics," Phys. Rev. Lett. 90(10), 108103 (2003).