



INSTITUTO TECNOLÓGICO DE BUENOS AIRES

Tesis de Doctorado en Ingeniería

PROPIEDADES DEL ESTIMADOR DE LA ENTROPÍA DE
PERMUTACIÓN Y SU APLICACIÓN EN PROBLEMAS DE
INGENIERÍA

Tesista: Ingeniero Francisco Traversaro Varela

Director: Dr. Ing. Francisco Oscar Redelico

Co-Director: Dr. Ing. Marcelo Risk

Ciudad Autónoma de Buenos Aires, 2018

Resumen

Las medidas provenientes de la Teoría de la Información evaluadas en una distribución de probabilidades adecuada son una herramienta muy potente para caracterizar la complejidad de un sistema dinámico. En particular se estudia la Entropía Informacional de Shannon evaluada en la Función de Distribución de Bandt y Pompe: la Entropía de Permutación. Esta medida es de cómputo rápido, no requiere pre-procesamiento de la señal, contiene información de la estructura de autocorrelaciones de la serie de tiempo y se basa sobre un supuesto de estacionariedad muy débil.

La Entropía de Permutación ha sido ampliamente utilizada en aplicaciones de Ingeniería. Sin embargo, hasta el momento el estudio de las propiedades estadísticas de su estimación ha sido poco desarrollado, relegando su utilización a estudios descriptivos de los sistemas dinámicos bajo estudio.

En esta Tesis, en una primera instancia se hace un estudio exhaustivo de la Función de Distribución de Bandt Pompe y las distintas metodologías utilizadas para calcularla. Luego se estudian las medidas de complejidad evaluadas en esta distribución y su comportamiento, y se resuelven problemáticas particulares: la aplicación de la Entropía de Permutación a series de tiempo con valores repetidos y la influencia de la distribución marginal de los datos en la estimación de la misma.

Finalmente esta Tesis propone una metodología estadística computacional, el *bootstrap paramétrico*, para obtener una aproximación de la distribución del estimador de la Entropía de Permutación y de esta manera poder hacer inferencia con esta medida de complejidad. Esto permite la construcción de un test estadístico que puede detectar cambios en la dinámica de un proceso mediante la Entropía de Permutación. Se hace un estudio final acerca de la influencia del ruido observacional en la estimación de la Entropía de Permutación mediante el uso de los tests estadísticos propuestos.

Abstract

The measures coming from the Theory of the Information evaluated in a suitable distribution of probabilities are a very powerful tool to characterize the complexity of a dynamic system. This Thesis studies Shannon's Informational Entropy evaluated in the Distribution Function of Bandt and Pompe: The Permutation Entropy. The computation of this measure is fast, does not require signal pre-processing, contains information on the structure of autocorrelations of the time series and is based on a very weak assumption of stationarity.

Permutation Entropy has been widely used in Engineering applications. However, up to now the study of the statistical properties of its estimation has been little developed, relegating its use to descriptive studies of the dynamic systems under study.

In a first instance an exhaustive study of the of Bandt & Pompe Distribution Function and the different methodologies used to calculate it, is made. Then other complexity measures evaluated in this distribution and their behavior are studied, and particular problems are solved: the application of the Permutation Entropy to time series with repeated values and the influence of the marginal distribution of the data in its estimation. Finally, this Thesis proposes a computational statistical methodology, the textit parametric bootstrap, to obtain an approximation of the distribution of the Permutation Entropy estimator and in this way to be able to make inference with this measure of complexity. This allows the construction of a statistical test that can detect changes in the dynamics of a process through the Permutation Entropy. A final study is made about the influence of observational noise in the estimation of Permutation Entropy using the proposed statistical tests.

Agradecimientos

Mis más sinceros agradecimientos a:

El Instituto Tecnológico de Buenos Aires (ITBA), por su financiamiento en el primer año de este proyecto.

El Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) por su financiamiento en los años subsiguientes.

A mi Director de Tesis, el Dr. Ing. Francisco O. Redelico, por iniciarme en la investigación científica, por el tiempo que me dedicó, y por la calidez humana con la que fue guiándome a lo largo de toda la Tesis.

A mi Co-Director de Tesis, el Dr. Ing. Marcelo Risk, por sus valiosos aportes.

Y finalmente le agradezco a mi Mamá, la Dra. Alicia Varela, en parte por sus tardes leyendo la Tesis para ayudarme en la redacción, pero principalmente porque de ella aprendí lo satisfactoria que puede ser la vida académica.

Dedicatoria

A mi Mamá y mi Papá, a ellos le debo todo.

A mi mujer, Denise, por su apoyo, su alegría y su optimismo.

Publicaciones en las que está basada esta Tesis

Esta Tesis doctoral está basada en las siguientes publicaciones:

Characterization of autoregressive processes using entropic quantifiers

Francisco Traversaro, Francisco O. Redelico.

**Physica A: Statistical Mechanics and its Applications Volume 490, 15
January 2018, Pages 13-23**

(Ver Capítulo 4)

*Confidence intervals and hypothesis testing for the Permutation Entropy with an
application to epilepsy*

Francisco Traversaro, Francisco O. Redelico.

**Communications in Nonlinear Science and Numerical Simulation
Volume 57, April 2018, Pages 388-401**

(Ver Capítulo 6)

*Bandt-Pompe symbolization dynamics for time series with tied values: a data-driven
approach*

Francisco Traversaro, Francisco O. Redelico, Marcelo Risk, Alejandro Frery and
Osvaldo A. Rosso.

Chaos: An Interdisciplinary Journal of Nonlinear Science, en prensa.

Comparing different approaches to compute Permutation Entropy with coarsened time series

Francisco Traversaro, Nicolás Quiroz, Francisco O. Redelico.

Presentada para revisión en Marzo 2018. Physica A: Statistical Mechanics and its Applications

(Ver Capítulo 5)

Classification of Normal and Pre-Ictal EEG Signals Using Permutation Entropies and a Generalized Linear Model as a Classifier

Francisco O. Redelico, Francisco Traversaro, María del Carmen García , Walter Silva, Osvaldo A. Rosso and Marcelo Risk.

Entropy 2017, 19(2), 72; doi:10.3390/e19020072

(Ver Capítulo 2)

Evaluation of the status of rotary machines by time causal Information Theory quantifiers

Francisco O. Redelico, Francisco Traversaro, Nicolás Oyarzabal, Ivan Vilaboa, Osvaldo A. Rosso.

Physica A: Statistical Mechanics and its Applications Volume 470, 15 March 2017, Pages 321-329

(Ver Capítulo 3)

Otras producciones científicas vinculadas a esta Tesis:

Big Data in the ICU: Experience in the Hospital Italiano de Buenos Aires

Astudillo, J., Zelechower, J., Traversaro, F., Redelico, F., Luna, D., Quiros, F., San Roman, E.

Studies in health technology and informatics, 245, 1319-1319. (2017)

Infrastructure for Big Data in the Intensive Care Unit

Zelechower, J., Astudillo, J., Traversaro, F., Redelico, F., Luna, D., Quiros, F., San Roman, E., Risk, M.

Studies in health technology and informatics, 245, 1346-1346. (2017)

An empirical evaluation of alternative methods of estimation for Permutation Entropy in time series with tied values

Francisco Traversaro, Marcelo Risk, Osvaldo A. Rosso, Francisco O. Redelico.

MEDYFINOL 2016, XIX Conference on Nonequilibrium Statistical Mechanics and Nonlinear Physics Valdivia, Chile. December 5th–9th, 2016. Book of Abstracts, Page 30.

Índice general

Resumen	II
Abstract	III
Agradecimientos	IV
Dedicatoria	V
Publicaciones en las que está basada esta Tesis	VI
Índice de figuras	XIV
Índice de cuadros	XVII
Lista de abreviaturas	XIX
1. Introducción General	1
1.1. Introducción	1
1.2. Objetivo de esta Tesis	4
1.3. Sinopsis	5
2. Entropía de Permutación: La entropía según la Función de Distribución de Probabilidades de Bandt y Pompe	8
2.1. Introducción	8
2.2. La simbolización propuesta por Bandt y Pompe	9
2.2.1. Permutando los rangos: Mapeo según Rangos	10

2.2.2. Permutando los índices cronológicos: Mapeo según el orden cronológico	11
2.3. La Función de Distribución de Probabilidades de Bandt y Pompe	14
2.4. La Entropía de Permutación	16
2.5. Extensiones de la Entropía de Permutación	18
2.5.1. Entropía de Permutación de Rényi	18
2.5.2. MinEntropía de Permutación	19
2.5.3. Entropía de Permutación de Tsallis	19
2.5.4. Entropía de Permutación Ponderada	20
2.6. Patrones Prohibidos y Patrones Faltantes	21
2.7. Aplicación: Clasificación de señales de EEG normales y pre-ictales usando las distintas Entropías de Permutación.	23
2.7.1. Introducción	23
2.7.2. Objetivo	25
2.7.3. Materiales y Métodos	26
2.7.4. Resultados	30
2.7.5. Conclusiones	36
3. Otras medidas de complejidad derivadas de la Función de Distribución de Probabilidades de Bandt y Pompe	38
3.1. Introducción	38
3.2. Medida de Complejidad Estadística	39
3.3. Medida de Información de Fisher	41
3.4. Aplicación: Evaluación del estado de máquinas rotativas mediante las distintas medidas de complejidad	44
3.4.1. Introducción	44
3.4.2. Objetivo	45
3.4.3. Preparación del experimento	45
3.4.4. Resultados	50

4. Influencia de la distribución marginal de los datos en la Entropía de Permutación	57
4.1. Introducción	57
4.2. Plano Informacional Entropía - Complejidad	58
4.3. Entropía de Shannon aplicada al histograma	59
4.4. Procesos Estocásticos Autorregresivos	60
4.4.1. Procesos Autorregresivos Gaussianos de orden 1: AR(1)	61
4.4.2. Procesos Autorregresivos Exponenciales de orden 1: NEARA(1)	61
4.4.3. Procesos Autorregresivos Uniforme de orden 1: UAR(1)	62
4.5. Resultados Numéricos	63
4.6. Conclusiones del Capítulo	68
5. Influencia de los datos repetidos en la Función de Distribución de Probabilidades de Bandt y Pompe	69
5.1. Introducción	69
5.2. El problema que representan los datos repetidos	71
5.3. Los patrones con valores repetidos como posibles estados del sistema: Extendiendo el Alfabeto Simbólico	73
5.3.1. Alfabeto extendido según mapeo cronológico	73
5.3.2. Alfabeto extendido según mapeo de rangos	74
5.4. Los valores repetidos como información faltante	76
5.4.1. Casos Completos	77
5.4.2. Imputación según el Orden de Aparición	77
5.4.3. Imputación al Azar	78
5.4.4. Imputación Basada en la Muestra	79
5.5. Aplicaciones	85
5.5.1. Simulación numérica: Mapas Caóticos	85
5.5.2. Números trascendentales	94
5.6. Conclusiones del Capítulo	99

6. Propiedades estadísticas del estimador de la Entropía de Permutación	101
6.1. Introducción	101
6.2. El Enfoque Bootstrap	103
6.2.1. Las probabilidades de transición de una secuencia de símbolos	105
6.2.2. El método bootstrap aplicado a la Entropía de Permutación	106
6.3. Simulación numérica	109
6.3.1. Diseño experimental	110
6.3.2. Resultados	112
6.4. Aplicación: datos EEG	116
6.5. Conclusiones del Capítulo	121
7. Influencia del ruido observacional en la estimación de la Entropía de Permutación	123
7.1. Introducción	123
7.2. Simulación numérica	124
7.2.1. Serie de tiempo proveniente de una dinámica caótica	125
7.2.2. Ruido observacional	126
7.2.3. Serie de tiempo proveniente de una dinámica caótica contaminada con ruido	126
7.3. Análisis descriptivo del efecto del ruido observacional en la estimación de la Entropía de Permutación	127
7.3.1. Efectos del largo de la serie en la estimación de la Entropía de Permutación	127
7.3.2. Efectos del ruido observacional en la estimación de la Entropía de Permutación	128
7.3.3. Análisis inferencial del efecto del ruido observacional en la estimación de la Entropía de Permutación	133
7.4. Conclusiones	141
8. Discusión, Conclusiones y Futuras Líneas de Investigación	143

8.1. Discusión	143
8.2. Conclusiones	149
8.3. Futuras líneas de investigación	149
Bibliografía	151
Apéndice A. Algoritmos bootstrap	160
A.1. Algoritmo 1	160
A.2. Algoritmo 2	161
A.3. Algoritmo 3	162
Apéndice B. Regla de Scott para la cantidad de intervalos de un histograma	163
Apéndice C. Códigos en R	165

Índice de figuras

2.1. Mapeo según rangos	12
2.2. Mapeo según el orden cronológico	13
2.3. El Área bajo la curva ROC (AUC) calculado mediante validación cruzada de 10 pliegues es graficado ($AUC \pm 1$ sd), versus el retardo τ para la clasificación de los EEG	32
2.4. Los cinco modelos de regresión logística presentados en el Cuadro 2.1	34
2.5. AUC de los distintos modelos para la clasificación de los EEG . . .	35
3.1. Banco de Vibración	46
3.2. Serie de tiempo correspondiente a una máquina rotativa balanceada.	48
3.3. Serie de tiempo correspondiente a una máquina rotativa desbalanceada.	49
3.4. Series de tiempo correspondientes a la máquina rotativa con distintos tipos de fallas abruptas.	51
3.5. Evolución de la Entropía de Permutación para los seis tipos de fallas abruptas en la máquina rotativa	53
3.6. Evolución de la Medida de la Complejidad Estadística para los seis tipos de fallas abruptas en la máquina rotativa	54
3.7. Evolución de la Medida de Información de Fisher para los seis tipos de fallas abruptas en la máquina rotativa	55
4.1. Localización de los procesos autorregresivos en el plano $\mathcal{H} \times \mathcal{C}$. . .	65
4.2. Las series de tiempo de los procesos autorregresivos y sus FDP de BP	66
4.3. Localización de los procesos autorregresivos en el plano $\mathcal{H} \times \mathcal{H}_a$. .	67

5.1. Ejemplo de aplicación de la metodología basada en la muestra para tratar con datos repetidos	82
5.2. Boxplot del error de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para cada una de las metodologías que intentan resolver el problema de los patrones repetidos.	88
5.3. Boxplot del error de aproximación absoluto $ \tilde{\mathcal{H}} - \mathcal{H} $ para cada una de las metodologías que intentan resolver el problema de los valores repetidos.	89
5.4. Boxplot del error de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para cada una de las metodologías que intentan resolver el problema de los valores repetidos para los procesos caóticos agrupados según el porcentaje de vectores con valores repetidos.	90
5.5. Idem a la Figura 5.4 pero para el error de aproximación absoluto	91
5.6. El error de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para todos los procesos caóticos simulados	92
5.7. Ídem a la Figura 5.6 pero para el error de aproximación absoluto	93
6.1. Diagrama esquemático del enfoque del bootstrap paramétrico.	104
6.2. Realización de los ruidos $1/f^k$ y su densidad espectral	111
6.3. Desviación estándar de $\hat{\mathcal{H}}$ de los ruidos $1/f^k$	114
6.4. El sesgo bootstrap para $\hat{\mathcal{H}}$ de los ruidos $1/f^k$	115
6.5. Histograma para un set de réplicas bootstrap	116
6.6. Intervalos de confianza del nivel de confianza del 90% para la Entropía de Permutación de las 10 señales EEG de la actividad cerebral para cada set	119
6.7. Test de hipótesis para la diferencia en la Entropía de Permutación entre las señales de EEG de voluntarios sanos despiertos con los ojos abiertos y las señales de EEG de voluntarios sanos despiertos con los ojos cerrados	120
6.8. Test de hipótesis para la diferencia en la Entropía de Permutación entre las señales de EEG extendido a todos los tipos de pacientes	121

7.1. Efecto del ruido observacional en la estimación de la Entropía de Permutación en una serie de tiempo contaminada con ruido para $m = 3$	129
7.2. Efecto del ruido observacional en la estimación de la Entropía de Permutación en una serie de tiempo contaminada con ruido para $m = 4$	130
7.3. Efecto del ruido observacional en la estimación de la Entropía de Permutación en una serie de tiempo contaminada con ruido para $m = 5$	131
7.4. Efecto del ruido observacional en la estimación de la Entropía de Permutación en una serie de tiempo contaminada con ruido para $m = 6$	132
7.5. Las curvas de potencia $1 - \beta$ en función del ruido agregado σ	135
7.6. El nivel de significación α en función del nivel de ruido agregado σ	139
B.1. La Entropía de Amplitud para los tres procesos estocásticos descorrelacionados.	164

Índice de cuadros

1.1. Número de artículos con referato donde la Entropía de Permutación ha sido utilizada de acuerdo a la base de datos bibliográfica de Scopus.	5
2.1. Mejores modelos para cada entropía ordenados según el valor de AUC en forma decreciente para la clasificación de los EEG	33
3.1. Media y error estándar para: Entropía de Permutación, $\hat{\mathcal{H}}$; Complejidad Estadística, $\hat{\mathcal{C}}$; Medida de Información de Fisher , $\hat{\mathcal{F}}$ para las distintas zonas de comportamiento de la máquina rotativa	54
4.1. Distribución Uniforme discreta para las perturbaciones a_t en el modelo UAR (1)	63
5.1. Cantidad de símbolos π_i para la dimensión de <i>embedding</i> m para cada alfabeto.	72
5.2. El alfabeto extendido según el orden cronológico y sus ambigüedades	75
5.3. Los diferentes vectores de <i>embedding</i> $X_t^{(3)}$ para la serie de tiempo $X_t = (2, 5, 1, 2, 7, 1, 1, 3, 1, 2, 4, 5, 1, 3, 2, 4, 4, 2, 2, 1, 0)$ y su mapeo a los símbolos π_i según las diferentes metodologías usadas para lidiar con valores repetidos.	83
5.4. Todos los patrones posibles para un vector de embedding $X_t^{(3)}$ y su mapeo según cada metodología presentada en este Capítulo.	84
5.5. Resultados para la expansión decimal del número π	96
5.6. Resultados para la expansión decimal del número e	97
5.7. Resultados para la expansión decimal del número $\sqrt{2}$	98

6.1. Intervalos de confianza de los ruidos $1/f^k$	113
7.1. Parámetros a ser estimados para cada dimensión de <i>embedding</i> para la estimación de la Entropía de Permutación	128
7.2. Las curvas de potencia $1 - \beta$ en función del ruido agregado σ	136
7.3. El nivel de significación α en función del nivel de ruido agregado σ	140

Lista de abreviaturas

Abreviatura	Significado
AUC	Área bajo la Curva ROC
CEMAT	Laboratorio de Materiales del ITBA
FDP	Función de Distribución de Probabilidades
FDP de BP	Función de Distribución de Probabilidades de Bandt y Pompe
HIBA	Hospital Italiano de Buenos Aires
IBE	Oficina Internacional para la Epilepsia
ILAE	Liga Internacional Contra la Epilepsia
ROC	Receiver Operating Characteristic Curve
UTIA	Unidad de Terapia Intensiva de Adultos

Capítulo 1

Introducción General

1.1. Introducción

El estudio de los sistemas dinámicos ha sido un área de continuo crecimiento en los últimos 50 años. El avance en este campo hizo que científicos en todas las disciplinas muestren interés en las técnicas analíticas y cualitativas desarrolladas en el mismo y las han implementado exitosamente para la resolución de problemas no lineales en áreas tan diversas como la Física, la Química y la Bioquímica, Economía, Medicina, Ecología y lo más importante, en lo a que a esta Tesis se refiere, en las distintas ramas de la Ingeniería.

La representación más común de los datos provenientes de un sistema dinámico es en forma de series temporales, y la manera de obtener estos datos se ha ido transformando desde las series de tiempo clásicas que consisten en unos pocos valores (de cientos a pocos miles de datos) tomados con gran cuidado, a datos generados automáticamente que consisten en millones de datos que generalmente no son chequeados en el momento y requieren un pre-procesamiento que puede ser complicado y que varía según el investigador.

Pero ¿qué es un sistema dinámico? En definitiva, un sistema dinámico consiste en un espacio de estados (o de fases) y una regla matemática describiendo la evolución en el tiempo de cualquier punto en dicho espacio. El estado de un sistema es un conjunto de variables que son consideradas importantes en relación al sistema

y el espacio de estados es el conjunto de todos los posibles valores que pueden tomar estas variables. Como la mayoría de los datos observados en la naturaleza vienen en forma de series de tiempo, se estudian las mismas como una trayectoria dentro del espacio de estados si es que la variable estudiada representa de alguna manera algún aspecto importante del sistema bajo estudio. A partir de estas trayectorias (series de tiempo) se intenta reconstruir el espacio de estados para caracterizar el sistema dinámico, que puede ser algo tan simple como una masa suspendida en un resorte o tan complejo como el cerebro humano. Es por esto que un aspecto importante de esta caracterización es determinar la complejidad de dicho sistema dinámico.

En otro contexto, a mediados del siglo pasado Claude E. Shannon se preguntaba si había alguna manera de medir la cantidad de información “producida” por un proceso y transmitida mediante un canal de comunicación, en un trabajo que daría comienzo a lo que se llamaría la Teoría de la Información ([Shannon \[2001\]](#)). Proponía pensar en un proceso que genera n símbolos posibles que serán enviados por dicho canal con probabilidades p_1, p_2, \dots, p_n respectivamente. Dichas probabilidades son conocidas pero es lo único que se conoce con respecto al símbolo que será enviado. ¿Habrá alguna medida global de incertidumbre, en relación a la transmisión de símbolos, que caracterice al canal de comunicación?

En este trabajo, Shannon llega a la conclusión de que si existe tal medida $H(p_1, p_2, \dots, p_n)$ debería tener las siguientes propiedades:

- H debe ser continua en los p_i
- Si todos los p_i son iguales, $p_i = 1/n$, entonces H debe ser una función monótona creciente en función de n , ya que con más símbolos equiprobables habrá más opciones, o lo que es lo mismo, más incertidumbre.
- Si se divide una opción entre dos opciones sucesivas, la H original debe ser la suma de los valores individuales de las H .

y demuestra que la única función H que cumple con estas tres propiedades es la siguiente:

$$H = -K \sum_i^n p_i \log p_i$$

En ese mismo artículo admite que la forma de H puede ser reconocida en ciertas formulaciones de mecánica estadística como entropía, donde cada p_i es la probabilidad de que un sistema se encuentre en la celda i de su estado de fases: citando textualmente “*H is then, for example, the H in Boltzmann’s famous H theorem*” (Shannon [2001]). De esta manera llama a su medida de incertidumbre como “Entropía”, simplemente por su similitud al funcional proveniente de la Teoría de la Mecánica Estadística, que luego se conocería como la Entropía de Información de Shannon.

Debido al rápido desarrollo que tuvo la Teoría de la Información, se intentó conectar la Entropía de Información de Shannon con la entropía termodinámica, pero recién 10 años después Edwin T. Jaynes es quien reformula la Mecánica Estadística sobre la base de la Teoría de la Información (Jaynes [1957a,b]). Sin embargo, esta metodología no siempre genera la función de distribución adecuada para caracterizar el sistema, dejando la puerta abierta a futuras investigaciones (Frieden [1990]).

Paralelamente, los trabajos de A.N Kolmogorov y J.G Sinai formalizaron el uso de la Teoría de Información para el estudio de sistemas dinámicos, representados como series de tiempo, que mostraron ser muy útiles para la caracterización de los mismos (Kolmogorov [1958]; Sinai [1959a,b]).

Mucho tiempo después, a principios de este siglo, Bandt y Pompe proponen una medida de complejidad vinculada a la Entropía Informacional de Shannon para el estudio de sistemas dinámicos que dieron en llamar la Entropía de Permutación (Bandt and Pompe [2002]). La metodología propuesta consiste en reemplazar los valores de la serie de tiempo por una secuencia de símbolos (π_i) cuyos valores

pertenecen a un conjunto de cardinalidad finita (o alfabeto) que de alguna manera preservan la autocorrelación existente entre los valores de la serie original. De esta simbolización se estima la distribución de probabilidades $p(\pi_i)$, llamada la Función de Distribución de Probabilidades de Bandt y Pompe, y se le calcula la Entropía Informacional de Shannon que da lugar a la Entropía de Permutación.

Esta medida es de cómputo rápido, no requiere pre-procesamiento de la señal y se basa sobre un supuesto de estacionariedad muy débil. Ha sido ampliamente utilizada en dinámicas no lineales (De Micco et al. [2008]; Keller and Simm [2005]; Masoller and Rosso [2011]; Rosso et al. [2010a]), y en menor medida en procesos estocásticos (Rosso et al. [2007a]; Sinn and Keller [2011]; Zunino et al. [2008a]). También ha tenido un gran impacto en distintas áreas de las ciencias aplicadas y la ingeniería tan variadas como Ingeniería Mecánica (Redelico et al. [2017b]; Yan et al. [2012]), Epilepsia (Olofsen et al. [2008]; Redelico et al. [2017a]), Anestesia (Jordan et al. [2008]), Cardiología (Frank et al. [2006]; Parlitz et al. [2012]), Finanzas (Matilla-García and Marín [2009]) y Cambio Climático (Carpi et al. [2013]), entre otras. Desde su publicación y hasta el final del 2017, este trabajo de Bandt y Pompe ha sido citado en 1617 artículos (ver Cuadro 4.1), de acuerdo a la base de datos bibliográfica de Scopus, y la evolución de las citas parece indicar que su impacto seguirá creciendo.

Por todo lo expuesto, resulta de interés científico el estudio en profundidad de la Entropía de Permutación, y de las distintas medidas de complejidad asociadas a la Función de Distribución de Probabilidades presentada por Bandt y Pompe.

1.2. Objetivo de esta Tesis

Cuando el cálculo de la Entropía de Permutación proviene de una serie de tiempo, que puede ser vista como una realización de un proceso generador de datos (sistema dinámico), este valor resulta ser una estimación de la Entropía de Permutación verdadera de dicho sistema. Si bien este estimador se ha usado con éxito de manera descriptiva en la caracterización de sistemas dinámicos, no se ha

Cuadro 1.1 Número de artículos con referato donde la Entropía de Permutación ha sido utilizada de acuerdo a la base de datos bibliográfica de Scopus.

Área de estudio	Cantidad de artículos publicados
Física y Astronomía	370
Ingeniería	295
Ciencia de los Materiales	71
Matemática	280
Ciencias de la Computación	207
Medicina	133
Neurociencia	62
Bioquímica	50
Química	36
Otros	113

encontrado en la bibliografía científica un estudio sobre sus propiedades estadísticas que permita hacer inferencia sobre la verdadera Entropía de Permutación del proceso generador de datos, quizás debido a que una derivación analítica de la *performance* estadística del estimador (si fuera de hecho posible) se ve dificultada por lo complicado del funcional H y por la estructura propia de la Función de Distribución de Probabilidades de Bandt y Pompe.

Por lo tanto, el objetivo de esta Tesis es formular un test de hipótesis para la Entropía de Permutación, que se logra mediante un método computacional: el *bootstrap paramétrico*.

1.3. Sinopsis

El desarrollo de esta Tesis refleja el trabajo realizado durante estos años para lograr el objetivo antes mencionado, y el orden de los Capítulos sigue de alguna manera el progreso en el cual se fueron realizando los distintos estudios:

- En el Capítulo 2 se estudian las distintas metodologías para transformar una serie de tiempo a valores reales en una secuencia de símbolos pertenecientes a un alfabeto finito, que cumplan con lo propuesto por Bandt y Pompe, y a

partir de ellos se construye la Función de Distribución de Probabilidades de Bandt y Pompe (FDP de BP) y se presenta a la Entropía de Permutación como la Entropía Informacional de Shannon aplicada a esta distribución. Luego se presentan diferentes funcionales de la entropía que han sido utilizados en la literatura, evaluados en la FDP de BP y se compara la capacidad de discriminación de estas distintas entropías para detectar cambios en las dinámicas subyacentes provenientes de señales de electroencefalogramas (EEG).

- En el Capítulo 3 se presentan otros funcionales evaluados en la FDP de BP, como la Medida de Complejidad Estadística y la Medida de Información de Fisher. Son medidas desarrolladas como complementarias a la Entropía de Permutación que procuran dar una percepción mas completa de la complejidad propia de un sistema y se estudian en una aplicación para el estudio de máquinas rotativas.
- Al ser importante la Función de Distribución de Probabilidades que describe un proceso, en el Capítulo 4 se compara la Entropía de Permutación con la Entropía de Amplitud, es decir la entropía evaluada en la función de distribución marginal de los datos, estimada mediante el histograma.
- En el Capítulo 5 se estudian las diversas metodologías existentes para estimar la Entropía de Permutación en series de tiempo cuyos datos son medidos con baja resolución, y se presenta una nueva metodología como posible solución, que pareciera superar en rendimiento a las ya existentes. Se comparan todas estas metodologías estudiando distintos procesos caóticos y las expansión decimal de números irracionales muy conocidos como π , e y $\sqrt{2}$.
- Una vez que se ha estudiado en profundidad las propiedades descriptivas de la Entropía de Permutación, en el Capítulo 6 se propone un método estadístico computacional, el *bootstrap*, que puede estimar de manera muy eficiente el sesgo y la varianza del estimador de la Entropía de Permutación permitiendo de esta manera desarrollar intervalos de confianza y tests estadísticos que

involucren a cualquier cuantificador que utilice como argumento la Función de Distribución de Probabilidades de Bandt y Pompe. Este método estadístico es novedoso ya que no utiliza un remuestreo como en las metodologías tradicionales de *bootstrap no paramétrico* para series de tiempo que tienen serias dificultades para mantener la autocorrelación dentro de la serie, sino que es un *bootstrap paramétrico* especialmente diseñado para la FDP de BP.

- Como los ruidos observacionales son propios de los datos de series de tiempos extraídos de un proceso mediante un instrumento de medición, en el Capítulo 7 se estudia la influencia de los ruidos observacionales en la estimación de la Entropía de Permutación utilizando las herramientas inferenciales desarrolladas en el Capítulo anterior.
- Finalmente se destina el Capítulo 8 a las discusiones generales surgidas de esta Tesis y a las posibles líneas de investigación a seguir en el futuro, de acuerdo a los resultados obtenidos.

Capítulo 2

Entropía de Permutación: La entropía según la Función de Distribución de Probabilidades de Bandt y Pompe

2.1. Introducción

La distribución de probabilidades utilizada para caracterizar un sistema es una elección sumamente importante para el estudio de dicho sistema. En este Capítulo se presenta la Función de Distribución de Probabilidades propuesta por Bandt y Pompe (FDP de BP) y su utilización como argumento en la Entropía Informacional de Shannon para dar lugar a la Entropía de Permutación. También se presentan otros funcionales de la entropía (extensiones de la entropía de Shannon) evaluados en la FDP de BP que han sido utilizados en la literatura para caracterizar sistemas dinámicos, como así también una aplicación en datos de EEG para comparar los funcionamientos de las distintas entropías.

Pero antes de presentar la Función de Distribución de Probabilidades propuesta por Bandt y Pompe es necesario estudiar como se propone transformar una serie de tiempo proveniente de un proceso generador de datos continuos a una secuencia

de símbolos pertenecientes a un conjunto finito mediante la cual se construirá la Función de Distribución de Probabilidad (FDP) deseada.

2.2. La simbolización propuesta por Bandt y Pompe

Sea una $\{x_t\}_{t \in T}$ una posible realización en forma de una serie tiempo de un proceso generador de datos $\{\mathcal{X}_t\}_{t \in T}$ a valores reales de largo T , asumiendo en un principio $P(\mathcal{X}_r = \mathcal{X}_s) = 0 \forall r \neq s$.

Como \mathcal{X}_t es una variable aleatoria continua, cada x_t tiene infinitos valores posibles dentro de un rango, y es de práctica común reemplazar la serie original por una secuencia de símbolos (o estados) $\{\pi_i \ i = 1 \dots N\}$ finitos y calcular la FDP de esta secuencia $\{\pi_i\}_{i \in T'}$ (Bandt and Pompe [2002]). Al conjunto de símbolos $S = \{\pi_1, \pi_2, \dots, \pi_N\}$ se lo denomina un alfabeto.

Sea $X_t^{(m)} = (x_t, x_{t+1}, \dots, x_{t+m-1})$ con $0 \leq t \leq T - m + 1$ el vector de *embedding* de largo m en la ubicación t de la serie de tiempo $\{x_t\}_{t \in T}$, donde a m se la denominará la dimensión de *embedding*.

Sea $S_{m \geq 2}$ el conjunto formado por todas las posibles permutaciones de orden m del conjunto $\mathfrak{I} = \{i_1, \dots, i_m\}$ donde $i_j \neq i_k \forall j \neq k$.

Entonces si $\pi_i \in S_m$, es de la forma $\pi_i = i_1 \dots i_m$ para cualquier permutación de los elementos de \mathfrak{I} . De esta manera $S_m = \{\pi_1, \dots, \pi_{m!}\}$ es un alfabeto, la cardinalidad de S_m es $m!$ y se llamara al elemento π_i un símbolo en el alfabeto S_m .

Se desea hacer un mapeo entre los infinitos posibles vectores de *embedding* $X_t^{(m)}$ a un grupo finito de símbolos pertenecientes al alfabeto S_m .

Este mapeo debe ser definido de una manera que preserve la relación relevante deseada entre los elementos $x_t \in X_t^{(m)}$ (un patrón, en el sentido de una estructura ordinal, definido de antemano), y todos los $t \in T$ que compartan este patrón deben ser mapeados unívocamente al mismo símbolo $\pi_i \in S_m$. Es decir, el procedimiento consta en convertir los patrones de los vectores de *embeddig* en símbolos de un

alfabeto finito, donde todos los vectores que compartan el mismo patrón, serán mapeados al mismo símbolo. En definitiva es un mapeo de **patrones** a **símbolos**.

En la literatura que abarca la Entropía de Permutación hemos encontrado dos maneras de definir este mapeo de patrones a símbolos:

1. **Permutando los rangos:** ordenando los rangos (definidos según la Ecuación 2.1) de los x_i en $X_t^{(m)}$ en orden cronológico (ver Figura 2.1).
2. **Permutando los índices cronológicos:** ordenando los índices i en $x_i \in X_t^{(m)}$ (ver Figura 2.2).

2.2.1. Permutando los rangos: Mapeo según Rangos

Para un t dado arbitrariamente, a los m valores reales $X_t^{(m)} = (x_t, x_{t+1}, \dots, x_{t+m-1})$ se los reemplazan por sus rangos, donde la función de rangos se define como:

$$R(x_{t+n}) = \sum_{k=0}^{m-1} \mathbb{1}(x_{t+k} \leq x_{t+n}) \quad (2.1)$$

donde $\mathbb{1}$ es la función indicadora (i.e $\mathbb{1}(Z) = 1$ si Z es verdadero 0 caso contrario); $x_{t+n}, x_{t+k} \in X_t^{(m)}$ y $1 \leq R(x_{t+n}) \leq m$. De esta manera $R(\min(x_{t+k})) = 1$ y $R(\max(x_{t+k})) = m$.

Esto significa que cada valor $x_t \in X_t^{(m)}$ es reemplazado por su rango:

$$X_t^{(m)} = (x_t, x_{t+1}, \dots, x_{t+m-1}) \rightarrow \pi_i = i_1 i_2 \dots i_m \in S_m$$

$$\text{donde } i_1 = R(x_t), i_2 = R(x_{t+1}), \dots, i_m = R(x_{t+m-1})$$

y $R(x_{t+k}); k = 0, \dots, m - 1$ está definido por la ecuación 2.1

En definitiva, con el mapeo según rangos simplemente convierte el valor $x_i \in X_t^{(m)}$ en su rango $R(x_i) \in \{1, 2, \dots, m\}$ manteniendo su ubicación en el tiempo, y esos m elementos forman el símbolo π_i en S_m . El alfabeto completo son todas las posibles $m!$ permutaciones de los rangos de un vector de *embedding*.

Por ejemplo, para el caso de $m = 3$ el alfabeto es:

$$S_3 = \{\pi_1 = 123, \pi_2 = 132, \pi_3 = 213, \pi_4 = 312, \pi_5 = 231, \pi_6 = 321\}$$

. Tomando como ejemplo la serie de siete valores , y dimensión de *embedding* $m = 3$ (Bandt and Pompe [2002]):

$$X_t = (4, 7, 9, 10, 6, 11, 3) ; T = 7 \quad (2.2)$$

- $X_1^{(3)} = (4, 7, 9)$ y $X_2^{(3)} = (7, 9, 10)$ se transforman en el símbolo $\pi = 123$ dado que
 $R(x_{t+1}) = \mathbf{1}$, $R(x_{t+2}) = \mathbf{2}$, $R(x_{t+3}) = \mathbf{3}$.
- $X_3^{(3)} = (9, 10, 6)$ y $X_5^{(3)} = (6, 11, 3)$ corresponden al símbolo $\pi = 231$ ya que
 $R(x_{t+1}) = \mathbf{2}$, $R(x_{t+2}) = \mathbf{3}$, $R(x_{t+3}) = \mathbf{1}$.
- $X_4^{(3)} = (10, 6, 11)$ corresponde al símbolo $\pi = 213$ ya que
 $R(x_{t+1}) = \mathbf{2}$, $R(x_{t+2}) = \mathbf{1}$, $R(x_{t+3}) = \mathbf{3}$.

La Figura 2.1 presenta una esquema de este mapeo para todas las alternativas para $m = 3$. Se puede observar que los índices del eje vertical están fijos, ordenados según la amplitud (i.e. rangos), y son mapeados al eje del tiempo. El símbolo resultante puede ser obtenido leyendo las etiquetas en el eje horizontal de izquierda a derecha (en orden cronológico). Este método es el utilizado en diferentes trabajos para construir el alfabeto simbólico (Bandt [2014]; Bandt and Shiha [2007]; Riedl et al. [2013]) .

2.2.2. Permutando los índices cronológicos: Mapeo según el orden cronológico

Nuevamente, para un t dado, los m valores $X_t^{(m)} = (x_t, x_{t+1}, \dots, x_{t+m-1})$ pueden ser ordenados en orden creciente según su amplitud. Para efectuar este mapeo:

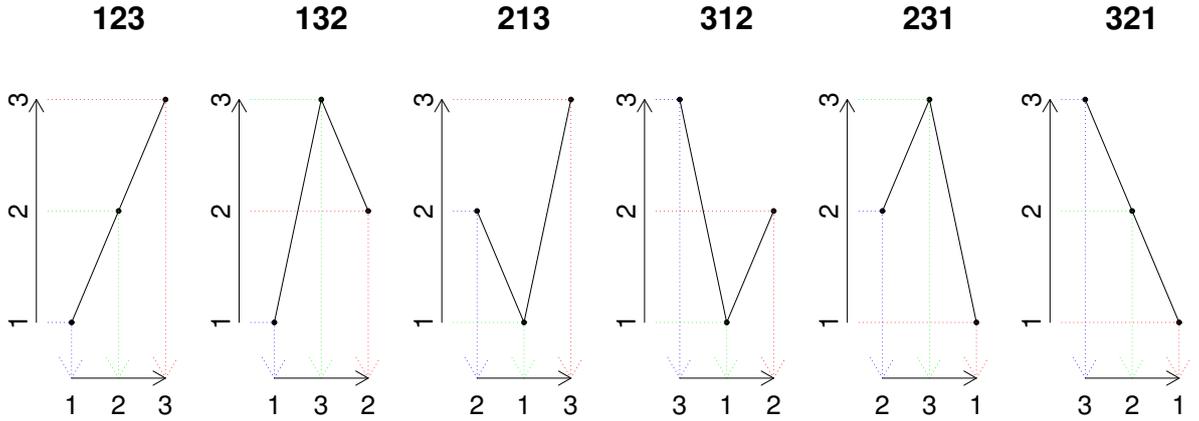


Figura 2.1 **Mapeo según rangos** Se muestran todos los símbolos posibles para $m = 3$. El mapeo según rangos simplemente convierte el valor x_i en $X_t^{(m)}$ en su rango $R(x_i) \in \{1, 2, \dots, m\}$ manteniendo su ubicación en el tiempo. Se puede observar que los índices del eje vertical están fijos, ordenados según la amplitud (i.e. rangos), y son mapeados al eje del tiempo. El símbolo resultante puede ser obtenido leyendo las etiquetas en el eje horizontal de izquierda a derecha (en orden cronológico).

$$X_t^{(m)} = (x_t, x_{t+1}, \dots, x_{t+m-1}) \rightarrow \pi_i = i_1 i_2 \dots i_m \in S_m$$

$$i_1 i_2 \dots i_m \text{ debe cumplir con } x_{t+i_1-1} < x_{t+i_2-1} < \dots < x_{t+i_m-1}$$

Con este procedimiento, los índices cronológicos son ordenados según su amplitud. El alfabeto completo son todas las posibles permutaciones de estos índices. Por ejemplo, para el caso de $m = 3$:

$$S_3 = \{\pi_1 = 123, \pi_2 = 132, \pi_3 = 213, \pi_4 = 231, \pi_5 = 312, \pi_6 = 321\}$$

Tomando la serie anterior (2.2) como ejemplo:

$$X_t = (4, 7, 9, 10, 6, 11, 3) ; T = 7 \tag{2.3}$$

- $X_1^{(3)} = (4, 7, 9)$ y $X_2^{(3)} = (7, 9, 10)$ representan la permutación $\pi = 123$ ya que $x_{t+1} < x_{t+2} < x_{t+3}$.

- $X_3^{(3)} = (9, 10, 6)$ y $X_5^{(3)} = (6, 11, 3)$ corresponden al símbolo $\pi = 312$ ya que $x_{t+3} < x_{t+1} < x_{t+2}$.
- $X_4^{(3)} = (10, 6, 11)$ corresponde al símbolo $\pi = 213$ dado que $x_{t+2} < x_{t+1} < x_{t+3}$.

Con este mapeo según el orden cronológico simplemente se mapea cada valor $x_i \in X_t^{(m)}$ ordenando su índice cronológico $t \in \{1, 2, \dots, m\}$ de acuerdo a la amplitud creciente de cada $x_i \in X_t^{(m)}$.

En la Figura 2.2 se observa una representación de este mapeo para todas las alternativas en $m = 3$. Se puede notar que los índices del eje del tiempo están fijos en orden cronológico, y estos son mapeados al eje vertical (el eje de la amplitud). El símbolo resultante puede ser obtenido leyendo las etiquetas del eje vertical de abajo hacia arriba (en la dirección de la amplitud creciente).

Este método es el utilizado en diferentes trabajos para construir el alfabeto simbólico (Bandt and Pompe [2002]; Bian et al. [2012]; Parlitz et al. [2012]).

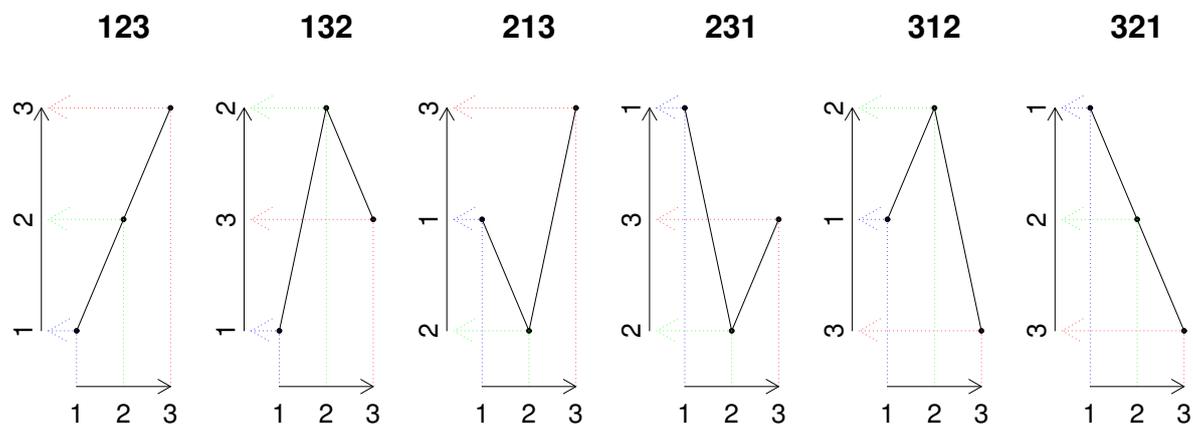


Figura 2.2 **Mapeo según el orden cronológico:** Se muestran todos los símbolos para $m = 3$. Con este mapeo según el orden cronológico simplemente se mapea cada valor x_i en $X_t^{(m)}$ ordenando su índice cronológico $t \in \{1, 2, \dots, m\}$ de acuerdo a la amplitud creciente de cada x_i en $X_t^{(m)}$. Se puede notar que los índices del eje del tiempo son fijos en el orden cronológico, y estos son mapeados al eje vertical (el eje de la amplitud). El símbolo resultante puede ser obtenido leyendo las etiquetas del eje vertical de abajo hacia arriba (en la dirección de la amplitud creciente).

Como el objetivo principal de estas metodologías es definir un set de símbolos diferentes $\pi_i \in S_m$ con una regla unívoca que realice el mapeo entre los vectores de *embedding* ($X_t^{(m)}$) a estos símbolos π_i , donde cada π_i represente a un único patrón, permutar los rangos de acuerdo a su aparición o permutar los índices cronológicos de acuerdo a los rangos tiene, en principio, la misma validez.

Las Figuras 2.1 y 2.2 revelan que estas metodologías difieren en 2 de los 6 símbolos para $m = 3$ (es decir difieren en el “nombre” de símbolo al que son transformados 2 de los 6 patrones). A medida que m crece, la diferencia entre los símbolos generados por estos dos mapeos se incrementa.

Se verá en la próxima Sección (Sección 2.3) que con las probabilidades de aparición de estos símbolos se definirá una FDP \mathbf{P} - que caracterizará al sistema dinámico en estudio y a esta FDP se le aplicarán cuantificadores $f(\mathbf{P})$ que resuman información importante sobre dicho proceso. Si bien las diferencias debidas al mapeo elegido para la simbolización no afectan a la entropía de permutación $\mathcal{H}(\mathbf{P})$ (ver Sección 2.4), cuando se propone extender el alfabeto (Bian et al. [2012], ver Capítulo 5) para calcular la entropía de permutación, la elección de dicho mapeo desempeña un rol importante.

De la misma manera cuando se le aplican a la FDP cuantificadores locales como la información de Fisher $\mathcal{F}(\mathbf{P})$ (Martin et al. [1999]; Rosso et al. [2015, 2010b]; Vignat and Bercher [2003], ver Capítulo 3) los resultados no son los mismos y dependen del tipo de mapeo utilizado. En esta Tesis se utilizará el mapeo según rangos salvo indicación contraria.

2.3. La Función de Distribución de Probabilidades de Bandt y Pompe

Usando algunas de las metodologías de simbolización expuestas en 2.2 a un proceso aleatorio $\{\mathcal{X}_t\}_{t \in T}$, de largo T , se le puede asociar una función de distribución de probabilidad de sus patrones, siempre y cuando el proceso cumpla con una estacionariedad muy débil, para $k \leq T$ la probabilidad de que $\mathcal{X}_t > \mathcal{X}_{t+k}$ no debe

depender de t .

Sea S una variable aleatoria tal que:

$$S : \Omega \rightarrow \mathfrak{S} = \{1, \dots, m!\} \in \mathbb{N}$$

donde $\Omega = S_m = \{\pi_1, \dots, \pi_{m!}\}$

y se la define de la siguiente manera:

$$S = \begin{cases} 1 & \text{si } s \text{ es del tipo } \pi_1 \\ 2 & \text{si } s \text{ es del tipo } \pi_2 \\ \dots & \\ m! & \text{si } s \text{ es del tipo } \pi_{m!} \end{cases} \quad (2.4)$$

Entonces se puede pensar que cada proceso aleatorio $\{\mathcal{X}_t\}_{t \in T}$ tiene una función de distribución probabilidades \mathbf{P} - asociada a sus patrones (mediante la FDP de la variable aleatoria S proveniente de los símbolos π_i) para cada orden m : $\mathbf{P} = \mathbb{P}(S = s)$. Esta función de distribución de probabilidades es la **Función de Distribución de Probabilidades de Bandt Pompe**

Para cada realización $\{x_t\}_{t \in T}$ una manera de estimar esta FDP de BP es realizar un mapeo de patrones a símbolos π_i según lo visto en la Sección anterior y calcular la frecuencia relativa de los mismos:

$$\hat{\mathbf{P}} = \widehat{\mathbb{P}(S = i)} = \frac{\#\{s | s \leq T - m + 1; s \text{ es del tipo } \pi_i\}}{T - m + 1}. \quad (2.5)$$

En esta expresión, el símbolo $\#$ representa “cardinalidad”.

Así,

$$\begin{aligned} \mathbb{P}(\widehat{S} = 1) &= p(\pi_1) = \frac{\#\{s | s \leq T - m + 1; s \text{ es del tipo } \pi_1\}}{T - m + 1} \\ &\dots \\ \mathbb{P}(\widehat{S} = m!) &= p(\pi_{m!}) = \frac{\#\{s | s \leq T - m + 1; s \text{ es del tipo } \pi_{m!}\}}{T - m + 1} \end{aligned}$$

y por lo tanto $\hat{\mathbf{P}} = \{p(\pi_1), \dots, p(\pi_{m!})\}$

Por lo tanto para la estimación de la FDP de BP es necesario elegir dos parámetros: un largo de patrón m (dimensión de *embedding*), y un retardo de tiempo entre valores τ . Este parámetro τ significa que el vector de *embedding* se puede definir como $X_i^{(m)} = (x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau})$, es decir tomar patrones con valores no consecutivos, separados por un cierto retardo τ . Bandt y Pompe sugieren trabajar con valores de $3 \leq m \leq 6$ y consideran específicamente $\tau = 1$ en su artículo base (Bandt and Pompe [2002]), y son los parámetros que se van a utilizar en esta Tesis, salvo indicación contraria para alguna aplicación en particular (ver Sección 2.7). Es claro que otros valores de τ pueden proveer información adicional (L. Zunino and Rosso [2012]), pero el análisis de las propiedades estadísticas de la Entropía de Permutación propuesto en esta Tesis se puede extender a otros valores de estos parámetros sin afectar el análisis ni los resultados obtenidos.

La definición en la Ecuación 2.5 estima la Función de Distribución de Probabilidades de Bandt y Pompe donde a cada símbolo π_i que representa un patrón dado en la serie de tiempo se le asocia una probabilidad.

2.4. La Entropía de Permutación

La entropía es una medida de desorden, de incertidumbre acerca de que un sistema se encuentre en un determinado estado, o de la cantidad de información que puede ser obtenida mediante observaciones de sistemas desordenados (Maszcyk and Duch [2008]). Dado una función de densidad de probabilidad $f(x)$ con $x \in \Delta \subset \mathbb{R}$ y $\int_{\Delta} f(x) dx = 1$, su *entropía de Shannon* S asociada

$$S[f] = - \int_{\Delta} f(x) \ln(f(x)) dx, \quad (2.6)$$

es una medida de carácter global que no es muy sensible a grandes cambios en la distribución que tomen lugar en pequeñas regiones (Shannon and Weaver [1949]). Extensiones del trabajo original de Shannon han derivado en distintas alternativas de medidas de información o entropías.

Sea ahora $\mathbf{P} = \{p_i; i = 1, \dots, N\}$, con $\sum_{i=1}^N p_i = 1$, una función de distribución de probabilidad de una variable aleatoria discreta, con N la cantidad de estados posibles del sistema bajo estudio. Para este caso se define una entropía de Shannon *normalizada* ($0 \leq \mathcal{H} \leq 1$) como:

$$\mathcal{H}(\mathbf{P}) = S(\mathbf{P})/S_{max} = \left\{ - \sum_{i=1}^N p_i \ln(p_i) \right\} / S_{max}, \quad (2.7)$$

donde el denominador $S_{max} = S(\mathbf{P}_e) = \ln N$ se obtiene por una distribución de probabilidad uniforme discreta (i.e $\mathbf{P}_e = \{p_i = 1/N, \forall i = 1, \dots, N\}$) y se define a $0 \cdot \ln(0)$ como 0.

Esta entropía puede ser vista como una medida de incertidumbre asociada al proceso físico descrito por \mathbf{P} . Por ejemplo, si $S(\mathbf{P}) = S_{min} = 0$, entonces $\mathcal{H}[\mathbf{P}] = 0$, se está en posición de predecir con total certidumbre cual de los resultados i va a tomar lugar ($p_i = 1, i = j$ y $p_i = 0 \forall i \neq j$). El conocimiento acerca del estado en el que se encuentra el proceso subyacente descrito por \mathbf{P} es máximo en esa instancia, y la incertidumbre es mínima. En contraste la incertidumbre es máxima para la distribución de probabilidad uniforme discreta dado que cada resultado exhibe la misma probabilidad de ocurrencia, donde $S(\mathbf{P}) = S_{max}$, y entonces $\mathcal{H}(\mathbf{P}) = 1$.

Cuando se usa la FDP de BP para describir un proceso físico se está teniendo en cuenta la estructura temporal de la serie de tiempo generada por dicho proceso. De este modo, esto permite descubrir detalles importantes concernientes a la estructura ordinal de la serie de tiempo (Rosso et al. [2012b, 2007b] y también produce información acerca de la correlación temporal (Rosso et al. [2013])).

De esta manera se define la *Entropía de Permutación* cuando en la Ecuación 2.7 se utiliza como \mathbf{P} , la FDP de BP.

2.5. Extensiones de la Entropía de Permutación

Shannon introdujo su entropía a fines de la década del 40 y al ser un concepto tan potente, se extendió a lo largo y a lo ancho de la comunidad científica, pero desde ese entonces otras entropías similares han aparecido en la literatura científica (Harremoës [2006]). Estas medidas de entropía comparten algunas, aunque no todas, las propiedades de la Entropía Informacional de Shannon. Los ejemplos mas importantes son la *Entropía de Rényi* (Rényi [1961]) y la *Entropía de Tsallis* (Tsallis [1988]). Otra medida interesante y que lleva su propio nombre a pesar de ser un caso particular de la entropía de Rényi es la MinEntropía, propuesta como un cuantificador mejorado para revelar la presencia de sutiles correlaciones temporales en series de tiempo (Zunino et al. [2015]).

2.5.1. Entropía de Permutación de Rényi

La Entropía de Rényi, que generaliza la Entropía Informacional de Shannon se define de la siguiente manera:

$$R_\alpha(\mathbf{P}) = \frac{1}{1-\alpha} \ln \left(\sum_{i=1}^{m!} p_i^\alpha \right), \quad (2.8)$$

donde el orden α ($\alpha \geq 0$ y $\alpha \neq 1$) es un parámetro de control y la Entropía Informacional de Shannon es recuperada cuando $\alpha \rightarrow 1$.

Si para el cálculo de \mathbf{P} en la ecuación 2.8 se utiliza la FDP de BP esta entropía es la *Entropía de Permutación de Rényi* (Mammone et al. [2015]).

La Entropía de Rényi esta vinculada a la distribución de probabilidades tomada de la serie de tiempo a través del parámetro α . Cuando α es alto se enfatiza la forma de distribución leptocúrtica, i. e. un pico mas filoso y colas mas ponderadas que una distribución Gaussiana (la entropía es más sensible a eventos que ocurren con mayor frecuencia), mientras que un α pequeño enfatiza las distribuciones platicúrticas, i.

e. picos más chatos y colas más livianas que la distribución Gaussiana (la entropía es mas sensible a eventos extremos). Es decir que con dicho parámetro se puede controlar la sensibilidad de la entropía a la curtosis de la FDP de BP.

2.5.2. MinEntropía de Permutación

En el límite cuando $\alpha \rightarrow \infty$, $R_\alpha(\mathbf{P})$ converge a $R_\infty(\mathbf{P})$ y se obtiene la MinEntropía.

$$R_\infty(\mathbf{P}) = -\ln \left(\max_{i=1, \dots, m} p(\pi_i) \right). \quad (2.9)$$

que utiliza sólo la probabilidad del evento más frecuente. Si para el cálculo de función de distribución de probabilidades \mathbf{P} se utiliza la FDP de BP se obtiene la *MinEntropía de Permutación*.

2.5.3. Entropía de Permutación de Tsallis

Existen una variedad de sistemas anómalos para los cuales la teoría de Boltmann-Gibbs puede exhibir algunas dificultades, como sistemas con autocorrelaciones de largo alcance (Pavón [1987]) y procesos no markovianos (Cáceres [1999]), entre otros. Para tratar con este tipo de sistemas se postuló una entropía no extensiva, llamada la Entropía de Tsallis (Tsallis [1988]), cuya forma funcional es:

$$S_q(\mathbf{P}) = -\frac{(1 - \sum_i p_i^q)}{1 - q}, \quad (2.10)$$

En el límite cuando $q \rightarrow 1$ la Entropía de Tsallis converge a la Entropía Informacional de Shannon.

Si para el cálculo de función de distribución de probabilidades \mathbf{P} se utiliza la FDP de BP se obtiene la *Entropía de Permutación de Tsallis*.

2.5.4. Entropía de Permutación Ponderada

Esta medida de entropía en particular está ideada específicamente para utilizarla utilizando la simbolización propuesta por Bandt y Pompe. Sin embargo, no utiliza la FDP de BP explicada en 2.3 sino una versión ponderada de la misma. Todas las entropías de permutación consideradas anteriormente toman en cuenta la amplitud de los valores de la serie de tiempo sólo para determinar si son mayores o menores con respecto a sus vecinos y de esta manera transformar los patrones o vectores de *embedding* a símbolos π_i . Es decir que la información de la magnitud de la diferencia relativa entre amplitudes dentro de un vector de *embedding* se pierde. Una versión de la Entropía de Permutación llamada *Entropía de Permutación Ponderada* se propone para incorporar esta información, que puede ser valiosa (Fadlallah et al. [2013]). Esta entropía le da un peso a cada vector de *embedding* de acuerdo a la siguiente medida de dispersión tomada de dicho vector:

$$w_t = \frac{1}{m} \sum_{i=1}^m (x_{t+i} - \bar{X}_t^{(m)})^2 \quad (2.11)$$

donde m es la dimensión de *embedding*, $X_t^{(m)} = \{x_t, x_{t+1}, \dots, x_{t+m-1}\}$ y $\bar{X}_t^{(m)}$ representa a la media en amplitud de $X_t^{(m)}$.

De modo que la estimación de las frecuencias relativas (o probabilidades ponderadas) para cada símbolo π_i se define como:

$$\hat{P}^w(\pi_i) = \frac{\sum_{t \leq (T-m+1)} (\{s_t | s_t \text{ es del tipo } \pi_i\}) \cdot w_t}{\sum_{k \leq T} s_k w_k} . \quad (2.12)$$

y la Entropía de Permutación Ponderada se obtiene como:

$$\mathcal{H}_w = - \sum_{i=1}^{m!} \hat{P}^w(\pi_i) \ln(\hat{P}^w(\pi_i)) \quad (2.13)$$

2.6. Patrones Prohibidos y Patrones Faltantes

La metodología de los patrones ordinales originados por los vectores de *embedding*, introducida por Bandt y Pompe ha proporcionado conocimientos nuevos muy importantes en lo que respecta la caracterización de series de tiempo teóricas y experimentales, en particular en el área sensible de la distinción entre dinámicas caóticas-deterministas y estocásticas (Amigó [2010]; Amigó et al. [2008, 2007]; Rosso et al. [2012a,b, 2007b, 2013]). En los trabajos de Amigó y sus colaboradores se estudiaron características locales de los componentes de la FDP de BP mostrando lo siguiente:

- Para series de tiempo generadas por dinámicas caóticas-deterministas no todos los posibles patrones ordinales pueden ser materializados, llamando a estos *patrones prohibidos*.
- El surgimiento de estos patrones prohibidos representa una propiedad particular de alguno de los elementos de la FDP de BP asociados a la serie de tiempo bajo estudio.
- La existencia de estos patrones ordinales prohibidos se vuelve un hecho persistente que puede ser considerado como una nueva propiedad dinámica.
- Para un largo de patrón fijo (o dimensión de *embedding*) m la cantidad de patrones prohibidos de una serie de tiempo, i.e patrones no observados, es independiente del largo de la serie, y presenta un comportamiento super-exponencial en función de m .

Algunos procesos estocásticos pueden presentar patrones prohibidos para series finitas (Carpi et al. [2010]; Rosso et al. [2012a,b]). Sin embargo, en el caso de procesos estocásticos como el ruido blanco, o correlacionados (ruidos con espectro de potencia $1/f^\alpha$, con $\alpha \geq 0$, Movimiento Browniano Fraccionario o ruido Gaussiano Fraccionario) puede ser numéricamente cerciorado que no emerge ningún patrón prohibido.

De hecho, para series de tiempo generadas por procesos estocásticos cuyos valores no están autocorrelacionados, cada patrón ordinal tiene la misma probabilidad de ocurrencia (Amigó [2010]; Amigó et al. [2007]; Carpi et al. [2010]), por lo que independientemente de la dimensión de *embedding* m , si la serie es lo suficientemente larga, todos los patrones ordinales aparecerán eventualmente. Para procesos estocásticos cuyos valores están autocorrelacionados la probabilidad de observar un patrón específico depende no sólo del largo de la serie T , sino también de estructura de esta correlación. La existencia de patrones ordinales no observados no califica a estos como *patrones prohibidos*, sino como *patrones faltantes*, y este fenómeno es debido a que la serie de tiempo tiene un largo finito. Una observación similar se puede hacer en el caso de datos reales que siempre poseen un componente estocástico, dada la omnipresencia del ruido dinámico (Cambanis et al. [1988]; Wold [1938]). Por lo tanto, la presencia de patrones faltantes puede ser tanto relacionada a procesos estocásticos o a procesos determinísticos contaminados con ruido (que es siempre el caso de las series de tiempo observadas).

Recientemente, se mostró que aún cuando la presencia es característica de dinámicas caóticas puras, una dimensión de *embedding* mínima m_{min} es necesaria para detectar su presencia en el caso de series de tiempo de largo finito (Rosso et al. [2013]). Resumiendo, un largo de patrón, m_{min} es necesario para detectar la presencia de patrones prohibidos, distintivos del determinismo (dinámicas caóticas). Este hecho no ha sido previamente mencionado en la literatura e ignorarlo puede ser fuente de interpretaciones erróneas.

2.7. Aplicación: Clasificación de señales de EEG normales y pre-ictales usando las distintas Entropías de Permutación.

2.7.1. Introducción

La epilepsia es una enfermedad neurológica en la que los pacientes sufren convulsiones espontáneas. La aparición de dos convulsiones, con condición idiopática desconocida, es necesaria para el diagnóstico de la epilepsia. Estas convulsiones son causadas por perturbaciones en la actividad eléctrica del cerebro. Como fue propuesto por la Liga Internacional contra la Epilepsia (ILAE) y la Oficina Internacional para la Epilepsia (IBE), una convulsión epiléptica es una ocurrencia transitoria de signos y/o síntomas debido a la actividad neuronal anormal, excesiva y sincrónica en el cerebro (Fisher et al. [2014, 2005]). La epilepsia se presenta en convulsiones y la aparición repentina y a menudo imprevista de las mismas representa uno de los aspectos más incapacitantes de la enfermedad. La identificación correcta de la presencia de actividad epiléptica, la caracterización de los patrones espacio-temporales de la actividad cerebral correspondiente y la predicción de la ocurrencia de convulsiones son retos importantes, y lograr esto podría mejorar significativamente la calidad de vida de los pacientes con epilepsia. El tiempo transcurrido entre las convulsiones en pacientes con epilepsia se denomina período *inter-ictal*.

En esta aplicación se aborda el problema de la clasificación automática, utilizando varias entropías basadas en en la distribución de Bandt y Pompe, entre dos tipos de señales de EEG: EEG de períodos libres de convulsiones de un paciente con epilepsia y EEG de una persona sana.

La *Entropía de Permutación* (Bandt and Pompe [2002]) se ha utilizado en varias aplicaciones para estudiar la actividad eléctrica del cerebro, más particularmente en la investigación sobre epilepsia.

A menudo, las crisis epilépticas se manifiestan en una altamente estereotipada secuencia ordenada de síntomas y signos con variabilidad limitada y por esta razón se conjeturó que este estereotipo puede implicar que la dinámica neuronal ictal podría tener características deterministas, y que ésto presumiblemente se realizaría en las regiones ictogénicas del cerebro (Schindler et al. [2011]). También una detección precisa de las transiciones de los estados normales a patológicos puede mejorar el diagnóstico y el tratamiento, por lo que la detección de dichos cambios dinámicos en las señales de EEG es importante en los estudios clínicos. Por lo tanto la Entropía de Permutación se utilizó para la detección de determinismo y cambios dinámicos en señales de EEG (Bruzzo et al. [2008]; Keller and Wittfeld [2004]; Li et al. [2007b]; Nicolaou and Georgiou [2012]; Ouyang et al. [2010, 2009]; Veisi et al. [2007]).

La Entropía de Permutación se utiliza también para identificar las diversas fases de la actividad epiléptica en las señales EEG intracraneales registradas en pacientes que padecen epilepsia intratable (Cao et al. [2004]). También se usó en la predicción y clasificación: se probó que para una población de ratas, esta entropía puede utilizarse no sólo para rastrear los cambios dinámicos de los datos del EEG, sino también para detectar con éxito los estados de pre-convulsión (Li et al. [2007a]). La tarea de clasificación es una actividad prominente en la investigación de la epilepsia (Zanin et al. [2012]), y es importante para fines de diagnóstico, ya que permite discriminar entre los registros electroencefalográficos normales y patológicos que son problemas no triviales.

La *Entropía de Rényi* se utilizó para fines de clasificación ya que permite diferenciar las señales de EEG focal y no focal (Kannathal et al. [2005]).

La *MinEntropía de Permutación* fue aplicada para discriminar entre señales de EEG de voluntarios sanos y señales de EEG de los períodos interictales de pacientes con epilepsia. Su poder discriminante se comparó con la habitual Entropía

de Permutación de una manera descriptiva ([Zunino et al. \[2015\]](#)).

La *Entropía de Tsallis* se usó en el estudio de la señal de EEG y ha mostrado que esta medida no extensiva de Tsallis puede discriminar mejor entre señales de EEG con distintas características que sus contrapartes de Shannon ([Capurro et al. \[1999\]](#)).

La *Entropía de Permutación Ponderada* se prueba en señales de EEG de un solo canal y multicanal ([Fadlallah et al. \[2013\]](#)). En otro estudio, se incluyeron en el estudio tres estados fisiológicos del EEG, ojos cerrados, ojos abiertos y tarea extraña visual para examinar la capacidad de esta entropía para identificar y discriminar diferentes estados fisiológicos [Vuong et al. \[2014\]](#).

En el contexto del *Machine Learning*, existen varios esfuerzos previos para discriminar entre EEG de voluntarios sanos y EEG de períodos interictales de pacientes con epilepsia: se usa una red neuronal recurrente con características de dominio de tiempo y frecuencia ([Vuong et al. \[2014\]](#)), un árbol de decisión combinado con una transformada de fourier rápida ([Polat and Güneş \[2007\]](#)) y una combinación de la transformada wavelet discreta con un modelo mixto experto ([Subasi \[2007\]](#)). En cuanto al uso de las entropías en este contexto se utiliza un sistema de inferencia *neurofuzzy* adaptativo dotado de entropías ([Kannathal et al. \[2005\]](#)) y la Entropía Aproximada junto con una red Elman ([Ocak \[2009\]](#)). La mayoría de ellos con alta precisión de separación entre las señales, más del 95 % de las señales correctamente clasificadas.

2.7.2. Objetivo

El objetivo de la presente Sección es cuantificar el potencial de la Entropía de Permutación y de las entropías presentadas en la Sección 2.5 como variables independientes en un modelo lineal generalizado para discriminar entre registros de EEG de pacientes sanos y de pacientes con epilepsia en un marco inferencial. El uso de un modelo de regresión logística dará nuevas perspectivas para fines de

clasificación. Es importante señalar que a diferencia de los artículos citados, se utilizó un método paramétrico de clasificación con las entropías como variables predictoras, que conduce a una interpretación de los parámetros estimados y también obtenemos una visión de la estructura causal de las series temporales mediante la mejor combinación de τ y m de la FDP de BP.

2.7.3. Materiales y Métodos

2.7.3.1. EEG data

Se utilizaron datos de series de tiempo EEG gratuitas del Departamento de Epileptología de la Universidad de Bonn, disponibles en: <http://www.texte.microsoft.com/spain/index.html>. Se registraron para el estudio dos conjuntos A y B que contenían cada uno 200 segmentos de EEG de un solo canal de 23,6 seg. de duración. Estos segmentos se seleccionaron y se cortaron a partir de registros de EEG multicanal continuos después de la inspección visual de artefactos, por ejemplo, debido a la actividad muscular o movimientos de los ojos. El conjunto A consistió en segmentos extraídos de grabaciones de EEG superficiales que se llevaron a cabo en cinco voluntarios sanos, relajados en un estado despierto. Este estado se denominará *normal* dentro de este documento. Los conjuntos B se originaron a partir de un archivo de EEG de diagnóstico prequirúrgico durante los períodos pre-ictal, es decir, períodos en los que no se detectan convulsiones en pacientes con epilepsia. Esta muestra será referida como *pre-ictal* dentro de este documento. Se recolectaron EEG epilépticos de electrodos intracraneales que fueron colocados en la zona epileptogénica correcta (Andrzejak et al. [2001a,b]). El conjunto de datos consta de 400 segmentos de datos, 200 pertenecientes a la condición normal y 200 pertenecientes a la condición pre-ictal, cuya longitud es 4097 puntos de datos con una frecuencia de muestreo de 173,61 Hz para cada grupo. Estos datos fueron analizados para clasificación en contextos muy diferentes, como la inteligencia artificial y la teoría de la información entre otros.

2.7.3.2. Modelos de Clasificación: Regresión Logística

La regresión logística es el método más popular para predecir los resultados binarios sobre la base de una o más variables predictoras, y el concepto matemático central que subyace a la regresión logística es el logit, o sea el logaritmo natural de $\rho/(1-\rho)$ donde $0 < \rho < 1$ (Hosmer Jr et al. [2013]). El caso más simple de regresión lineal es para un predictor continuo X y una variable explicada dicotómica Y . La gráfica de tales datos da como resultado dos líneas paralelas, que son difíciles de describir mediante una recta de regresión de cuadrados mínimos, y se ajusta mucho mejor con una forma en S (a menudo referida como sigmoidal). La aplicación de la transformación logit a la probabilidad de que la variable dependiente sea 1 dado un vector de observaciones x , resuelve este problema de ajuste, y el modelo no requiere que el error sea normal o constante a través del rango de datos. Dado que la variable de respuesta Y es binaria, la describiremos como una variable aleatoria que toma el valor 0 o el valor 1, dependiendo de si la observación tiene un atributo presente o no. Por ejemplo, $Y = 1$ si el paciente presenta una enfermedad determinada y $Y = 0$ en caso contrario. En una ecuación de regresión logística simple con una única variable predictora, X , denotamos por $\rho(x)$ la probabilidad de que la variable de respuesta Y sea igual a 1 (la enfermedad está presente) dado que $X = x$.

$$\text{logit}(\rho(x)) = \ln\left(\frac{\rho(x)}{1-\rho(x)}\right) = \alpha + \beta x, \quad (2.14)$$

$$\rho(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}. \quad (2.15)$$

Donde α es la ordenada al origen del modelo lineal transformado y el valor del coeficiente β determina la relación entre X y el logit de ρ . La interpretación de este coeficiente es la siguiente:

$$\frac{\rho(x)}{1-\rho(x)} = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = e^{\beta x}. \quad (2.16)$$

Por lo que si el valor de la variable X aumenta en 1 punto, y el valor de β es positivo, el ratio entre estas dos probabilidades aumenta en un factor de e^β .

Siguiendo el ejemplo médico, este ratio representa cuanto más es la probabilidad de que el paciente esté enfermo en relación con la probabilidad de que el paciente no tenga esa enfermedad dado el valor $X = x$. Si el modelo tiene buenas propiedades de pronóstico, la regresión logística puede utilizarse como una herramienta de clasificación. Un excelente método para probar el poder de clasificación de una regresión logística es la curva característica de funcionamiento del receptor (ROC).

2.7.3.3. Curva ROC: Performance del clasificador

Inicialmente, sólo se considera un proceso de clasificación de dos categorías. Cada observación se asigna a un elemento del conjunto $\{P, N\}$, siendo las mismas etiquetas de clase positiva y negativa respectivamente. Un modelo de clasificación (clasificador) es un mapeo de instancias a clases predichas. Como se indicó anteriormente en la Sección 2.7.3.2, la regresión logística puede ser un excelente clasificador. Cada instancia Y , u observación, es 1 o 0, es decir, tiene la enfermedad o no. Las predicciones producidas por el modelo, $\rho(x)$ son continuas, por lo que se necesita un umbral o un punto de corte (c) para tener un clasificador de dos clases. Si $\rho(x) \geq c$, entonces Y se predice como P , y la predicción es N de lo contrario. Para simplificar, la clase de predicción va a estar en el conjunto $\{P, N\}$.

Ahora, dado un clasificador y una instancia hay cuatro resultados posibles: *TP True Positive* cuando la instancia es positiva y se clasifica como positiva; *FN False Negative*, la instancia es positiva y se clasifica como negativa; *FP False Positive* la instancia es negativa y se clasifica como positiva; Y finalmente *TN True Negative* cuando la instancia es negativa y se clasifica como negativa. De esta manera definimos:

$$\text{Sensitividad} = \frac{TP}{TP + FN} , \tag{2.17}$$

$$\text{Especificidad} = \frac{TN}{TN + FP} . \tag{2.18}$$

Así, la sensibilidad refleja el poder del modelo para identificar correctamente la clase positiva y la especificidad, el poder de identificar a la clase negativa. Otro

indicador de rendimiento útil es la precisión del modelo y se define simplemente como el porcentaje de observaciones bien clasificadas en el conjunto de datos,

$$\frac{TP + TN}{N_{obs}} \quad (2.19)$$

Donde N_{obs} representa el número de observaciones.

Dado que la clasificación como P o N depende del punto de corte elegido c , estas medidas de desempeño del clasificador también van a depender de c . En una curva ROC, la sensibilidad se representa en función de la tasa de falsos positivos (1 - Especificidad) para diferentes puntos de corte. Cada punto de la curva ROC representa un par de sensibilidad / especificidad correspondiente a un umbral de decisión particular c . Una prueba con discriminación perfecta (sin superposición en las dos distribuciones) tiene un gráfico ROC que pasa por la esquina superior izquierda (Especificidad = 1, Sensibilidad = 1). Esto lleva a la conclusión de que cuanto más cerca la curva ROC esté en la esquina superior izquierda, mayor será la precisión global de la prueba. Como la curva ROC es una representación bidimensional del rendimiento del clasificador, un único valor escalar sería útil para comparar los clasificadores. Un método común es calcular el área bajo la curva, AUC ([Bradley \[1997\]](#)). Dado que el AUC es una porción del área de la cuadrícula de la unidad su área va a ser menor que 1. La suposición al azar produce la línea entre (0, 0) y (1, 1), cualquier clasificador realista debería tener un AUC mayor de 0,5. En términos generales, cualquier área entre 0,8 y 0,9 significa una buena *performance*, y cualquier área entre 0,9 y 1 representa una excelente *performance*.

2.7.3.4. El enfoque de la validación cruzada

Cuando se utiliza el conjunto total de las observaciones para estimar el indicador deseado de rendimiento del clasificador del modelo, en este caso el AUC, se produce un ajuste excesivo a los propios datos y este indicador de rendimiento se sobrestima. Con el fin de evitar ésto se desarrollaron varios métodos que consisten en mantener un subconjunto del set de datos libre del proceso de ajuste, y luego aplicar el

modelo a dicho subconjunto para estimar las AUC. El método consiste en dividir aleatoriamente el conjunto disponible de observaciones en dos partes, el conjunto de entrenamiento y el conjunto de validación. El modelo se ajusta con el conjunto de entrenamiento, y este modelo ajustado se utiliza en el conjunto de validación para calcular el AUC o la tasa de *performance* deseada. Una mejora de este método simple es dividir el conjunto de observaciones en k subconjuntos *-pliegues-* de tamaño aproximadamente igual, usando los primeros $k - 1$ *pliegues* para ajustar el modelo y aplicar el modelo ajustado al remanente para estimar el AUC. Luego, repetir esta metodología k veces usando cada *pliegue* exactamente una vez como el conjunto de validación, y el resto como el conjunto de entrenamiento. Para cada vez, el AUC estimado es independiente de los valores usados para ajustar el modelo. El promedio de estos valores de k AUC (clasificador de rendimiento) es el AUC resultante para el modelo propuesto. Los valores típicos de k son $k = 5$ y $k = 10$. Para este ejemplo se usará $k = 10$.

2.7.4. Resultados

Para discriminar entre las señales EEG *normales* y *pre-ictales* se utiliza la regresión logística. Se ajustó un modelo de regresión logística para cada estimación de las entropías presentadas en este Capítulo:

- Entropía de Permutación, $\hat{\mathcal{H}} = \mathcal{H}(\hat{\mathbf{P}})$
- MinEntropía de Permutación, $\hat{R}_\infty = R_\infty(\hat{\mathbf{P}})$
- Entropía de Permutación de Rényi, $\hat{R}_\alpha = R_\alpha(\hat{\mathbf{P}})$
- Entropía de Permutación de Tsallis, $\hat{S}_q = S_q(\hat{\mathbf{P}})$
- Entropía de Permutación Ponderada, \mathcal{H}_w

cada una como variable explicativa para clasificar las señales de EEG como *normal* ($Y = 1$) o *pre-ictal* ($Y = 0$), y donde $\hat{\mathbf{P}}$ es la estimación de la FDP de BP (ver ecuación 2.5).

Se evaluó cada combinación de longitud de vector de *embedding* $m = \{3, 4, 5, 6\}$ y retardo de tiempo $\tau = \{1, 2, 3, 4, 5\}$ para calcular las entropías. Para la Entropía de Permutación de Rényi se varía el parámetro α de 0,25 a 7,5 con incrementos de 0,25 (Mammone et al. [2015]) y para la Entropía de Permutación de Tsallis se varía el parámetro q de 0,1 a 3 con incrementos de 0,1 (Plastino A. [2005]).

En la Figura 2.3 se representa el AUC calculada mediante la validación cruzada de 10 pliegues ($AUC \pm 1 SD$), en función del retardo τ . La Figura está dividida por las diferentes entropías y por la dimensión m para una mejor comprensión.

Se puede ver que independientemente de la entropía utilizada, el mejor modelo para clasificar es cuando se calcula el PDF de BP con la dimensión $m = 3$ y el retardo $\tau = 5$ a excepción de la MinEntropía de Permutación ya que el modelo para $m = 4$ y $\tau = 4$ es ligeramente mejor. Cuando el retardo de tiempo aumenta, todas las entropías funcionan mejor como clasificadoras para distinguir las señales EEG normales de las pre-ictales. En la Entropía de Permutación de Rényi y la Entropía de Permutación de Tsallis la influencia del parámetro en el desempeño de clasificación disminuye a medida que el τ aumenta, esto se evidencia en la disminución de la dispersión entre los valores de AUC para cada m y τ .

La Entropía de Permutación como clasificadora entre señales de EEG de *normal* y *pre-ictal* tiene la correlación más fuerte con $\beta = -336$, con un p-valor cercano a cero (Cuadro 2.1). Esto significa que para cada 1/1000 que esta entropía sube, el cociente entre la probabilidad de clasificar la señal EEG como *pre-ictal* y la probabilidad de clasificar la señal EEG como *normal*, disminuye 28%. En otras palabras, pequeños incrementos en la Entropía de Permutación afectan significativamente la probabilidad de detectar estados *pre-ictales*. La adición de ruido a una señal aumenta la entropía, por lo que las señales de EEG ruidosas darían como resultado una menor sensibilidad (es decir, la habilidad de detectar señales de EEG *pre-ictales*). Por otro lado, la MinEntropía de Permutación tiene la correlación más débil con un $\beta = -13,29$, lo que significa que para cada 1/1000 que la MinEntropía de Permutación se mueve hacia arriba, el ratio entre probabilidades disminuye sólo 1,2% y sigue siendo un excelente clasificador. Este comportamiento

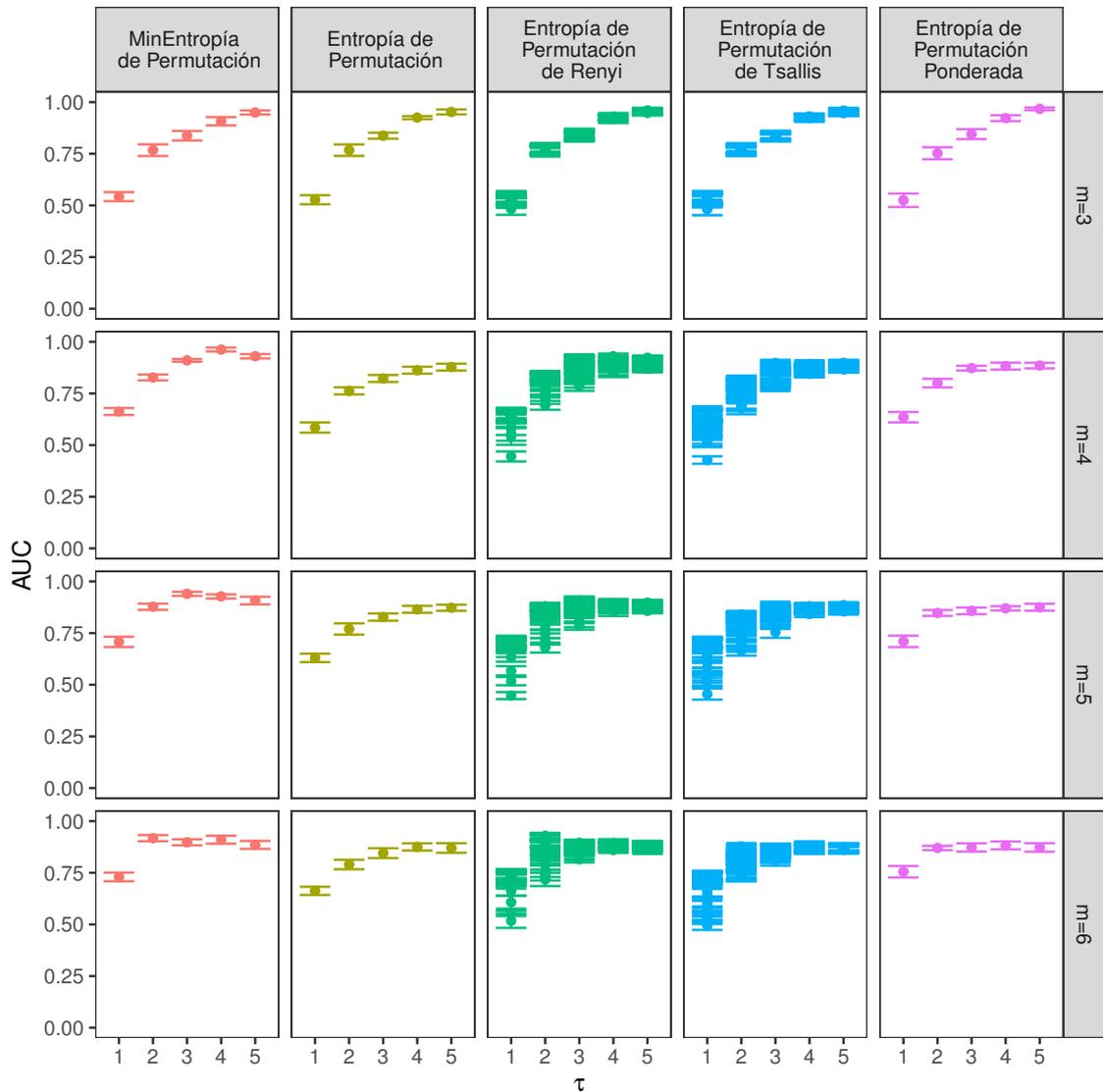


Figura 2.3 El Área bajo la curva ROC (AUC) calculado mediante validación cruzada de 10 pliegues ($AUC \pm 1\text{ sd}$), versus el retardo τ . Para un mejor entendimiento la Figura está separada por la dimensión de *embedding* m y por las distintas entropías calculadas. Se puede ver que independientemente de la entropía utilizada, el mejor modelo para clasificar es cuando se calcula el PDF de BP con la dimensión $m = 3$ y el retardo $\tau = 5$ a excepción de la MinEntropía de Permutación ya que el modelo para $m = 4$ y $\tau = 4$ es ligeramente mejor. Cuando el retardo de tiempo aumenta, todas las entropías funcionan mejor como clasificadoras para distinguir las señales EEG normales de las pre-ictales

Entropía	AUC	Precisión	Sensitividad	Especificidad	Coefficiente de regresión	p-valor
\mathcal{H}_w	0.9675	0.97	0.985	0.95	-107.98	$3.57 \cdot 10^{-13}$
$R_\infty(\hat{\mathbf{P}})$	0.9575	0.965	0.975	0.94	-13.298	$1.86 \cdot 10^{-12}$
$\mathcal{H}(\hat{\mathbf{P}})$	0.955	0.950	0.975	0.935	-335.99	$2.76 \cdot 10^{-12}$
$R_\alpha(\hat{\mathbf{P}})$	0.955	0.950	0.97	0.94	-121.78	$1.26 \cdot 10^{-12}$
$S_q(\hat{\mathbf{P}})$	0.955	0.945	0.97	0.94	-203.07	$2.78 \cdot 10^{-12}$

Cuadro 2.1 Mejores modelos para cada entropía ordenados según el valor de AUC en forma decreciente para la clasificación de los EEG. Esta tabla presenta los mejores modelos para cada entropía ordenados según el valor de AUC en forma decreciente. Para la Entropía de Permutación de Rényi y para Entropía de Permutación de Tsallis el parámetro se elige también según el desempeño de clasificación (teniendo en cuenta que hay un modelo para cada valor del parámetro), que es la Entropía de Permutación de Rényi con $\alpha = 2, 75$ y la Entropía de Permutación de Tsallis con $q = 1, 1$. La entropía que tiene el mejor rendimiento de clasificación es la Entropía de Permutación Ponderada seguida por la MinEntropía de Permutación por menos de una desviación estándar, y las entropías restantes tienen un rendimiento similar en términos del AUC (Figura 2.5). Todas las entropías tienen un desempeño excelente y similar en términos de precisión, por lo que el error de clasificación general es pequeño para el punto de corte $c = 0, 5$, y para este c , mirando la Especificidad y la Sensibilidad, los modelos son más precisos para clasificar las señales EEG *pre-ictales* como tales (tasa de verdaderos positivos) que para clasificar los *normales* correctamente.

en un modelo de clasificación indica robustez porque pequeños incrementos en el valor de la entropía no afectan a la clasificación. Esto es importante porque todas las entropías están normalizadas, es decir, en la misma escala.

La Figura 2.4 muestra los cinco modelos presentados en el Cuadro 2.1. La clase real de las observaciones se trazan como círculos negros. La curva de cada gráfica representa la probabilidad de que la señal sea una señal EEG *pre-ictal* según el modelo, en función del valor de la entropía. Cuando esta probabilidad es mayor que $c = 0, 5$, esta observación se clasifica como *pre-ictal* (cruces azules) y cuando es menor que $c = 0, 5$ como *normal* (cruces rojas). Los modelos con el coeficiente β más alto (en valor absoluto) tienen una pendiente más pronunciada en la curva en forma de S que conduce a una clasificación más sensible a los cambios. Por todo lo expresado anteriormente, la MinEntropía de Permutación es el modelo más robusto, seguido por el modelo que usa la Entropía de Permutación Ponderada con indicadores similares de efectividad en la clasificación.

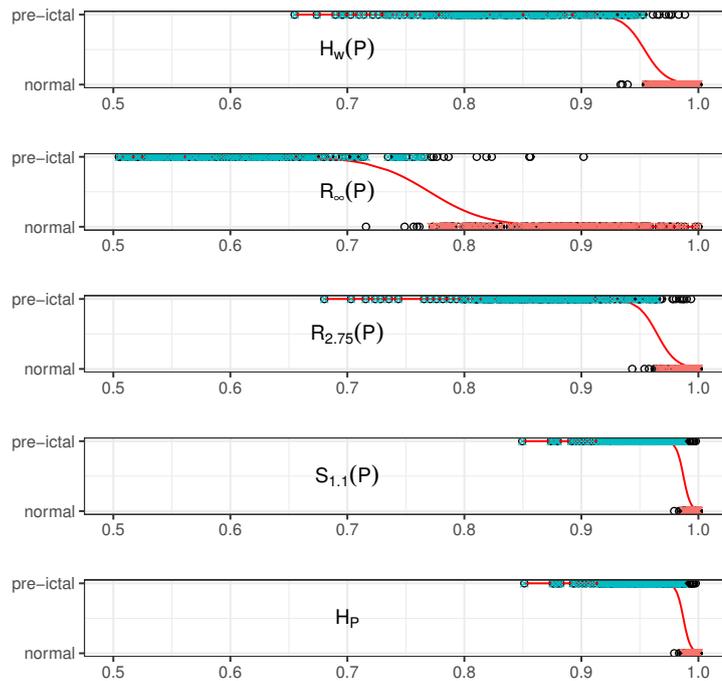


Figura 2.4 **Los cinco modelos de regresión logística.** Los modelos presentados en el Cuadro 2.1, con la variable explicativa en el eje x (las diferentes entropías) y en el eje y la probabilidad de que la señal de EEG sea *pre-ictal*. La curva en cada gráfico representa la probabilidad de que la señal de EEG provenga de un paciente en estado *pre-ictal*, según cada modelo, en función del valor de la entropía. Cuando esta probabilidad es mayor que $c = 0,5$, esta observación se clasifica como señal de EEG *pre-ictal* (cruces azules) y cuando es menor que $c = 0,5$ como *normal* (cruces rojas). La clase real de las observaciones se trazan como círculos negros. Los modelos con el coeficiente β más alto (en valor absoluto) tienen una pendiente más pronunciada en la curva en forma de S que conduce a una clasificación que es más sensible a pequeños cambios en el valor de la entropía. El modelo que usa a la MinEntropía de Permutación como clasificador es el modelo más robusto seguido por el modelo que usa la Entropía de Permutación Ponderada.

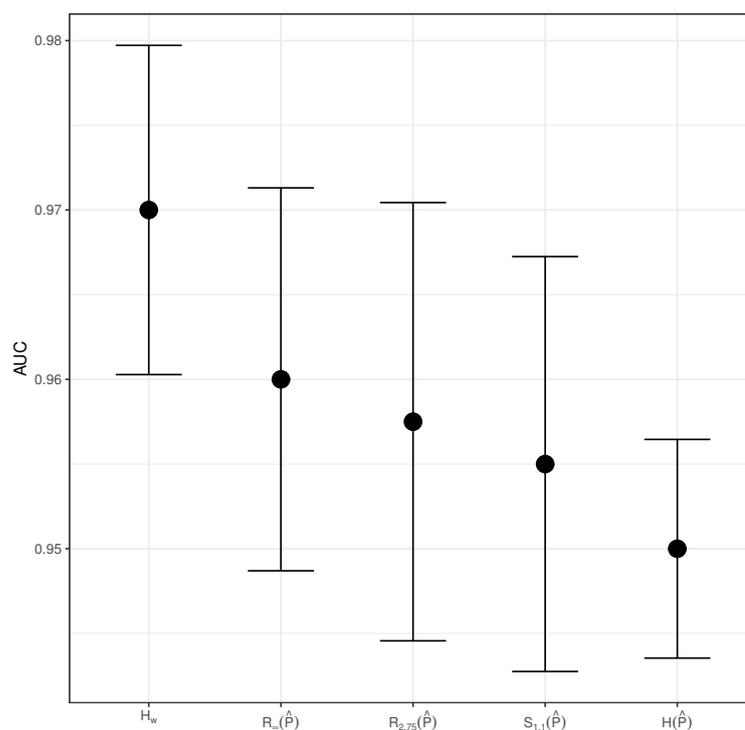


Figura 2.5 **AUC de los distintos modelos para la clasificación de los EEG.** La entropía que tiene el mejor rendimiento en términos de la AUC es la Entropía de Permutación Ponderada, seguida de la MinEntropía de Permutación por aproximadamente un desvío standard. Las restantes entropías tienen una *performance* similar.

Estos valores de c pueden ser cambiados de acuerdo a cada problemática por lo que a efectos de comparación se utiliza el AUC porque tiene en cuenta todos los valores posibles de c .

El Cuadro 2.1 presenta los mejores modelos para cada entropía ordenados según el valor de AUC en forma decreciente. Para la Entropía de Permutación de Rényi y para la Entropía de Permutación de Tsallis el parámetro se elige también según el desempeño de clasificación (teniendo en cuenta que hay un modelo para cada valor del parámetro), que es la Entropía de Permutación de Rényi con $\alpha = 2,75$ y la Entropía de Permutación de Tsallis con $q = 1,1$. La entropía que tiene el mejor rendimiento de clasificación es la Entropía de Permutación Ponderada seguida por la MinEntropía de Permutación por menos de una desviación estándar, y las entropías restantes tienen un rendimiento similar en términos del AUC (Figura 2.5). Todas las entropías tienen un desempeño excelente y similar en términos de precisión, por lo que el error de clasificación general es pequeño para el punto de corte $c = 0,5$, y para este c , mirando la Especificidad y la Sensibilidad, los modelos son más precisos para clasificar las señales EEG *pre-ictales* como tales (tasa de verdaderos positivos) que para clasificar los *normales* correctamente.

2.7.5. Conclusiones

El valor de AUC obtenido es el mayor que encontramos dentro de la literatura escrita hasta el momento, entre los manuscritos descritos en la Sección 2.7.1. Usar las entropías provenientes de la Función de Distribución de Probabilidades de Bandt y Pompe permite tener una percepción de la estructura causal de la serie de tiempo.

En resumen, en esta aplicación comparamos el potencial de clasificación de varias entropías competidoras: Entropía de Permutación \mathcal{H} , Entropía de Permutación Ponderada \mathcal{H}_w , MinEntropía de Permutación R_∞ , Entropía de Permutación de Rényi R_α , y Entropía de Permutación de Tsallis S_q . Todos estos cuantificadores son excelentes como clasificadores. Para este caso, la Entropía de Permutación Ponderada \mathcal{H}_w resulta en el mejor clasificador, por lo que tener en cuenta la amplitud

de los patrones podría mejorar el rendimiento de la clasificación. Sin embargo, como la diferencia entre los AUC no es significativa, algunas consideraciones deben ser tenidas en cuenta para seleccionar una u otra entropía, como el ruido o artefactos presentes en la señal. La MinEntropía de Permutación es más robusta contra el ruido que la Entropía de Permutación (Zunino et al. [2015]). Esto puede ser muy útil en presencia de ruidos en las señales de EEG. La Entropía de Permutación Ponderada retiene más información de la señal que sus contrapartes y podría ser un buen candidato para la clasificación automática de señales de EEG. La excelente *performance* de las Entropía de Permutación de Tsallis y de Rényi como clasificadoras indican una posible línea de investigación acerca de las dinámicas subyacentes en estas señales.

Capítulo 3

Otras medidas de complejidad derivadas de la Función de Distribución de Probabilidades de Bandt y Pompe

3.1. Introducción

En este Capítulo se presentan dos medidas de complejidad, provenientes de la Teoría de la Información, asociadas a la Función de Distribución de Bandt y Pompe que son complementarias a la Entropía de Permutación \mathcal{H} y permiten una mejor caracterización del sistema dinámico bajo estudio: la Medida de Complejidad Estadística \mathcal{C} y la Medida de Información de Fisher \mathcal{F} . Estas tres medidas de complejidad (\mathcal{H} , \mathcal{C} y \mathcal{F}) son utilizadas en una aplicación de la Ingeniería Mecánica para caracterizar el estado de trabajo de máquinas rotativas bajo distintos estados de funcionamiento.

3.2. Medida de Complejidad Estadística

Como se argumentó en el Capítulo anterior, la entropía puede ser vista como una medida de incertidumbre (y como contrapartida de información) asociada al proceso físico descrito por \mathbf{P} , donde $\mathbf{P} = \{p_j, j = 1, \dots, N\}$ es una distribución de probabilidades de una variable aleatoria asociada a los posibles N estados que puede tomar el sistema bajo estudio.

Por otro lado, siguiendo el trabajo de Lopez-Ruiz (Lopez-Ruiz et al. [1995]), un sistema se dice complejo cuando no sigue ciertos patrones considerados como simples. Si bien esta definición parece una tautología, permite empezar a entender el término si se examinan algunos modelos o idealizaciones teóricas.

Tomemos un cristal perfecto: está perfectamente *ordenado* y la distribución de probabilidades \mathbf{P} están acumuladas en un sólo estado que representa la simetría perfecta ($\mathbf{P} = \{p_i = 1 \text{ si } i \text{ es el estado de simetría } p_j = 0 \forall j \neq i\}$). Se puede considerar a este sistema como un sistema con complejidad *cero* y también que la *información* contenida en este sistema es mínima, es decir que muy poca *información* es necesaria para describir completamente este sistema.

Ahora tomemos un gas ideal aislado: está completamente *desordenado* y el sistema se puede encontrar en cualquiera de sus posibles estados con la misma probabilidad ($\mathbf{P} = \{p_j = 1/N, j = 1, \dots, N\}$), pero también a este sistema se lo puede considerar de complejidad *cero* aunque la *información* necesaria para describir el sistema es máxima, porque todos los estados contribuyen en igual medida a la *información* contenida en dicho sistema.

De estos sistemas no complejos idealizados, que se ubican en extremos opuestos de *orden* e *información*, se puede concluir que una definición de medida de complejidad no se puede hacer sólo a partir de estos términos, que en lo que respecta a esta Tesis están cuantificados a través de la entropía del sistema bajo estudio. En este trabajo (Lopez-Ruiz et al. [1995]) proponen una medida de complejidad como una interacción entre la distancia a la distribución equiprobable (o de equilibrio) de los estados accesibles del sistema y la entropía, y es la que se tomó para definir a la Medida de Complejidad Estadística asociada al proceso

físico descrito por \mathbf{P} (con \mathbf{P} la distribución de probabilidades de Bandt y Pompe) que se describe a continuación (Rosso et al. [2010a]).

Como primer paso se define una medida de desequilibrio como una distancia \mathcal{D} entre una distribución de probabilidad \mathbf{P} y la distribución equiprobable \mathbf{P}_e :

$$\mathcal{Q}(\mathbf{P}, \mathbf{P}_e) = Q_0 \cdot \mathcal{D}(\mathbf{P}, \mathbf{P}_e) \quad (3.1)$$

donde Q_0 es una constante de normalización ($0 \leq Q \leq 1$), cuyo valor es igual a la inversa del valor máximo posible para la distancia \mathcal{D} . Este desequilibrio \mathcal{Q} da indicios de la estructura de este sistema, siendo diferente de cero si existen estados más privilegiados dentro de los estados accesibles. Y finalmente se define a la Medida de Complejidad Estadística como la interacción entre este desequilibrio \mathcal{Q} y la cantidad de estados disponibles reflejada en la cantidad de información cuantificada por la entropía del sistema \mathcal{H} , mediante el funcional propuesto por Lopez-Ruiz:

$$\mathcal{C}(\mathbf{P}) = \mathcal{H}(\mathbf{P}) \cdot \mathcal{Q}(\mathbf{P}) \quad (3.2)$$

Queda entonces definir la distancia \mathcal{D} a la distribución de equilibrio que se va a utilizar en la ecuación 3.2 y elegir la entropía \mathcal{H} para el funcional descrito en la Ecuación 3.2.

La elección primaria de la distancia fue la distancia euclidiana, o de norma 2 (Lopez-Ruiz et al. [1995]). En un trabajo posterior se propone la entropía relativa de Kullback-Leiber (Kullback [1997]) y finalmente se recomienda adoptar como distancia para definir al equilibrio la divergencia de Jensen-Shannon, la cual es la entropía relativa de Kullback-Leiber simetrizada (Rosso et al. [2006]):

$$\mathcal{J}(\mathbf{P}, \mathbf{P}) = S((\mathbf{P} + \mathbf{P}_e)/2) - S(\mathbf{P})/2 - S(\mathbf{P}_e)/2 \quad (3.3)$$

quedando la Ecuación 3.2:

$$\mathcal{Q}(\mathbf{P}, \mathbf{P}) = Q_0 \cdot \mathcal{J}(\mathbf{P}, \mathbf{P}) \quad (3.4)$$

La entropía elegida para definir la Medida de Complejidad Estadística puede ser cualquiera de las formas funcionales descritas en el Capítulo 2 (i.e Shannon, Rényi, Minentropía o Tsallis). Si por ejemplo se utiliza la Entropía Informacional de Shannon, queda así definida la Medida de Complejidad Estadística de Shannon.

Finalmente, cuando se usa en \mathbf{P} la Función de Distribución de Probabilidades de Bandt Pompe para describir el proceso, y \mathcal{H} como la Entropía de Permutación, queda definida la *Complejidad Estadística de Permutación de Shannon -C-* que es la que se va a usar de acá en adelante en este trabajo y se la llamará simplemente Medida de Complejidad Estadística.

3.3. Medida de Información de Fisher

El número de información de Fisher es un concepto de la estadística inferencial presentado hace casi 100 años por uno de los fundadores de la misma, R.A. Fisher (RA Fisher [1922]).

En este artículo, Fisher expone que uno de los principales problemas a resolver de la estadística es reducir una gran cantidad de datos a unas relativamente pocas cantidades que contengan la información relevante del todo. Propone construir una población hipotética e infinita, en la cual los datos obtenidos provienen de una muestra de esta población. La ley de distribución de esta hipotética población está definida por relativamente pocos parámetros que son suficientes para describirla exhaustivamente en lo que respecta a la cualidades en las que se está interesado estudiar.

Supongamos que X es una variable aleatoria cuya distribución pertenece a la familia de distribuciones discreta o continua con densidad $p(x, \theta)$, con $\theta \in \Theta$, donde Θ es un conjunto abierto de \mathbb{R} . Sea $\mathbf{x} = \{x_1, \dots, x_n\}$ la muestra, o el conjunto de datos independientes observados provenientes de la distribución $p(x, \theta)$, entonces $\hat{\theta}$,

es decir el estimador del parámetro que caracteriza a la distribución, debiera ser el que maximiza la verosimilitud de la muestra obtenida.

Para encontrar este máximo en $p(\mathbf{x}, \theta)$ basta encontrar el máximo en el $\ln p(\mathbf{x}, \theta)$ que convierte a las multiplicaciones de $p(x_i, \theta)$ provenientes de la independencia entre las muestras en sumas más simples de manejar.

Por lo tanto para maximizar la función $\ln p(\mathbf{x}, \theta)$ simplemente se la deriva y se la iguala a cero.

$$\frac{\partial \ln p(\mathbf{x}, \hat{\theta})}{\partial \theta} = 0 \quad (3.5)$$

Cuando se trabaja con un vector de parámetros $\boldsymbol{\theta}$ a cada $\frac{\partial \ln p(\mathbf{x}, \theta_i)}{\partial \theta_i}$ se lo denomina *score* y bajo condiciones muy generales se puede probar que su esperanza es cero:

$$E_{\theta_i} \left(\frac{\partial \ln p(\mathbf{x}, \theta_i)}{\partial \theta_i} \right) = 0 \quad (3.6)$$

Las varianzas de estos *scores* pueden ser vistas como la calidad del proceso de estimación del parámetro (RA Fisher [1922]). Por lo tanto se define al número de información de Fisher como:

$$I(\theta_i) = V_{\theta_i} \left(\frac{\partial \ln p(\mathbf{x}, \theta_i)}{\partial \theta_i} \right) \quad (3.7)$$

Luego el teorema de Rao-Cramer explicitó una cota inferior para la varianza del estimador,

$$V_{\theta_i}(\hat{\theta}_i) \geq \frac{1}{I(\theta_i)} \quad (3.8)$$

Si $\hat{\theta}_i$ es un estimador de θ_i , su varianza debe ser mayor o igual que $\frac{1}{I(\theta_i)}$. Luego se puede esperar que cuanto mayor sea $I(\theta_i)$ (como $\frac{1}{I(\theta_i)}$ será menor) existe la

posibilidad de encontrar estimadores con menor varianza y por lo tanto más precisos. De ahí el nombre de número de información que se le da a $I(\theta_i)$. Es decir cuanto mayor es $I(\theta_i)$, mejores estimadores de θ se pueden encontrar, y por lo tanto se puede decir que X brinda más información sobre θ (Yohai and Boente [2004]).

Frieden estudió esta medida para la utilización en aplicaciones físicas (Frieden [2004]) y se define a la medida de información de Fisher asociada a una distribución de probabilidades \mathbf{P} análogamente a la Ecuación 3.3

$$\mathcal{F}(\mathbf{P}) = \int \left(\frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta} \right)^2 p(\mathbf{x}, \theta) dx \quad (3.9)$$

Notar que la Ecuación 3.3 es la definición de la varianza de $\left(\frac{\partial \ln p(\mathbf{X}, \theta)}{\partial \theta} \right)$, dado que $E_\theta \left(\frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta} \right) = 0$.

La Medida de Información de Fisher como cuantificador de un proceso físico descrito por \mathbf{P} puede ser vista como una medida del estado de desorden de dicho fenómeno. Es una medida de *información local*, siendo sensible a pequeñas perturbaciones de la distribución de probabilidades (Olivares et al. [2012a]), a diferencia de la entropía que es una medida global de \mathbf{P} .

Tomando como punto de partida la ecuación 3.3 y aplicando el método de diferencias finitas se obtiene la información de Fisher discreta (Olivares et al. [2012b]), es decir la expresión de la información de Fisher para el caso de una función de distribución de probabilidades discreta \mathbf{P} , donde $\mathbf{P} = \{p_i, i = 1, \dots, N\}$:

$$\mathcal{F}(\mathbf{P}) = F_0 \sum_{i=1}^{N-1} [(p_{i+1})^{1/2} - (p_i)^{1/2}]^2. \quad (3.10)$$

donde F_0 es una constante de normalización,

$$F_0 = \begin{cases} 1 & \text{si } p_{i^*} = 1 \text{ para } i^* = 1 \text{ o } i^* = N \text{ y } p_i = 0 \forall i \neq i^* \\ 1/2 & \text{caso contrario} \end{cases}. \quad (3.11)$$

Nuevamente, cuando se usa en \mathbf{P} la FDP de BP para describir el proceso, queda definida la Medida de Información de Fisher \mathcal{F} que es la que se va a usar de acá en adelante en este trabajo. Este cuantificador tiene la desventaja de depender del orden de los símbolos π_i en el alfabeto S_m , que no tienen un orden preestablecido y puede dar distintos resultados según se haga la simbolización propuesta por Bandt y Pompe con el mapeo según rangos o el mapeo según el orden cronológico.

3.4. Aplicación: Evaluación del estado de máquinas rotativas mediante las distintas medidas de complejidad

3.4.1. Introducción

Normalmente, el estado de las máquinas rotativas no puede evaluarse directamente. En su lugar, las señales vibratorias generadas a partir de máquinas rotativas han sido a menudo medidas y luego analizadas para evaluar el estado de trabajo de la máquina (Yan et al. [2012]). El procesamiento de la señal para su estudio en una etapa temprana puede prevenir fallos inesperados evitando posibles percances como costo de la vida humana, interrupción de la producción y otras pérdidas financieras, y pudiendo permitir un aumento de hasta 30 por ciento en la productividad de la operación (Henríquez et al. [2014]). Tanto la fricción, los golpes, el juego entre piezas móviles, como las fracturas en el banco de soporte o los sujetadores rotos pueden ocurrir durante la vida útil de la máquina (Sheng et al. [2006]). Estas son todas fuentes de vibraciones no estacionarias y no lineales y por lo tanto, el estudio tradicional de las vibraciones a través de métodos lineales puede no ser eficaz para detectar cambios dinámicos en la señal.

El uso de medidas de complejidad basadas en la Teoría de la Información ha dado lugar a resultados interesantes sobre las características de la dinámicas no lineales, mejorando la comprensión de sus series de tiempo. En particular, la combinación de la Medida de Complejidad Estadística y la Entropía de Permutación

permite una buena distinción entre dinámica estocástica y caótica y no requieren la condición de estacionariedad (Rosso et al. [2012a,b, 2007b, 2013]).

3.4.2. Objetivo

En particular, lo que interesa estudiar es el comportamiento de la transición entre una dinámica proveniente de una máquina rotativa balanceada a una dinámica desbalanceada para poder detectar posibles fallas. Para caracterizar la transición de una máquina rotativa entre un estado balanceado y uno desbalanceado se propone el uso de las medidas de complejidad estudiadas en esta Tesis:

- Entropía de Permutación, \mathcal{H}
- Medida de Complejidad Estadística, \mathcal{C}
- Medida de Información de Fisher, \mathcal{F}

3.4.3. Preparación del experimento

Una máquina rotativa está formada por una serie de componentes que interactúan entre sí cuando la máquina está operativa. La señal vibratoria obtenida en este estado es compleja debido a que en la medición hay fuertes ruidos de interferencia. También, la vibración resultante se vuelve más compleja cuando hay fallas en un componente (Yan et al. [2012]). Se han realizado una serie de experimentos, en el Laboratorio de Materiales (CEMAT), en el Instituto Tecnológico de Buenos Aires (ITBA) para obtener dichas señales vibratorias. Un banco específicamente diseñado (ver Figura 3.1) para simular diferentes condiciones de falla en una máquina rotativa se utilizó para las mediciones.

Un motor eléctrico marcado como (4) en la Figura 3.1 está conectado a un eje a través de una junta universal. El eje está montado en dos cojinetes (5) y (9) y se le han acoplado dos discos de aluminio perforados (7) y (8) en los que pueden ser aplicadas cargas excéntricas desbalanceadas. Las fijaciones de los cojinetes están atornillados a una base que puede estar desalineada. El VFD (2) permite regular la velocidad del motor. El dispositivo de medición (1) o VDL es un DSP Logger

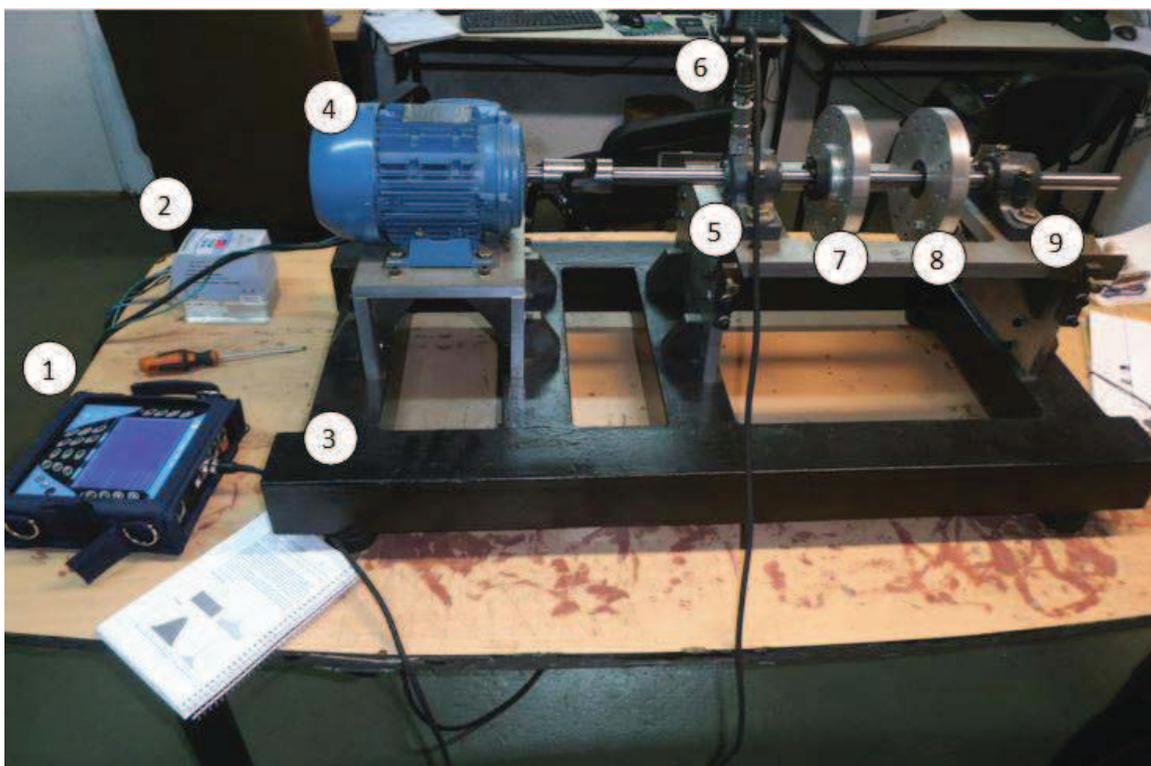


Figura 3.1 **Banco de Vibración.** Componentes Principales: (1) VDL (Vibration Data Logger). (2) VFD (Vibration Frequency Drive) (3) Banco. (4) Motor eléctrico. (5) Cojinete 1. (6) Acelerómetro. (7) Disco Balanceado 1. (8) Disco Balanceado 2. (9) Cojinete 2

MX 300 con un acelerómetro (6) con 10 mV/g de sensibilidad. Está configurado para medir el desplazamiento, la velocidad y la aceleración y está acoplado a un soporte de cojinete mediante un bulón sin cabeza. El motor se puso en marcha y los valores de las muestras se tomaron a 20 Hz .

Las mediciones de aceleración (series temporales) se realizaron primero en condiciones de vacío con el eje desalineado y finalmente se cambió uno de los cojinetes con otro que estaba erosionado en el anillo exterior.

Las muestras de las señales de aceleración provenientes de la máquina fueron extraídas bajo diferentes condiciones:

- **Funcionamiento en vacío:** La máquina está funcionando en condiciones normales. El eje está balanceado. Esta condición está etiquetada como Tipo 0. Una señal típica en estas condiciones se muestra en la Figura 3.2. La dinámica en esta condición es bastante regular con variaciones prácticamente periódicas.
- **Eje desbalanceado:** el eje del motor lleva dos discos de aluminio perforados en los que pueden ser aplicadas cargas excéntricas (se le agregan tuercas y/o tornillos) para separar el centro de gravedad del sistema del eje de rotación. Como consecuencia de esto se producen vibraciones en la máquina. Una señal típica en estas condiciones se muestra en la Figura 3.3.

Se proponen los siguientes estados desbalanceados:

- **Tipo 1:** dos masas adicionales, localizadas en las perforaciones externas de modo coaxial.
- **Tipo 2:** dos masas adicionales, localizadas en las perforaciones externas, con un desplazamiento de fase 90° .
- **Tipo 3:** una única masa localizada en la perforación externa de un solo disco.
- **Tipo 4:** una única masa localizada en la perforación interna de un solo disco.

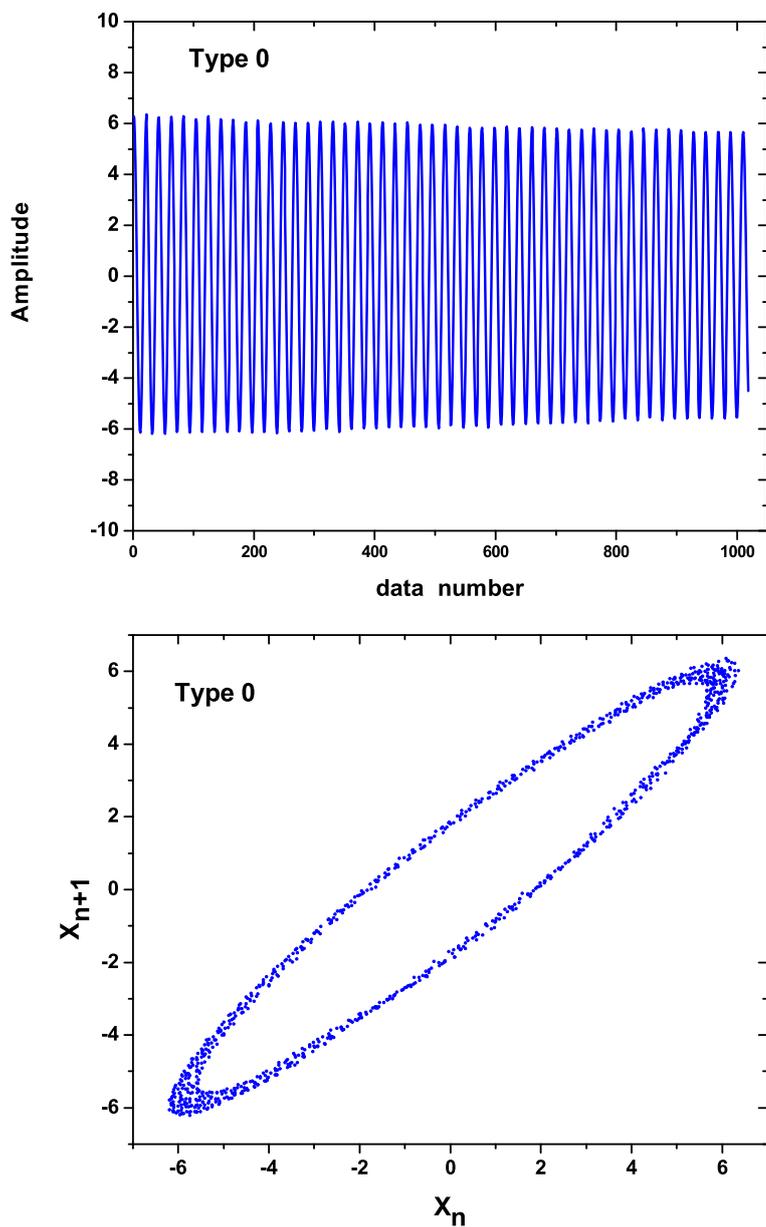


Figura 3.2 Serie de tiempo correspondiente a una máquina rotativa balanceada. A esta condición se la denota como de Tipo 0.

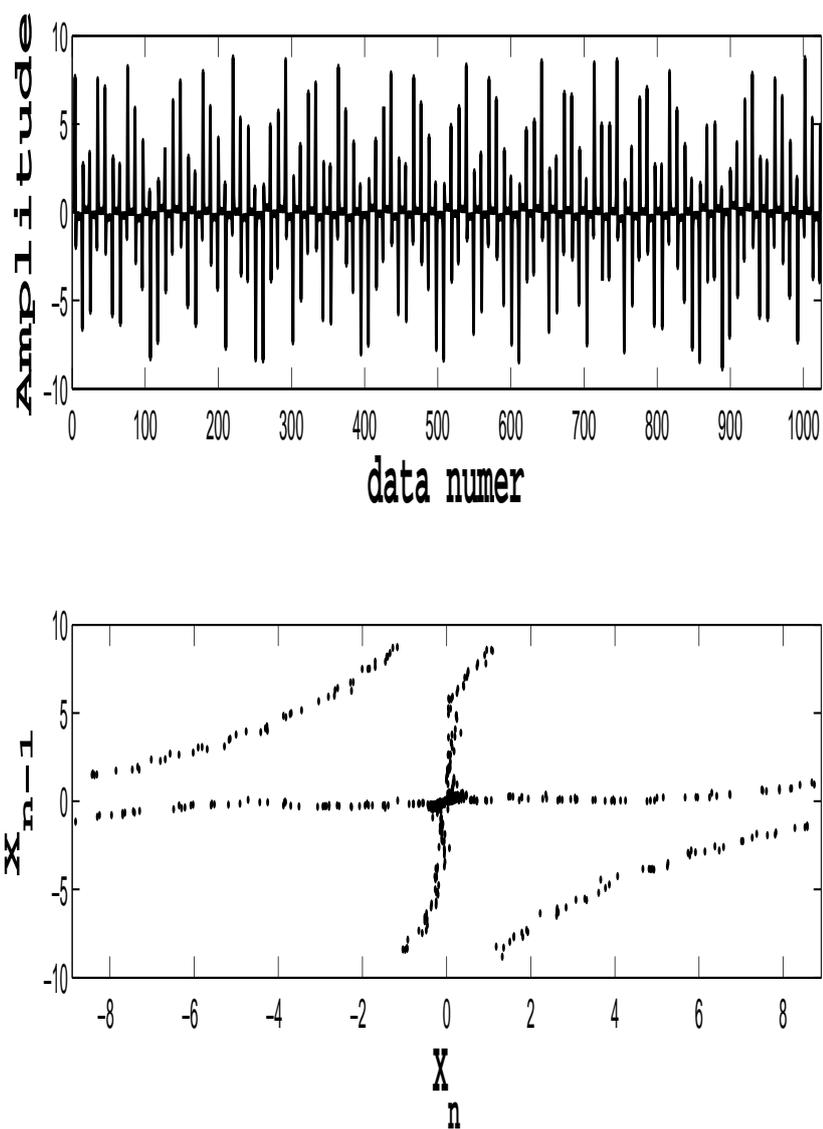


Figura 3.3 Serie de tiempo correspondiente a una máquina rotativa desbalanceada. A esta condición se la denota como de Tipo 1.

- Tipo 5: como el Tipo 3 pero con la masa localizada en la perforación con un desplazamiento de fase 90° .
- Tipo 6: como el Tipo 4 pero con la masa localizada en la perforación con un desplazamiento de fase 90° .

Todos los ensayos fueron realizados a una frecuencia de 20 Hz y el acelerómetro fue siempre colocado en el primer cojinete, el más cercano al motor. Para cada estado, series de tiempo de 1024 puntos fueron tomadas. Seis señales diferentes (series de tiempo) fueron generadas conectando una señal de Tipo 0 con las distintas señales del Tipo 1-6, para obtener distintas fallas del tipo abrupto. En la Figura 3.4 se presentan las seis series de tiempo a ser analizadas. En todas las Figuras, la línea vertical punteada representa el instante que falla (punto 1024) a partir del cual la dinámica de la máquina rotativa cambia de balanceada a desbalanceada.

3.4.4. Resultados

Las seis señales que se muestran en la Figura 3.4 fueron analizadas usando las estimaciones de las medidas de complejidad asociadas a la FDP de BP.

- La Entropía de Permutación, $\hat{\mathcal{H}} = \mathcal{H}(\hat{\mathbf{P}})$
- la Medida de Complejidad Estadística, $\hat{\mathcal{C}} = \mathcal{C}(\hat{\mathbf{P}})$
- la Medida de Información de Fisher, $\hat{\mathcal{F}} = \mathcal{F}(\hat{\mathbf{P}})$

Las distintas medidas fueron evaluadas en ventanas de longitud $N = 512$ datos, desplazándolas de a un punto con un solapamiento $\delta = 1$ dato. Para la simbolización al alfabeto de Bandt y Pompe se tomó una dimensión de *embedding* de $m = 4$ y retardo de tiempo $\tau = 1$ (ver Capítulo 2).

La evolución en el tiempo de las tres medidas de complejidad se muestran en las Figuras 3.5 a 3.7. Cada punto representa el valor de la medida en la correspondiente ventana de tiempo. En estas tres Figuras se pueden ver tres zonas distinguibles de comportamiento. La primera puede ser asociada al comportamiento balanceado (señal pura de Tipo 0), que corresponde a las ventanas de tiempo 1 a 512, la

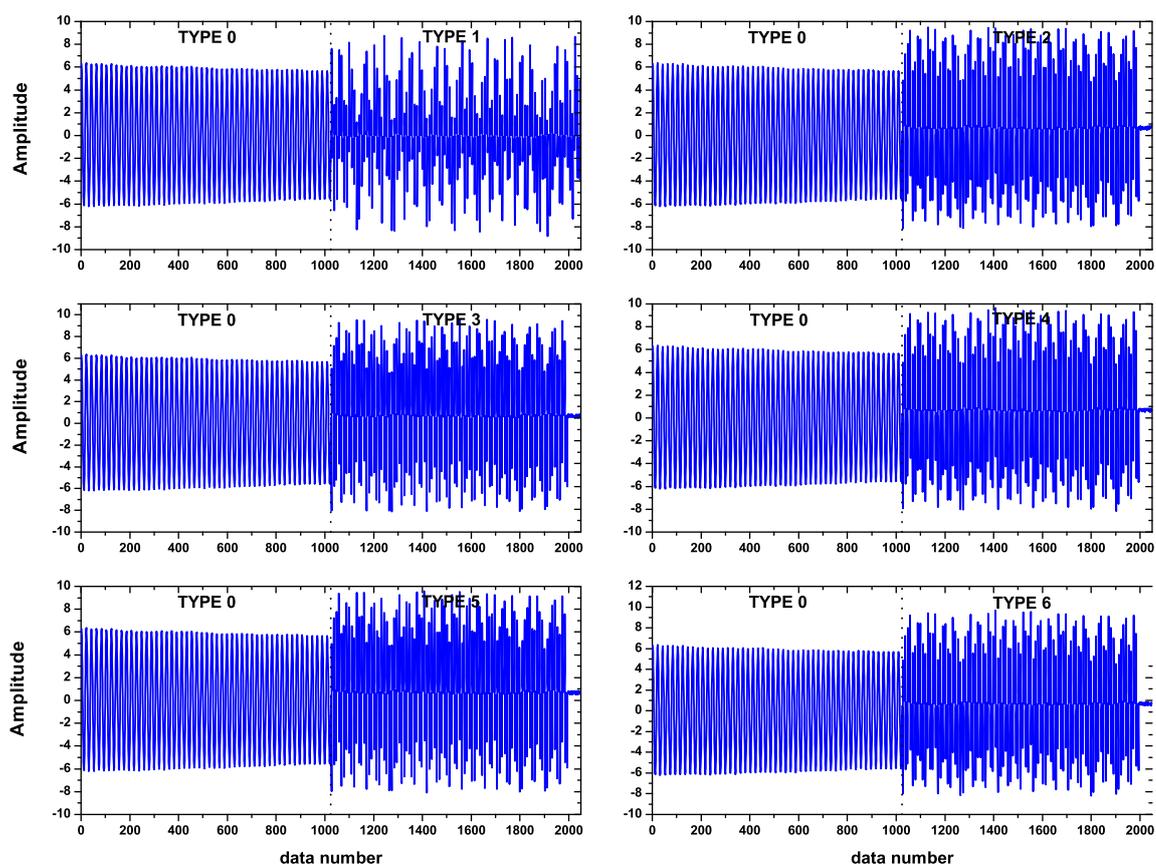


Figura 3.4 Series de tiempo correspondientes a la máquina rotativa con distintos tipos de fallas abruptas. La línea vertical punteada representa el instante que falla (punto 1024) a partir del cual la dinámica de la máquina rotativa cambia de balanceada a desbalanceada.

segunda es una zona de transición entre las ventanas 513 a 1024, y una tercera zona asociada al comportamiento desbalanceado (señales pura del Tipo 1 al Tipo 6) entre las ventanas 1025 a 1537.

La Entropía de Permutación puede detectar efectivamente cambios en la dinámica de una máquina rotativa en concordancia con estudios previamente realizados (Yan et al. [2012]). Lo mismo es válido para las otras medidas de complejidad basadas en la FDP de BP, la Medida de Complejidad Estadística y la Medida de Información de Fisher. Todas las señales analizadas presentaron un comportamiento pseudo-periódico o periódico ruidoso. En particular este comportamiento es más evidente en el caso de la señal balanceada (señal Tipo 0, ver Figura 3.2) que es caracterizada por una entropía baja, una complejidad media y una medida de información también media. Este comportamiento es compatible con el grado medio de orden en el sistema reflejado en la señal al mirar la Figura 3.2.

En contraste, un comportamiento funcional desbalanceado (señales del Tipo 1 a 6) es caracterizado por un alto valor de $\hat{\mathcal{H}}$, un bajo valor de $\hat{\mathcal{C}}$ y también un bajo valor de $\hat{\mathcal{F}}$. Es decir que un comportamiento ruidoso y un sistema desordenado pueden ser asociados a comportamientos desbalanceados.

El Cuadro 3.1 presenta las medias y las desviaciones estándar para los valores de las tres medidas de complejidad. Para la señal de Tipo 0 se tomaron los valores correspondientes a las ventanas de tiempo 1 a 512 y para las señales de Tipo 1 a 6 se tomaron los valores de las ventanas de tiempo 1025 a 1537.

Se puede notar que el comportamiento de estas tres medidas para cada tipo de señal es bastante estable, como se puede ver en las Figuras 3.5 a 3.7 y también en el Cuadro 3.1.

En estas Figuras (3.5 a 3.7) se puede observar que la zona de transición (entre las ventanas 513 a 1024) está caracterizada por un rápido incremento del valor de $\hat{\mathcal{H}}$ y una rápida disminución del valor de $\hat{\mathcal{C}}$ y $\hat{\mathcal{F}}$, comportamiento compatible con el incremento del grado de desorden inducido por el desbalanceo.

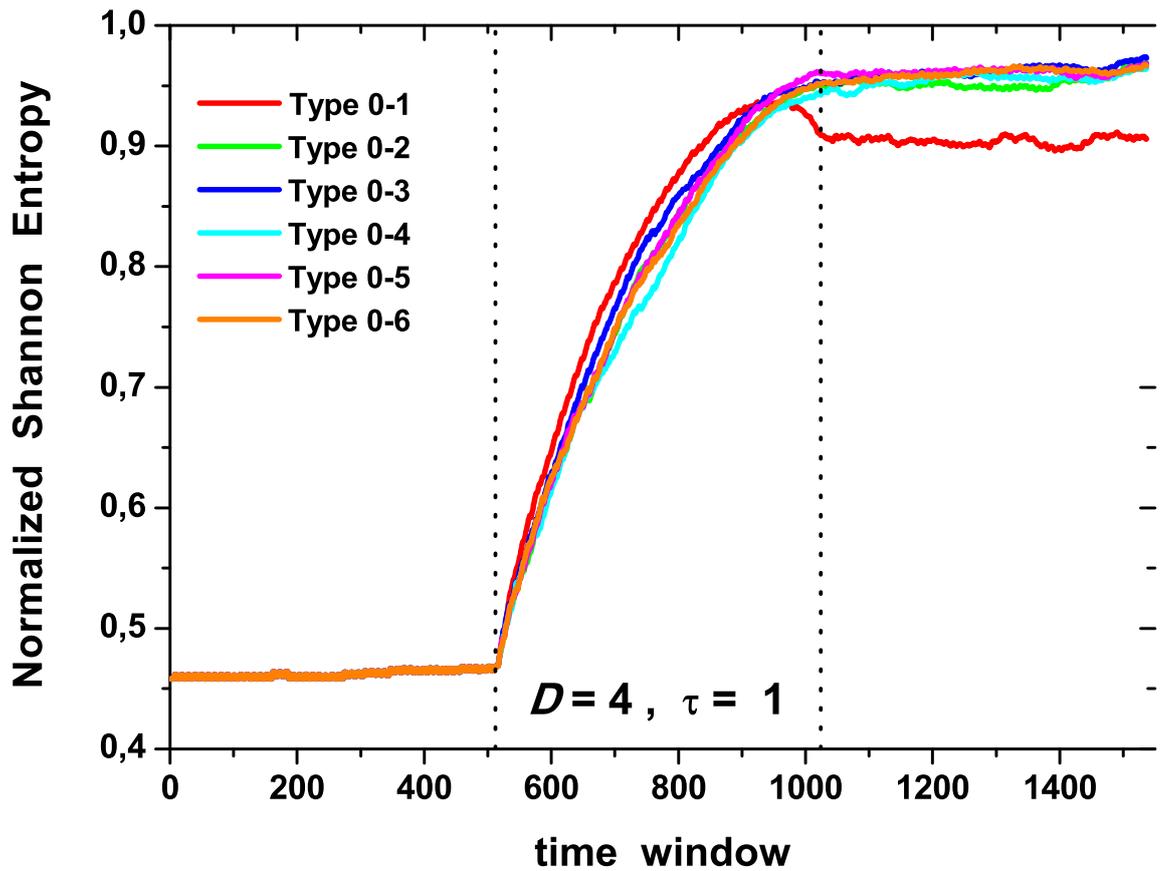


Figura 3.5 Evolución en el tiempo de la Entropía de Permutación \hat{H} , en ventanas de tiempo solapadas de longitud $N = 512$ observaciones y solapamiento $\delta = 1$ observación, para las seis señales presentadas en la Figura 3.4. La FDP de BP fue evaluada para una dimensión de *embedding* $m = 4$ y retardo de tiempo $\tau = 1$. Las líneas verticales representan los cambios de estado de los tres comportamientos diferentes: Tipo 0 puro, transición y Tipo 1-6 respectivamente.

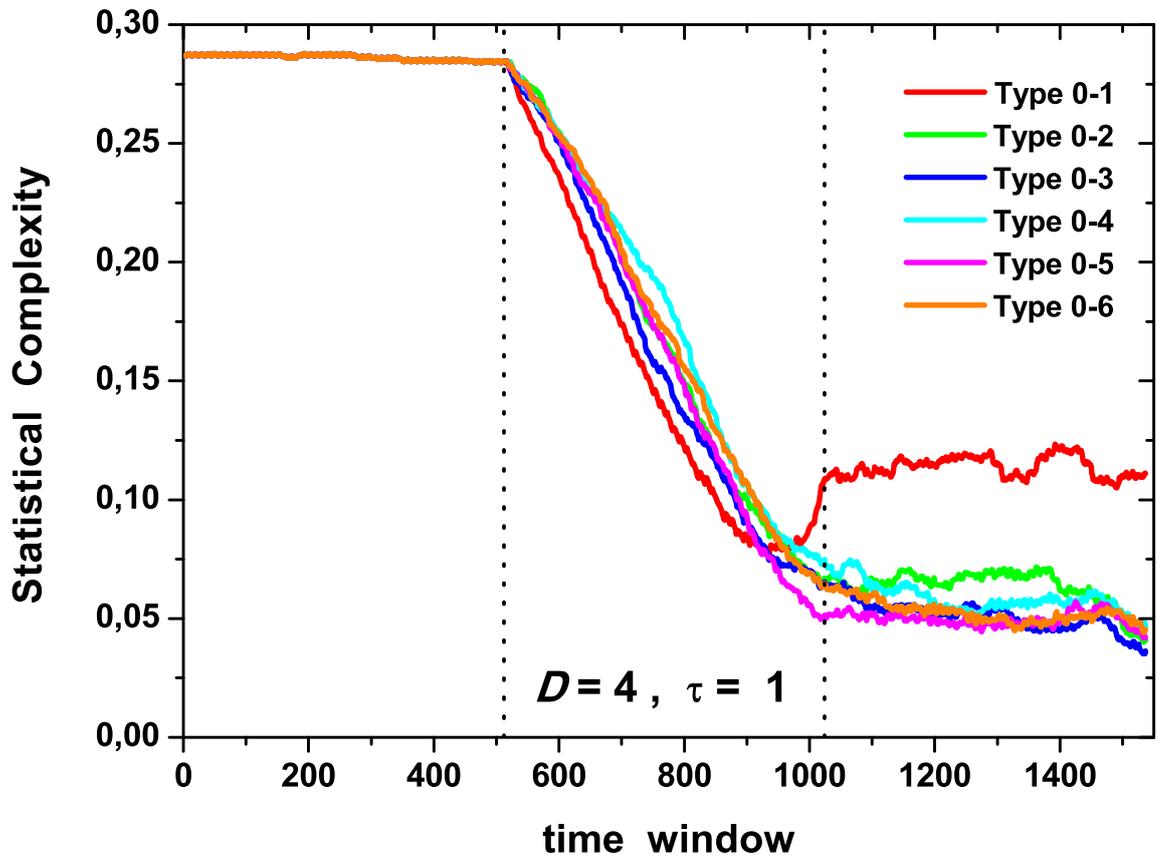


Figura 3.6 Evolución de la Medida de la Complejidad Estadística para los seis tipos de fallas abruptas en la máquina rotativa. Idem a Figura 3.5 para la Complejidad Estadística, \hat{C} .

	\hat{H}	SE(\hat{H})	\hat{C}	SE(\hat{C})	\hat{F}	SE(\hat{F})
Tipo 0	0.46212	0.00307	0.28622	0.00116	0.37958	0.00941
Tipo 1	0.90437	0.00341	0.11385	0.00431	0.16037	0.00753
Tipo 2	0.95265	0.00521	0.06359	0.00709	0.09828	0.01232
Tipo 3	0.96190	0.00464	0.05119	0.00617	0.08094	0.01133
Tipo 4	0.95495	0.00501	0.05880	0.00607	0.07898	0.00720
Tipo 5	0.96145	0.00225	0.04979	0.00298	0.06860	0.00420
Tipo 6	0.96002	0.00388	0.05278	0.00452	0.07883	0.00541

Cuadro 3.1 **Media y error estándar para las distintas medidas de complejidad.** Media y error estándar para: Entropía de Permutación, \hat{H} ; Complejidad Estadística, \hat{C} ; Medida de Información de Fisher, \hat{F} para las distintas zonas de comportamiento de la máquina rotativa, con la FDP de BP evaluada para $m = 4$ y $\tau = 1$, de las ventanas 1 a 512 para la señal de Tipo 0 y de las ventanas 1025 a 1537 para las señales de Tipo 1 a 6.

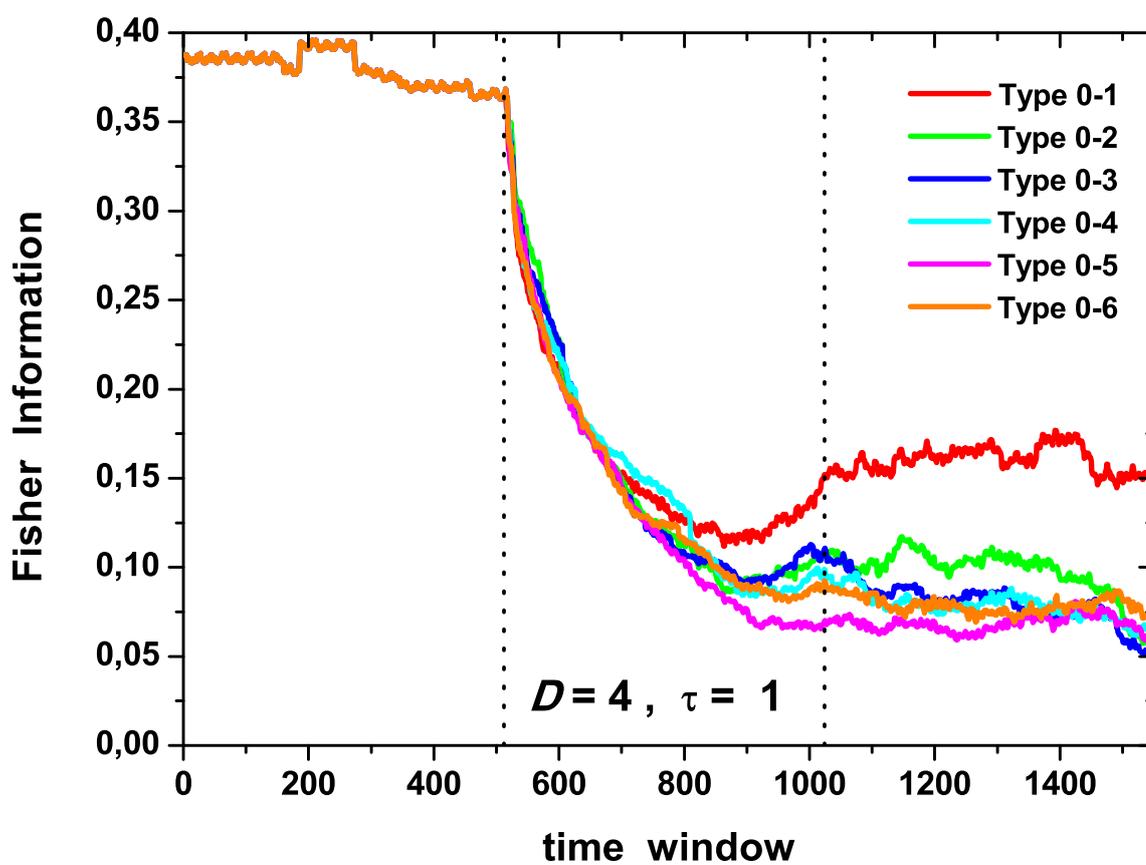


Figura 3.7 Evolución de la Medida de Información de Fisher para los seis tipos de fallas abruptas en la máquina rotativa. Idem a Figura . 3.5 para la Medida de Información de Fisher, $\hat{\mathcal{F}}$.

Teniendo en cuenta estos hechos, se puede proponer un método simple y rápido para la detección del cambio de comportamiento de un motor en funcionamiento, mediante el monitoreo de estas medidas de complejidad. Es interesante notar que la inclusión de la Medida de Información de Fisher para la caracterización de las señales vibratorias de las máquinas rotativas puede ayudar a estudiar cambios locales en la FDP extraída de la señal y puede ser un factor importante en la detección de la razón de la falla, discriminando cambios locales de cambios globales en dicha FDP extraída de la señal.

Capítulo 4

Influencia de la distribución marginal de los datos en la Entropía de Permutación

4.1. Introducción

Como se menciona en los Capítulos 2 y 3, el uso de cuantificadores basados en la Teoría de la Información ha llevado a resultados interesantes con respecto a la caracterización de dinámicas no lineales, mejorando la comprensión de sus series temporales asociadas. En este Capítulo se introducirá la representación de la *Entropía de Permutación* \mathcal{H} - (Capítulo 2) y la *Medida de Complejidad Estadística* \mathcal{C} - (Capítulo 3) en un plano informacional $\mathcal{H} \times \mathcal{C}$ que es muy útil para distinguir entre dinámicas caóticas y estocásticas.

Una vez hecho este análisis preliminar de la dinámica, uno puede empezar a preguntarse acerca de como se distribuyen marginalmente los datos y si estos cuantificadores nos dan alguna información acerca de dicha distribución. Varios trabajos han sido presentados tratando de responder dicha pregunta: se estudió el comportamiento de \mathcal{H} y \mathcal{C} para el ruido no Gaussiano del tipo $\frac{1}{k}$ y el Movimiento Browniano Fraccionario (Rosso et al. [2010a, 2007b]) y se estudiaron curvas teóricas de \mathcal{H} para el Movimiento Browniano Fraccionario y ruido Fraccionario Gaussiano

(Zunino et al. [2008a]). También se aplicó el plano informacional $\mathcal{H} \times \mathcal{C}$ en procesos estocásticos no Gaussianos en general (Rosso et al. [2007b]), y \mathcal{H} se aplicó en series de tiempo Gaussianas (Rosso et al. [2010a]).

De esta manera surge la siguiente pregunta: ¿La metodología de Bandt y Pompe es útil para caracterizar la distribución de probabilidad marginal de un proceso estocástico?. La respuesta a esta pregunta también dará información acerca de si la Entropía de Permutación se ve afectada por la distribución marginal de los datos.

Esto se aborda en este Capítulo mediante la simulación de procesos autorregresivos Gaussianos y no Gaussianos de orden 1 y comparando la *Entropía de Permutación* \mathcal{H} con la entropía de Shannon derivada del histograma de los datos (que se denominará Entropía de Amplitud \mathcal{H}_a), mediante un novedoso plano $\mathcal{H} \times \mathcal{H}_a$, que dará tanto información acerca de la estructura de autocorrelación como de la distribución marginal de los datos. En estos procesos simulados tanto la estructura de correlación asociada como la distribución de probabilidad están bien establecidas y dicha estructura de correlación es fácilmente manipulable a través del parámetro de correlación. Aunque los procesos autorregresivos Gaussianos son bien conocidos (Box et al. [2015]), este no es el caso de los procesos autorregresivos no Gaussianos. Sin embargo, últimamente una serie de trabajos en distintos campos de la ciencia y tecnología tan diversos como generadores de números pseudoaleatorios (Lawrance [1992]), modelado de transacciones financieras espaciadas irregularmente (Engle and Russell [1998]), modelado de la volatilidad de la tasa de cambio entre monedas (Hafner [2013]), estudios del sistema nervioso (Farashi [2015]) y análisis de señales de voz (Ishizuka et al. [2005]) entre otros, incitan su estudio.

4.2. Plano Informacional Entropía - Complejidad

La Medida de Complejidad Estadística es una función no trivial de la entropía ([Rosso et al., 2010a]) debido a que es la interacción de la cantidad de información contenida en un sistema descrito por P y la distancia al equilibrio P_e . De este modo,

dado un valor de la entropía \mathcal{H}_0 la complejidad se encuentra dentro de dos cotas, una mínima C_{min} (dado \mathcal{H}_0) y una máxima C_{max} (dado \mathcal{H}_0) que en el caso de utilizar como P a la FDP de BP dependen de la dimensión de *embedding* m elegida para el cálculo. De esta manera la Medida de Complejidad Estadística provee información adicional importante relacionada a la estructura de correlaciones entre los componentes del sistema físico. La localización de sistemas dinámicos en el plano entropía-complejidad ($\mathcal{H} \times \mathcal{C}$) ha sido utilizado en el estudio y visualización de cambios en la dinámica del sistema ([Rosso et al., 2007b]).

4.3. Entropía de Shannon aplicada al histograma

Anteriormente se definió la entropía de una función de densidad de probabilidad continua como:

$$S(f) = - \int_{\Delta} f \ln(f) dx , \quad (4.1)$$

Y cuando $\mathbf{P} = \{p_i; i = 1, \dots, N\}$, con $\sum_{i=1}^N p_i = 1$, es una FDP discreta, con N cantidad de estados posibles del sistema en estudio, se definió en el Capítulo 2 la Entropía Informacional de Shannon normalizada como:

$$\mathcal{H}(\mathbf{P}) = S(\mathbf{P})/S_{max} = \left\{ - \sum_{i=1}^N p_i \ln(p_i) \right\} / S_{max} , \quad (4.2)$$

donde el denominador $S_{max} = S(\mathbf{P}_e) = \ln N$ es el obtenido mediante la distribución uniforme discreta $\mathbf{P}_e = \{p_i = 1/N, \forall i = 1, \dots, N\}$. De esta manera, $0 \leq \mathcal{H} \leq 1$.

Sea una $\{x_t\}_{t \in T}$ una posible realización de un proceso generador de datos $\{\mathcal{X}_t\}_{t \in T}$ en forma de una serie tiempo a valores reales de largo T . Como no hay una única manera de computar el set \mathbf{P} para cada proceso, una posibilidad es estimar la FDP mediante un histograma de las amplitudes de sus datos, y pensar en cada intervalo de este histograma como un posible estado del sistema bajo estudio. La primera medida es definir la cantidad de intervalos del histograma, y si bien en la literatura hay diferentes reglas posibles para su determinación, en esta Tesis se

utilizará la regla de Scott (ver el Apéndice B para el fundamento de esta decisión). Habiendo definido la cantidad de intervalos N , se puede definir una función de distribución de probabilidades $\mathbf{P}_a = \{p_i; i = 1, \dots, N\}$, siendo p_i la probabilidad de que \mathcal{X}_t pertenezca al intervalo i .

Dada una realización $\{x_t\}_{t \in T}$, la estimación natural de estos p_i es la siguiente:

$$\hat{p}_i = \frac{\sum_{l=1}^{T-m+1} \mathbb{1}(x_t \text{ pertenece al intervalo } i)}{T - m + 1}, \quad (4.3)$$

quedando $\hat{\mathbf{P}}_a = \{\hat{p}_i; i = 1, \dots, N\}$ como la estimación de \mathbf{P}_a

Si se aplica en la Ecuación 4.2 la función de distribución de probabilidades \mathbf{P}_a , queda definida la Entropía Informacional de Shannon aplicada al histograma, o *Entropía de Amplitud* \mathcal{H}_a .

Se puede ver que la ubicación temporal de los datos no afecta el cálculo de esta entropía, mientras que la Entropía de Permutación no tiene en cuenta la amplitud de los datos excepto por el valor relativo con respecto a sus vecinos.

4.4. Procesos Estocásticos Autorregresivos

Un proceso estocástico lineal general se describe de tal manera que se supone que la serie de tiempo es generada por una sumatoria lineal de perturbaciones aleatorias (Box et al. [2015]). Un modelo ampliamente utilizado es el autorregresivo de orden p ,

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \dots + \phi_p z_{t-p} + a_t \quad (4.4)$$

donde el valor actual de z_t es expresado como una sumatoria finita de los valores previos del proceso $\{z_{t-1}, z_{t-2}, \dots, z_{t-p}\}$ y una perturbación a_t , tal que para todo t tiene media 0 y varianza finita σ^2 .

Si z_t es estacionario en el sentido amplio (i.e la media μ y la varianza σ_z^2 no dependen del tiempo), la autocorrelación entre z_t y z_s sólo depende del retardo entre t y s . De esta manera la autocorrelación puede ser expresada como función del desfase entre observaciones ($\nu = t - s$):

$$R(\nu) = \frac{E(z_t - \mu)E(z_{t-\nu} - \mu)}{\sigma_z^2} \quad (4.5)$$

Por simplicidad $R(\nu) = \rho_\nu$ por ejemplo $R(1) = \rho(r_t, r_{t+1}) = \rho_1$.

Se simularán tres tipos de procesos autorregresivos: el Normal o Gaussiano, el Exponencial y el Uniforme. La diferencia entre ellos reside en la forma de la perturbación a_t , hecha de tal manera que la distribución marginal de z_t sea la deseada. Por una cuestión de simplicidad los procesos serán de orden $p = 1$ obteniendo el proceso autorregresivo de primer orden AR(1).

4.4.1. Procesos Autorregresivos Gaussianos de orden 1: AR(1)

Para el proceso Gaussiano, la Ecuación 4.4 toma la forma:

$$z_t = \phi_1 z_{t-1} + a_t \quad (4.6)$$

donde el valor actual de z_t es expresado como una sumatoria finita de los valores previos del proceso $\{z_{t-1}, z_{t-2}, \dots, z_{t-p}\}$ y una perturbación a_t , independientes e idénticamente distribuidas (i.i.d.) con una distribución marginal Normal de media 0 y varianza σ^2 . Para este proceso, $\rho_1 = \phi_1$.

4.4.2. Procesos Autorregresivos Exponenciales de orden 1: NEARA(1)

Muchas series de tiempo a valores positivos tienen una distribución marginal Exponencial (Farashi [2015]; Hafner [2013]). Cuando la variable aleatoria z_t tiene una distribución marginal Exponencial de parámetro λ , el proceso autorregresivo lineal de la Ecuación 4.4 toma la forma:

$$z_t = a_t + \begin{cases} \beta \cdot z_{t-1} & \text{c.p } \alpha \\ 0 & \text{c.p } 1 - \alpha \end{cases} \quad (4.7)$$

con

$$a_t = \begin{cases} e_t & \text{c.p } \frac{1-\beta}{1-(1-\alpha)\beta} \\ (1-\alpha) \cdot \beta \cdot e_t & \text{c.p } \frac{\alpha\beta}{1-(1-\alpha)\beta} \end{cases} \quad (4.8)$$

donde “c.p” significa con probabilidad, $\alpha > 0$ y $\beta > 0$ son parámetros tal que $\rho_1 = \alpha\beta$, siempre y cuando α y β no sean ambos iguales a 1. a_t tiene una distribución Exponencial partida, donde los $e_t \{t = 0, 1, 2..\}$ son variables aleatorias independientes e idénticamente distribuidas Exponenciales con parámetro $\lambda > 0$ de tal manera que z_t es Exponencial de parámetro λ .

Notar que si α y β son iguales a cero, los z_t son Exponenciales independientes. Como α y β son no negativas, con este método las autocorrelaciones $\rho_k = (\alpha\beta)^k$ son positivas y geoméricamente decrecientes. Esto es distinto al proceso Gaussiano AR(1) donde ρ_1 puede ser negativo. Para extender los modelos exponenciales a estas posibilidades, dos secuencias z_t y z'_t se construyen cruzadas, las cuales incluyen variables antagónicas, desarrollando el modelo NEARA(1) ([Lawrance and Lewis \[1981\]](#)).

4.4.3. Procesos Autorregresivos Uniforme de orden 1: UAR(1)

Otro modelo autorregresivo de orden 1 no Gaussiano de interés es aquel que su distribución marginal es la Uniforme. El modelo responde a:

$$z_t = \frac{1}{k} z_{t-1} + a_t \quad (4.9)$$

con $k \geq 2$. Se demostró que z_t tendría una distribución marginal uniforme $(0, 1)$ si las perturbaciones a_t son independientes con la distribución presentada en el Cuadro 4.1 donde $k \in \{2, 3, \dots\}$ y $\rho_1 = 1/k$ ([Chernick \[1981\]](#)). Este modelo se llama

UAR(1) y tiene la autocorrelacion de retardo r : $\rho_r = \rho^r = (1/k)^r$. Si ρ_1 es $-1/k$ se obtienen resultados similares para series correlacionadas negativamente

a_t	0	1/k	2/k	...	(k-1)/k
$P(a_t)$	1/k	1/k	1/k	...	1/k

Cuadro 4.1 **Distribución Uniforme discreta para las perturbaciones a_t en el modelo UAR (1)** $k \in \{2, 3, \dots\}$.

4.5. Resultados Numéricos

Los procesos presentados en la Sección previa fueron simulados variando el valor de la autocorrelación. El código utilizado para la programación de las simulaciones de los procesos autorregresivos no Gaussianos se presenta en el Apéndice C, y fue desarrollado especialmente para esta Tesis.

Para el proceso Gaussiano nueve series de tiempo autocorrelacionadas positivamente y nueve negativamente fueron simuladas con

$$\rho_1 = \pm \{0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9\}$$

y además para $\rho_1 = 0$, el bien conocido ruido blanco Gaussiano. Todos estos procesos fueron simulados de tal manera que los a_t tengan $\sigma = 1$ y $\mu = 0$.

Los procesos Exponenciales fueron simulados con α y β tal que

$$\rho_1 = \{-0,2, -0,1, 0,125, 0,25, 0,5, 0,75\}$$

más el ruido descorrelacionado con distribución exponencial.

Los procesos autorregresivos Uniformes fueron simulados para

$$\rho_1 = \pm \{1/2, 1/3, \dots, 1/9\}$$

más el ruido descorrelacionado con distribución uniforme. Todas las series tienen un largo de $T = 10^6$

A todos ruidos se les calculó:

- la Entropía Permutación, $\hat{\mathcal{H}} = \mathcal{H}(\hat{\mathbf{P}})$ (Capítulo 2)
- la Medida de Complejidad Estadística, $\hat{\mathcal{C}} = \mathcal{C}(\hat{\mathbf{P}})$ (Sección 3.2)
- la Entropía de Amplitud , $\hat{\mathcal{H}}_a = \mathcal{H}(\hat{\mathbf{P}}_a)$ (sección 4.3).

Donde $\hat{\mathbf{P}}$ es la estimación de la FDP de BP y $\hat{\mathbf{P}}_a$ es la estimación de la FDP derivada del histograma. Primero se analizó la ubicación de estos procesos en el plano $\mathcal{H} \times \mathcal{C}$ (Figura 4.1) Dicha localización planar está en concordancia con las otras previamente reportadas (Rosso et al. [2007b]). Hay que resaltar que la localización de los procesos autorregresivos Gaussianos y no Gaussianos coincide. Las series con autocorrelación cero (secuencias de números aleatorios no correlacionados) Uniformes, Exponenciales o Gaussianas, están localizadas en la misma región $(\mathcal{H}, \mathcal{C}) = (1, 0)$, con máxima entropía y mínima complejidad. A medida que aumenta la autocorrelación (en valor absoluto) aumenta $\hat{\mathcal{C}}$ y disminuye $\hat{\mathcal{H}}$. Los tres procesos son mayormente indistinguibles en el plano informacional $\mathcal{H} \times \mathcal{C}$, especialmente para valores chicos de autocorrelación. Por lo tanto la distribución marginal de probabilidad de los datos no influye en la localización del proceso en el plano informacional $\mathcal{H} \times \mathcal{C}$. Esto se debe a que la metodología para estimar los $p(\pi_i)$ de la FDP de BP tiene en cuenta la estructura de correlaciones de la serie de tiempo pero pierde la información acerca de la distribución de probabilidad marginal del proceso generador de datos.

Este hecho puede ser observado en la Figura 4.2 donde se presentan los tres histogramas para la FDP de BP de los ruidos no correlacionados ($\rho_k = 0 \quad \forall k$). Como se puede observar los tres histogramas son indistinguibles, fundando la sospecha de que \mathcal{H} no es capaz de discriminar entre distribuciones de probabilidad marginales distintas, o visto desde otro punto de vista, la distribución marginal de los datos no afecta a la estimación de la Entropía de Permutación.

Para obtener una mejor caracterización de la serie de tiempo de una manera simple mediante cuantificadores, se presenta el novedoso plano $\mathcal{H} \times \mathcal{H}_a$ (Figura 4.3). En este plano, los procesos estocásticos simulados están claramente diferenciados

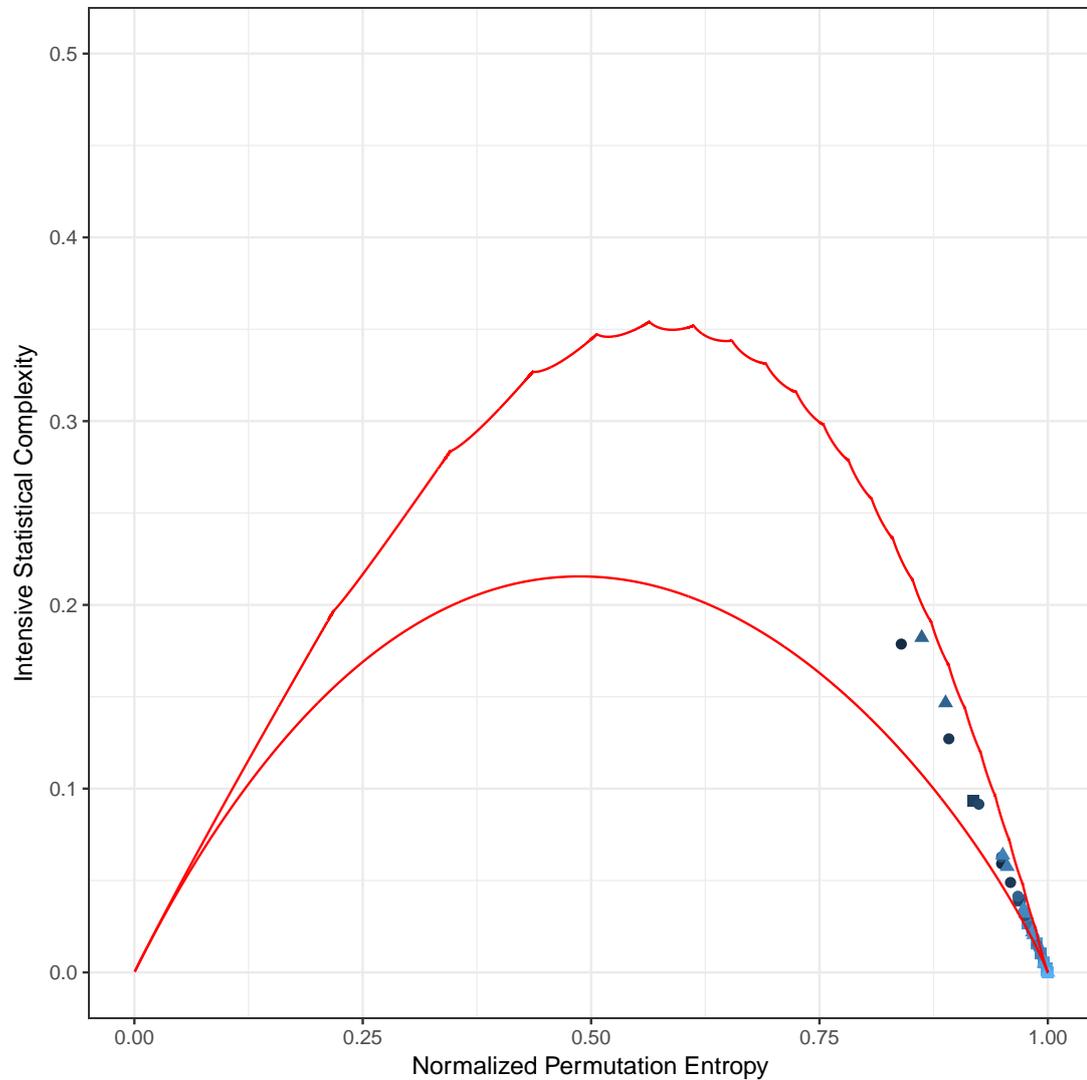


Figura 4.1 **Localización de los procesos autorregresivos en el plano $\mathcal{H} \times \mathcal{C}$ con $m = 4, \tau = 1$** , cuadrados: Exponenciales con parámetro $\lambda = 1$, círculos: Normales Estándar (Gaussianos) y triángulos: Uniformes $[0, 1]$. Los colores más oscuros representan mayores valores absolutos de autocorrelación. Se presentan valores promedios de 10 simulaciones con distintas semillas. Las curvas rojas continuas representan la curvas de máxima y mínima Complejidad Estadística, \mathcal{C}_{max} y \mathcal{C}_{min} , para una Entropía de Permutación dada \mathcal{H} . La localización en este plano depende del tipo de dinámica del proceso (estocástico o determinístico) y de la estructura de correlación del mismo. Se puede ver que la distribución de probabilidades marginal de los datos no es un factor en la ubicación en este plano por lo que se presenta el nuevo plano $\mathcal{H} \times \mathcal{H}_a$ en la Figura 4.3.

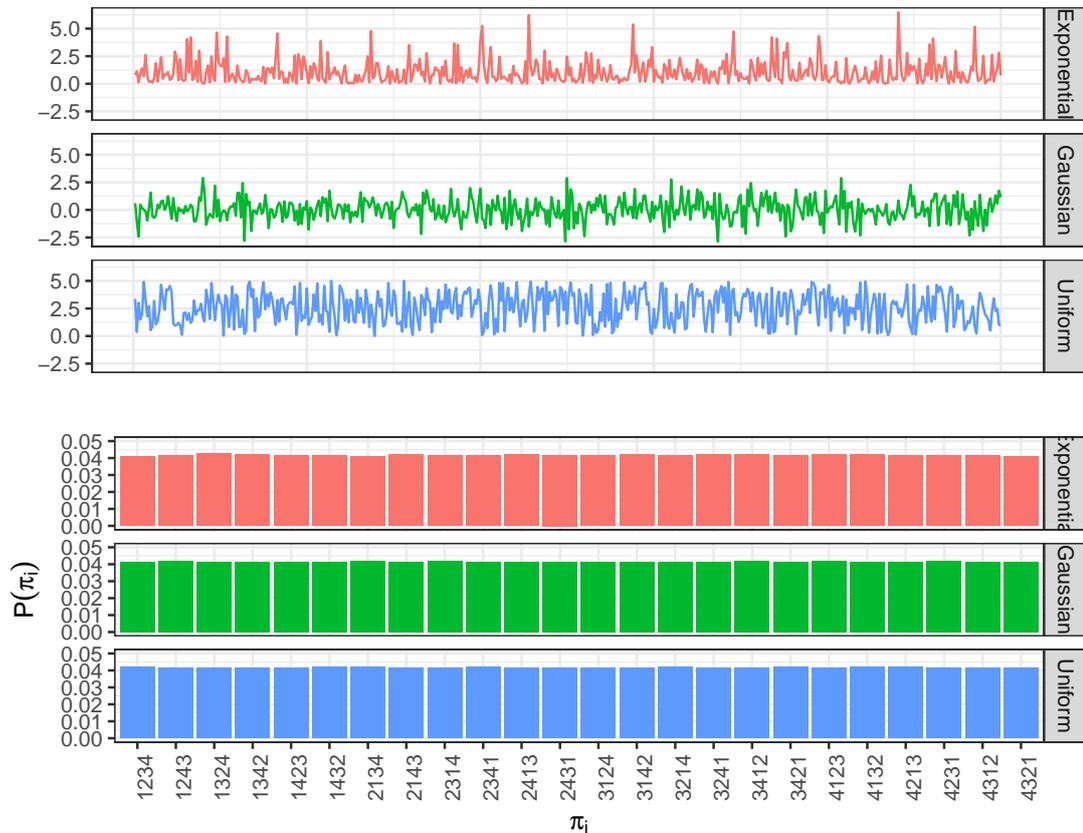


Figura 4.2 Las series de tiempo de los procesos autorregresivos y sus FDP de BP. Las series de tiempo con autocorrelación 0 simuladas y las estimaciones de sus funciones de distribución de probabilidad usando la metodología de Bandt y Pompe para las tres familias de procesos estocásticos, en rojo: la Exponencial de parámetro $\lambda = 1$, en verde: Normal Estándar y en azul claro: la Uniforme[0, 1]. Las tres FDP de BP son indistinguibles

tanto en el eje x (\mathcal{H}) como en el eje y (\mathcal{H}_a). Cuanto mayor es el coeficiente de autocorrelación en valor absoluto (más estructura de correlación), menor es la Entropía de Permutación $\hat{\mathcal{H}}$ mientras que la Entropía de Amplitud $\hat{\mathcal{H}}_a$ se mantiene constante. Por otro lado cuanto más simétrica y menos curtosis tiene la distribución de probabilidad (más se asemeja a una uniforme) mayor es $\hat{\mathcal{H}}_a$. Por lo tanto la ubicación de una serie de tiempo en este plano puede dar información acerca del proceso generador de datos, no sólo de la estructura de correlación temporal (\mathcal{H}) sino también de la distribución de probabilidades marginal(\mathcal{H}_a).

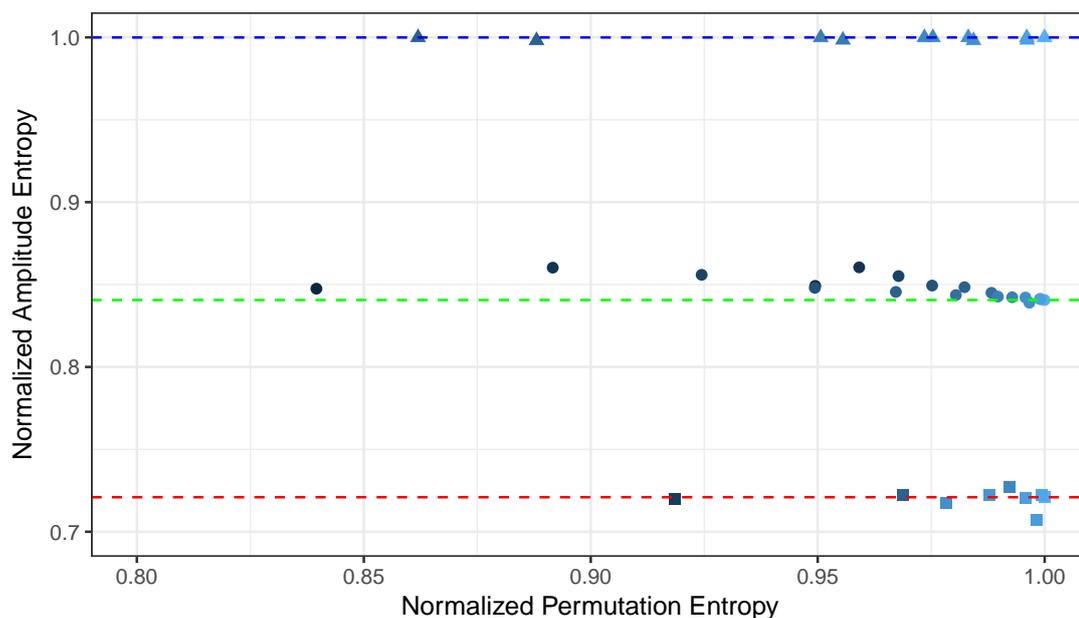


Figura 4.3 **Localización de los procesos autorregresivos en el plano $\mathcal{H} \times \mathcal{H}_a$** , cuadrados: Exponenciales con parámetro $\lambda = 1$, círculos: Normales Estándar y triángulos: Uniformes $[0, 1]$. Los colores más oscuros representan mayores valores absolutos de autocorrelación. Las líneas punteadas indican la entropía teórica para distribución dada (Exponencial en rojo, Normal Estándar en verde y Uniforme en azul). Los procesos están claramente diferenciados tanto en el eje x $-\mathcal{H}$ - como en el eje y $-\mathcal{H}_a$ -. La Entropía de Amplitud $-\mathcal{H}_a$ - depende principalmente de la forma de la distribución de probabilidades marginal de los datos y en menor medida de cómo fue construido el histograma, es por esto la variación entre los procesos. Sin embargo esta perturbación no interfiere en la discriminación de los procesos con distribuciones marginales diferentes.

4.6. Conclusiones del Capítulo

Se introdujo un nuevo plano informacional, $\mathcal{H} \times \mathcal{H}_a$, que es muy simple y rápido de computar. Este plano complementa al plano $\mathcal{H} \times \mathcal{C}$ dando información global acerca de la familia de distribuciones obteniendo una descripción más completa del proceso generador de datos.

Capítulo 5

Influencia de los datos repetidos en la Función de Distribución de Probabilidades de Bandt y Pompe

5.1. Introducción

En el artículo donde se introdujo la Entropía de Permutación se estableció como condición para su estimación que los datos de la serie de tiempo provengan de un proceso generador de datos continuos, y por lo tanto la probabilidad de que dos números sean iguales dentro de un vector de *embedding* es cero (Bandt and Pompe [2002]). Las metodologías de simbolización propuestas en el Capítulo 2 se basan en este supuesto y no pueden ser utilizadas en caso de empates en los rangos ($R(x_s) = R(x_r)$ para $r \neq s$ donde x_s y x_r pertenecen al mismo vector de *embedding* $X_t^{(m)}$). En el improbable caso de que existan valores repetidos, se ha sugerido o bien ignorar estos patrones que los contengan, ya que su cantidad sería insignificante y

no alteraría la distribución de probabilidades asociada, o simplemente agregar una pequeña perturbación aleatoria.

Desafortunadamente, datos experimentales digitalizados con una relativa baja resolución pueden tener una cantidad no despreciable de vectores de *embedding* con valores repetidos y consecuentemente la estimación de la FDP de BP puede verse significativamente afectada. Se mostró que las distintas maneras de abordar esta problemática generan sesgos en la estimación de la FDP de BP y por lo tanto, la manera de tratar datos con posibles igualdades en sus valores debe ser estudiada con cuidado ([Zunino et al., 2017]).

Las series de tiempo con valores repetidos pueden ocurrir por dos razones. O bien la serie de tiempo proviene de un proceso generador de datos discretos, por ejemplo un proceso de Poisson autorregresivo donde la probabilidad marginal es de una familia de distribuciones de Poisson (McKenzie [1988]), o la serie de tiempo viene de un proceso generador de datos continuos pero debido a una falta de precisión del artefacto de medición se obtiene una versión truncada de los datos.

Muchos artículos han sugerido modificaciones en la estimación de la Entropía de Permutación para tratar con estos datos repetidos. Una manera es extendiendo el alfabeto simbólico presentado por Bandt y Pompe para incluir estos patrones (Bian et al. [2012]) y la otra es estableciendo nuevas reglas artificiales para ordenar los datos (Parlitz et al. [2012]).

En esta Tesis se propone una nueva metodología para resolver este problema. Esta metodología propuesta, la Imputación Basada en la Muestra, usa la información de la propia serie de tiempo para tratar con los valores repetidos dentro de los vectores de *embedding*. Se asume que hay una serie de tiempo subyacente (no observada) a valores continuos y los datos observados son una versión no fiel (de acá en adelante se la denominará versión corrupta) de dicha serie. Con esto en mente se propone reconstruir los verdaderos vectores de *embedding* subyacentes usando la información presente en los vectores de *embedding* con datos corruptos, mediante una adecuada distribución de probabilidades. De este modo es que se puede ver como una Imputación Basada en la Muestra. También se muestra

numéricamente que si bien esta metodología fue concebida para datos adulterados debido al proceso de medición provenientes de un proceso generador de datos continuos, funciona notablemente bien cuando los datos son discretos desde su origen, como se puede ver en la Sección 5.5.2 de los números trascendentales.

Este Capítulo tiene un doble objetivo, el primero es hacer un exploración exhaustiva de todos los métodos existentes en la literatura que intentan lidiar con esta problemática y el segundo es introducir este nuevo método basado en la muestra para tratar con datos repetidos. Se compararán luego todas las metodologías presentadas mediante dos simulaciones:

- Se tomarán series de tiempo provenientes de mapas caóticos conocidos con una precisión de 15 decimales, que no presentan vectores de *embedding* con componentes repetidas y luego, a cada serie de tiempo se le truncarán sus valores a 1 decimal emulando un instrumento de medición de baja resolución.
- Las series de tiempo provenientes de la expansión decimal de tres números irracionales conocidos: π , $\sqrt{2}$ y e , que presentan vectores de *embedding* con componentes repetidas que son una característica propia de la serie de tiempo analizada.

5.2. El problema que representan los datos repetidos

Con la condición $P(x_r = x_s) = 0 \forall r \neq s$ establecida por Bandt y Pompe al presentar la Entropía de Permutación todos los vectores de *embedding* $X_t^{(m)}$ tienen m valores únicos (sin repetirse), y los dos mapeos, **según los rangos** y **según el orden cronológico** (ver Capítulo 2) conducen al mismo resultado para el cálculo de la Entropía de Permutación para cualquier realización de una serie de tiempo $\{\mathcal{X}_t\}_{t \in T}$.

Como se indicó en la Introducción de este Capítulo, en muchas de las series de tiempo reales esta condición no se cumple, por lo que una cantidad significativa

de los vectores de *embedding* $X_t^{(m)}$ de esta serie de tiempo pueden tener valores repetidos y ninguno de los mapeos presentados puede ser usado para transformar a estos $X_t^{(m)}$ en algún $\pi_i \in S_m$. Por esta razón, diferentes metodologías para tratar con este problema fueron desarrolladas.

En esencia hay dos estrategias para lidiar con el problema de los valores repetidos en los datos cuando se quiere estimar la Entropía de Permutación. La primera asume que el proceso es de hecho continuo, por lo que los vectores de *embedding* con estas repeticiones presentan en realidad información faltante, y derivan de vectores originales sin valores repetidos. La segunda no hace esta suposición y extiende el alfabeto para incorporar estos patrones a la dinámica propia del proceso e ignora la restricción impuesta $i_k \neq i_j \forall i \neq j$ para $\pi_i = (i_1, i_2, \dots, i_m) \in S_m$ y permite que $i_k = i_j$ para $i \neq j$. Estos alfabetos tienen más símbolos que el original ($m!$) y varían sustancialmente de acuerdo a si son construidos con el mapeo según rangos o el mapeo según el orden cronológico (Cuadro 5.1).

Alfabeto	m=3	m=4	m=5	m=6
Original	6	24	120	720
Extendido según mapeo cronológico	13	73	501	4051
Extendido según mapeo de rangos	13	75	541	4683

Cuadro 5.1 Cantidad de símbolos π_i para la dimensión de *embedding* m para cada alfabeto estudiado en esta Tesis. Mientras que para el alfabeto regular, la cantidad es $m!$, el número de símbolos del alfabeto extendido según el mapeo cronológico es mucho mayor (ver [Bian et al., 2012] para la fórmula explícita), y mayor aún es la cantidad del alfabeto extendido según el rango que responde a la fórmula del número de Bell ordenado B_m .

5.3. Los patrones con valores repetidos como posibles estados del sistema: Extendiendo el Alfabeto Simbólico

Si los vectores de *embedding* que tienen componentes con valores iguales son característicos del proceso generador de datos, un mapeo de los valores iguales en $\{x_t\}_{t \in T}$ a una similar representación de un símbolo π puede ser tenido en consideración, ampliando alguno de los alfabetos presentados en el Capítulo 2.

5.3.1. Alfabeto extendido según mapeo cronológico

Este alfabeto se presentó para utilizarlo en el cálculo de lo que dieron en llamar *Entropía de Permutación Modificada* (Bian et al. [2012]). Como en el alfabeto original que se mapea según el orden cronológico, los valores de $X_t^{(m)}$ se ordenan de manera creciente: $x_{t+i_{j_1}-1} \leq x_{t+i_{j_2}-1} \leq \dots \leq x_{t+i_{j_m}-1}$.

Normalmente, cuando no hay valores iguales, x_{t+i_*-1} es representado por i_* en el símbolo π_i .

Sin embargo, cuando se producen igualdades, dichos valores iguales son mapeados al mismo símbolo, que es el índice más pequeño i_* de ellos. Por ejemplo, si $x_{t+i_{j_1}-1} = x_{t+i_{j_2}-1}$ y $i_{j_1} < i_{j_2}$, ambos $x_{t+i_{j_1}-1}$ y $x_{t+i_{j_2}-1}$ son representados por i_{j_1} en el símbolo π_i . El correspondiente símbolo de patrón $X_t^{(m)}$ es definido como: $\pi_i = (i_1 \ i_2 \ \dots \ i_{j_1} \ i_{j_1} \ \dots \ i_m)$.

Por ejemplo, si tomamos la serie:

$$X_t = (2, 5, 1, 2, 7, 1, 1, 3, 1) \quad T = 9 \quad (5.1)$$

y tomamos el vector de *embedding* $X_1^{(5)} = (2, 5, 1, 2, 7)$, éste se mapea al símbolo $\pi = 31125$. Mientras que $X_6^{(3)} = (1, 1, 3)$ y $X_6^{(4)} = (1, 3, 1, 1)$ se mapean a los símbolos $\pi = 113$ y $\pi = 1112$ respectivamente.

Este alfabeto modificado, mapeado según el orden cronológico resulta en un alfabeto con más símbolos que el original por lo que caracteriza más estados posibles

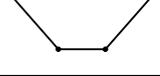
que el alfabeto de Bandt y Pompe (ver [Bian et al., 2012] para la fórmula explícita de la cantidad de símbolos para cada m).

Para estimar la distribución de probabilidades asociada al proceso, se calculan las frecuencias de aparición de cada símbolo y su correspondiente probabilidad estimada de aparición de manera análoga a la Ecuación 2.5 del Capítulo 2.

Una característica fundamental de un mapeo de patrones (cada vector de *embedding* identifica un patrón) a símbolos es que es una transformación 1 a 1. Es decir, partiendo de la secuencia de símbolos se debería reconstruir sin ambigüedades la secuencia de patrones de la cual proviene. El alfabeto extendido según el orden cronológico no cumple con esta importante premisa. Si bien para la dimensión de *embedding* $m = 3$ la transformación es 1 a 1, para $m = 4$ hay dos casos donde distintos patrones en $X_t^{(m)}$ son mapeados al mismo símbolo, imposibilitando cualquier reconstrucción (ver Cuadro 5.2). A medida que la dimensión de *embedding* m crece, la cantidad de casos en donde los patrones son mapeados al mismo símbolo también crece. De hecho, esta cantidad puede ser cuantificada en el Cuadro 5.1 mediante la diferencia entre la cantidad de símbolos que posee el alfabeto extendido según rangos (presentado en la próxima Sección) y este alfabeto extendido. De esta manera para esta Tesis, si lo que se desea es extender el alfabeto simbólico propuesto por Bandt y Pompe, se utilizará el alfabeto extendido según mapeo de rangos presentado en la siguiente Sección.

5.3.2. Alfabeto extendido según mapeo de rangos

Una manera de solucionar el problema que presenta el alfabeto extendido de la Sección anterior de no proveer una transformación 1 a 1 de patrones a símbolos es usar el alfabeto extendido según el mapeo de rangos. Nuevamente como se indicó en el Capítulo 2 para el alfabeto original según rangos, $X_t^{(m)} = (x_t, x_{t+1}, \dots, x_{t+m-1})$ puede ser mapeado $\pi_i = (R(x_{t+1}) R(x_{t+2}) \dots R(x_{t+m})) \in S_m$. Si existen valores iguales, estos valores adquieren el mismo rango. Tomaremos la serie anterior como ejemplo:

Patrón	Alfabeto extendido según mapeo cronológico	Alfabeto extendido según mapeo de rangos	$X_t^{(4)}$
	$\pi = 1122$	$\pi = 1313$	$X_i^{(4)}$
		$\pi = 1331$	$X_j^{(4)}$
	$\pi = 2211$	$\pi = 3131$	$X_k^{(4)}$
		$\pi = 3113$	$X_l^{(4)}$

Cuadro 5.2 El alfabeto extendido según el orden cronológico y sus ambigüedades. Una característica fundamental de un mapeo de valores a símbolos, es que es una transformación 1 a 1. Es decir, partiendo de la secuencia de símbolos se debería reconstruir sin ambigüedades la secuencia de patrones de la cual proviene. El alfabeto extendido según el orden cronológico propuesto por [Bian et al. \[2012\]](#) no cumple con esta importante premisa. Si bien para la dimensión de *embedding* $m = 3$ la transformación es 1 a 1, para $m = 4$ hay dos casos donde distintos patrones en $X_t^{(4)}$ son mapeados al mismo símbolo, imposibilitando cualquier reconstrucción. A medida que la dimensión de *embedding* m crece, la cantidad de casos en donde los patrones son mapeados al mismo símbolo crece. De hecho, esta cantidad puede ser cuantificada en el Cuadro 5.1 mediante la diferencia entre la cantidad de símbolos que posee el alfabeto extendido según rangos (presentado en la sección 5.3.2) y este alfabeto extendido.

$$X_t = (2, 5, 1, 2, 7, 1, 1, 3, 1) \quad T = 9$$

Tomemos el vector de *embedding* $X_1^{(5)} = (2, 5, 1, 2, 7)$, éste se transforma en el símbolo $\pi = 24125$. Mientras que $X_6^{(3)} = (1, 1, 3)$ y $X_7^{(3)} = (1, 3, 1, 1)$ se mapean a los símbolos $\pi = 113$ y $\pi = 1411$ respectivamente. El problema combinatorio de determinar el número de ordenamientos $-w(m)$ - de m candidatos cuando se permiten los empates en los rangos (en este caso provenientes de valores repetidos dentro de un vector de *embedding*) se denomina *número de Bell ordenado* (Good [1975]) y se resuelve como:

$$w(m) = \sum_{r=0}^{\infty} \frac{r^m}{2^{r+1}} \quad (5.2)$$

Los resultados de la Ecuación 5.2 para m de 3 a 6 se pueden ver en el Cuadro 5.1. Si bien este alfabeto soluciona el problema mencionado al principio, tiene muchos más símbolos que el alfabeto original según rangos propuesto por Bandt y Pompe.

Para estimar la distribución de probabilidades asociada al proceso, se calculan las frecuencias de aparición de cada símbolo y su correspondiente probabilidad estimada de aparición de manera análoga a la Ecuación 2.5 del Capítulo 2.

5.4. Los valores repetidos como información faltante

Si se asume que la serie viene de un proceso generador de datos continuos, entonces los vectores de *embedding* con componentes iguales pueden ser pensados como vectores con *información perdida* (datos faltantes) que derivan de vectores de *embedding* originales sin estas repeticiones.

Los datos faltantes son un problema común en todos los campos de estudio y hay varios métodos para manejar este inconveniente: el primero es eliminar todas las observaciones incompletas (y trabajar solo con los casos completos) y el segundo

es imputar a estos datos perdidos un valor apropiado para retener la información que sí contiene esta observación (Donders et al. [2006]). Para todos los ejemplos que siguen se va a usar el alfabeto según rangos.

5.4.1. Casos Completos

Este método fue originalmente sugerido por Bandt y Pompe y es análogo a análisis de casos completos de la teoría estadística. Es simplemente eliminar los vectores de *embedding* que contengan datos repetidos.

Por ejemplo para $m = 3$ la serie $X_t = (2, 5, 1, 2, 7, 1, 1, 3, 1)$

$$X_1^{(3)} = (2, 5, 1) \rightarrow \pi = 231,$$

$$X_2^{(3)} = (5, 1, 2) \rightarrow \pi = 312,$$

$$X_3^{(3)} = (1, 2, 7) \rightarrow \pi = 123,$$

$$X_4^{(3)} = (2, 7, 1) \rightarrow \pi = 231,$$

$$X_5^{(3)} = (7, 1, 1) \text{ se elimina,}$$

$$X_6^{(3)} = (1, 1, 3) \text{ se elimina,}$$

$$X_7^{(3)} = (1, 3, 1) \text{ se elimina.}$$

Con los símbolos que representan a los vectores de *embedding* que no fueron eliminados se estima la FDP de BP análogamente a la Ecuación 2.5 del Capítulo 2. De esta manera se pierde mucha información del proceso subyacente.

5.4.2. Imputación según el Orden de Aparición

Es una de las técnicas más usadas para tratar con patrones con componentes iguales. Simplemente establece que si dos componentes de un vector de *embedding* son iguales, $x_{t_1} = x_{t_2}$ y $t_1 < t_2$ entonces $x_{t_1} < x_{t_2}$. Siguiendo el ejemplo anterior,

$$X_1^{(3)} = (2, 5, 1) \rightarrow \pi = 231,$$

$$X_2^{(3)} = (5, 1, 2) \rightarrow \pi = 312,$$

$$X_3^{(3)} = (1, 2, 7) \rightarrow \pi = 123,$$

$$X_4^{(3)} = (2, 7, 1) \rightarrow \pi = 231,$$

son mapeados de la misma manera que antes ya que no tienen valores repetidos,

$$X_5^{(3)} = (7, 1, 1) \rightarrow \pi = 312,$$

$$X_6^{(3)} = (1, 1, 3) \rightarrow \pi = 123,$$

$$X_7^{(3)} = (1, 3, 1) \rightarrow \pi = 132,$$

ya que se asume que en caso de empate en rangos, el valor que aparece primero según el orden temporal tiene el rango más pequeño.

Con los símbolos que representan a los vectores de *embedding* se estima la FDP de BP análogamente a la Ecuación 2.5 del Capítulo 2. Se ha usado ampliamente (Cao et al. [2004]; Matilla-García and Marín [2009]; Parlitz et al. [2012]; Saco et al. [2010]; Zunino et al. [2008b]) pero en ningún caso se da una explicación del por qué de este ordenamiento artificial.

5.4.3. Imputación al Azar

Bandt y Pompe también recomiendan romper estas igualdades entre valores añadiendo una pequeña perturbación aleatoria. La amplitud de esta perturbación debe ser establecida de modo de no modificar las relaciones ordinales en el vector de *embedding*, excepto por aquellos componentes del vector que presenten componentes iguales.

Esto es equivalente numéricamente a mapear a un vector de *embedding* con componentes de igual valor a cualquiera de los posibles vectores “originales” (sin repeticiones) que pudieran haber derivado en esta observación alterada, dándole un peso w igual a cada posible símbolo compatible de tal manera de que los pesos de los símbolos de dicho vector sumen 1.

Siguiendo el mismo ejemplo,

$$X_1^{(3)} = (2, 5, 1) \rightarrow \pi = 231 \quad w = 1,$$

$$X_2^{(3)} = (5, 1, 2) \rightarrow \pi = 312 \quad w = 1,$$

$$X_3^{(3)} = (1, 2, 7) \rightarrow \pi = 123 \quad w = 1,$$

$$X_4^{(3)} = (2, 7, 1) \rightarrow \pi = 231 \quad w = 1,$$

son mapeados de la misma manera que antes ya que no tienen valores repetidos,

$$X_5^{(3)} = (7, 1, 1) \rightarrow \begin{cases} \pi = 312 & w = 1/2 \\ \pi = 321 & w = 1/2 \end{cases}$$

estas son las dos únicas opciones ya que la primer componente del vector no debe ser alterada por esta perturbación, y con la misma lógica,

$$X_6^{(3)} = (1, 1, 3) \rightarrow \begin{cases} \pi = 123 & w = 1/2 \\ \pi = 213 & w = 1/2 \end{cases}$$

$$X_7^{(3)} = (1, 3, 1) \rightarrow \begin{cases} \pi = 132 & w = 1/2 \\ \pi = 231 & w = 1/2 \end{cases}$$

Por ejemplo, para $m = 3$, si los tres valores del vector de *embedding* son iguales, por ejemplo $X_k^{(3)} = (7, 7, 7)$ es mapeado a todos los símbolos con ponderación $w = 1/6$.

La frecuencia relativa de los símbolos para estimar la FDP de BP se hace análogamente a la Ecuación 2.5, pero en este caso se tiene en cuenta los pesos de los mismos para su cómputo.

5.4.4. Imputación Basada en la Muestra

La imputación al azar sugiere que independientemente de la serie de tiempo bajo estudio, como los vectores de *embedding* con valores repetidos son el resultado de una observación errónea de vectores sin componentes repetidas en la serie original, deberían ser mapeados a alguno de estos símbolos compatibles con la misma ponderación.

En muchas situaciones, técnicas simples para tratar con información faltante (como *casos completos* o *imputación al azar*) en otros campos de estudio producen resultados sesgados (Donders et al. [2006]), y existen otras técnicas más sofisticadas con las que se obtienen muchos mejores resultados. Con estas técnicas, a la

información faltante de un sujeto se le imputa un valor predicho por las otras características conocidas de dicho sujeto.

En este trabajo se propone un método similar a la imputación al azar, pero en vez de agregar una perturbación aleatoria que mapea con igual ponderación a cualquier símbolo compatible, estas ponderaciones o pesos w , son originadas de acuerdo a una FDP previamente conocida y no son necesariamente iguales. La FDP propuesta como distribución *a priori* es que resulta de computar los $\hat{p}(\pi_i) \forall i$ con la metodología de Casos Completos (ver Sección 5.4.1).

El proceso consiste en ocho pasos

1. Definir la dimensión de *embedding* m , y definir a $S_m = \{\pi_i\}$ como todas las $m!$ posibles permutaciones de $(1, 2, \dots, m)$.
2. Mapear todos los vectores de *embedding* $X_t^{(m)} \forall t$ a su correspondiente $\pi_i \in S_m$ de acuerdo al mapeo según rangos (Capítulo 2).
3. Si hay empate en los rangos de $X_t^{(m)}$ para algún t , eliminar dicho vector (Sección 5.4.1).
4. Calcular las frecuencias relativas $\hat{p}^*(\pi_i) = \hat{p}(\pi_i)$ (Capítulo 2).
5. Repetir el procedimiento para los $X_t^{(m)} \forall t$ a su correspondiente $\pi_i \in S_m$ de acuerdo al mapeo según rangos, pero no eliminar los vectores $X_t^{(m)}$ que posean valores repetidos.
6. Si el vector $X_t^{(m)}$ no tiene valores repetidos, mapearlo al correspondiente símbolo $\pi_i \in S_m$ con ponderación $w(\pi_i) = 1$
7. Si el vector $X_t^{(m)}$ tiene valores repetidos, romper los empates añadiendo pequeñas perturbaciones de manera similar a la *Imputación al Azar* (Sección 5.4.3). Identificar todos los k patrones compatibles $\pi_k \in S_m$ y asignarles a cada uno un peso $w(\pi_k) = \hat{p}^*(\pi_k) / \sum_k (\hat{p}^*(\pi_k))$ (notar que las probabilidades $\hat{p}^*(\pi_k)$ son las obtenidas en el paso 4).

8. Calcular las nuevas probabilidades $\hat{p}(\pi_i), i = 1 \dots m!$ resultantes de los vectores de *embedding* pero considerando las ponderaciones obtenidas en los pasos 6 y 7 ($w(\pi_i) = 1$ si el vector no contiene valores repetidos y $w(\pi_i) \neq 1$ caso contrario).

Como ejemplo de esta metodología de imputación basada en la muestra tomemos la serie

$$X_t = (2, 5, 1, 2, 7, 1, 1, 3, 1, 2, 4, 5, 1, 3, 2, 4, 4, 2, 2, 1, 0) \quad (5.3)$$

que se muestra en la figura 5.1.

De manera comparativa el Cuadro 5.3 hace un mapeo de cada $X_t^{(3)}$ de esta serie de tiempo para cada metodología presentada en este Capítulo.

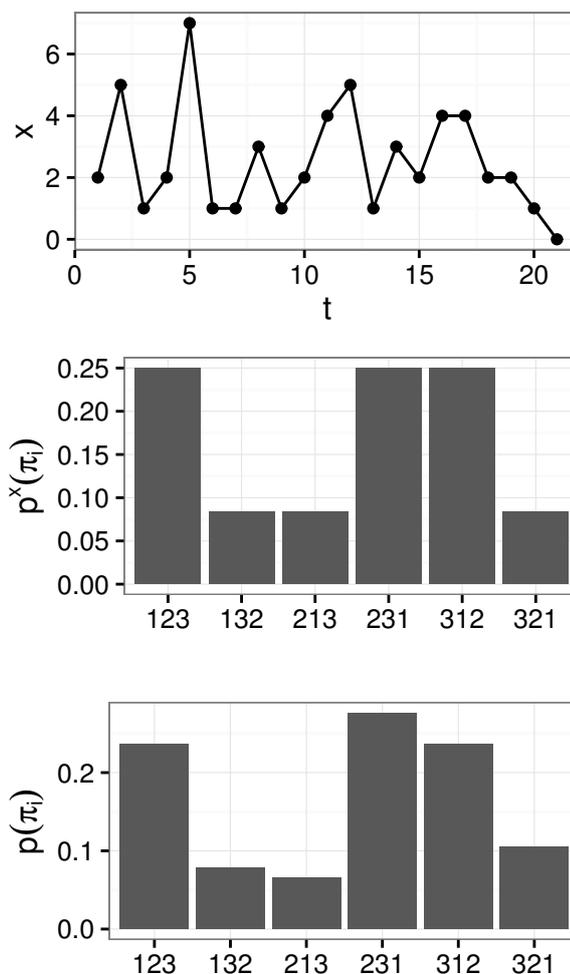


Figura 5.1 **Ejemplo de aplicación de la metodología basada en la muestra para tratar con datos repetidos.** El gráfico de arriba muestra la serie de tiempo $X_t = (2, 5, 1, 2, 7, 1, 1, 3, 1, 2, 4, 5, 1, 3, 2, 4, 4, 2, 2, 1, 0)$, el gráfico del medio computa la FDP de BP para una dimensión de *embedding* $m = 3$ usando la metodología de Casos Completos (Sección 5.4.1). Notar que usa los 12 patrones que no presentan datos repetidos. El tercer gráfico muestra la distribución resultante de los 19 vectores $X_t^{(3)}$ utilizando la metodología de Imputación Basada en la Muestra.

$X_t^{(3)}$			π_i					
			Ext. s/ Ord. Cron.	Ext. s/ Rangos	Casos Comp.	Imp s/ Ord.	Imp. al Azar	Imp. Basada en Muestra.
2	5	1	312	231	231	231	231 $w = 1$	231 $w = 1$
5	1	2	231	312	312	312	312 $w = 1$	312 $w = 1$
1	2	7	123	123	123	123	123 $w = 1$	123 $w = 1$
2	7	1	312	231	231	231	231 $w = 1$	231 $w = 1$
7	1	1	221	311	X	312	312 $w = \frac{1}{2}$ 321 $w = \frac{1}{2}$	312 $w = \frac{3}{4}$ 321 $w = \frac{1}{4}$
1	1	3	113	113	X	123	123 $w = \frac{1}{2}$ 213 $w = \frac{1}{2}$	123 $w = \frac{3}{4}$ 213 $w = \frac{1}{4}$
1	3	1	112	131	X	132	132 $w = \frac{1}{2}$ 231 $w = \frac{1}{2}$	132 $w = \frac{3}{4}$ 231 $w = \frac{1}{4}$
3	1	2	231	312	312	312	312 $w = 1$	312 $w = 1$
2	4	5	123	123	123	123	123 $w = 1$	123 $w = 1$
4	5	1	312	231	231	231	231 $w = 1$	231 $w = 1$
5	1	3	231	312	312	312	312 $w = 1$	312 $w = 1$
1	3	2	132	132	132	132	132 $w = 1$	132 $w = 1$
3	2	4	213	213	213	213	213 $w = 1$	213 $w = 1$
2	4	4	122	122	X	123	123 $w = \frac{1}{2}$ 132 $w = \frac{1}{2}$	123 $w = \frac{3}{4}$ 132 $w = \frac{1}{4}$
4	4	2	311	221	X	231	231 $w = \frac{1}{2}$ 321 $w = \frac{1}{2}$	231 $w = \frac{3}{4}$ 321 $w = \frac{1}{4}$
4	2	2	221	311	X	312	312 $w = \frac{1}{2}$ 321 $w = \frac{1}{2}$	312 $w = \frac{3}{4}$ 321 $w = \frac{1}{4}$
2	2	1	311	221	X	231	231 $w = \frac{1}{2}$ 321 $w = \frac{1}{2}$	231 $w = \frac{3}{4}$ 321 $w = \frac{1}{4}$
2	1	0	321	321	321	321	321 $w = 1$	321 $w = 1$

Cuadro 5.3 Los diferentes vectores de *embedding* $X_t^{(3)}$ para la serie de tiempo $X_t = 2, 5, 1, 2, 7, 1, 1, 3, 1, 2, 4, 5, 1, 3, 2, 4, 4, 2, 2, 1, 0$ y su mapeo a símbolos π_i según las diferentes metodologías usadas para lidiar con valores repetidos. Extender el alfabeto en cualquiera de sus dos versiones implica que los valores repetidos en $X_t^{(3)}$ son patrones que de alguna manera representan la dinámica del proceso, por lo que son representados por símbolos π_i que tienen empates. Las cuatro metodologías de la derecha asumen que los vectores $X_t^{(3)}$ con datos repetidos son casos particulares de observaciones con información faltante. *Casos Completos* elimina esos patrones y calcula las probabilidades $\hat{p}^*(\pi_i)$ sin ellos. La *Imputación según el Orden de Aparición* asume que el primer valor que aparece es el menor en caso de empate. Finalmente la *Imputación al Azar* y la *Imputación Basada en la Muestra* mapean a los vectores $X_t^{(3)}$ con valores repetidos a un símbolo compatible. Mientras que la primera lo hace con la misma ponderación, la segunda tiene en cuenta la estructura subyacente del proceso y usa las probabilidades $\hat{p}^*(\pi_i)$ calculada en los Casos Completos para el cálculo de dichos pesos w .

PATTERN	Ext. s/ Orden C.	Ext. s/ Rangos	Casos Comp.	Imp s/ Ord.	Imp. al Azar	Imp. Basada en la Muestra
	π_i	π_i	π_i	π_i	π_i con w	π_i con w
	111	111	X	123	a todos $w = 1/6$	a todos $w = p^*(\pi_i)$
	112	131	X	132	$132_{w=1/2}$ $231_{w=1/2}$	$132_{w=p^*(132)/\{p^*(132)+p^*(231)\}}$ $231_{w=p^*(231)/\{p^*(132)+p^*(231)\}}$
	113	113	X	123	$123_{w=1/2}$ $213_{w=1/2}$	$123_{w=p^*(123)/\{p^*(123)+p^*(213)\}}$ $213_{w=p^*(213)/\{p^*(123)+p^*(213)\}}$
	122	122	X	123	$123_{w=1/2}$ $132_{w=1/2}$	$123_{w=p^*(123)/\{p^*(123)+p^*(132)\}}$ $132_{w=p^*(132)/\{p^*(123)+p^*(132)\}}$
	123	123	123	123	$123_{w=1}$	$123_{w=1}$
	132	132	132	132	$132_{w=1}$	$132_{w=1}$
	211	212	X	213	$213_{w=1/2}$ $312_{w=1/2}$	$213_{w=p^*(213)/\{p^*(213)+p^*(312)\}}$ $312_{w=p^*(312)/\{p^*(213)+p^*(312)\}}$
	213	213	213	213	$213_{w=1}$	$213_{w=1}$
	311	221	X	312	$312_{w=1/2}$ $321_{w=1/2}$	$312_{w=p^*(312)/\{p^*(312)+p^*(321)\}}$ $321_{w=p^*(321)/\{p^*(312)+p^*(321)\}}$
	312	231	312	312	$312_{w=1}$	$312_{w=1}$
	221	311	X	231	$231_{w=1/2}$ $321_{w=1/2}$	$231_{w=p^*(231)/\{p^*(231)+p^*(321)\}}$ $321_{w=p^*(321)/\{p^*(231)+p^*(321)\}}$
	231	312	231	231	$231_{w=1}$	$231_{w=1}$
	321	321	321	321	$321_{w=1}$	$321_{w=1}$

Cuadro 5.4 Todos los patrones posibles para un vector de embedding $X_t^{(3)}$ y su mapeo según cada metodología presentada en este Capítulo.

5.5. Aplicaciones

5.5.1. Simulación numérica: Mapas Caóticos

En esta Sección las metodologías presentadas en el presente Capítulo para tratar con series de tiempo cuyos vectores de *embedding* presentan datos repetidos son evaluadas usando simulaciones de procesos caóticos. Para obtener un set reproducible de datos se simularon todos los mapas caóticos presentados en Rosso et al. [2013] usando idénticos parámetros y valores iniciales. Todas esas series presentaban o bien ningún vector con datos repetidos, o bien una cantidad despreciable de los mismos. A estas series las denominaremos las series “originales”. Luego, a cada una de estas series se le truncaron sus valores a una resolución de un decimal, simulando una versión corrupta debido a una medición de baja resolución. Debido a esta baja resolución, estas versiones corruptas tienen una cantidad significativa de vectores de *embedding* con datos repetidos para todo m . A estas series las denominaremos las series “truncadas”.

La simulación consiste en 39 series de tiempo generadas por diferentes procesos caóticos, cada una de largo $T = 100000$. Todas estas series presentaron o bien ningún vector $X_t^{(m)}$ con datos repetidos, o bien una cantidad despreciable. Para cada una de esas series se calculó la entropía de permutación $\mathcal{H} = \mathcal{H}(\mathbf{P})$ para las dimensiones de *embedding* $m = \{3, 4, 5, 6\}$. Notar que al ser cada serie determinística, la Entropía de Permutación \mathcal{H} de las series “originales” para cada m es un cuantificador fijo (para esos parámetros y valores iniciales) y se puede pensar como el valor a ser aproximado mediante el cálculo de la Entropía de Permutación según las distintas metodologías presentadas en este Capítulo, aplicadas a las series de tiempo truncadas:

- Alfabeto extendido: Extender el alfabeto simbólico y computar la Entropía de Permutación con este método. Para representar esta estrategia se utiliza el alfabeto extendido según rangos (Sección 5.3.2).
- Casos completos: Eliminar los vectores de *embedding* que presenten componentes repetidas (Sección 5.4.1) y computar la Entropía de Permutación.

- Imputación según Orden de Aparición: Identificar los vectores *embedding* que presenten componentes repetidas y romper los empates en dicho vector de acuerdo al orden de aparición en dicho vector (Sección 5.4.2). Computar la Entropía de Permutación con la FDP de BP resultante.
- Imputación al azar: Identificar los vectores *embedding* que presenten componentes repetidas y romper los empates en dicho vector con una perturbación aleatoria infinitesimal (Sección 5.4.3). Computar la Entropía de Permutación con la FDP de BP resultante.
- Imputación Basada en la muestra: Identificar los vectores *embedding* que presenten componentes repetidas y mapearlos a un símbolo compatible teniendo en cuenta la estructura subyacente del proceso usando las probabilidades $\hat{p}^*(\pi_i)$ calculadas en los Casos Completos para el cálculo de dichos pesos w . (Sección 5.4.4). Computar la Entropía de Permutación con la FDP de BP resultante.

A las Entropías de Permutación calculadas con cada metodología se las denominará $\tilde{\mathcal{H}} = \mathcal{H}(\tilde{\mathbf{P}})$.

A cada serie de tiempo truncada se le registra:

- η^* , la cantidad de vectores $X_t^{(m)}$ con valores repetidos
- $\eta(\%) = \eta^*/(T + m - 1)$ el porcentaje de vectores $X_t^{(m)}$ con valores repetidos sobre el total de vectores de *embedding*

Esta simulación pretende cuantificar la *performance* de cada estrategia para lidiar con datos repetidos mediante el error de aproximación ($\tilde{\mathcal{H}} - \mathcal{H}$) y error de aproximación absoluto ($|\tilde{\mathcal{H}} - \mathcal{H}|$) para cada serie de tiempo truncada proveniente de cada uno de los procesos caóticos, para cada dimensión de *embedding* $m = \{3, 4, 5, 6\}$.

En la Figura 5.2 se muestran los boxplots (Tukey [1977]) de los errores de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para cada una estas metodologías que intentan resolver el

problema de los valores repetidos en la estimación de la Entropía de Permutación, para las dimensiones de *embedding* $m = \{3, 4, 5, 6\}$. Se puede observar que la metodología de Imputación al Azar siempre sobrestima la Entropía de Permutación dado que es el equivalente a agregarle un pequeño ruido blanco a la serie original. Extender el alfabeto siempre subestima la Entropía de Permutación ya que agrega estados ficticios. Se puede ver que la Imputación Basada en la Muestra tiene menos dispersión en sus errores que la metodología de Casos Completos ya que recupera información perdida. La metodología de Imputación según el Orden de Aparición tiene un desempeño similar al de la Imputación Basada en la Muestra para esta aplicación.

En la Figura 5.3 se muestra el error de aproximación absoluto $|\tilde{\mathcal{H}} - \mathcal{H}|$ para todas las metodologías. Se puede observar, de manera similar a la Figura 5.2 que extender el alfabeto no se aproxima a la entropía verdadera de los procesos \mathcal{H} , teniendo un pobre desempeño. La Imputación al Azar incurre en errores muy grandes para muchos de los procesos caóticos simulados. La Imputación según el Orden de Aparición y la Imputación Basada en la Muestra tienen un buen desempeño, superando a la metodología de Casos Completos.

En la Figura 5.4 se muestra el error de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para cada una de las metodologías que intentan resolver el problema de los valores repetidos, para la dimensión de *embedding* $m = 6$ para los procesos caóticos agrupados según distintos niveles de η (%). Para pequeños valores de η (%) todas las metodologías tienen un buen desempeño salvo extender el alfabeto. Al aumentar el valor de η (%) el error de aproximación aumenta, debido a que se perdió mucha información sobre el proceso, para todas las metodologías excepto para extender el alfabeto que disminuye su error, aunque de igual manera su desempeño es el peor. En la Figura 5.5 se repite el procedimiento pero para el error absoluto $|\tilde{\mathcal{H}} - \mathcal{H}|$.

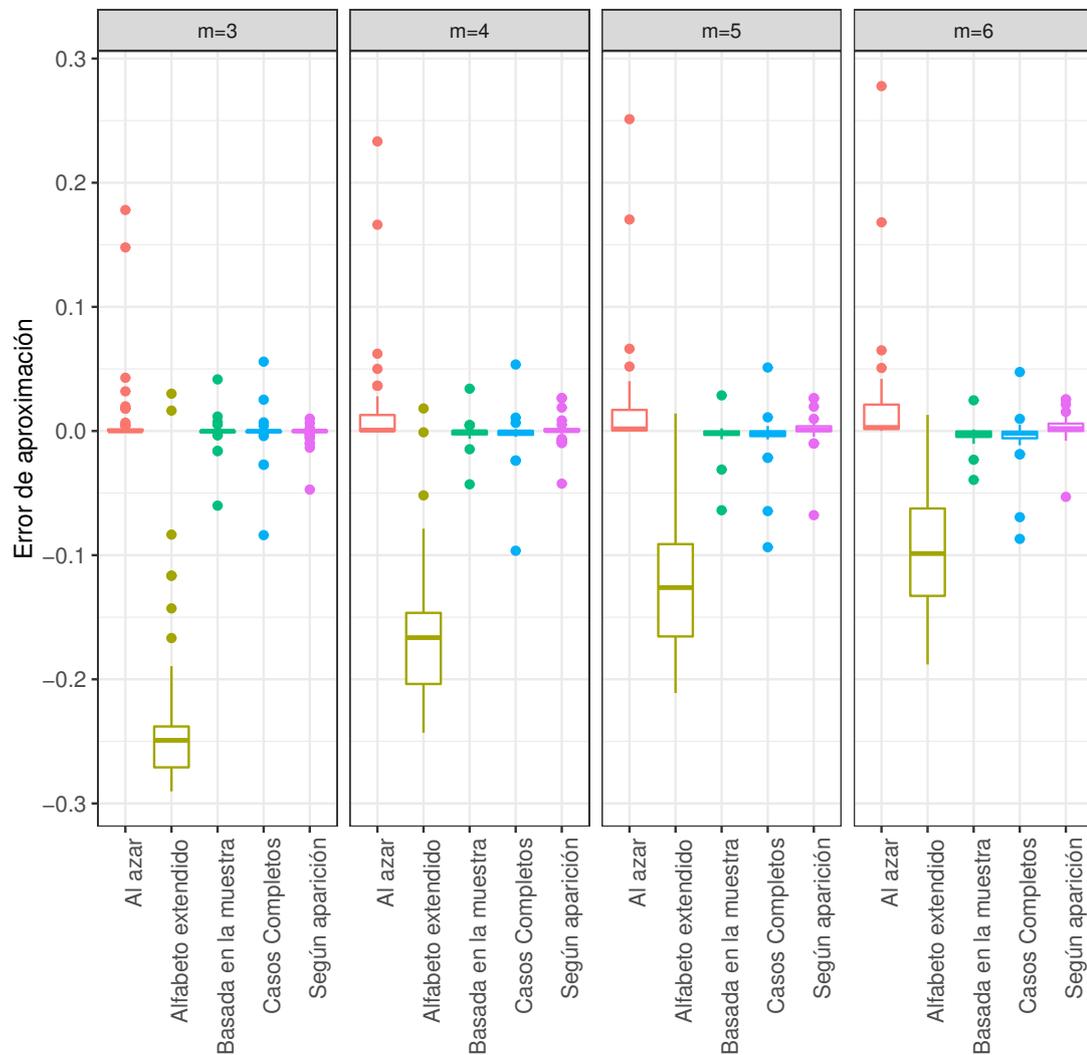


Figura 5.2 **Boxplot del error de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para las dimensiones de *embedding* $m = \{3, 4, 5, 6\}$.** Se puede observar que la metodología de Imputación al Azar siempre sobrestima la Entropía de Permutación dado que es el equivalente a agregarle un pequeño ruido blanco a la serie original. Extender el alfabeto siempre subestima la Entropía de Permutación ya que agrega estados ficticios. Se puede ver que la Imputación Basada en la Muestra tiene una dispersión menor que la metodología de Casos Completos ya que recupera información perdida. La metodología de Imputación según el Orden de Aparición tiene un desempeño similar al de la Imputación Basada en la Muestra para esta aplicación.

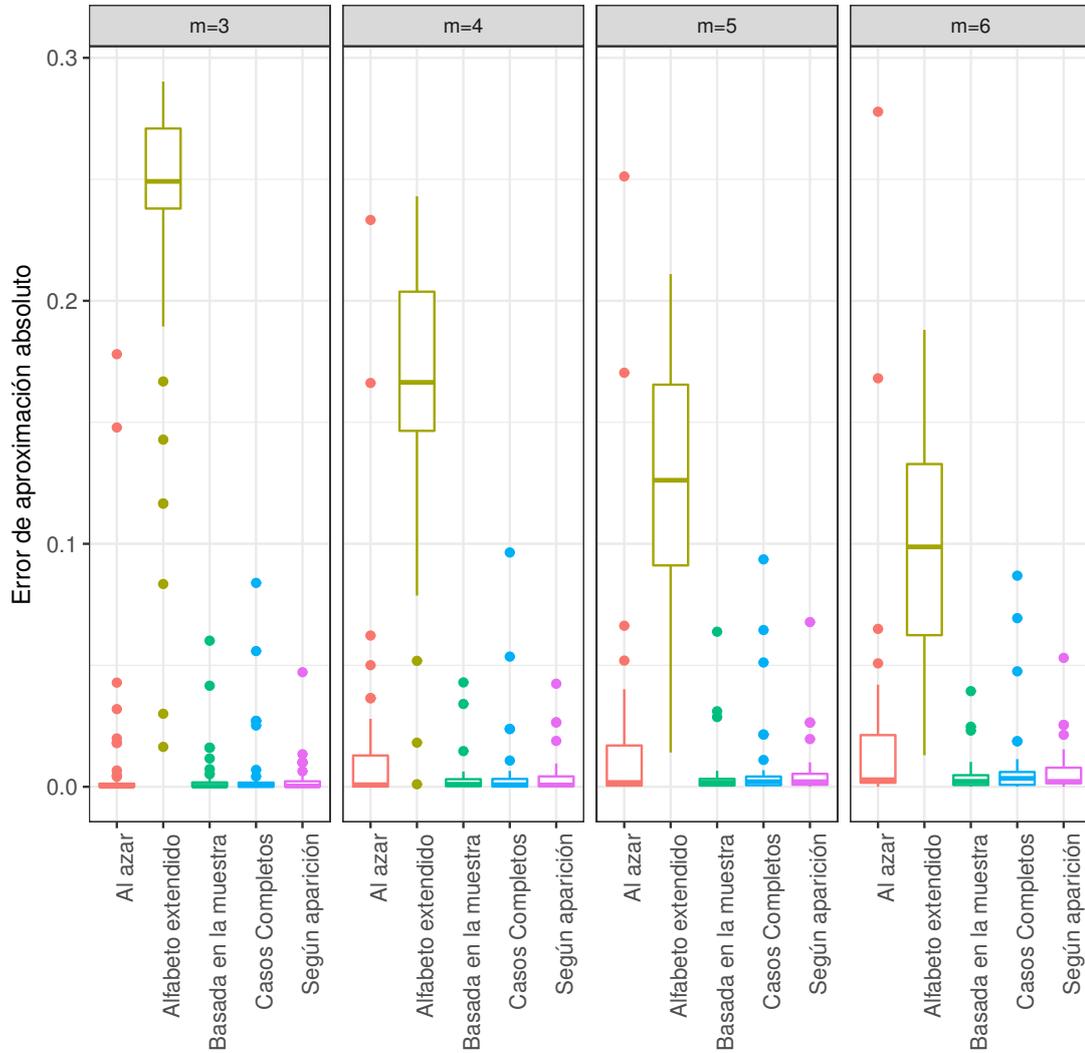


Figura 5.3 **Boxplot del error de aproximación absoluto $|\tilde{\mathcal{H}} - \mathcal{H}|$ para las dimensiones de *embedding* $m = \{3, 4, 5, 6\}$.** Se puede observar, de manera similar a la Figura 5.2 que extender el alfabeto no se aproxima a la entropía verdadera de los procesos \mathcal{H} , teniendo un pobre desempeño. La Imputación al Azar incurre en errores muy grandes para mucho de los procesos caóticos simulados. La Imputación según el Orden de Aparición y la Imputación Basada en la Muestra tienen un buen desempeño, superando a la metodología de Casos Completos.

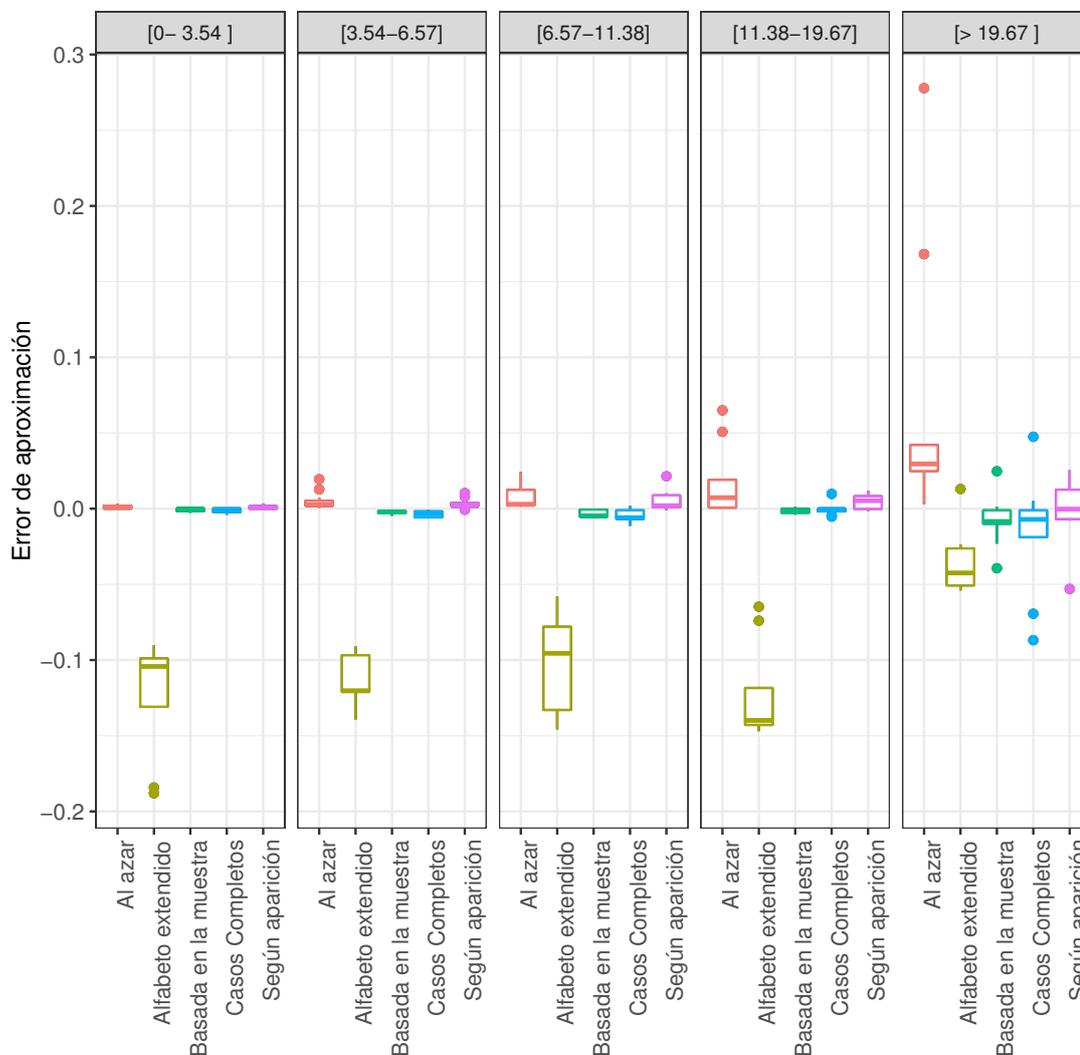


Figura 5.4 **Boxplot del error de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para cada una de las metodologías que intentan resolver el problema de los valores repetidos, para la dimensión de *embedding* $m = 6$, con los procesos caóticos agrupados para distintos niveles de η (%).** Para pequeños valores de η (%) todas las metodologías tienen un buen desempeño salvo extender el alfabeto. Al aumentar el valor de η (%) el error de aproximación aumenta, debido a que se perdió mucha información sobre el proceso, para todas las metodologías excepto para extender el alfabeto que disminuye su error, aunque de igual manera su desempeño es el peor.

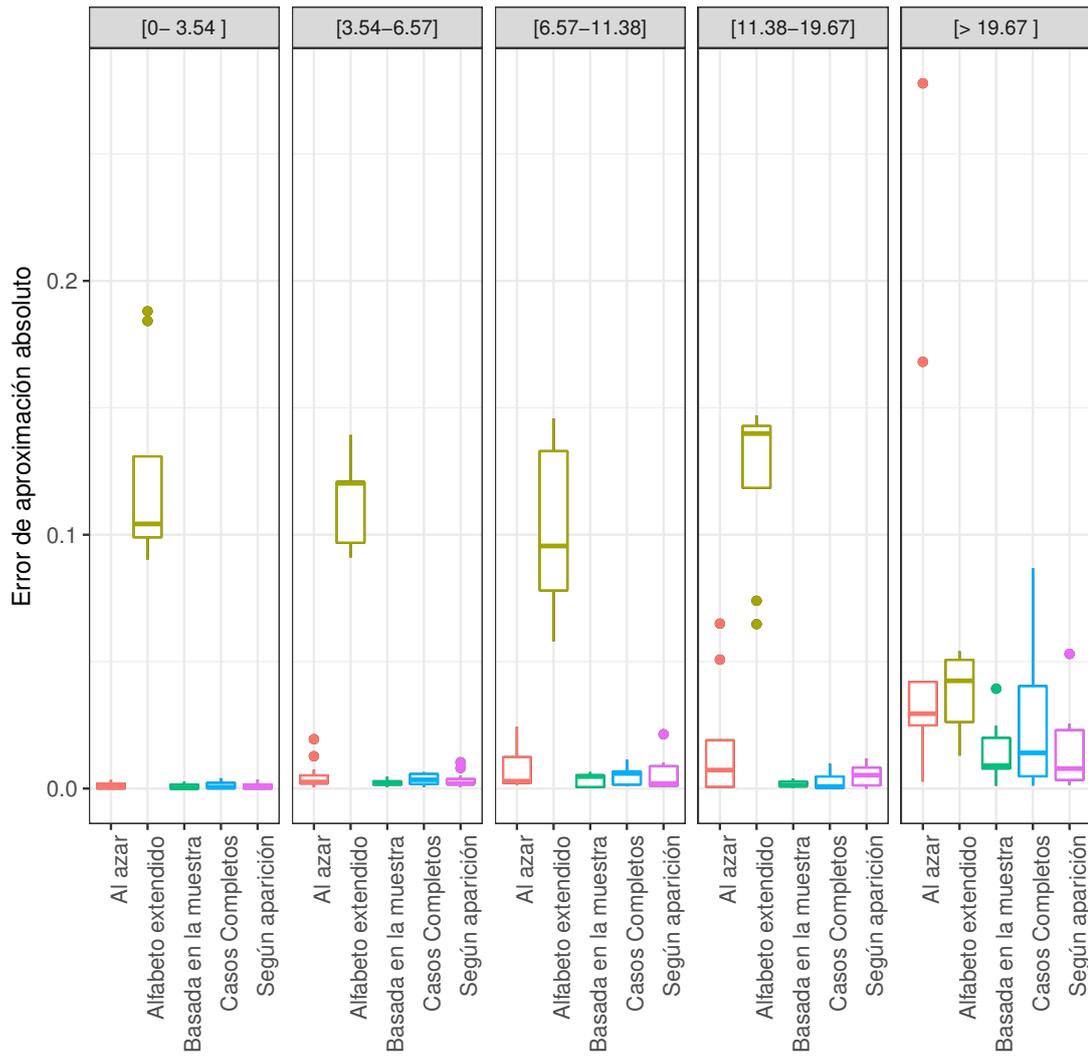


Figura 5.5 Ídem a la Figura 5.4 pero para el error de aproximación absoluto $|\hat{\mathcal{H}} - \mathcal{H}|$.

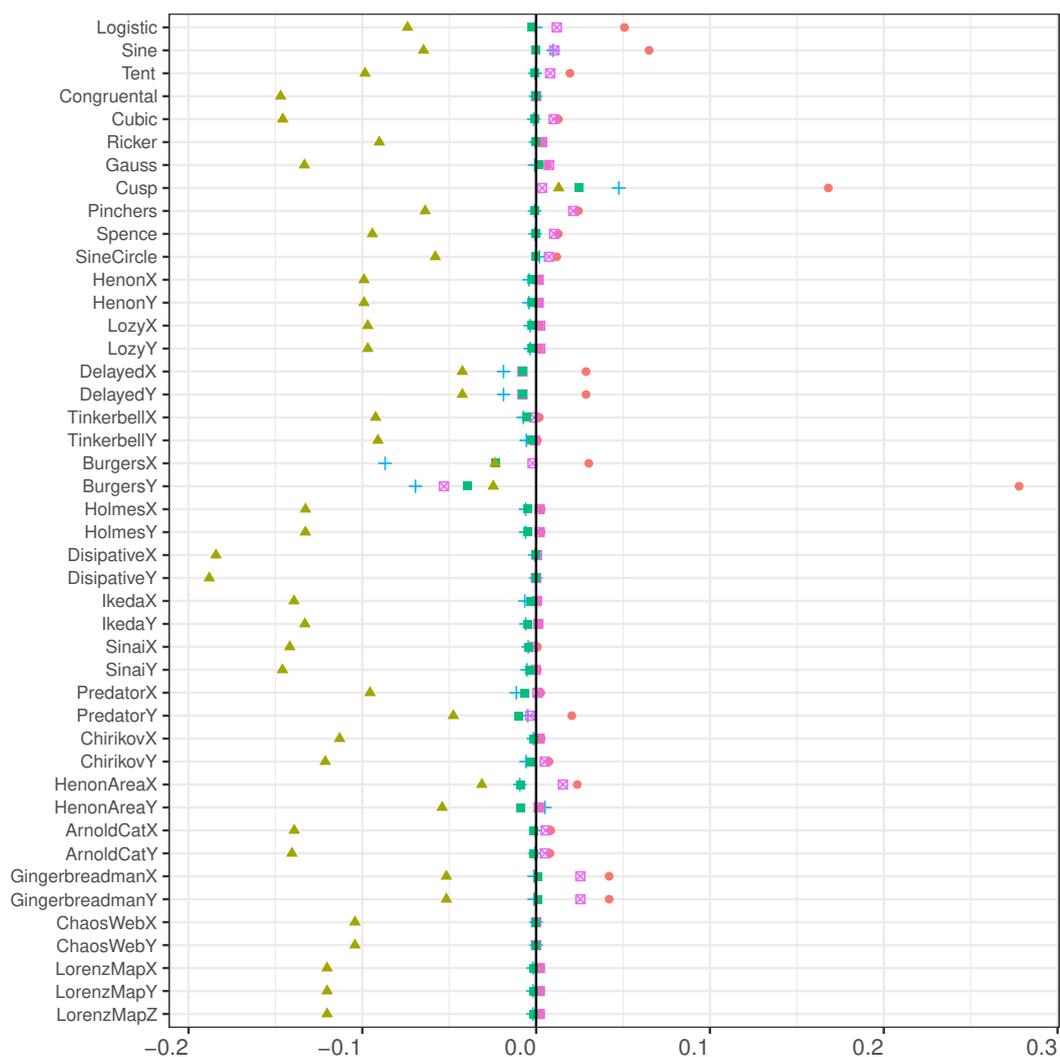


Figura 5.6 El error de aproximación $\tilde{\mathcal{H}} - \mathcal{H}$ para todos los procesos caóticos simulados. En verde oscuro: la metodología de extender el alfabeto, en verde claro: la Imputación Basada en la Muestra, en rojo: la Imputación al Azar y en violeta: la Imputación según el Orden de Aparición.

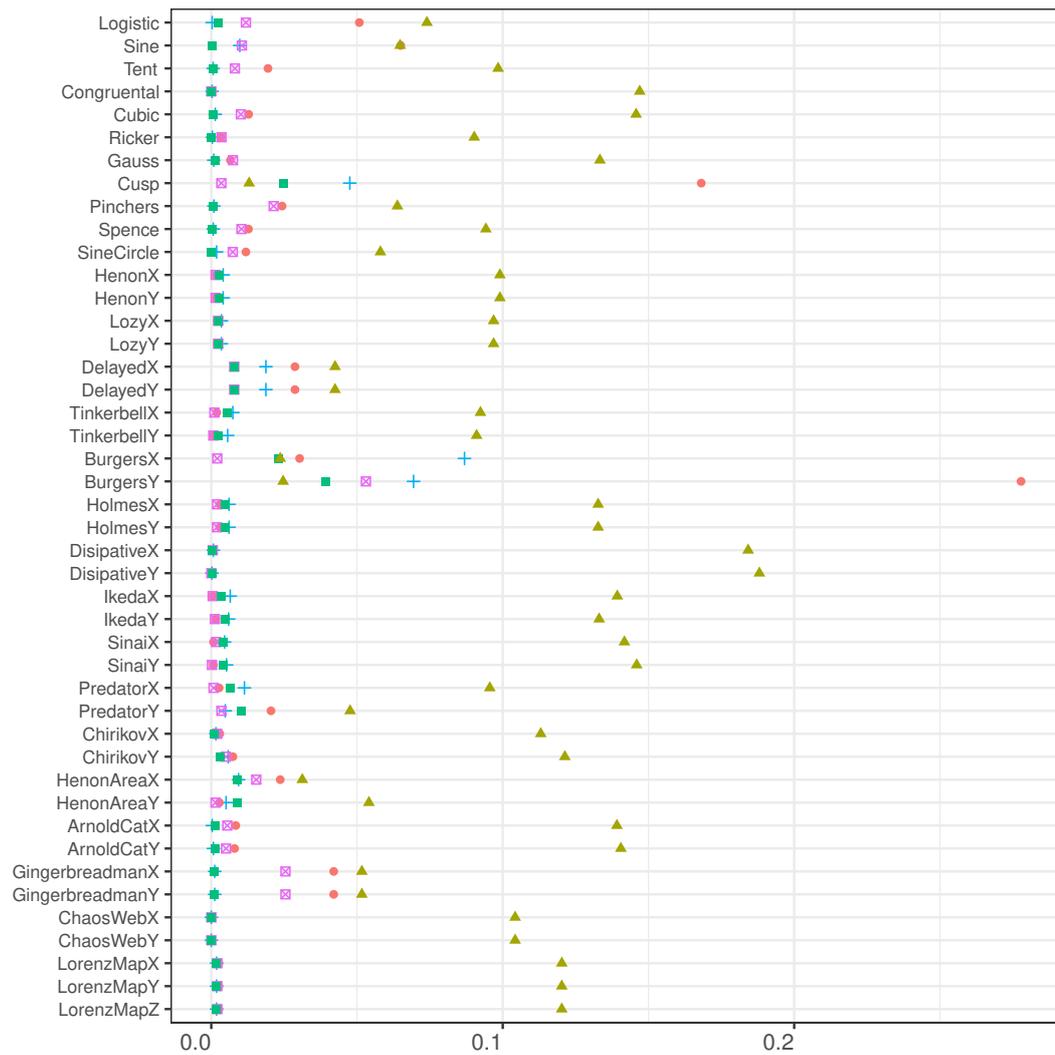


Figura 5.7 Ídem a la Figura 5.6 pero para el error de aproximación absoluto.

5.5.2. Números trascendentales

El interés de esta aplicación se centra en estudiar la aleatoriedad de la expansión decimal de números irracionales usando la Entropía de Permutación, con las distintas metodologías presentadas en este Capítulo, de las secuencias temporales obtenidas al recolectar los primeros 10^6 dígitos de la expansión decimal de los números π , e y $\sqrt{2}$. Si bien las igualdades dentro de las componentes de los vectores de *embedding* provenientes de estas series no son “información faltante”, es de interés mostrar el desempeño de la Entropía de Permutación calculada con la metodología de la Imputación Basada en la Muestra.

En particular, para estos tres números irracionales se consideran las series de largo $T = \{5 \times 10^3, 10^4, 10^5, 10^6\}$ de los primeros dígitos y se evalúan las Entropías de Permutación resultantes para una dimensión de *embedding* $m = 3, 4, 5, 6$ y retardo de tiempo $\tau = 1$.

Los Cuadros 5.5, 5.6 y 5.7 presentan los resultados obtenidos para π , e y $\sqrt{2}$ respectivamente. En estos Cuadros se reportan: la cantidad de vectores de *embedding* de dimensión m que presentan componentes con igual valor $-n^*$; el porcentaje de vectores $X_t^{(m)}$ con valores repetidos sobre el total de vectores de *embedding* $-\eta(\%)$ - de la serie de tiempo de largo T , y los valores de la Entropía de Permutación $\tilde{\mathcal{H}} = \mathcal{H}(\tilde{\mathbf{P}})$ para las dimensiones de *embedding* $m = \{3, 4, 5, 6\}$, evaluada según la FDP de BP usando los métodos de Casos Completos, Imputación según el Orden de Aparición, Imputación al Azar, Imputación Basada en la Muestra y la Entropía de Permutación evaluada según la FDP que se obtiene al extender el alfabeto. También se reportó para cada caso el número de patrones faltantes o prohibidos (ver Capítulo 2, Sección 2.6).

Notar que si bien η^* es una función creciente con respecto al largo de la serie de tiempo T y de la dimensión de *embedding* m , el porcentaje $\eta(\%)$ es prácticamente independiente de T para un m fijo. Más aún, se puede ver por los resultados expuestos en los Cuadros 5.5, 5.6 y 5.7 que este valor de $\eta(\%)$ presenta valores similares para los tres números racionales estudiados, comportamiento que puede

estar ligado a que los dígitos usados en la expansión decimal son del 0 al 9 y en teoría están dispuestos de forma completamente aleatoria.

Cuando se observan los resultados obtenidos para $\tilde{\mathcal{H}}$ evaluada con la FDP de BP resultante de la Imputación según el Orden de Aparición, en particular para valores de $m \geq 4$ e independientemente del largo de la serie T se puede advertir una aparente estructura temporal compleja en los datos. Notablemente, estos valores son similares para las expansiones decimales de los tres números irracionales. Este efecto es una consecuencia de la alta frecuencia de vectores de *embedding* con empates en sus rangos (Zunino et al. [2017]).

La Entropía de Permutación evaluada según la FDP obtenida cuando se extiende el alfabeto muestra una estructura de correlaciones aún más compleja para dimensiones de *embedding* $m \leq 4$. El mismo cuantificador $\tilde{\mathcal{H}}$, cuando se aplica a la FDP de BP resultante de la Imputación Basada en la Muestra no muestra ningún tipo de estructura de correlación en los datos, indicando que la expansión decimal de los números irracionales tiene un comportamiento completamente aleatorio, sin autocorrelaciones, concordante con los resultados obtenidos utilizando una metodología de *Visibility Graph* (Luque et al. [2009]). Se puede notar también que resultados similares se obtienen cuando se utiliza la FDP de BP con los Casos Completos, pero sin embargo hay una fuerte reducción en la cantidad de vectores de *embedding* resultantes implicando que series de tiempo más largas son necesarias para alcanzar la ausencia de patrones faltantes o prohibidos, como en el caso de $m = 6$.

Expansión decimal del número π

T	m	n^*	η [%]	Ext. s/rangos		Casos Completos		Imp. s/ Orden		Imp. al Azar		Imp. Basada en la Muestra	
				\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$
5×10^3	3	1332	26,64	0,933630713	0	0,999983033	0	0,993241941	0	0,999978965	0	0,999973708	0
	1×10^4	2714	27,14	0,935493651	0	0,999956462	0	0,991949033	0	0,99995944	0	0,99993972	0
	1×10^5	27911	27,91	0,939271256	0	0,999988701	0	0,99199081	0	0,999994868	0	0,999991366	0
	1×10^6	279616	27,96	0,939425633	0	0,999999289	0	0,991634948	0	0,999999689	0	0,999999483	0
5×10^3	4	2423	48,46	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$
	1×10^4	4901	49,01	0,970003434	0	0,999083838	0	0,990384203	0	0,999694589	0	0,999248336	0
	1×10^5	49535	49,53	0,972000423	0	0,999628671	0	0,989598352	0	0,9982839	0	0,999639847	0
	1×10^6	495599	49,56	0,973739377	0	0,999950456	0	0,989752703	0	0,999983808	0	0,999995625	0
5×10^3	5	3470	69,40	0,97396508	0	0,999993518	0	0,989473294	0	0,999998438	0	0,999995625	0
	1×10^4	6960	69,60	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$
	1×10^5	69915	69,91	0,980598544	0	0,990692144	0	0,985338717	0	0,998567329	0	0,99422764	0
	1×10^6	697637	69,76	0,986061657	0	0,995754267	0	0,985619302	0	0,999385206	0	0,997368567	0
5×10^3	6	4241	84,82	0,989922188	0	0,999518569	0	0,987094212	0	0,999937429	0	0,999703403	0
	1×10^4	8485	84,85	0,990327823	0	0,99995402	0	0,986937835	0	0,999993333	0	0,999973337	0
	1×10^5	85139	85,14	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$
	1×10^6	849015	84,90	0,93273125	1686	0,912393855	259	0,972623954	12	0,994719588	0	0,9487366453	0
5×10^3	6	8485	84,85	0,965861198	640	0,956891046	93	0,977989801	1	0,997588542	0	0,971986907	46
	1×10^4	85139	85,14	0,994241993	1	0,996105459	0	0,984057686	0	0,999746689	0	0,997787493	0
	1×10^5	849015	84,90	0,996752409	0	0,999642595	0	0,984054591	0	0,999974682	0	0,999802363	0
	1×10^6												

Cuadro 5.5 **Resultados para la expansión decimal del número π** . Se reporta la cantidad de vectores de *embedding* de dimensión m que presentan componentes con valores iguales n^* , el porcentaje de estos vectores en relación a la cantidad de patrones observados η , para la serie de tiempo de largo T y los valores obtenidos de $\tilde{\mathcal{H}}$, evaluados con las distintas metodologías presentadas en este Capítulo: extendiendo el alfabeto según rangos, Casos Completos, Imputación según el Orden de Aparición, la Imputación al Azar y la Imputación Basada en la Muestra, como también el número de patrones prohibidos o faltantes ($NPPF$) en cada caso.

Expansión decimal del número e														
T	m	n^*	η [%]	Ext. s/rangos		Casos Completos		Imp. s/ Orden		Imp. al Azar		Imp. Basada en la Muestra		
				\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	
5×10^3	3	1404	28,08	0,939855611	0	0,999908248	0	0,991383284	0	0,999965621	0	0,999938519	0	
	1×10^4	2834	28,34	0,940270056	0	0,999961926	0	0,991654981	0	0,999984091	0	0,999974069	0	
	1×10^5	28113	28,11	0,940035094	0	0,999993385	0	0,99119218	0	0,999994613	0	0,9999932	0	
	1×10^6	279055	27,91	0,939113326	0	0,999998017	0	0,991678797	0	0,999998969	0	0,999998368	0	
5×10^3	4	2463	49,26	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	
	1×10^4	4962	49,62	0,972004023	0	0,998405105	0	0,988365038	0	0,999503388	0	0,998945166	0	
	1×10^5	49615	49,62	0,973473842	0	0,999112427	0	0,988967037	0	0,999808044	0	0,999575734	0	
	1×10^6	494965	49,50	0,97419896	0	0,999936691	0	0,988818611	0	0,99981488	0	0,999959068	0	
5×10^3	5	3468	69,36	0,973720241	0	0,999991569	0	0,989603095	0	0,999997053	0	0,99999471	0	
	1×10^4	6977	69,77	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	
	1×10^5	69851	69,85	0,980691878	0	0,991670342	0	0,984363339	0	0,998444901	0	0,994357956	0	
	1×10^6	5	696451	69,65	0,985934034	0	0,99563841	0	0,98552295	0	0,999276066	0	0,997387771	0
		5	696451	69,65	0,990221357	0	0,999628731	0	0,986185864	0	0,999938628	0	0,999784352	0
5×10^3	6	4230	84,60	0,990199385	0	0,999956277	0	0,98705778	0	0,999992739	0	0,999973223	0	
	1×10^4	8489	84,89	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	
	1×10^5	85102	85,10	0,932508591	1683	0,92148706	240	0,971655603	9	0,994550881	0	0,949427685	65	
	1×10^6	6	847870	84,79	0,965907509	617	0,960524765	85	0,97767609	0	0,997490386	0	0,972847146	52
		6	847870	84,79	0,99431399	1	0,996291932	0	0,9827214	0	0,999768651	0	0,997957587	0
	1×10^6	6	847870	84,79	0,996702185	0	0,999628684	0	0,984099807	0	0,999976963	0	0,999791651	0

Cuadro 5.6 Resultados para la expansión decimal del número e

Expansión decimal del número $\sqrt{2}$													
T	m	n^*	η [%]	Ext. s./rangos		Casos Completos		Imp. s./ Orden		Imp. al Azar		Imp. Basada en la Muestra	
				\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$	\mathcal{H}	$NPPF$
5×10^3	3	1410	28,20	0,941886378	0	0,999747074	0	0,992048353	0	0,999872825	0	0,999800555	0
	1×10^4	2717	27,17	0,936903686	0	0,999929546	0	0,99226800	0	0,999961546	0	0,999939491	0
	1×10^5	27847	27,84	0,93910470	0	0,999989334	0	0,991574560	0	0,99994327	0	0,999990761	0
	1×10^6	280251	28,02	0,93969118	0	0,999997982	0	0,991660474	0	0,99999129	0	0,999998336	0
5×10^3	4	2524	50,48	0,974147301	0	0,998989711	0	0,990015459	0	0,999614340	0	0,999250923	0
	1×10^4	4905	49,05	0,972247139	0	0,999548691	0	0,99057375	0	0,999849033	0	0,999656636	0
	1×10^5	49489	49,49	0,973523101	0	0,999920019	0	0,989687644	0	0,999973675	0	0,999930760	0
	1×10^6	496216	49,62	0,974042837	0	0,999991682	0	0,989554206	0	0,999997243	0	0,999993702	0
5×10^3	5	3554	71,08	0,982050700	3	0,991309030	0	0,986103863	0	0,998617420	0	0,994522113	0
	1×10^4	6926	69,26	0,985511540	0	0,996496298	0	0,987507165	0	0,999397201	0	0,997836410	0
	1×10^5	69760	69,76	0,989721649	0	0,999569330	0	0,987054171	0	0,999924013	0	0,999713923	0
	1×10^6	698095	69,81	0,990350555	0	0,999960754	0	0,98697109	0	0,999991625	0	0,999976071	0
5×10^3	6	4297	85,94	0,933328020	1667	0,909654948	273	0,973946014	12	0,995114392	0	0,949352450	52
	1×10^4	8472	84,72	0,965147391	653	0,959981670	86	0,980467820	1	0,997786442	0	0,9729313798	59
	1×10^5	84956	84,96	0,994149140	0	0,996167476	0	0,98381474	0	0,999750330	0	0,9977802752	0
	1×10^6	849625	84,96	0,996745451	0	0,999609416	0	0,984012768	0	0,999973801	0	0,999792631	0

Cuadro 5.7 Resultados para la expansión decimal del número $\sqrt{2}$.

5.6. Conclusiones del Capítulo

En este Capítulo se expusieron todas las metodologías encontradas en la literatura para tratar con series de tiempo que en sus vectores de *embedding* presentan componentes con valores iguales. Extender el alfabeto para incluir estos vectores de *embedding* con empates en los rangos probó no ser una buena herramienta para tratar con esta problemática. El alfabeto simbólico extendido con el mapeo según el orden cronológico expuesto por [Bian et al. \[2012\]](#) no es una transformación 1 a 1 por lo que los patrones no pueden ser reconstruidos inambiguamente a partir de la secuencia de símbolos. Este problema teórico se soluciona extendiendo el alfabeto con el mapeo según rangos, aunque en las aplicaciones se muestra que no tiene un buen desempeño.

Una mejor estrategia para solucionar el problema que se presenta en este Capítulo es tratar a los vectores con valores repetidos como información faltante, o corrupta, de un vector “original” sin empates en sus rangos. Eliminar todos los patrones que presenten empates (Casos Completos) y calcular la Entropía de Permutación resultante es una buena aproximación ya que no altera la estructura original de la serie, pero precisa que el largo de la serie T sea lo suficientemente grande como para recuperar la información perdida debido a la eliminación de datos.

La Imputación al Azar sobrestima la Entropía de Permutación ya que le añade un ruido blanco para romper las igualdades, y también enmascara a los patrones prohibidos o faltantes, que pueden ayudar a describir la dinámica de un proceso.

La Imputación según el Orden de Aparición refleja, en la Entropía de Permutación resultante, una complejidad en la estructura temporal de autocorrelaciones ficticia en series de tiempo completamente aleatorias.

Finalmente, la Entropía de Permutación calculada con la FDP de BP aproximada con la Imputación Basada en la Muestra no presenta dinámicas ajenas a la propia serie de tiempo, recupera información proveniente de los patrones con empates, y no enmascara los patrones perdidos o faltantes, por lo que se recomienda su uso para series de tiempo tomadas mediante un instrumento de baja resolución

y se alienta el estudio de su desempeño para series de tiempo provenientes de un proceso generador de datos discretos.

Capítulo 6

Propiedades estadísticas del estimador de la Entropía de Permutación

6.1. Introducción

Las propiedades estadísticas clásicas de los cuantificadores, en cuanto estimadores, usados en dinámicas no lineales para caracterizar series de tiempo han sido poco estudiadas según lo encontrado en la literatura. La investigación hecha sobre la distribución del máximo exponente de Lyapunov y la dimensión de correlación ([Boeing \[2016\]](#)) es uno de los pocos estudios encontrados sobre este tema.

Esta falta de estudio se puede deber a que no hay una teoría acerca de las distribuciones de probabilidad sobre estos estimadores y quedan las técnicas de remuestreo, en especial el *bootstrap* ([Efron and Tibshirani \[1986\]](#)), como las más potentes para llevar a cabo esta tarea. Por ejemplo, se presenta una metodología para calcular la distribución empírica de los exponentes de Lyapunov basada en una técnica tradicional de remuestreo de bootstrap, proveyendo un test formal

de la existencia de caos bajo la hipótesis nula (Gençay [1996]). Sin embargo se muestra que esta técnica de bootstrap parece fallar en proveer límites confiables para las estimaciones de los exponentes de Lyapunov y se concluye que el bootstrap tradicional de remuestreo no puede ser aplicado para estimar estadísticos ergódicos multiplicativos (Ziehmann et al. [1999]).

Por otro lado, un procedimiento de bootstrap por bloques es usado para detectar exponentes de Lyapunov positivos, en series de tiempo financieras (Brzozowska-Rup and Orłowski [2004]). Sin embargo, las series de tiempo generadas con este remuestreo presentan artificios que son causados por la característica de este método de unir los bloques al azar, por lo que la correlación se preserva dentro, pero no entre los bloques.

En cuanto a la dinámica de la representación simbólica de las series de tiempo, las probabilidades generadas usando la metodología de Bandt y Pompe son calculadas analíticamente para procesos Gaussianos con largo de palabra igual a tres, pero reconocen que para largos mayores esto no es plausible, y por esa razón un método computacional sería necesario para estimar el sesgo y la varianza en la estimación de la Entropía de Permutación (Bandt and Shiha [2007]).

El objetivo de este Capítulo es proponer un método de remuestreo diferente, *el bootstrap paramétrico*, para estimar el sesgo y la varianza de la Entropía de Permutación de una única serie de tiempo y usar esta metodología para construir los intervalos de confianza y un test de hipótesis que pueda diferenciar mediante la Entropía de Permutación, series de tiempo provenientes de distintas dinámicas.

El método consiste en simular muestras de secuencias según la simbolización de Bandt y Pompe, provenientes de una serie de tiempo mediante un bootstrap paramétrico. Estas muestras son generadas por un proceso generador de datos que sigue un modelo probabilístico con las probabilidades de transición extraídas de la secuencia de símbolos proveniente de la serie de tiempo original.

Para mostrar los resultados de este novedoso método se hizo una simulación para una familia de series de tiempo conocidas, los ruidos $1/f^\alpha$. Se tomó un α y, para esta serie de tiempo, se calculó la varianza y la distribución del estimador de

la Entropía de Permutación mediante el método de Montecarlo, que consiste en repetir el experimento n veces y de esta manera estimar la distribución empírica del estimador de la Entropía de Permutación. Si n es lo suficientemente grande dicha distribución empírica se puede tomar como la verdadera distribución del estimador de la Entropía de Permutación para las series de tiempo provenientes de estos ruidos. Estos resultados se comparan con los obtenidos con la metodología del bootstrap paramétrico aplicada sobre una única realización de esta serie de tiempo. Este experimento se repite para cada valor de α .

Para ilustrar el método presentado en un problema real, se aplicó el bootstrap paramétrico para comparar señales de EEG normales con señales EEG pre-ictales.

6.2. El Enfoque Bootstrap

El bootstrap es un método estadístico computacional que sirve para asignar medidas de precisión a las estimaciones de los parámetros deseados (Efron and Tibshirani [1986]).

Si \mathcal{H} es una característica desconocida del modelo Ψ , un estimador $\hat{\mathcal{H}}$ puede ser extraído de la muestra generada por Ψ en un único experimento. Una manera de obtener la distribución de $\hat{\mathcal{H}}$ es repetir el experimento una gran cantidad de veces y aproximar la verdadera distribución de $\hat{\mathcal{H}}$ por la distribución empírica obtenida. En la mayoría de las situaciones prácticas, este método no es posible porque el experimento no es reproducible, o es impracticable por cuestiones de costo.

El espíritu de la metodología bootstrap es estimar la distribución muestral de un estadístico, con los datos de la muestra como analogía del “experimento real” que motivó la distribución muestral.

Supongamos que un modelo probabilístico desconocido Ψ genera un proceso aleatorio $\mathcal{X} = \{\mathcal{X}_t\}_{t=1}^T$ y sea $\hat{\theta}_\Psi(\mathcal{X}, T)$ el estadístico de interés que estima el verdadero valor de $\theta = f(\Psi)$. Con la muestra observada $\mathbf{x} = \{x_t\}_{t=1}^T$ se obtiene un estimador del modelo $\hat{\Psi}$.

Lo novedoso es que ahora repetimos el experimento utilizando al modelo estimado $\hat{\Psi}$ como un proceso generador de datos bootstrap $\mathcal{X}^* = \{\mathcal{X}_t^*\}_{t=1}^T$. Este proceso

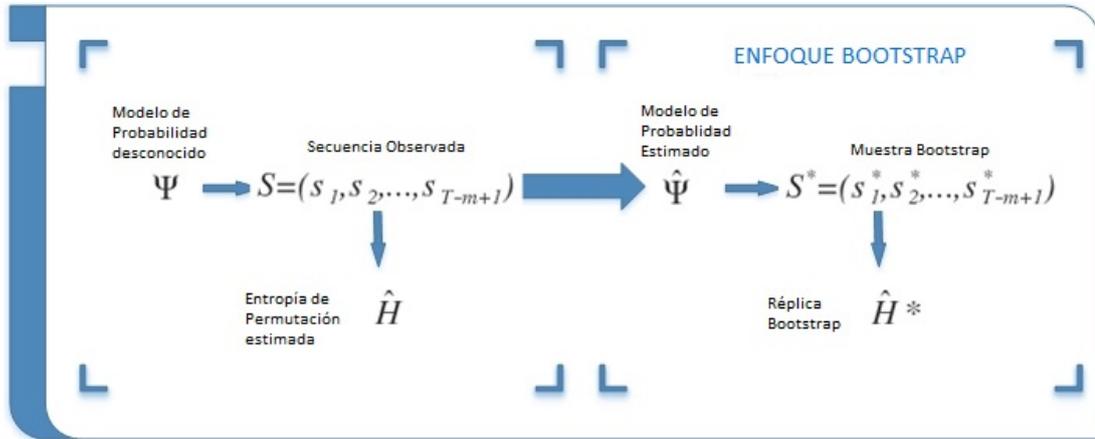


Figura 6.1 **Diagrama esquemático del enfoque del bootstrap paramétrico.** Un modelo probabilístico desconocido $\Psi = \Psi(P^{ij})$, genera una secuencia observada S de la cual se estima \hat{H} , por lo que el enfoque bootstrap sugiere estimar $\hat{\Psi} = \Psi(\hat{P}^{ij})$ y obtener las correspondientes muestras S^* de las cuales se estiman a su vez \hat{H}^*

genera la cantidad deseada de réplicas bootstrap $\mathbf{x}^* = \{x_t^*\}_{t=1}^T$ y para cada una se calcula $\hat{\theta}_{\hat{\Psi}}^*(\mathbf{x}^*, T)$ obteniendo la distribución de $\hat{\theta}_{\hat{\Psi}}^*(\mathcal{X}^*, T)$ que es el estimador bootstrap de la distribución de $\hat{\theta}_{\Psi}(\mathcal{X}, T)$. Con dicha estimación podemos obtener la varianza, el sesgo y los intervalos de confianza para nuestro parámetro desconocido.

Formalmente, la metodología bootstrap está basada en el principio “plug-in” (Efron and Tibshirani [1993]). El parámetro de interés puede ser escrito en función del modelo de probabilidad $\theta = f(\Psi)$. Como este modelo de probabilidad es desconocido, el estimador “plug-in” del parámetro es definido como $\hat{\theta} = f(\hat{\Psi})$. Por lo tanto el bootstrap propone que se generen las muestras con este modelo de probabilidad estimado que es cercano en algún sentido al verdadero Ψ .

Si se tiene información acerca de Ψ además de los datos, el $\hat{\Psi}$ elegido debe contener esta información. Si suponemos que los datos $\mathcal{X} = \{\mathcal{X}_t\}_{t=1}^T$ provienen de un cierto proceso regido por un modelo probabilístico Ψ que depende de un número finito de parámetros $\Xi = \{\xi\}_{i=1}^k$, entonces se puede expresar a este modelo como $\Psi = f(\Xi)$. Estos parámetros podrían ser estimados en una manera tradicional de estadística paramétrica, como la Máxima Verosimilitud obteniendo $\hat{\Xi} = \{\hat{\xi}\}_{i=1}^k$, y equivalentemente $\hat{\Psi} = f(\hat{\xi})$.

De esta manera, las muestras bootstrap $\mathbf{x}^* = \{x_1^*, x_2^* \dots x_n^*\}$ provienen de un proceso regido por un modelo probabilístico $\hat{\Psi}$ (Fig. 6.1). Estas muestras bootstrap emulan en todo sentido a las muestras originales incluyendo la correlación entre los valores.

6.2.1. Las probabilidades de transición de una secuencia de símbolos

Como se mostró en el Capítulo 2, usando la metodología propuesta por Bandt y Pompe, la dinámica de un proceso $\{\mathcal{X}_t\}_{t \in T}$ con $\mathcal{X}_t \in \mathbb{R}$ es representada por una secuencia de símbolos de $m!$ estados: $\{S_t\}_{t \in (T-m+1)}$ con $S_t \in S_m = \{\pi_1, \pi_2 \dots \pi_m\}$ para todos los posibles $m \geq 3$. La realización de esta secuencia de símbolos puede ser pensada como producida por un modelo probabilístico con una probabilidad de transición fija P^{ij} (i.e. la probabilidad de pasar de un símbolo π_i a un símbolo π_j para $1 \leq i \leq m!$ y $1 \leq j \leq m!$) denotado $\Psi(P^{ij})$.

Se puede ver en la Figura 6.1 que $\Psi = \Psi(P^{ij})$ genera una secuencia observada S de la cual se estima $\hat{\mathcal{H}}$, por lo que se necesita una estimación \hat{P}^{ij} para estimar el modelo $\hat{\Psi} = \Psi(\hat{P}^{ij})$ que va a ser usado para obtener las replicas bootstrap $\hat{\mathcal{H}}^*$.

La frecuencia relativa

$$\hat{P}_T(\pi_i) = \frac{n_i}{T - m + 1} \quad (6.1)$$

es un estimador "*tan bueno como sea posible para una serie finita de valores*" (Bandt and Pompe [2002]) de $P(\pi_i)$, con n_i la cantidad de veces que π_i es observado en la serie simbólica de largo $T - m + 1$. El subíndice T en $\hat{P}_T(\pi_i)$ refuerza la noción de la dependencia del estimador con respecto al largo de la serie original T .

Con el mismo espíritu, se definen las probabilidades de transición de la serie simbólica como:

$$P^{ij} = P(s_{t+1} = \pi_j | s_t = \pi_i) \quad 1 \leq i \leq j \leq m! \quad (6.2)$$

y el estimador P^{ij}

$$\hat{P}_T^{ij} = \begin{cases} \frac{n_{ij}}{n_i} & \text{if } n_i \geq 0 \\ 0 & \text{caso contrario} \end{cases} \quad (6.3)$$

donde n_{ij} es el número de transiciones observadas desde π_i a π_j en la serie simbólica de largo $T - m + 1$ y n_i el numero de veces que π_i es observada en dicha serie. Notar que $n_i = \hat{P}(\pi_i)(T - m + 1)$

Por lo tanto, por la ley de probabilidad total:

$$P(\pi_j) = \sum_{i=1}^{m!} P(\pi_i)P^{ij} \quad (6.4)$$

Si se llama $\mathbf{P}(\pi)$ al vector $(m!)$ -dimensional que contiene a $P(\pi_i)$ en cada coordenada (i.e $\mathbf{P}(\pi) = (P(\pi_1), P(\pi_2), \dots, P(\pi_{m!}))$), entonces $\mathbf{P}(\pi)$ es determinado por P^{ij} , llegando a la conclusión que el estimador $\hat{\mathbf{P}}_T(\pi)$ es determinado por la estimación de \hat{P}_T^{ij} .

6.2.2. El método bootstrap aplicado a la Entropía de Permutación

La Entropía de Permutación está definida en la ecuación 2.7, por lo tanto debido al principio “plug-in”, el estimador natural es:

$$\hat{\mathcal{H}}_T = \left\{ - \sum_{i=1}^N \hat{P}_T(\pi_i) \ln(\hat{P}_T(\pi_i)) \right\} / \ln(m!) \quad (6.5)$$

En la Sección 6.2.1 se mostró que las probabilidades $\mathbf{P}(\pi)$ están totalmente definidas por las probabilidades de transición P^{ij} , por lo que estas últimas pueden ser tomadas como parámetros del modelo probabilístico Ψ .

Siguiendo el esquema de la Figura 6.1 se tiene :

$$\Psi = \Psi(P^{ij}) \longrightarrow \mathbf{S} = (s_1, s_2, \dots, s_{T-m+1}) \longrightarrow \hat{\mathcal{H}}_T$$

El modelo probablístico con probabilidades de transición desconocidas P^{ij} genera la secuencia de símbolos observada \mathbf{S} , y con esta secuencia, se obtiene la estimación de la Entropía de Permutación.

En el enfoque bootstrap:

$$\hat{\Psi} = \Psi(\hat{P}_T^{ij}) \longrightarrow \mathbf{S}^* = (s_1^*, s_2^* \dots, s_{T-m+1}^*) \longrightarrow \hat{\mathcal{H}}_T^*$$

$\hat{\Psi}$ genera a \mathbf{S}^* mediante una simulación, generando la réplica bootstrap $\hat{\mathcal{H}}_T^*$. Se puede repetir esta simulación tantas veces como sea posible.

Computar B réplicas bootstrap de la Entropía de Permutación para una serie de tiempo $\{x_t\}_{t \in T}$ es simple: dada una serie de tiempo de longitud T , se elige un largo de palabra m y un lag τ para efectuar el mapeo desde $\{x_t\}_{t \in T}$ a $\{S_t\}_{t \in (T-m+1)}$ como se definió en el Capítulo 2.

Con esta secuencia: se computan $\hat{P}_T(\pi_i)$, (ecuación 6.1), \hat{P}_T^{ij} (ecuación 6.3) y se estima $\hat{\mathcal{H}}_T$ (ecuación 6.5).

Se empieza por elegir con probabilidad $\hat{P}_T(\pi)$ un estado inicial $s_1^*(1) = \pi_k$ y luego se elige al azar con probabilidad \hat{P}_T^{kj} (notar que k es fijo con el valor del estado previo) el próximo estado simulado $s_2^*(b)$. Se repite este último paso $T - m + 1$ veces para obtener la simulación $\mathbf{S}^*(\mathbf{1}) = (s_1^*(1), s_2^*(1) \dots, s_{T-m+1}^*(1))$.

Con esta réplica bootstrap de la secuencia de símbolos se estima $\hat{\mathcal{H}}_T^*(b)$ (Ecuación 6.1). Se repite B veces para obtener $\hat{H}_T^*(b)$ $b = 1 \dots B$. Con el conjunto $\hat{\mathcal{H}}_T^*(b)$ $b = 1 \dots B$ se tienen las réplicas bootstrap necesarias para estimar el sesgo, la varianza, los intervalos de confianza, o el test presentado en la próxima Sección.

En el Apéndice A se muestra el pseudocódigo para simular las réplicas bootstrap de la Entropía de Permutación y en el Apéndice C se muestra el código de R utilizado.

Teniendo las B réplicas bootstrap de $\hat{\mathcal{H}}_T^*$:

$$\hat{\mathcal{H}}_T^*(1), \hat{\mathcal{H}}_T^*(2) \dots \hat{\mathcal{H}}_T^*(B)$$

El desvío estándar bootstrap de $\hat{\mathcal{H}}_T^*$ es la estimación del desvío estándar de $\hat{\mathcal{H}}_T$:

$$\hat{\sigma}_B(\hat{\mathcal{H}}_T) = \hat{\sigma}(\hat{\mathcal{H}}_T^*) \quad (6.6)$$

y se define como

$$\hat{\sigma}(\hat{\mathcal{H}}_T^*) = \sqrt{\frac{1}{B-1} \sum_{i=1}^B (\hat{\mathcal{H}}_T^*(i) - \hat{\mathcal{H}}_T^*(\bullet))^2} \quad (6.7)$$

donde

$$\hat{\mathcal{H}}_T^*(\bullet) = \frac{1}{B} \sum_{i=1}^B \hat{\mathcal{H}}_T^*(i) \quad (6.8)$$

Definimos el sesgo bootstrap de $\hat{\mathcal{H}}_T^*$ como:

$$\text{Bias}(\hat{\mathcal{H}}_T^*) = \hat{\mathcal{H}}_T^*(\bullet) - \hat{\mathcal{H}}_T \quad (6.9)$$

Finalmente, el error cuadrático medio bootstrap como:

$$\text{MSE}(\hat{\mathcal{H}}_T^*) = \text{Var}(\hat{\mathcal{H}}_T^*) + \text{Bias}^2(\hat{\mathcal{H}}_T^*) \quad (6.10)$$

6.2.2.1. Intervalos de confianza

Los intervalos de confianza de nivel de significación α de \mathcal{H} se definen por los percentiles del δ de bootstrap. Para cada réplica bootstrap $\hat{\mathcal{H}}_T^*(b)$ se computa la diferencia $\delta^*(b)$ entre esa réplica y la media de todas las réplicas bootstrap. Luego se eligen los percentiles $(\frac{\alpha}{2})$ y $(1 - \frac{\alpha}{2})$ de la distribución de los δ^* y se le suman a la estimación original, $\hat{\mathcal{H}}_T$, corrigiendo por el sesgo, y el intervalo de confianza al $(1 - \alpha)100\%$ es:

$$\left[\max(2\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_T^*(\bullet) + \delta_{\frac{\alpha}{2}}^*, 0), \min(2\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_T^*(\bullet) + \delta_{(1-\frac{\alpha}{2})}^*, 1) \right] \quad (6.11)$$

En el Apéndice A se muestra el pseudocódigo para construir los intervalos de confianza para la Entropía de Permutación.

6.2.2.2. Test de Hipótesis

Con el mismo espíritu, se puede construir un intervalo de confianza para la diferencia entre la Entropía de Permutación de dos series de tiempo diferentes. En la estadística inferencial existe una relación directa entre intervalos de confianza y tests de hipótesis. Un intervalo de confianza bilateral de nivel de confianza $(1 - \alpha)$ en la diferencia entre dos medidas puede ser usado para determinar si dichas medidas son significativamente (en términos estadísticos) diferentes entre sí simplemente verificando si el *cero* pertenece al intervalo o no.

$$H_0 : \Delta = \mathcal{H}_1 - \mathcal{H}_2 = 0$$

Si $0 \notin (1 - \alpha)100\% CI(\Delta)$

se rechaza H_0 y

$$\mathcal{H}_1 \neq \mathcal{H}_2$$

El procedimiento para efectuar este test se muestra con mayor detalle en el Apéndice A

6.3. Simulación numérica

Para exponer y probar el método propuesto en series de tiempo generales, se simularon series de tiempo provenientes de sistemas dinámicos conocidos: los ruidos $1/f^\alpha$. Todas las series fueron simuladas para diferentes largos $-T-$ de manera de obtener una mejor evaluación de las propiedades estadísticas de $\hat{\mathcal{H}}_T$ de acuerdo con la Ecuación 6.5.

Como se expresó anteriormente, una manera de obtener la distribución de $\hat{\mathcal{H}}$ es repetir el experimento donde se generan los datos una gran cantidad de veces y aproximar la distribución de $\hat{\mathcal{H}}$ por la distribución empírica obtenida. Mientras que para experimentos reales ésto puede ser inaplicable, para series de tiempo simuladas se puede hacer sencillamente mediante una simulación de Montecarlo

Una vez obtenidas las n réplicas de $\hat{\mathcal{H}}_T = \{\hat{\mathcal{H}}_T(1), \dots, \hat{\mathcal{H}}_T(n)\}$, la desviación estándar es estimada mediante

$$\hat{\sigma}(\hat{\mathcal{H}}_T) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\mathcal{H}}_T(i) - \hat{\mathcal{H}}_T(\bullet))^2} \quad (6.12)$$

donde

$$\hat{\mathcal{H}}_T(\bullet) = \frac{1}{n} \sum_{i=1}^n \hat{\mathcal{H}}_T(i) \quad (6.13)$$

6.3.1. Diseño experimental

Los ruidos $1/f^\alpha$ refieren a una señal con densidad espectral $S(f)$ con la forma $S(f) = k \frac{1}{f^\alpha}$ donde k es una constante, α es un parámetro dependiente de la señal y f es la frecuencia (Kasdin [1995]). Es un modelo estocástico que aparece frecuentemente en la naturaleza (ver Kasdin [1995] y sus referencias)

Se simularon ruidos $1/f^\alpha$ con $\alpha = \{-1, 0, 1, 2\}$. La Figura 6.2 muestra un ejemplo de estas señales.

Un proceso de ruido blanco ($\alpha = 0$) generará una curva con una potencia constante en el espectro. El caso $\alpha = 1$ o *ruido rosa* es el caso canónico y de mucho interés ya que muchos de los procesos de la naturaleza tienen un exponente cercano a 1.0 (Caloyannides [1974]; Dutta and Horn [1981]; Kobayashi and Musha [1982]; Novikov et al. [1997]; Voss and Clarke [1975]). Una caminata aleatoria continua (Ruido Browniano o ruido rojo, $\alpha = 2$) presentará una curva del tipo $(1/f^2)$ en $S(f)$. Para simular estos procesos estocásticos, se usó el algoritmo propuesto por Timmer and Koenig [1995].

Para cada $\alpha = \{-1, 0, 1, 2\}$ se simularon 1000 réplicas para cada $T = \{60, 100, 120, 400, 600, 2000, 3600, 5000, 10000, 20000, 50000\}$.

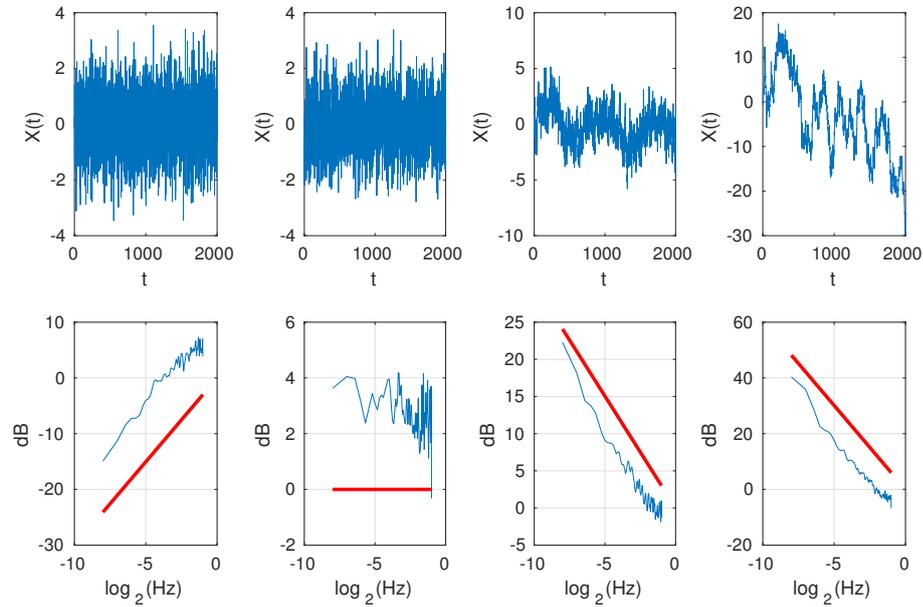


Figura 6.2 **Realización de los ruidos $1/f^k$ y su densidad espectral.** (arriba) Una realización de los ruidos $1/f^k$ ($T=2000$), de izquierda a derecha: $k = -1, k = 0, k = 1, k = 2$. (abajo) Densidad espectral de los ruidos $1/f^k$.

Para cada $m = \{3, 4, 5, 6\}$ se obtienen $\hat{\mathcal{H}}_i(T, m)$ donde $\{i = 1, \dots, 1000\}$ junto su desvío $\hat{\sigma}(\hat{\mathcal{H}}_T)$.

Luego, para cada $\alpha = \{-1, 0, 1, 2\}$ una única serie de tiempo fue simulada para cada $T = \{60, 100, 120, 400, 600, 2000, 3600, 5000, 10000, 20000, 50000\}$. En cada serie se utilizó el algoritmo 1 para obtener 1000 réplicas bootstrap, $\hat{\mathcal{H}}_i^*(T, m)$ para $\{b = 1 \dots 1000\}$ para cada $m = \{3, 4, 5, 6\}$. Para estas muestras bootstrap se analizó el sesgo (Ecuación 6.9), el desvío estándar (Ecuación 6.7) y MSE (Ecuación 6.10).

En lo que respecta a los intervalos de confianza, como para set de 1000 réplicas bootstrap se obtiene un único intervalo de confianza, se repite este paso 50 veces para obtener el Cuadro 6.1 que indica el nivel de confianza aproximado obtenido por este método junto con la amplitud media de estos intervalos.

6.3.2. Resultados

La intención del experimento de simulación es mostrar que la distribución obtenida mediante el bootstrap es cercana, en todo sentido relevante, a la distribución obtenida mediante la repetición del experimento original (distribución empírica), y puede ser usada cuando el experimento no puede ser exactamente replicado.

La Figura 6.3 muestra una comparación entre el desvío estándar de las réplicas bootstrap, con las hechas mediante la simulación de Montecarlo ($\hat{\sigma}_B(\hat{\mathcal{H}}_T)$ y $\hat{\sigma}(\hat{\mathcal{H}}_T)$ respectivamente) para los distintos procesos estocásticos. Se encuentran algunas discrepancias para valores pequeños de T (debido a que posiblemente no hay suficientes datos para una buena estimación de las probabilidades de transición), pero a partir de un cierto valor de T , en todos los casos, la desviación estándar coincide.

En la Figura 6.4 se puede ver que para cada m y α , el sesgo de la estimación bootstrap tiende a cero a medida que T aumenta. De esta manera, se puede decir que el estimador bootstrap de la Entropía de Permutación es un estimador asintóticamente insesgado. Con este último resultado y con el hecho que la desviación estándar tienda a cero cuando T aumenta, este estimador bootstrap pareciera ser un estimador consistente en media cuadrática. Y más aún, para valores grandes de T , el estimador bootstrap es tan eficiente como la estimación producida mediante la repetición del experimento.

En la Figura 6.5, para un valor de $\alpha = 1$, y para el mayor largo simulado $T = 50000$, se presenta el histograma de un set de réplicas bootstrap junto con el histograma del estimador obtenido por la simulación de Montecarlo, en diferentes escalas para cada m . En esta figura se puede apreciar la similitud de las formas de los histogramas. Cabe notar que la diferencia en la posición se debe a que las muestras bootstrap dependen de sólo una de las estimaciones de la Entropía de Permutación (que son aleatorias), pero esto no afecta las conclusiones inferenciales que pueden ser obtenidas a partir de este estimador

m	α	$\hat{\mathcal{H}}(T, m)$	Fallo por menor	Fallo por mayor	Amplitud media
3	-1	0.995831848	0	0.04	0.00222
4	-1	0.989083439	0.02	0.04	0.00420
5	-1	0.983495069	0.04	0.04	0.00500
6	-1	0.97547007	0.02	0	0.00555
3	0	0.99990292	0	0	0.00057
4	0	0.999679839	0	0	0.00080
5	0	0.998800463	0	0	0.00134
6	0	0.994503528	0	0	0.00235
3	1	0.991622896	0.02	0.02	0.00340
4	1	0.983385433	0.02	0.02	0.00493
5	1	0.97600538	0.02	0.04	0.00591
6	1	0.966355927	0	0.06	0.00657
3	2	0.943233315	0.08	0.06	0.00959
4	2	0.90634703	0.04	0.02	0.01273
5	2	0.878452628	0.02	0.02	0.01413
6	2	0.853327039	0.04	0.02	0.01482

Cuadro 6.1 **Intervalos de confianza con un nivel de confianza del 90 % para cada m y cada α de los ruidos $1/f^k$.** Se computó cuántas veces el valor real de \mathcal{H} -se utilizó la media de $\hat{\mathcal{H}}(T, m)$ - quedó fuera de los intervalos de confianza, es decir menor al límite inferior - *Fallo por menor* - o mayor al límite superior - *Fallo por mayor*-. Para el ruido blanco el nivel de confianza es mayor que el 90 % (de hecho es siempre acertado) pero para otros valores de α el nivel de confianza general se encuentra entre un 85 % y un 95 %.

Para un estudio más meticuloso del estimador bootstrap se calcularon cincuenta intervalos de confianza con un nivel de confianza del 90 % para cada m y cada α y se computó cuántas veces el valor real de \mathcal{H} (de hecho se utilizó la media de $\hat{\mathcal{H}}(T, m)$ que es el estimador de máxima verosimilitud) quedó fuera de los intervalos de confianza. Los resultados se muestran en el Cuadro 6.1. Para el ruido blanco el nivel de confianza es mayor que el 90 % propuesto, pero para otros valores de α el nivel de confianza general se encuentra entre un 85 % y un 95 %.

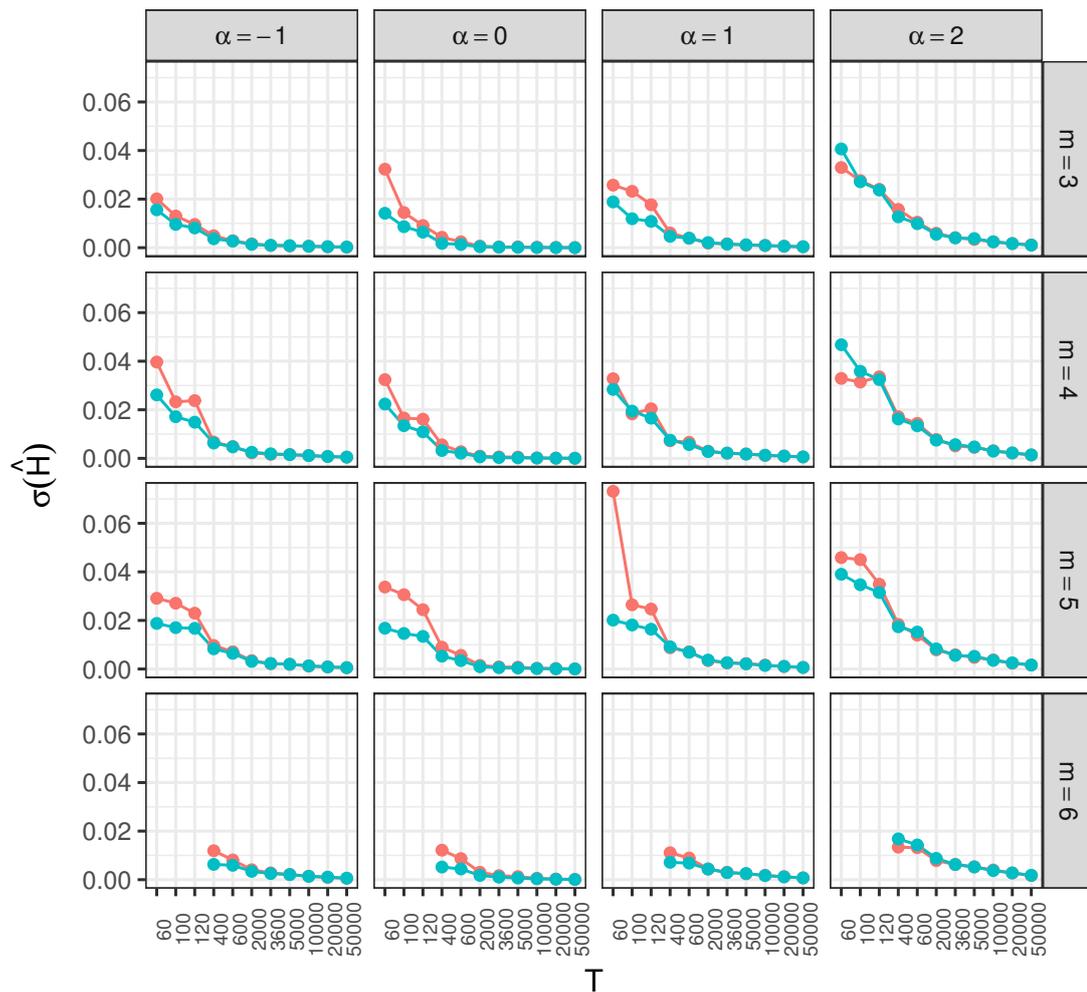


Figura 6.3 **Desviación estándar de $\hat{\mathcal{H}}$ en función de T .** Comparación de los desvíos estándar de $\hat{\mathcal{H}}$ de las réplicas bootstrap en rojo y las réplicas simuladas por Montecarlo en azul. Se encuentran algunas discrepancias para valores pequeños de T (debido a que posiblemente no hay suficientes datos para una buena estimación de las probabilidades de transición), pero a partir de un cierto valor de T , en todos los casos, la desviación estándar coincide.

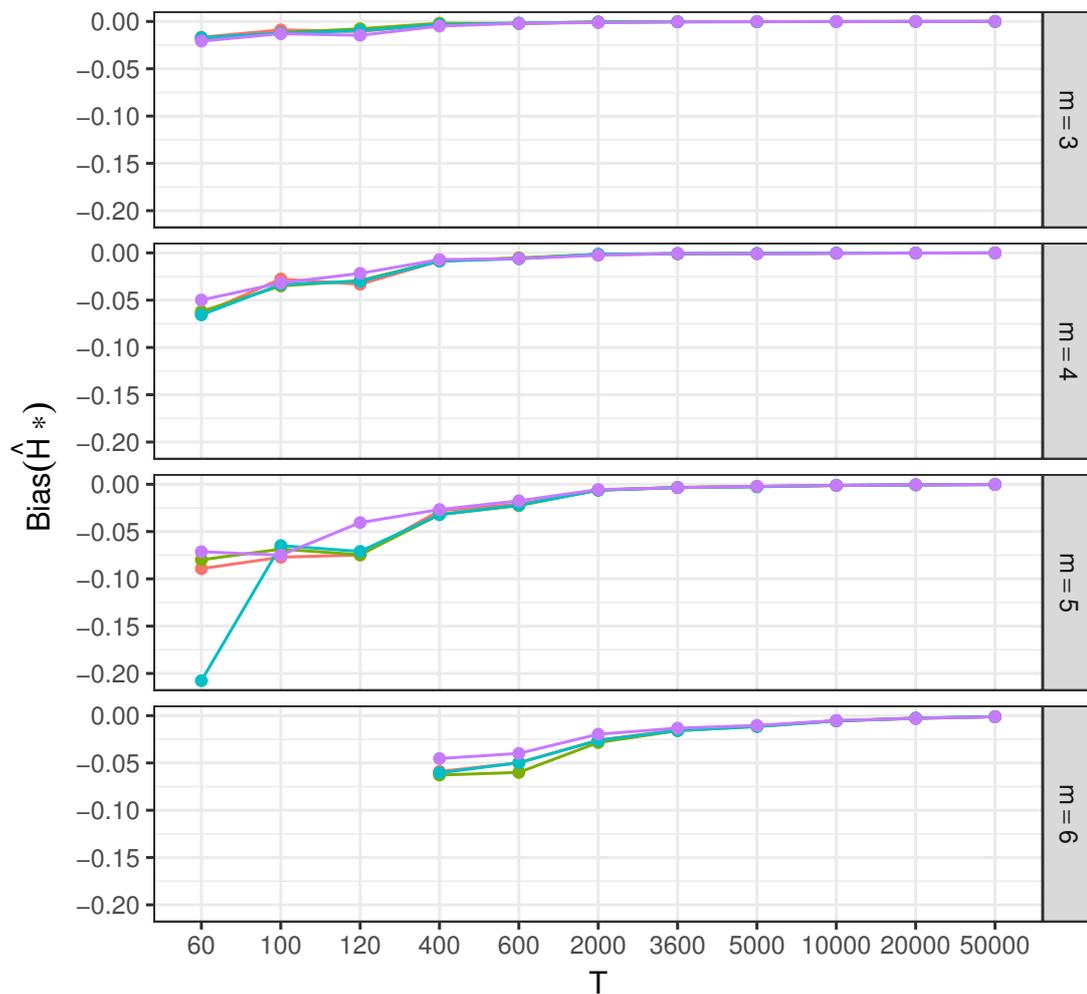


Figura 6.4 El sesgo bootstrap para diferentes valores de α en función de T . Se puede ver que para cada m y α , el sesgo de la estimación bootstrap tiende a cero a medida que T aumenta. De esta manera, se puede decir que el estimador bootstrap de la Entropía de Permutación es un estimador asintóticamente insesgado. Con este último resultado y con el hecho que la desviación estándar tienda a cero cuando T aumenta, este estimador bootstrap pareciera ser un estimador consistente en media cuadrática.

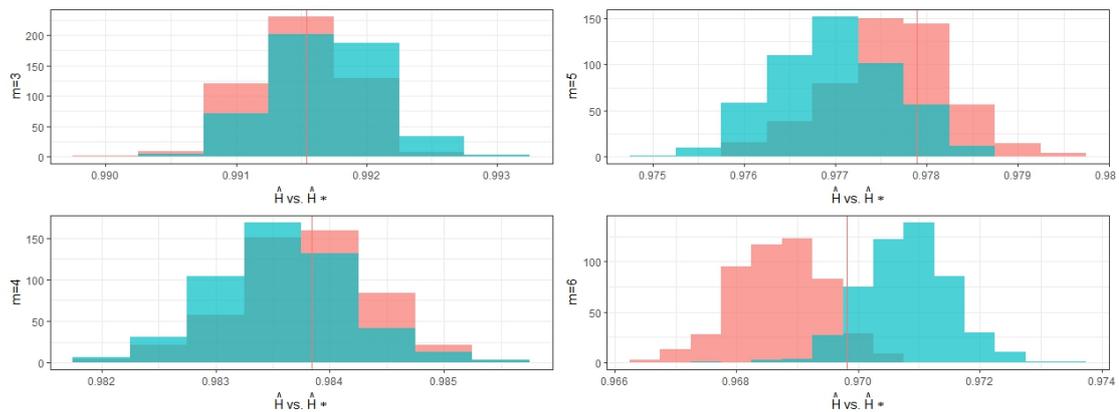


Figura 6.5 **Histograma para un set de réplicas bootstrap.** Para un valor de $\alpha = 1$, y para el mayor largo simulado $T = 50000$, se presenta el histograma de un set de réplicas bootstrap (en rojo) junto con el histograma del estimador obtenido por la simulación de Montecarlo (en azul), en diferentes escalas para cada m . En esta Figura se puede apreciar la similitud de las formas de los histogramas. Cabe notar que la diferencia en la posición se debe a que las muestras bootstrap dependen de sólo una de las estimaciones de la Entropía de Permutación (que son aleatorias), pero ésto no afecta las conclusiones inferenciales que pueden ser obtenidas de este estimador.

6.4. Aplicación: datos EEG

Para ilustrar la utilización de los intervalos de confianza y test de hipótesis propuestos en un contexto real, se presenta cómo pueden describir la incertidumbre en la estimación de la Entropía de Permutación dentro de una única observación de una señal de EEG. Más precisamente como una primera aplicación, se analizaron cuatro conjuntos diferentes de EEG para pacientes sanos y pacientes con epilepsia que han sido previamente analizados por [Andrzejak et al. \[2001b\]](#) (disponibles en <http://www.meb.unibonn.de/epileptologie/science/physik/eegdata.html>).

La información consiste en 100 segmentos de datos (de los cuales se eligieron 10 al azar), cuya longitud es de 4097 observaciones con una frecuencia de muestreo de 173.61 Hz., de actividad cerebral para diferentes grupos: se grabaron EEG de superficie para cinco pacientes sanos despiertos con los ojos abiertos (set A) y cerrados (set B), también se grabaron EEG intracraneales para cinco pacientes con epilepsia durante intervalos libres de convulsiones desde fuera (set C) y desde dentro (set D) de la zona generadora de convulsiones. Los detalles acerca de las

técnicas de grabación se pueden buscar en el artículo original de [Andrzejak et al. \[2001b\]](#).

El objetivo es comparar, mediante un procedimiento inferencial, la Entropía de Permutación de 4 sets diferentes de pacientes.

El enfoque clásico a este problema es muestrear varias señales de EEG de cada tipo, y realizar un test de comparación de medias con un test t si se presume normalidad, o mediante un test no paramétrico, como el de Mann Whitney-Wilcoxon. Las comparaciones múltiples se pueden hacer mediante un análisis de la varianza y un posterior test de Tukey, o con su equivalente no paramétrico, el test de rangos de Kruskal-Wallis.

Con este enfoque se pueden sacar conclusiones acerca de la relación entre las medias de la Entropía de Permutación de las señales de EEG de los distintos sets, pero no se puede comparar las dinámicas (mediante la Entropía de Permutación) entre señales individuales de esos sets distintos. Con el enfoque bootstrap presentado en esta Tesis, se puede establecer si la dinámica de alguna señal particular en el set A, por ejemplo, es significativamente diferente a alguna otra señal en particular del set B. Y se podría repetir este procedimiento entre todas las señales de todos los sets.

Se construyeron, como primera medida intervalos de confianza del nivel de confianza del 90% para la Entropía de Permutación de las 10 señales de EEG de la actividad cerebral para cada set (Figura 6.6) con el fin de tener una visualización de la posición del estimador de la Entropía de Permutación junto con su variabilidad. No está de más notar que el solapamiento entre intervalos no necesariamente significa que no hay diferencias significativas entre dos Entropías de Permutación ([Payton et al. \[2003\]](#)). Para llegar a esa conclusión se debe hacer un test de hipótesis para la diferencia.

Por lo tanto, como segundo paso se realizó un test para la diferencia en la Entropía de Permutación entre las señales de EEG de voluntarios sanos despiertos con los ojos abiertos (set A, en filas) y las señales de EEG de voluntarios sanos

despiertos con los ojos cerrados. Los resultados se pueden visualizar en la Figura 6.7.

Para obtener un nivel de significación global de 0.1 ($\alpha = 0,1$) para los 100 tests realizados, cada test individual para la diferencia en la Entropía de Permutación de una única señal del set A versus una única señal del set B (cada celda) fue llevado a cabo con un nivel de significación de 0.001 ($\alpha = 0,001$). 8 tests individuales fueron rechazados de un total de 100 indicando que hay diferencias entre el set A y el set B, pero 7 de esos rechazos pertenecen al mismo paciente cuya medición pudo haber sido afectada por un factor externo.

En la Figura 6.8, el mismo análisis es extendido a todos los tipos de pacientes. Todos los test globales indican que hay diferencias entre los sets, pero entre el set A y el set B la diferencia es principalmente debida a un sólo paciente; todas las Entropías de Permutación de las señales de los EEG de los sets A y B son diferentes a las de los sets C y D, excepto por una. Por el contrario, entre las entropías de permutación de las señales de los sets C y D las diferencias entre los individuos están distribuidas entre significativas y no significativas.

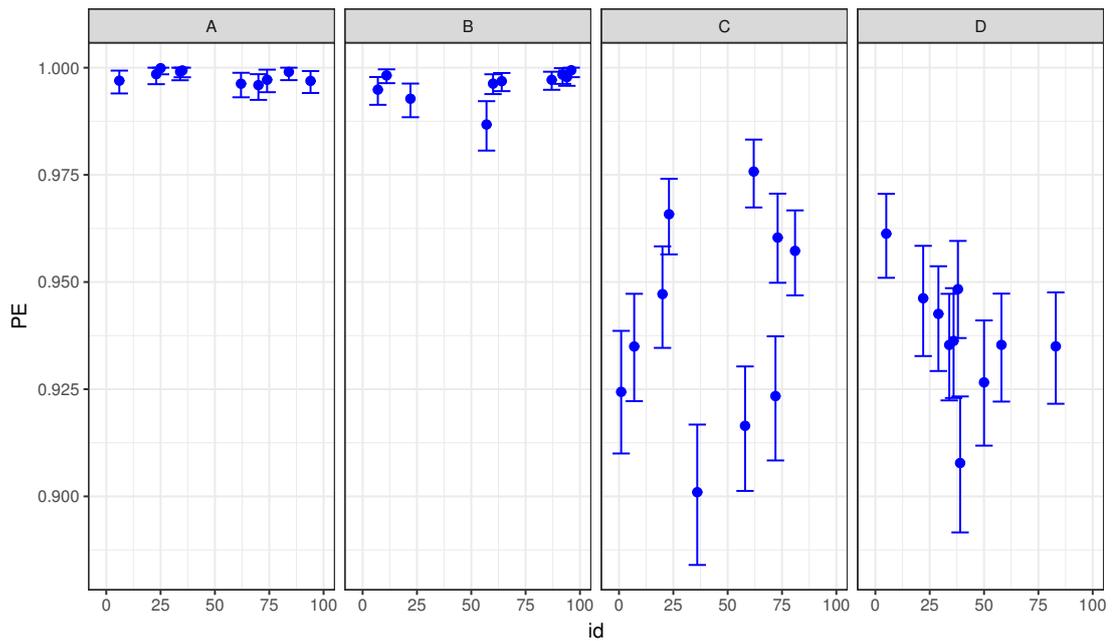


Figura 6.6 **Intervalos de confianza del nivel de confianza del 90 % para la Entropía de Permutación de las 10 señales EEG de la actividad cerebral para cada set.** Despiertos con los ojos abiertos (set A) y cerrados (set B), pacientes con epilepsia durante intervalos libres de convulsiones desde fuera (set C) y desde dentro (set D) de la zona generadora de convulsiones. No está de más notar que el solapamiento entre intervalos no necesariamente significa que no hay diferencias significativas entre dos Entropías de Permutación. Para llegar a esa conclusión se debe hacer un test de hipótesis para la diferencia.

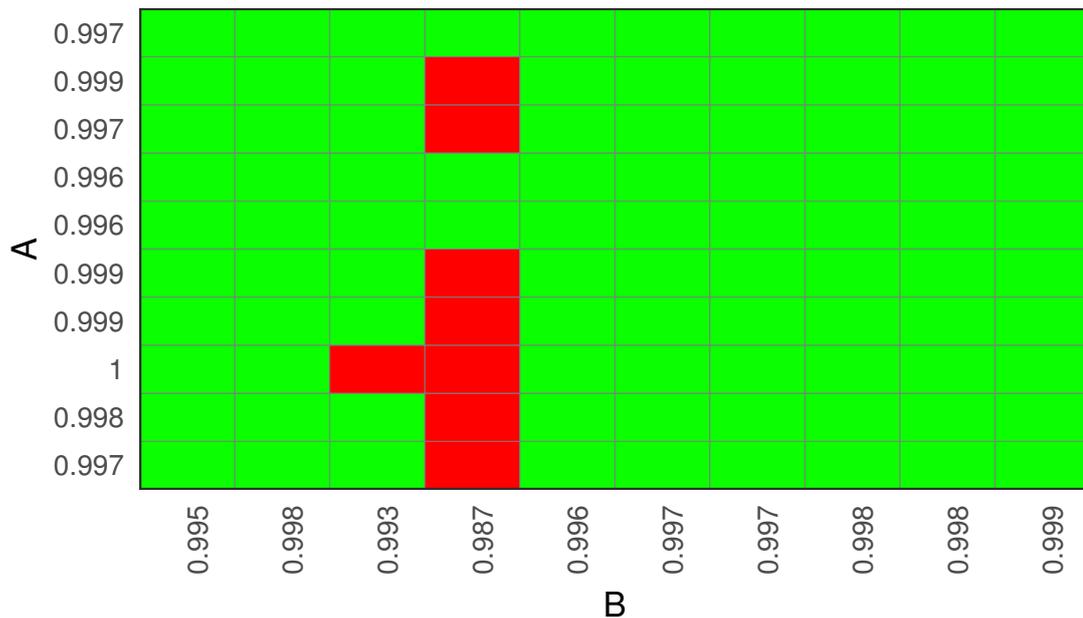


Figura 6.7 Test de hipótesis para la diferencia en la Entropía de Permutación entre las señales de EEG de voluntarios sanos despiertos con los ojos abiertos (set A, en filas) y las señales de EEG de voluntarios sanos despiertos con los ojos cerrados (set B, en columnas). Para obtener un nivel de significación global de 0.1 ($\alpha = 0,1$) para los 100 tests realizados, cada test individual para la diferencia en la Entropía de Permutación de una única señal del set A versus una única señal del set B (cada celda) fue llevado a cabo con un nivel de significación de 0.001 ($\alpha = 0,001$). 8 tests individuales fueron rechazados de un total de 100 indicando que hay diferencias entre el set A y el set B, pero 7 de esos rechazos pertenecen al mismo paciente cuya medición pudo haber sido afectada por un factor externo.

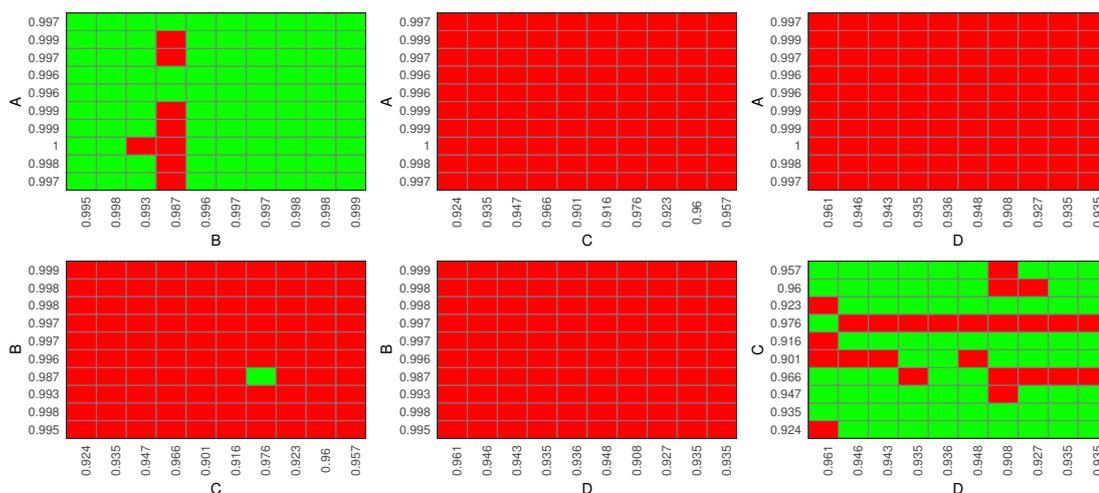


Figura 6.8 El mismo análisis de la Figura 6.7 es extendido a todos los tipos de pacientes. Todos los test globales indican que hay diferencia entre los sets, pero entre el set A y el set B la diferencia es principalmente debida a un sólo paciente; todas las Entropías de Permutación de las señales de los EEG de los sets A y B son diferentes a las de los sets C y D, excepto por una. Por el contrario, entre las Entropías de Permutación de las señales de los sets C y D las diferencias entre los individuos están distribuidas entre significativas y no significativas.

6.5. Conclusiones del Capítulo

La repetición del experimento que produce los datos para calcular el estadístico $\hat{\mathcal{H}}$ es la base para obtener su distribución. La repetición de este experimento con las mismas condiciones, infinitas veces, daría la distribución real de esta variable aleatoria, por lo que un número suficientemente grande de repeticiones constituyen la mejor estimación posible de la distribución del estadístico $\hat{\mathcal{H}}$. En la práctica es imposible repetir un experimento no controlado, por lo que el espíritu del bootstrap es realizar un experimento análogo con los datos que se tienen a mano.

En muchas situaciones, se desea comparar las dinámicas de dos procesos por medio de la Entropía de Permutación de sus series de tiempo. La pregunta que surge es la siguiente:

$$¿\mathcal{H}_1 = \mathcal{H}_2?$$

Esta pregunta no puede ser respondida con estimadores puntuales ($\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2$), porque éstos son variables aleatorias continuas y van a ser diferentes con probabilidad 1.

La pregunta que uno puede proponer es si la diferencia entre estas entropías es estadísticamente significativa o no, y eso sólo puede responderse si existe alguna medida de la variabilidad para esta variable aleatoria continua, $\hat{\Delta} = \hat{\mathcal{H}}_1 - \hat{\mathcal{H}}_2$. No se ha encontrado en la literatura este tipo de medida de variabilidad para la Entropía de Permutación como la propuesta en este Capítulo.

El estimador bootstrap propuesto para la Entropía de Permutación es asintóticamente insesgado y consistente en media cuadrática y para largos de serie lo suficientemente grandes es tan eficiente como el que se obtendría mediante el enfoque frecuentista si se pudiera repetir el experimento.

En resumen, se presenta una metodología basada en la computación para obtener una medida de precisión para el estimador de la Entropía de Permutación. Hasta el momento en la literatura sólo se han usado estadísticas descriptivas para caracterizar este cuantificador y, si el objetivo es obtener conclusiones que se extiendan mas allá de las muestras, no se han encontrado métodos inferenciales. Este método se presta también para realizar estadísticas inferenciales referidas a la Entropía de Permutación o mismo para cualquier otra medida de complejidad o información que surja de la Función de Distribución de Probabilidad propuesta por Bandt y Pompe.

Capítulo 7

Influencia del ruido observacional en la estimación de la Entropía de Permutación

7.1. Introducción

La Entropía de Permutación \mathcal{H} se presenta como particularmente útil en presencia de ruido dinámico y ruido observacional ([Bandt and Pompe \[2002\]](#)). En ese mismo artículo, se comenta que las medidas de complejidad presentadas hasta ese momento ([Abarbanel \[2012\]](#); [Ding et al. \[1993\]](#); [Grassberger and Procaccia \[1983\]](#)) funcionan notablemente bien para series de tiempo simuladas de sistemas dinámicos pero que la mayoría de ellas falla cuando se le añade ruido a las series.

Para series de tiempo observadas, la eliminación del ruido requiere un pre-procesamiento cuidadoso de los datos y un ajuste minucioso de los parámetros a utilizar llevando a complicaciones en el análisis, y los resultados obtenidos no pueden ser reproducidos sin un específico detalle del método de eliminación de ruido utilizado. El enfoque presentado por Bandt y Pompe se introduce como una solución posible a esta problemática y en el artículo donde la presentan muestran el comportamiento de la Entropía de Permutación al agregarle un ruido Gaussiano a una señal determinística, en particular al mapa logístico, concluyendo que dicho

ruido observacional causa sólo un pequeño aumento en la Entropía de Permutación. En otro estudio, se hace un análisis descriptivo del efecto del ruido dinámico en la Entropía de Permutación pero sin llegar a conclusiones acerca de la robustez del estimador de la Entropía de Permutación al ruido (Quintero-Quiroz et al. [2015]).

El objetivo de este Capítulo es profundizar sobre los efectos del ruido observacional en el cálculo de la Entropía de Permutación con los siguientes tres objetivos particulares:

- Hacer un análisis descriptivo del efecto del ruido observacional en la estimación de la Entropía de Permutación.
- Hacer un análisis inferencial para determinar el umbral de ruido observacional a partir del cual el test de hipótesis presentado en la Sección 6.2.2.2 del Capítulo 6 puede diferenciar entre las Entropías de Permutación de una serie totalmente determinística y de la misma serie con un cierto nivel de ruido. σ .
- Hacer una análisis inferencial utilizando el mismo test de hipótesis para observar si series determinísticas con el mismo nivel de ruido σ presentan diferencias estadísticamente significativas en la estimación de sus Entropías de Permutación.

El análisis descriptivo del efecto del ruido observacional en la estimación de la Entropía de Permutación arrojó resultados interesantes acerca de cómo el largo de la serie T afecta dicha estimación, que si bien no era uno de los objetivos particulares, su estudio es de interés principalmente para determinar un largo mínimo T_{min} que debe tener la serie de tiempo para utilizar esta medida correctamente.

7.2. Simulación numérica

La idea de la simulación numérica es simular una serie de tiempo X^* , proveniente de un proceso determinístico sin ruido, de largo T^* (lo suficientemente grande como para suponer que $T^* \rightarrow \infty$) para calcularle FDP de BP (\mathbf{P}_x). La Entropía de Permutación resultante $-\mathcal{H} = \mathcal{H}(\mathbf{P}_x)$ - para cada dimensión de *embedding* m , se

puede pensar como la Entropía de Permutación “verdadera” de este proceso para cada dimensión.

De esta serie simulada, se elige un tramo X completamente al azar de largo T , donde $T \lll T^*$, como una muestra de la serie de tiempo completa X^* . Para este tramo se estima la FDP de BP ($\hat{\mathbf{P}}_x$) y la Entropía de Permutación $-\hat{\mathcal{H}}_x = \mathcal{H}(\hat{\mathbf{P}}_x)$ - correspondiente.

Luego a este tramo se le agrega un ruido observacional descorrelacionado Gaussiano N de media 0 y varianza σ^2 , obteniendo la serie Y contaminada con ruido: $Y = X + N$.

A esta serie Y se le estima la FDP de BP $\hat{\mathbf{P}}_y$ y la Entropía de Permutación $-\hat{\mathcal{H}}_y = \mathcal{H}(\hat{\mathbf{P}}_y)$ -, que es la Entropía de Permutación de la muestra afectada por un ruido observacional.

Con estas simulaciones se va a realizar un análisis descriptivo de $\hat{\mathcal{H}}_x$ y $\hat{\mathcal{H}}_y$ para cada largo de serie T y cada nivel de ruido σ , y luego un análisis inferencial para determinar si el nivel de ruido σ influye significativamente en la estimación de la Entropía de Permutación.

7.2.1. Serie de tiempo proveniente de una dinámica caótica

El proceso elegido para esta simulación está dado por la ecuación logística:

$$x_{t+1} = x_t + 4 \cdot (1 - x_t) \quad (7.1)$$

donde la serie $X^* = \{x_t\}_{t=1}^{T^*}$ presenta una dinámica caótica dado el parámetro $r=4$, y se simuló para un largo $T^* = 12 \times 10^7$, con un valor inicial $x_0 = 0, 1$. A esta serie se le calculó la Entropía de Permutación $\mathcal{H} = \mathcal{H}(\mathbf{P}_x)$ para la dimensión de *embedding* $m = \{3, 4, 5, 6\}$ que dado el largo de la serie, se puede pensar como la Entropía de Permutación verdadera del proceso.

A continuación se eligió un numero k al azar tal que $1 \leq k \leq (T^* - T + 1)$ y se tomó el fragmento de la serie $X = \{x_t\}_{t=k}^{T+k-1}$ de largo T . A este fragmento de la serie se le estimó la Entropía de Permutación $-\hat{\mathcal{H}}_x = \mathcal{H}(\hat{\mathbf{P}}_x)$ - para cada dimensión de *embedding* $m = \{3, 4, 5, 6\}$.

Para cada largo de serie $T = \{30, 120, 600, 3600\}$ se repitió este procedimiento 100 veces, obteniendo los estimadores $\hat{\mathcal{H}}_x^{(i)}$; $i = 1, 2, \dots, 100$; para cada T .

7.2.2. Ruido observacional

El ruido observacional simulado N es el Gaussiano de media 0 y varianza σ^2 descorrelacionado:

$$N = \{\epsilon_t\}_{t=1}^T, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \text{ donde } \rho(\epsilon_r, \epsilon_s) = 0 \quad \forall r \neq s \text{ y } \rho(\epsilon_r, \epsilon_r) = 1 \quad (7.2)$$

donde ρ es la función de correlación y los valores utilizados para σ son los siguientes:

$$\sigma = \{10^{-6}, 5 \cdot 10^{-6}, 10^{-5}, 5 \cdot 10^{-5}, 10^{-4}, 5 \cdot 10^{-4}, 0,001, 0,005, 0,01, 0,05, 0,1, 0,25, 0,5, 1, 2,5, 5\} \quad (7.3)$$

7.2.3. Serie de tiempo proveniente de una dinámica caótica contaminada con ruido

A cada fragmento $X = \{x_t\}_{t=k}^{T+k-1}$ de largo T se le adicionó el ruido blanco observacional Gaussiano siendo la serie resultante:

$$Y = X + N$$

$$\{y_t\}_j^{j+T} = \{x_t\}_{t=k}^{T+k-1} + \{\epsilon_t\}_j^{j+T} \quad (7.4)$$

Entonces Y es un fragmento de la serie logística elegido al azar contaminado con un ruido blanco observacional Gaussiano de varianza σ^2 . A cada una de estas series Y se les estimó la Entropía de Permutación $-\hat{\mathcal{H}}_y = \mathcal{H}(\hat{\mathbf{P}}_y)$, obteniendo los estimadores $\hat{\mathcal{H}}_y^{(i)}$; $i = 1, 2, \dots, 100$ para cada T y para cada nivel de ruido σ .

7.3. Análisis descriptivo del efecto del ruido observacional en la estimación de la Entropía de Permutación

En las Figuras 7.1, 7.2, 7.3 y 7.4 se muestran los *boxplots* para las entropías $\hat{\mathcal{H}}_y^{(i)}$; $i = 1, 2, \dots, 100$ para $m = 3, 4, 5, 6$ respectivamente en función del nivel de ruido σ , y separadas por el largo de la serie T para su mejor comprensión. La línea horizontal negra en cada gráfico corresponde a la Entropía de Permutación $\mathcal{H} = \mathcal{H}(\mathbf{P}_x)$ calculada para $T = T^*$ sin ruido agregado, que se asumió como la verdadera Entropía de Permutación proveniente del proceso caótico. La Entropía de Permutación correspondiente a $\sigma = 0$ es la que corresponde a la muestra X sin contaminación de ruido, es decir $\hat{\mathcal{H}}_x$.

7.3.1. Efectos del largo de la serie en la estimación de la Entropía de Permutación

El primer objetivo de este Capítulo es hacer un análisis descriptivo del efecto del ruido observacional en la estimación de la Entropía de Permutación. Antes de proceder al estudio del efecto del ruido se encontraron resultados interesantes acerca de como afecta el largo T de la serie a la estimación de $\hat{\mathcal{H}}_x$:

En la literatura científica se sugiere utilizar un largo mínimo de $5m!$ (Amigó et al. [2008]) pero se puede observar que con esta regla se pueden obtener resultados sesgados: para cada dimensión de *embedding* hay un T_{min} a partir del cual la mediana de los $\hat{\mathcal{H}}_x$ se acerca al valor original de la Entropía de Permutación \mathcal{H} .

Para $m = 3$ esto ocurre a partir de $T = 120$, para $m = 4$ para un largo mayor a $T = 600$, y para $m = 5$ y $m = 6$ ya es necesario un largo mínimo de $T = 3600$. Esto se debe que para estimar la Entropía de Permutación para una dimensión de *embedding* m es necesario estimar $\hat{\mathbf{P}} = \{p(\pi_1), \dots, p(\pi_{m!})\}$, es decir $m! - 1$ parámetros, como se puede ver en el Cuadro 7.1.

m	3	4	5	6
Parámetros	5	23	119	719

Cuadro 7.1 Parámetros a ser estimados para cada dimensión de *embedding* para la estimación de la Entropía de Permutación

Si bien las series de tiempo obtenidas de procesos reales con las que se trabaja usualmente son de largos $T \gg 5m!$, es de uso común el uso de ventanas deslizantes de ancho L (i.e se estima $\hat{\mathcal{H}}$ para cada $\{x_t\}_k^{k+L-1} \forall k \in \{1, \dots, T - L + 1\}$) para mostrar la evolución de la Entropía de Permutación en el tiempo y es necesario determinar un ancho mínimo L_{min} para estas ventanas. Si bien este es un problema abierto pareciera que este L_{min} debiera ser bastante mayor a $5m!$.

7.3.2. Efectos del ruido observacional en la estimación de la Entropía de Permutación

Se analizará el efecto del ruido en la estimación $\hat{\mathcal{H}}_y$ a partir de largo T_{min} correspondiente (especificado en la Sección anterior) para cada dimensión de *embedding* m con el fin de aislar el efecto del ruido del efecto del largo de la serie.

En las Figuras 7.1 a 7.4 se pueden apreciar tres zonas: una zona de crecimiento lento de $\hat{\mathcal{H}}_y$ para valores pequeños de σ , dominando la dinámica determinística; una zona de transición al ruido, de crecimiento rápido de la estimación de la Entropía Permutación; y para valores grandes de σ la Entropía de Permutación se acerca a su valor máximo $\mathcal{H}_{max} = 1$, es decir una dinámica dominada por el ruido, en consonancia con otros estudios previamente realizados (Quintero-Quiroz et al. [2015]).

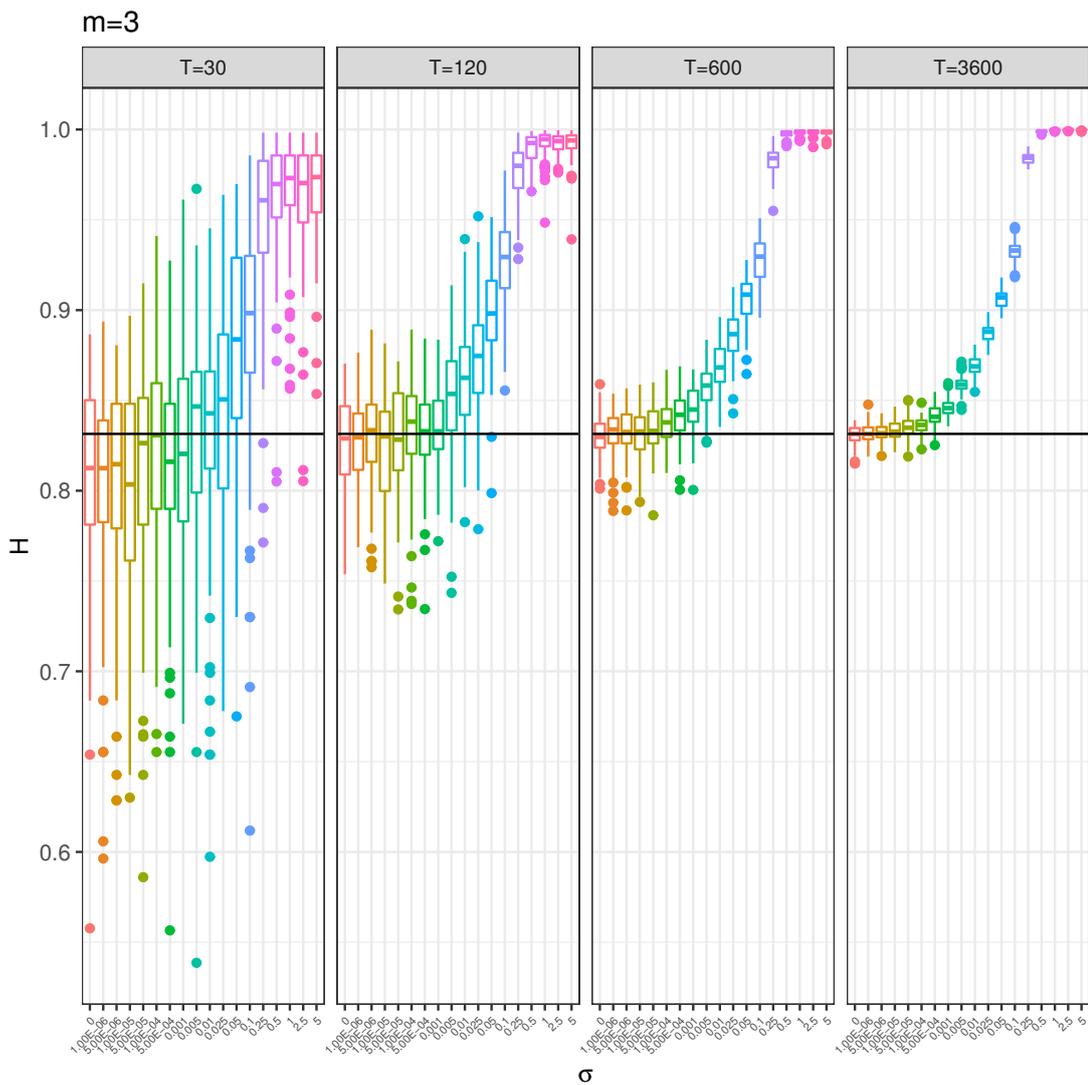


Figura 7.1 Efecto del ruido observacional en la estimación de la Entropía de Permutación en una serie de tiempo contaminada con ruido para $m = 3$. Hay un T_{min} a partir del cual la mediana de los \hat{H}_x se acerca al valor original de la Entropía de Permutación H y para $m = 3$ esto ocurre a partir de $T = 120$

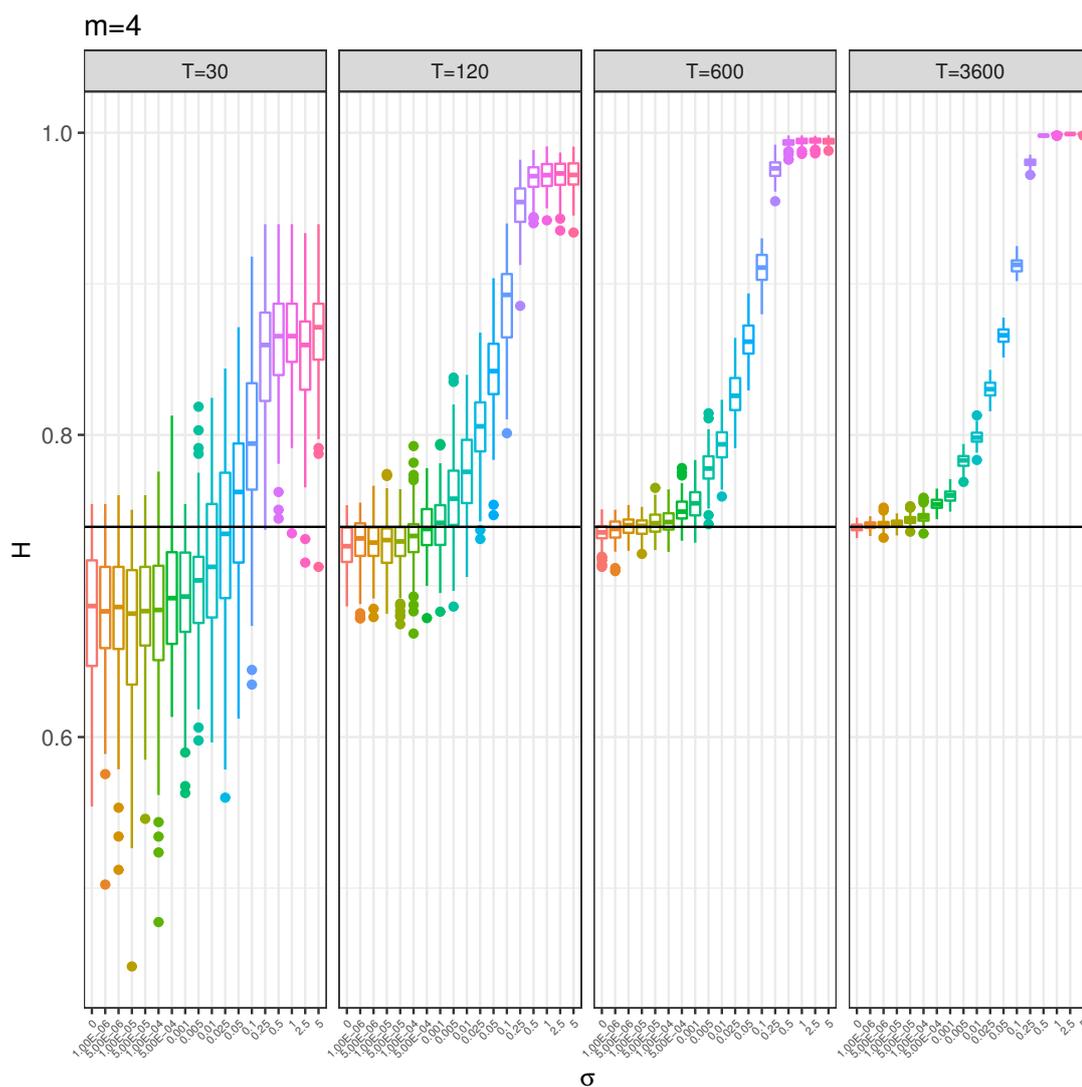


Figura 7.2 Idem a la Figura 7.1 para $m = 4$. Se puede observar que el T_{min} en este caso es $T = 600$

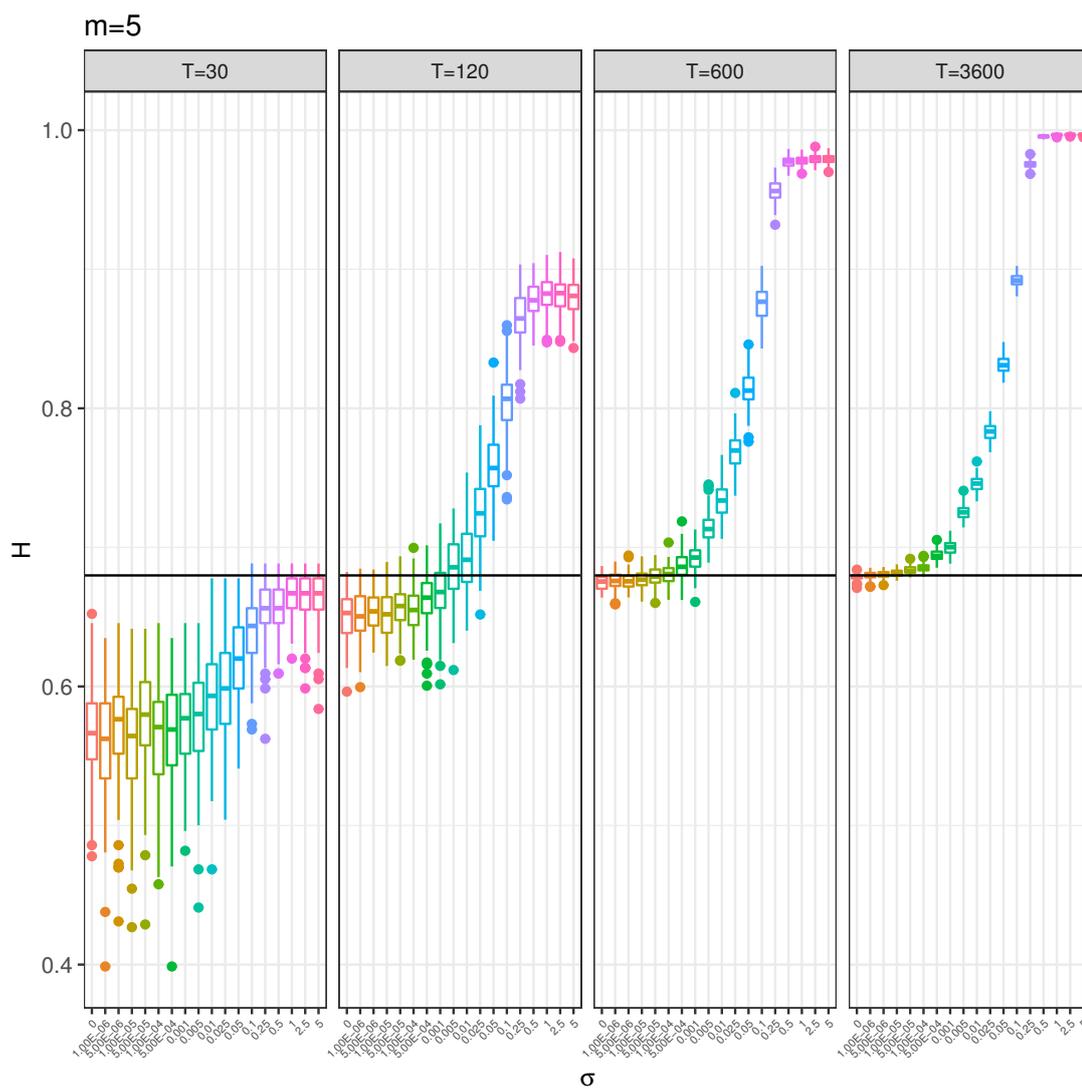


Figura 7.3 Idem a la Figura 7.1 para $m = 5$. Se puede observar que el T_{min} en este caso está entre $T = 600$ y $T = 3600$.

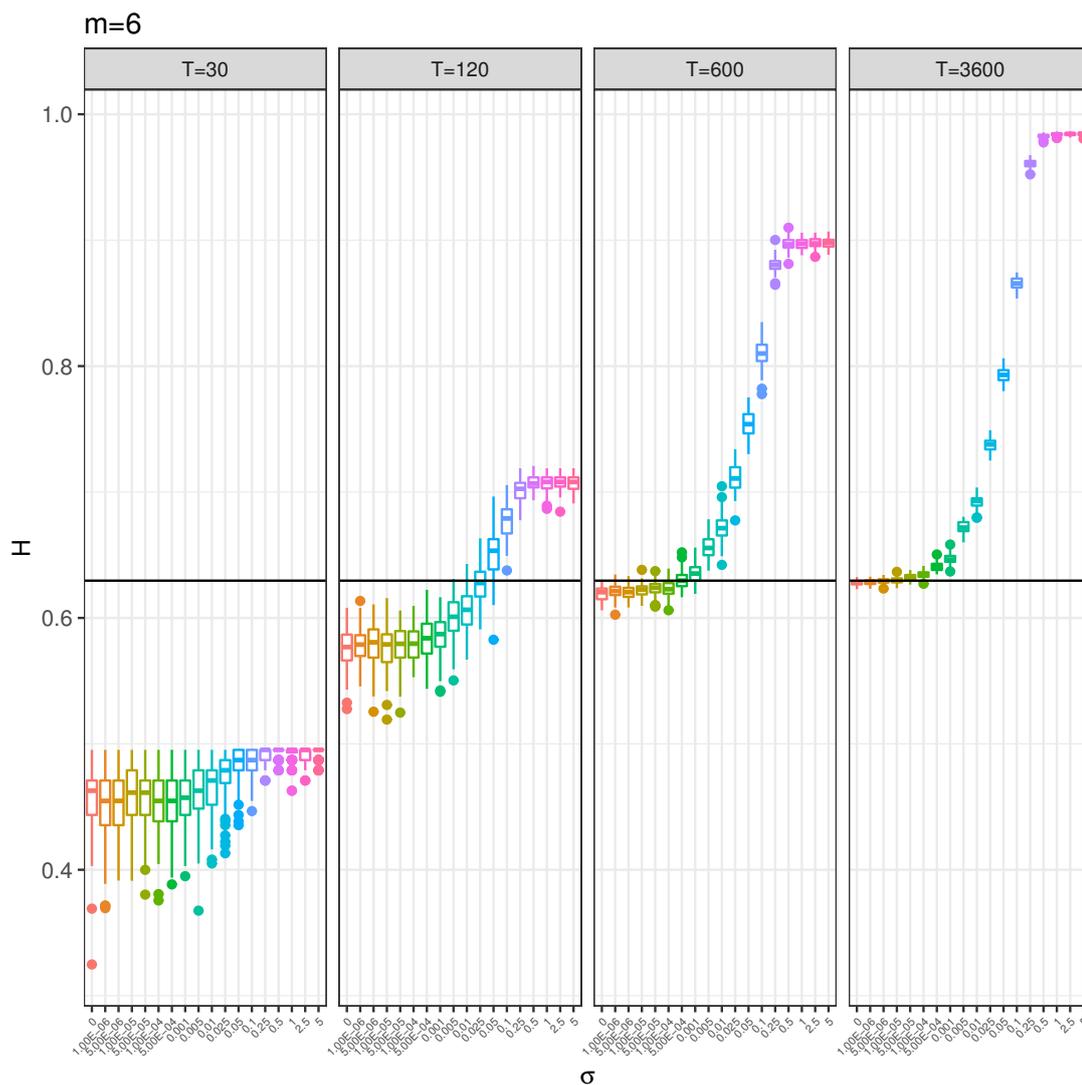


Figura 7.4 Idem a la Figura 7.1 para $m = 6$. Se puede observar que el T_{min} en este caso es $T = 3600$

7.3.3. Análisis inferencial del efecto del ruido observacional en la estimación de la Entropía de Permutación

El segundo objetivo es hacer un análisis inferencial para determinar el umbral de ruido observacional a partir del cual el test de hipótesis presentado en la Sección 6.2.2.2 del Capítulo 6 puede diferenciar entre las Entropías de Permutación de una serie totalmente determinística y de la misma serie con un cierto nivel de ruido.

Como se mostró en la Sección anterior, el ruido observacional influye en la estimación de la Entropía de Permutación. Se pudo observar que para un largo de serie dado T , a medida que aumenta dicho ruido (aumenta σ), aumenta el valor de la estimación de la Entropía de Permutación. Se diferenciaron tres zonas: una dominada por la dinámica determinística, una zona de transición de crecimiento rápido de la Entropía de Permutación y una zona estable de ruido dominante con Entropía de Permutación máxima. Una pregunta que intenta responder esta Sección es a partir de qué nivel de ruido (cuantificado por su desvío σ) un test de hipótesis puede diferenciar la entropía \mathcal{H}_x de una señal pura, o sin ruido, de la entropía $\mathcal{H}_y^{(\sigma)}$ de una señal a la que se le agregó dicho ruido observacional de nivel σ .

Se va a utilizar el test de hipótesis presentado en el Capítulo 6:

$H_0 : \Delta = \mathcal{H}_x - \mathcal{H}_y^{(\sigma)} = 0$, y si se rechaza H_0 , se puede decir que \mathcal{H}_x es distinto a $\mathcal{H}_y^{(\sigma)}$ con una probabilidad α de cometer un error.

Para cada largo de serie $T = \{120, 200, 600, 3600\}$, para cada dimensión de *embedding* $m = \{3, 4, 5, 6\}$ y para cada nivel de ruido σ (acorde a 7.3), se repitió este test 1000 veces, y se computó el ratio de test rechazados sobre el total de test realizados. Este ratio se puede ver como una estimación de la potencia del test $(1 - \beta)$ para detectar ruido en una señal.

En la Figura 7.5 se pueden observar las curvas de potencia para los distintos m , separadas para largo T . En el Cuadro 7.2 se pueden ver los valores de $1 - \beta$

en función del largo T y dimensión de *embedding* m que se utilizaron para la realización de esta Figura.

Para valores pequeños de T , la potencia del test toma valores altos (i.e. $1 - \beta \geq 0,8$) sólo cuando la serie está contaminada con ruidos con valores altos de σ , pero a medida que el largo de la serie crece, el test se vuelve más potente y, se puede observar que para un largo de serie de $T = 3600$ el test puede detectar con mucha eficacia la adición de un pequeño nivel de ruido.

En un test estadístico clásico de medias existe un punto crítico de la curva de potencia que es cuando $1 - \beta(\mu_1) = 0,5$. Este punto indica que si el valor real del parámetro fuera μ_1 la probabilidad de cometer un error se equipara con la probabilidad de acertar, al rechazar la hipótesis nula. Para muchas aplicaciones, este valor puede ser visto como un punto de equilibrio.

Haciendo una analogía para este test puede ser de importancia analizar el valor de $\sigma = \sigma_c$ donde $1 - \beta(\sigma_c) = 0,5$. Este punto crítico σ_c se puede ver como el punto a partir del cual el test de hipótesis tiene más probabilidad de diferenciar correctamente la entropía \mathcal{H}_x de una señal pura, o sin ruido, de la entropía $\mathcal{H}_y^{(\sigma)}$ de una señal a la que se le agregó dicho ruido observacional de nivel $\sigma > \sigma_c$. En la Figura 7.5, las líneas horizontales en negro marcan $1 - \beta = 0,5$, y su intersección con las curvas de potencia estimadas para cada m es $1 - \beta(\sigma_c) = 0,5$. A medida que aumenta T este valor σ_c disminuye, y la Entropía de Permutación detecta este cambio de dinámicas para valores muy pequeños de ruido agregado.

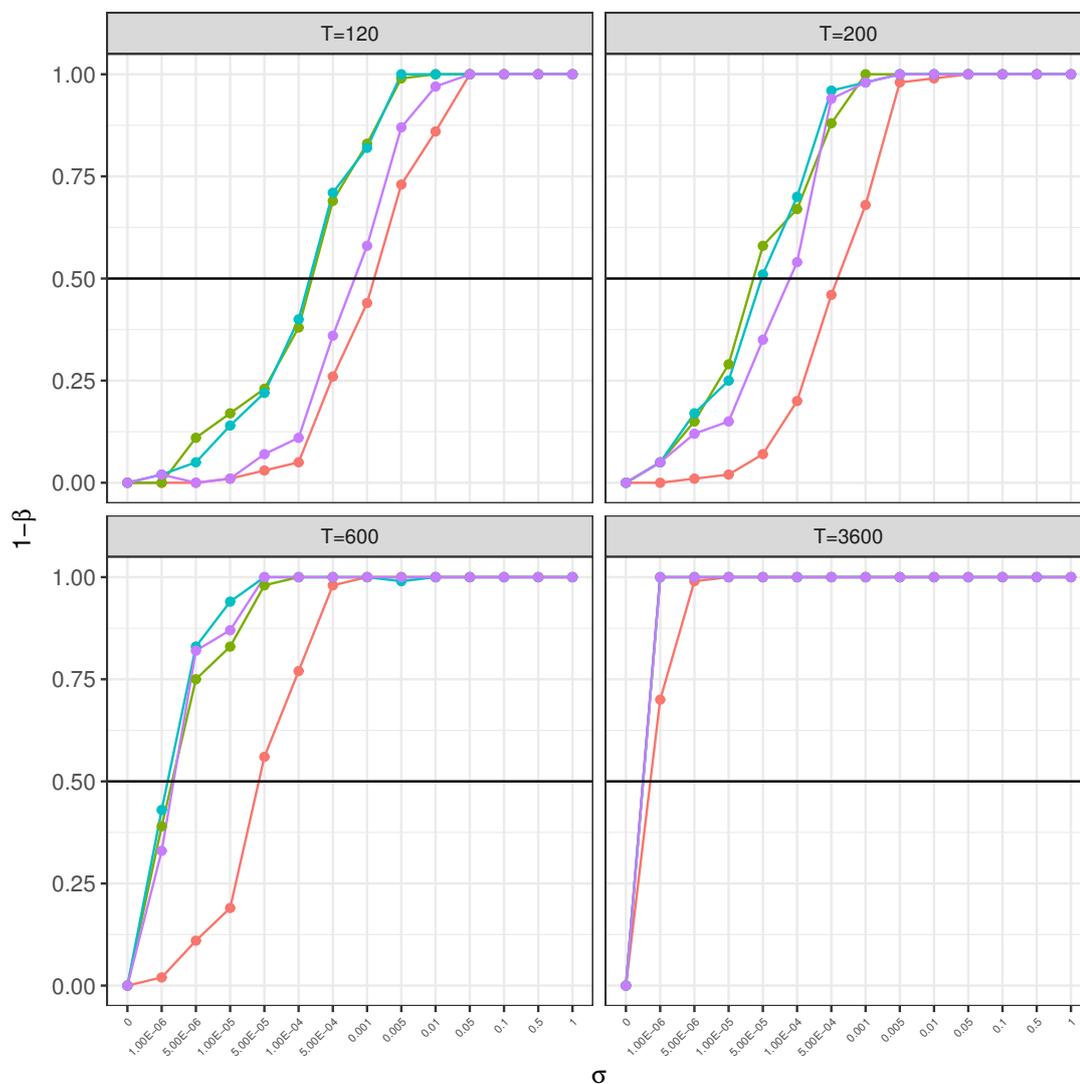


Figura 7.5 Las curvas de potencia $(1 - \beta)$ en función del ruido agregado σ . Se presentan las curvas para cada m : $m = 3$ en rojo, $m = 4$ en verde, $m = 5$ en azul y $m = 6$ en violeta. Para valores pequeños de T , la potencia del test toma valores altos (i.e. $1 - \beta \geq 0,8$) sólo cuando la serie está contaminada con ruidos con valores altos de σ , pero a medida que el largo de la serie crece, el test se vuelve más potente y, se puede observar que para un largo de serie de $T = 3600$ el test puede detectar con mucha eficacia la adición de un pequeño nivel de ruido.

		$1-\beta$												
σ		10^{-06}	$5x10^{-06}$	10^{-05}	$5x10^{-05}$	10^{-04}	$5x10^{-04}$	10^{-03}	$5x10^{-03}$	10^{-02}	$5x10^{-02}$	10^{-01}	$5x10^{-01}$	1
$T = 120$	$m = 3$	0	0	0.01	0.03	0.05	0.26	0.44	0.73	0.86	1	1	1	1
	$m = 4$	0	0.11	0.17	0.23	0.38	0.69	0.83	0.99	1	1	1	1	1
	$m = 5$	0	0.02	0.05	0.22	0.4	0.71	0.82	1	1	1	1	1	1
	$m = 6$	0	0.02	0	0.01	0.07	0.11	0.36	0.58	0.87	0.97	1	1	1
$T = 200$	$m = 3$	0	10^{-06}	$5x10^{-06}$	10^{-05}	$5x10^{-05}$	10^{-04}	$5x10^{-04}$	0.001	0.005	0.01	0.05	0.1	0.5
	$m = 4$	0	0	0.01	0.02	0.07	0.2	0.46	0.68	0.98	0.99	1	1	1
	$m = 5$	0	0.05	0.15	0.29	0.58	0.67	0.88	1	1	1	1	1	1
	$m = 6$	0	0.05	0.17	0.25	0.51	0.7	0.96	0.98	1	1	1	1	1
$T = 600$	$m = 3$	0	10^{-06}	$5x10^{-06}$	10^{-05}	$5x10^{-05}$	10^{-04}	$5x10^{-04}$	0.001	0.005	0.01	0.05	0.1	0.5
	$m = 4$	0	0.02	0.11	0.19	0.56	0.77	0.98	1	1	1	1	1	1
	$m = 5$	0	0.39	0.75	0.83	0.98	1	1	1	1	1	1	1	1
	$m = 6$	0	0.43	0.83	0.94	1	1	1	0.99	1	1	1	1	1
$T = 3600$	$m = 3$	0	10^{-06}	$5x10^{-06}$	10^{-05}	$5x10^{-05}$	10^{-04}	$5x10^{-04}$	0.001	0.005	0.01	0.05	0.1	0.5
	$m = 4$	0	0.7	0.99	1	1	1	1	1	1	1	1	1	1
	$m = 5$	0	1	1	1	1	1	1	1	1	1	1	1	1
	$m = 6$	0	1	1	1	1	1	1	1	1	1	1	1	1

Cuadro 7.2 Valores utilizados para graficar las curvas de potencia $1 - \beta$ en función del ruido agregado σ de la Figura 7.5

El tercer objetivo es hacer un análisis inferencial utilizando el mismo test de hipótesis para observar si series determinísticas con el mismo nivel de ruido σ presentan diferencias estadísticamente significativas en la estimación de sus Entropías de Permutación.

Esta situación es la que ocurre generalmente en las aplicaciones del mundo real donde la obtención de los datos generados por un proceso se realiza con el mismo artefacto de medición que se puede suponer que le añade el mismo nivel de ruido observacional a todas las series registradas por el mismo.

Por lo tanto la segunda pregunta a responder, y quizás la más importante, es si un test de hipótesis detecta cambios en la Entropía de Permutación debido a que la señal es ruidosa y no debido a que la dinámica subyacente del proceso generador de datos se ha modificado.

Se va a utilizar nuevamente el test de hipótesis presentado en el Capítulo 6, pero con la siguiente hipótesis nula:

$$H_0 : \Delta = \mathcal{H}_{y_1}^{(\sigma)} - \mathcal{H}_{y_2}^{(\sigma)} = 0,$$

donde Y_1 e Y_2 son dos series de tiempo provenientes de la ecuación logística (es decir con la misma dinámica subyacente) con el mismo nivel de ruido.

Nuevamente, para cada largo de serie T , para cada dimensión de *embedding* m y para cada nivel de ruido σ , se repitió este test 1000 veces, y se computó el ratio de test rechazados sobre el total de test realizados. En este caso, este ratio representa una estimación del nivel de significación α del test, ya que el rechazo del test se puede ver como un error, indicando que los cambios en la estimación de la Entropía de Permutación se pueden deber a efectos de la medición y no a la dinámica propia del proceso.

En la Figura 7.6 se muestra el nivel de significación α en función del nivel de ruido agregado σ . En el Cuadro 7.3 se pueden ver los valores de α en función del largo T y dimensión de *embedding* m que se utilizaron para la realización de esta Figura.

Se puede ver cómo el nivel de significación para valores pequeños de σ es muy bajo $\alpha < 0,05$, es decir que el test no detecta diferencias en la Entropía de Permutación para señales con la misma dinámica que presenten un bajo nivel de ruido. En la zona de transición al ruido, este error aumenta, pero se mantiene en un nivel aceptable $\alpha \approx 0,1$, y finalmente, en la zona dominada por el ruido el nivel de significación vuelve a disminuir mostrando de manera acertada que ambas señales son prácticamente ruido.

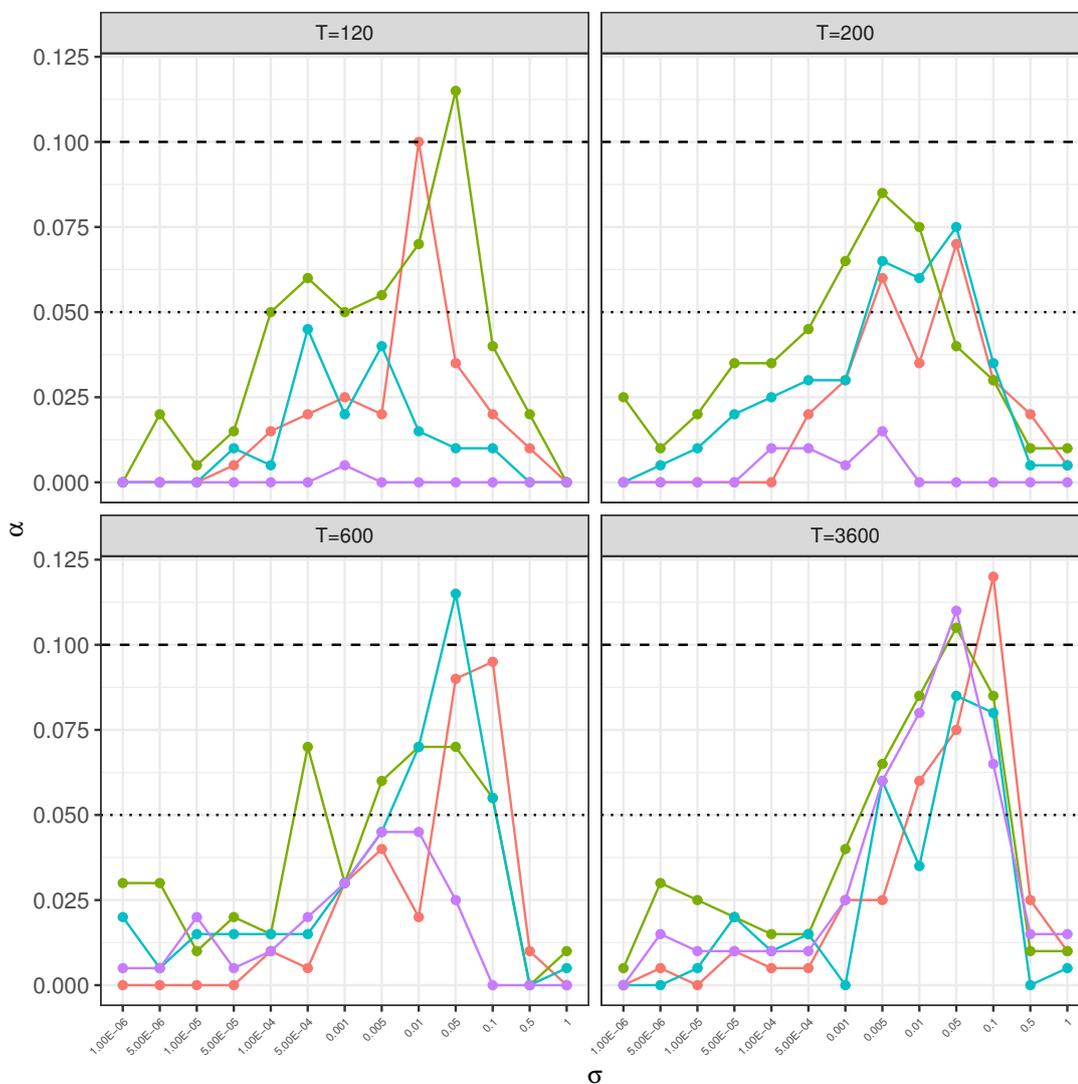


Figura 7.6 El nivel de significación α en función del nivel de ruido agregado σ . Se presentan las curvas para cada m : $m = 3$ en rojo, $m = 4$ en verde, $m = 5$ en azul y $m = 6$ en violeta. Se puede ver cómo el nivel de significación para valores pequeños de σ es muy bajo $\alpha < 0,05$, es decir que el test no detecta diferencias en la Entropía de Permutación para señales con la misma dinámica que presenten un bajo nivel de ruido. En la zona de transición al ruido, este error aumenta, pero se mantiene en un nivel aceptable $\alpha \approx 0,1$, y finalmente, en la zona dominada por el ruido el nivel de significación vuelve a disminuir mostrando de manera acertada que ambas señales son prácticamente ruido.

		Error de tipo I (α)												
σ		10^{-06}	$5x10^{-06}$	10^{-05}	$5x10^{-05}$	10^{-04}	$5x10^{-04}$	$0,001$	$0,005$	$0,01$	$0,05$	$0,1$	$0,5$	1
$T = 120$	$m = 3$	0	0	0,005	0,015	0,02	0,025	0,02	0,1	0,035	0,02	0,01	0	
	$m = 4$	0	0,02	0,005	0,015	0,05	0,06	0,05	0,055	0,07	0,115	0,04	0,02	
	$m = 5$	0	0	0	0,01	0,005	0,045	0,02	0,04	0,015	0,01	0	0	
	$m = 6$	0	0	0	0	0	0,005	0	0	0	0	0	0	
$T = 200$	σ	10^{-06}	$5x10^{-06}$	10^{-05}	$5x10^{-05}$	10^{-04}	$5x10^{-04}$	$0,001$	$0,005$	$0,01$	$0,05$	$0,1$	$0,5$	1
	$m = 6$	0	0	0	0	0,02	0,03	0,06	0,035	0,07	0,03	0,02	0,005	
	$m = 6$	0,025	0,01	0,02	0,035	0,045	0,065	0,085	0,075	0,04	0,03	0,01	0,01	
	$m = 6$	0	0,005	0,01	0,02	0,025	0,03	0,03	0,065	0,06	0,075	0,035	0,005	
$T = 600$	$m = 6$	0	0	0	0	0,01	0,01	0,005	0,015	0	0	0	0	
	$m = 6$	0	0	0	0,01	0,005	0,03	0,04	0,02	0,09	0,095	0,01	0	
	$m = 6$	0,03	0,03	0,01	0,02	0,015	0,07	0,03	0,06	0,07	0,07	0,055	0	
	$m = 6$	0,02	0,005	0,015	0,015	0,015	0,03	0,045	0,07	0,115	0,055	0	0,005	
$T = 3600$	$m = 6$	0,005	0,005	0,02	0,005	0,01	0,02	0,03	0,045	0,045	0,025	0	0	
	σ	10^{-06}	$5x10^{-06}$	10^{-05}	$5x10^{-05}$	10^{-04}	$5x10^{-04}$	$0,001$	$0,005$	$0,01$	$0,05$	$0,1$	$0,5$	1
	$m = 3$	0	0,005	0	0,01	0,005	0,005	0,025	0,025	0,06	0,075	0,12	0,025	
	$m = 4$	0,005	0,03	0,025	0,02	0,015	0,015	0,04	0,065	0,085	0,105	0,085	0,01	
$m = 5$	$m = 5$	0	0	0,005	0,02	0,01	0,015	0	0,06	0,035	0,085	0,08	0	
	$m = 6$	0	0,015	0,01	0,01	0,01	0,01	0,025	0,06	0,08	0,11	0,065	0,015	

Cuadro 7.3 El nivel de significación α en función del nivel de ruido agregado σ .

7.4. Conclusiones

La estimación de la Entropía de Permutación de una serie de tiempo se ve afectada tanto por el ruido observacional σ añadido a la serie como por el largo T del tramo analizado.

El ruido observacional que presenta una señal proveniente de una dinámica tiende a equiparar las probabilidades de aparición de cada símbolo del alfabeto simbólico, aumentando su entropía. Si dicho ruido observacional es muy pequeño comparado con la amplitud de los valores de la serie, no influye en la relación ordinal original de los valores de un patrón y por ende, no modifica su mapeo al símbolo correspondiente, conservando la entropía original del proceso. A medida que aumenta el ruido, dicha relación se puede ver modificada e incluso presentar patrones que no existirían de otro modo en dicha serie original, aumentando la frecuencia de aparición de símbolos originalmente poco probables y disminuyendo la frecuencia de aparición de los símbolos más probables ya que existe la condición $\sum_{i=1}^{m!} p(\pi_i) = 1$. Esto hace que la FDP de BP sea más cercana a la distribución equiprobable, aumentando su entropía.

En cuanto al largo de la serie T , el estimador de la Entropía de Permutación presenta la misma problemática que un estimador clásico cuando la cantidad de datos disponible es pequeña. Para cada estimación $\hat{\mathcal{H}}$ de dimensión de *embedding* m es preciso estimar $m! - 1$ parámetros por lo que el largo de la serie debe ser mucho mayor a este valor ($T \gg (m! - 1)$).

Un test de hipótesis con la Entropía de Permutación puede detectar la adición de ruido a una serie determinística para valores pequeños de σ si el largo de la serie es lo suficientemente grande, pero rara vez en una aplicación real se tiene una serie no contaminada.

Más importante es remarcar que cuando dos series provenientes del mismo proceso medidas con el mismo instrumento de medición (con el mismo nivel de ruido) son comparadas mediante la Entropía de Permutación, el test de hipótesis presentado en esta Tesis comete errores pequeños. Por lo tanto en el caso de que el test detecte una diferencia entre ambas Entropías de Permutación, hay una alta

probabilidad de que se deba a que hay un cambio en la dinámica subyacente y no al ruido observacional.

Capítulo 8

Discusión, Conclusiones y Futuras Líneas de Investigación

8.1. Discusión

Hace poco más de 15 años, Bandt y Pompe desarrollaron una herramienta para caracterizar una dinámica proveniente de una serie de tiempo (o señal) que no requiere de pre-procesamiento y que sólo exige una estacionariedad muy débil que es generalmente satisfecha en casos experimentales, cuyo algoritmo es de fácil cómputo y ha rendido frutos en el estudio de los sistemas dinámicos que surgen en las distintas disciplinas de la ciencia, y en particular de la Ingeniería. Esta herramienta se basa en calcular las frecuencias de aparición de ciertas estructuras (o patrones) basadas en una simple función de autocorrelación ordinal, presentes en la serie de tiempo, y a partir de estas frecuencias, estimar la Función de Distribución de Probabilidades de estos patrones, que a lo largo de toda esta Tesis dimos en llamar la FDP de BP.

A partir de esta estimación de la FDP de BP, surgieron diversos cuantificadores basados en la Teoría de la Información, siendo el principal y de mayor utilización la Entropía Informacional de Shannon que dio lugar a la Entropía de Permutación $-\mathcal{H}$ -, que mide el grado de desorden o incertidumbre asociado a esta distribución.

Como consecuencia de los logros obtenidos por este cuantificador, otras medidas de entropía que utilizan la FDP de BP, que son presentadas en el Capítulo 2, han sido implementadas en el estudio de señales con diversos resultados. La Entropía de Tsallis $-S_q$ - y la Entropía de Rényi $-R_\alpha$ - son algunos de estos casos, extensiones de la Entropía de Shannon para sistemas dinámicos con propiedades particulares, que evaluados en la FDP de BP han dado resultados muy interesantes. Sin embargo, su utilización no se ha extendido y un estudio mas exhaustivo de las mismas puede llevar a descripciones más acertadas en ciertos sistemas que las obtenidas por la misma Entropía de Permutación. La contrapartida de estas entropías es que requieren de un parámetro de ajuste o afinación que no necesariamente es obvio para cada señal a analizar. Más recientemente en el tiempo, se desarrolló la Entropía de Permutación Ponderada $-\mathcal{H}_w$ - que intenta extraer más información que la simple autocorrelación ordinal, ponderando cada patrón con una medida de dispersión asociada a sus amplitudes, que ha sido un buen instrumento para el estudio de señales, principalmente provenientes de EEG. Si bien es un algoritmo que ha dado resultados en este tipo de señales, el marco teórico en el que se basa no es claro y no ha sido bien estudiado hasta el momento.

Como complemento de una medida de incertidumbre como la Entropía de Permutación, surgieron dos cuantificadores que utilizan la FDP de BP: una medida global de complejidad, la Medida de Complejidad Estadística $-\mathcal{C}$ -, y una medida local de información, la Medida de Información de Fisher $-\mathcal{F}$ -.

La Medida de Complejidad Estadística mide la interacción entre el desequilibrio (una distancia a una medida de equilibrio) y la Entropía de Permutación. Es un aporte importante ya que permite diferenciar las distintas dinámicas en un plano informacional $\mathcal{H} \times \mathcal{C}$.

Por otro lado, la Medida de Información de Fisher hace posible cuantificar cambios en la dinámica de un sistema frente a cambios en el espacio de parámetros del mismo, gracias a que es una medida sensible a los cambios locales de la FDP de BP. El principal problema de esta medida, que todavía no está resuelto, es que es dependiente del orden de los elementos del alfabeto simbólico, de donde proviene la

FDP de BP. No sólo el orden de los símbolos propuesto es arbitrario, sino como se pudo ver en el Capítulo 2, este orden también depende de la metodología utilizada para la transformación de vectores *embedding* a símbolos.

Los Capítulos 2, 3 y 4 son el resultado de la primera etapa de la Tesis, un estudio exhaustivo de las propiedades descriptivas de los cuantificadores provenientes de la estimación de la FDP de BP.

El primer período de esta etapa se realizó en las instalaciones del Instituto Tecnológico de Buenos Aires (ITBA) donde se tuvo acceso a experimentos con máquinas rotativas y se estudió cómo los distintos estimadores de estas medidas de complejidad - $\hat{\mathcal{H}}$, $\hat{\mathcal{C}}$ y $\hat{\mathcal{F}}$ - caracterizaban el funcionamiento de las máquinas en estado balanceado y distintos estados desbalanceados. En el segundo período de esta etapa se trabajó en el Hospital Italiano de Buenos Aires (HIBA), donde se centró el estudio en el uso del estimador la Entropía de Permutación - $\hat{\mathcal{H}}$ - y sus extensiones - $\hat{\mathcal{H}}$, \hat{S}_q , \hat{R}_α , \hat{R}_∞ , y \mathcal{H}_w - para la detección automática de actividad anormal en la dinámica eléctrica del cerebro mediante el estudio de señales de EEG.

Habiendo estudiado cómo los estimadores de las distintas medidas de complejidad caracterizaban un sistema dinámico surgió la siguiente pregunta: ¿La Entropía de Permutación puede distinguir entre series de tiempo provenientes de procesos generadores de datos con distintas distribuciones marginales? Debido a la pérdida de información acerca de la amplitud de los valores provenientes de utilizar un método ordinal para su simbolización, la respuesta pareciera ser que no.

En el Capítulo 4 se trata esta problemática mediante la simulación de procesos estocásticos autorregresivos, que si bien está muy estudiada y documentada para las series de tiempo Gaussianas, no es el caso de las series de tiempo autorregresivas Uniformes y Exponenciales cuya simulación no es trivial. Mediante los resultados obtenidos, se confirmó que la Entropía de Permutación no distingue (y por ende no se ve afectada) por la distribución marginal de los datos, sino que caracteriza únicamente su estructura de correlación. Cuando en el estudio de la serie de tiempo sea de relevancia dicha distribución marginal, se recomienda usar la Entropía de

Amplitud $-\mathcal{H}_a$ -, es decir el funcional de la Entropía de Shannon evaluado en la Función de Distribución de Probabilidades empírica proveniente del histograma, mediante un plano $\mathcal{H} \times \mathcal{H}_a$ que permite distinguir tanto la estructura de correlación como la distribución marginal de la que provienen los datos. La Entropía de Amplitud requiere, sin embargo, un parámetro adicional que no es un problema menor, porque es todavía un problema abierto inclusive en la estadística clásica, que es definir la cantidad de intervalos a utilizar para el cálculo del histograma. Si bien en esta Tesis se aplica la regla de Scott (el fundamento de esta decisión se puede encontrar en el Apéndice B), existen muchas otras reglas que pudieran dar buenos resultados para distintas series de tiempo.

Los resultados obtenidos en el estudio de las señales de EEG mediante la Entropía de Permutación alentaron su uso en el estudio de señales de signos vitales para un modelado predictivo del comportamiento de pacientes en la Unidad de Terapia Intensiva de Adultos (UTIA) (Astudillo et al. [2017]; Zelechower et al. [2017]). Dichas señales de signos vitales están tomadas mediante un monitor conectado al paciente cuya resolución es de dos dígitos decimales, resultando en que el análisis de estas series de tiempo mediante la Entropía de Permutación se ve dificultado al no cumplir uno de los supuestos en los que se basa esta medida de complejidad: un vector de *embedding* debe tener probabilidad 0 de contener componentes iguales.

Es decir, esta problemática surge cuando se precisa estimar la Entropía de Permutación para caracterizar un sistema dinámico cuya serie temporal es adquirida mediante un instrumento de medición de baja resolución que resulta en una alta frecuencia de vectores de *embedding* con componentes iguales, situación en la cual no se puede efectuar la simbolización propuesta por Bandt y Pompe, ya que no se cumple con uno de los presupuestos presentados en la teoría original. Para lidiar con esta problemática se han encontrado en la literatura distintas metodologías que intentan aplicar desde nuevas reglas de ordenamiento, pasando por agregar un pequeño ruido observacional y también eliminar estos vectores de *embedding* con componentes iguales y hasta definir un nuevo alfabeto extendido. En el Capítulo 5 se

analizan en profundidad estas metodologías existentes y se presenta una propuesta superadora, la Imputación Basada en la Muestra. Mediante la simulación de 39 mapas caóticos se muestra el buen desempeño de esta metodología y se recomienda su uso para estas series de tiempo cuyos datos se presentan con una baja resolución. Y si bien la Entropía de Permutación no fue ideada para caracterizar series de tiempo a valores discretos, es frecuente encontrar artículos donde se extiende su uso a dichas series. La nueva metodología presentada muestra ser eficaz también en estos casos.

El principal objetivo de esta Tesis fue hallar una medida de precisión del estimador de la Entropía de Permutación que permita construir un intervalo de confianza asociado a dicha estimación y desarrollar un test de hipótesis potente que pudiera dictaminar si dos series de tiempo difieren de manera significativa (desde el punto de vista estadístico) en cuanto a su Entropía de Permutación.

El Capítulo 6 es la consecuencia del estudio exhaustivo de la Función de Distribución de Probabilidades de Bandt y Pompe y los cuantificadores de la Teoría de la Información evaluados en dicha FDP de BP, realizados en los primeros Capítulos, y da como resultado el logro del objetivo principal de esta Tesis: una medida de precisión para el estimador de la Entropía de Permutación que permite hacer inferencias con este cuantificador. No hay un método analítico hasta el momento para encontrar la función de distribución de este estimador y los métodos tradicionales de bootstrap no paramétricos aplicados a series de tiempo han mostrado no ser útiles debido a que dichas metodologías, por su estructura de funcionamiento, fallan en preservar la autocorrelación de los valores de la serie, aspecto fundamental en el que se basa la Entropía de Permutación. El novedoso método presentado en esta Tesis se inspira en el modelo de las cadenas de Markov para pensar que la secuencia simbólica, derivada de la serie de tiempo bajo estudio, está caracterizada en su totalidad por la matriz de las probabilidades de transición entre los distintos símbolos del alfabeto propuesto por Bandt y Pompe. De esta manera, se modela a la dinámica mediante un proceso generador de datos que tiene como parámetros a estas probabilidades de transición. Con el método de

bootstrap paramétrico, una estimación de este modelo da lugar a la simulación de réplicas del estimador de la Entropía de Permutación que permiten obtener una distribución que es en todas las maneras relevantes, análoga a la distribución empírica de dicho estimador. Esta herramienta estadística computacional permite hacer inferencias acerca de la verdadera Entropía de Permutación proveniente de un sistema dinámico y da lugar a la construcción de test de hipótesis e intervalos de confianza para esta medida.

El cumplimiento de este objetivo permite estudiar el comportamiento del estimador de la Entropía de Permutación frente al ruido observacional omnipresente en todas las series de tiempo provenientes de un experimento real desde un punto de vista inferencial. Por lo tanto, en el Capítulo 7 se estudia en un principio de manera descriptiva cómo afecta el ruido observacional a la estimación de la Entropía de Permutación, y se muestra que agregar ruido siempre aumenta el valor de esta estimación. En cuanto al largo de la serie T , el estimador de la Entropía de Permutación presenta la misma problemática que un estimador clásico cuando la cantidad de datos disponible es pequeña y es necesario un largo mínimo de la serie de tiempo para lograr una estimación confiable, que depende de la dimensión de *embedding* elegida. Luego se muestra que un test de hipótesis con la Entropía de Permutación puede detectar la adición de ruido a una serie determinística para valores pequeños de σ si el largo de la serie es lo suficientemente grande. Si bien este resultado es interesante desde el punto de vista teórico, en la práctica se suelen comparar dos series provenientes del mismo proceso generador de datos, medidos con el mismo instrumento de medición (con el mismo nivel de ruido). Por lo tanto desde un punto de vista práctico el resultado más interesante es que en el caso de que un test detecte una diferencia entre Entropías de Permutación provenientes de series de tiempo obtenidas con el mismo instrumento de medición, hay una alta probabilidad de que se deba a que hay un cambio en la dinámica subyacente y no debido al ruido observacional.

8.2. Conclusiones

Las medidas de complejidad han mostrado ser una herramienta importante en el estudio de los sistemas dinámicos. En particular en esta Tesis se estudian las medidas de complejidad provenientes de la Función de Distribución de Bandt y Pompe. Se realizó un estudio meticuloso de la FDP de BP y las distintas formas de calcularlas según la bibliografía existente.

Esta Tesis presenta herramientas novedosas para la solución de problemáticas particulares en la estimación de la Entropía de Permutación y logra el objetivo de encontrar una metodología para hacer inferencias de manera exitosa con respecto a esta medida que se puede extrapolar a cualquier medida de complejidad derivada de la Función de Distribución de Bandt y Pompe. Finalmente se pudo mostrar de manera inferencial que la Entropía de Permutación es robusta al ruido cuando se estudian series de tiempo provenientes de un proceso generador de datos, extraídas de un mismo instrumento de medición.

8.3. Futuras líneas de investigación

Esta Tesis fue pensada y escrita de manera tal de representar un instrumento bibliográfico básico que sirva de punto de partida para los profesionales que en un futuro se incorporen al grupo de investigación.

En las Discusiones de este Capítulo se mencionaron distintas áreas de estudio que quedaron abiertas, como por ejemplo establecer un orden para los símbolos del alfabeto de Bandt y Pompe que permitan un único estimador de la Medida de Información de Fisher, y también hacer un estudio más amplio de los distintos funcionales de la entropía aplicados a la FDP de BP, como la Entropía de Permutación de Tsallis y la Entropía de Permutación de Rényi.

Por otro lado, las líneas de investigación que surgen inmediatamente de esta Tesis y son el próximo paso a seguir son las siguientes:

- El análisis de la dinámica subyacente en las señales de EEG mediante la Entropía de Permutación se ha hecho hasta el momento de manera univariada.

Sin embargo, durante el proceso de medición de estas señales se graban para un mismo paciente hasta 128 canales simultáneamente. El estudio de una posible Entropía de Permutación Multivariada que tenga en cuenta la autocorrelación una señal y también las correlaciones entre las señales de los distintos canales es un desafío interesante para estudiar.

- El método bootstrap paramétrico presentado se inspira en las cadenas de Markov, por lo tanto un estudio de las herramientas utilizadas en esta teoría puede dar una visión mas completa de los cuantificadores evaluados en la Función de Distribución de Bandt y Pompe.

Bibliografía

- Abarbanel, H. (2012). *Analysis of observed chaotic data*. Springer Science & Business Media.
- Amigó, J. (2010). *Permutation complexity in dynamical systems: ordinal patterns, permutation entropy and all that*. Springer Science & Business Media.
- Amigó, J. M., Zambrano, S., and Sanjuan, M. A. F. (2008). Combinatorial detection of determinism in noisy time series. *EPL (Europhysics Letters)*, 83(6):60005.
- Amigó, J. M., Zambrano, S., and Sanjuán, M. A. F. (2007). True and false forbidden patterns in deterministic and random dynamics. *EPL (Europhysics Letters)*, 79(5):50001.
- Andrzejak, R., Widman, G., Lehnertz, K., Rieke, C., David, P., and Elger, C. (2001a). The epileptic process as nonlinear deterministic dynamics in a stochastic environment: an evaluation on mesial temporal lobe epilepsy. *Epilepsy research*, 44(2):129–140.
- Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001b). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907.
- Astudillo, J., Zelechower, J., Traversaro, F., Redelico, F., Luna, D., Quiros, F., Risk, M., and San, E. R. (2017). Big data in the icu: Experience in the hospital italiano de buenos aires. *Studies in health technology and informatics*, 245:1319–1319.
- Bandt, C. (2014). Autocorrelation type functions for big and dirty data series. *ArXiv e-prints*.
- Bandt, C. and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102.
- Bandt, C. and Shiha, F. (2007). Order patterns in time series. *Journal of Time Series Analysis*, 28(5):646–665.
- Bian, C., Qin, C., Ma, Q. D., and Shen, Q. (2012). Modified permutation-entropy analysis of heartbeat dynamics. *Physical Review E*, 85(2):021906.
- Boeing, G. (2016). Visual analysis of nonlinear dynamical systems: Chaos, fractals, self-similarity and the limits of prediction. *Systems*, 4(4):37.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

- Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.
- Bruzzo, A. A., Gesierich, B., Santi, M., Tassinari, C. A., Birbaumer, N., and Rubboli, G. (2008). Permutation entropy to detect vigilance changes and preictal states from scalp eeg in epileptic patients. a preliminary study. *Neurological Sciences*, 29(1):3–9.
- Brzozowska-Rup, K. and Orłowski, A. (2004). Application of bootstrap to detecting chaos in financial time series. *Physica A: Statistical Mechanics and its Applications*, 344(1):317–321.
- Cáceres, M. O. (1999). Non-markovian processes with long-range correlations: fractal dimension analysis. *Brazilian Journal of Physics*, 29:125 – 135.
- Caloyannides, M. (1974). Microcycle spectral estimates of 1/f noise in semiconductors. *Journal of Applied Physics*, 45(1):307–316.
- Cambanis, S., Hardin, C. D., and Weron, A. (1988). Innovations and wold decompositions of stable sequences. *Probability theory and related fields*, 79(1):1–27.
- Cao, Y., Tung, W.-w., Gao, J., Protopopescu, V. A., and Hively, L. M. (2004). Detecting dynamical changes in time series using the permutation entropy. *Physical Review E*, 70(4):046217.
- Capurro, A., Diambra, L., Lorenzo, D., Macadar, O., Martín, M., Mostaccio, C., Plastino, A., Perez, J., Rofiman, E., Torres, M., et al. (1999). Human brain dynamics: the analysis of eeg signals with tsallis information measure. *Physica A: Statistical Mechanics and its Applications*, 265(1):235–254.
- Carpi, L. C., Saco, P. M., Figliola, A., Serrano, E., and Rosso, O. A. (2013). Analysis of an el nino-southern oscillation proxy record using information theory quantifiers. *Concepts and Recent Advances in Generalized Information Measures and Statistics*, page 3.
- Carpi, L. C., Saco, P. M., and Rosso, O. (2010). Missing ordinal patterns in correlated noises. *Physica A: Statistical Mechanics and its Applications*, 389(10):2020 – 2029.
- Chernick, M. R. (1981). A limit theorem for the maximum of autoregressive processes with uniform marginal distributions. *The Annals of Probability*, pages 145–149.
- De Micco, L., González, C., Larrondo, H., Martin, M., Plastino, A., and Rosso, O. (2008). Randomizing nonlinear maps via symbolic dynamics. *Physica A: Statistical Mechanics and its Applications*, 387(14):3373–3383.
- Ding, M., Grebogi, C., Ott, E., Sauer, T., and Yorke, J. A. (1993). Plateau onset for correlation dimension: When does it occur? *Physical Review Letters*, 70(25):3872.
- Donders, A. R. T., van der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087 – 1091.

- Dutta, P. and Horn, P. (1981). Low-frequency fluctuations in solids: 1 f noise. *Reviews of Modern physics*, 53(3):497.
- Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, pages 1127–1162.
- Fadlallah, B., Chen, B., Keil, A., and Principe, J. (2013). Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. *Phys. Rev. E*, 87:022911.
- Farashi, S. (2015). Spike sorting method using exponential autoregressive modeling of action potentials. *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, 8(12):864–870.
- Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., Engel, J., Forsgren, L., French, J. A., Glynn, M., et al. (2014). Ilae official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482.
- Fisher, R. S., Boas, W. v. E., Blume, W., Elger, C., Genton, P., Lee, P., and Engel, J. (2005). Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, 46(4):470–472.
- Frank, B., Pompe, B., Schneider, U., and Hoyer, D. (2006). Permutation entropy improves fetal behavioural state classification based on heart rate analysis from biomagnetic recordings in near term fetuses. *Medical and Biological Engineering and Computing*, 44(3):179.
- Frieden, B. R. (1990). Fisher information, disorder, and the equilibrium distributions of physics. *Physical Review A*, 41(8):4265.
- Frieden, B. R. (2004). *Science from Fisher information: a unification*. Cambridge University Press.
- Gençay, R. (1996). A statistical framework for testing chaotic dynamics via lyapunov exponents. *Physica D: Nonlinear Phenomena*, 89(3-4):261–266.
- Good, I. (1975). The number of orderings of n candidates when ties are permitted. *Fib. Quart*, 13:11–18.
- Grassberger, P. and Procaccia, I. (1983). Characterization of strange attractors. *Physical review letters*, 50(5):346.
- Hafner, C. (2013). *Nonlinear time series analysis with applications to foreign exchange rate volatility*. Springer Science & Business Media.

- Harremoës, P. (2006). Interpretations of rényi entropies and divergences. *Physica A: Statistical Mechanics and its Applications*, 365(1):57–62.
- Henríquez, P., Alonso, J. B., Ferrer, M. A., and Travieso, C. M. (2014). Review of automatic fault diagnosis systems using audio and vibration signals. *IEEE Transactions on systems, man, and cybernetics: Systems*, 44:642–652.
- Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.
- Ishizuka, K., Solvang, H. K., and Nakatani, T. (2005). Speech signal analysis with exponential autoregressive model. In *ICASSP (1)*, pages 225–228.
- Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical review*, 106(4):620.
- Jaynes, E. T. (1957b). Information theory and statistical mechanics. ii. *Physical review*, 108(2):171.
- Jordan, D., Stockmanns, G., Kochs, E. F., Pilge, S., and Schneider, G. (2008). Electroencephalographic order pattern analysis for the separation of consciousness and unconsciousness: an analysis of approximate entropy, permutation entropy, recurrence rate, and phase coupling of order recurrence plots. *The Journal of the American Society of Anesthesiologists*, 109(6):1014–1022.
- Kannathal, N., Choo, M. L., Acharya, U. R., and Sadasivan, P. (2005). Entropies for detection of epilepsy in eeg. *Computer methods and programs in biomedicine*, 80(3):187–194.
- Kasdin, N. J. (1995). Discrete simulation of colored noise and stochastic processes and 1/f/spl alpha//power law noise generation. *Proceedings of the IEEE*, 83(5):802–827.
- Keller, K. and Sinn, M. (2005). Ordinal analysis of time series. *Physica A: Statistical Mechanics and its Applications*, 356(1):114–120.
- Keller, K. and Wittfeld, K. (2004). Distances of time series components by means of symbolic dynamics. *International Journal of Bifurcation and Chaos*, 14(02):693–703.
- Kobayashi, M. and Musha, T. (1982). 1/f fluctuation of heartbeat period. *IEEE transactions on Biomedical Engineering*, (6):456–457.
- Kolmogorov, A. N. (1958). A new metric invariant of transient dynamical systems and automorphisms in lebesgue spaces. In *Dokl. Akad. Nauk SSSR (NS)*, volume 119, page 2.
- Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.
- L. Zunino, M. C. S. and Rosso, O. A. (2012). Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach. *PHYSICAL REVIEW E*, 86.

- Lawrance, A. (1992). Uniformly distributed first-order autoregressive time series models and multiplicative congruential random number generators. *Journal of Applied Probability*, pages 896–903.
- Lawrance, A. and Lewis, P. (1981). A new autoregressive time series model in exponential variables (near (1)). *Advances in Applied Probability*, pages 826–845.
- Li, H., Heusdens, R., Muskulus, M., and Wolters, L. (2007a). Analysis and synthesis of pseudo-periodic job arrivals in grids: A matching pursuit approach. In *Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid'07)*, pages 183–196. IEEE.
- Li, X., Ouyang, G., and Richards, D. A. (2007b). Predictability analysis of absence seizures with permutation entropy. *Epilepsy research*, 77(1):70–74.
- Lopez-Ruiz, R., Mancini, H. L., and Calbet, X. (1995). A statistical measure of complexity. *Physics Letters A*, 209(5-6):321–326.
- Luque, B., Lacasa, L., Ballesteros, F., and Luque, J. (2009). Horizontal visibility graphs: Exact results for random time series. *Physical Review E*, 80(4):046103.
- Mammone, N., Duun-Henriksen, J., Kjaer, T. W., and Morabito, F. C. (2015). Differentiating interictal and ictal states in childhood absence epilepsy through permutation rényi entropy. *Entropy*, 17(7):4627–4643.
- Martin, M., Pennini, F., and Plastino, A. (1999). Fisher’s information and the analysis of complex signals. *Physics Letters A*, 256(2–3):173 – 180.
- Masoller, C. and Rosso, O. A. (2011). Quantifying the complexity of the delayed logistic map. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 369(1935):425–438.
- Maszczyk, T. and Duch, W. (2008). Comparison of shannon, renyi and tsallis entropy used in decision trees. In *International Conference on Artificial Intelligence and Soft Computing*, pages 643–651. Springer.
- Matilla-García, M. and Marín, M. R. (2009). Detection of non-linear structure in time series. *Economics Letters*, 105(1):1–6.
- McKenzie, E. (1988). Some arma models for dependent sequences of poisson counts. *Advances in Applied Probability*, pages 822–835.
- Nicolaou, N. and Georgiou, J. (2012). Detection of epileptic electroencephalogram based on permutation entropy and support vector machines. *Expert Systems with Applications*, 39(1):202–209.
- Novikov, E., Novikov, A., Shannahoff-Khalsa, D., Schwartz, B., and Wright, J. (1997). Scale-similar activity in the brain. *Phys. Rev. E*, 56:R2387–R2389.
- Ocak, H. (2009). Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy. *Expert Systems with Applications*, 36(2):2027–2036.

- Olivares, F., Plastino, A., and Rosso, O. A. (2012a). Ambiguities in bandt–pompe’s methodology for local entropic quantifiers. *Physica A: Statistical Mechanics and its Applications*, 391(8):2518–2526.
- Olivares, F., Plastino, A., and Rosso, O. A. (2012b). Contrasting chaos with noise via local versus global information quantifiers. *Phys. Lett A*, 376:1577–1583.
- Olofsen, E., Sleight, J., and Dahan, A. (2008). Permutation entropy of the electroencephalogram: a measure of anaesthetic drug effect. *British journal of anaesthesia*, 101(6):810–821.
- Ouyang, G., Dang, C., Richards, D. A., and Li, X. (2010). Ordinal pattern based similarity analysis for eeg recordings. *Clinical Neurophysiology*, 121(5):694–703.
- Ouyang, G., Li, X., Dang, C., and Richards, D. A. (2009). Deterministic dynamics of neural activity during absence seizures in rats. *Physical Review E*, 79(4):041146.
- Parlitz, U., Berg, S., Luther, S., Schirdewan, A., Kurths, J., and Wessel, N. (2012). Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics. *Computers in biology and medicine*, 42(3):319–327.
- Pavón, D. (1987). Thermodynamics of superstrings. *General Relativity and Gravitation*, 19(4).
- Payton, M. E., Greenstone, M. H., and Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? *Journal of Insect Science*, 3(1):34.
- Plastino A., Rosso, O. A. (2005). Entropy and statistical complexity in brain activity. *Europhysics News*, 36(6):224–228.
- Polat, K. and Güneş, S. (2007). Classification of epileptiform eeg using a hybrid system based on decision tree classifier and fast fourier transform. *Applied Mathematics and Computation*, 187(2):1017–1026.
- Quintero-Quiroz, C., Pigolotti, S., Torrent, M., and Masoller, C. (2015). Numerical and experimental study of the effects of noise on the permutation entropy. *New Journal of Physics*, 17(9):093002.
- RA Fisher, M. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368.
- Redelico, F. O., Traversaro, F., García, M. d. C., Silva, W., Rosso, O. A., and Risk, M. (2017a). Classification of normal and pre-ictal eeg signals using permutation entropies and a generalized linear model as a classifier. *Entropy*, 19(2):72.
- Redelico, F. O., Traversaro, F., Oyarzabal, N., Vilaboa, I., and Rosso, O. A. (2017b). Evaluation of the status of rotary machines by time causal information theory quantifiers. *Physica A: Statistical Mechanics and its Applications*, 470:321–329.
- Rényi, A. (1961). On measures of entropy and information. Technical report, HUNGARIAN ACADEMY OF SCIENCES Budapest Hungary.
- Riedl, M., Müller, A., and Wessel, N. (2013). Practical considerations of permutation entropy. *The European Physical Journal Special Topics*, 222(2):249–262.

- Rosso, O., Carpi, L., Saco, P., Ravetti, M. G., Larrondo, H., and Plastino, A. (2012a). The amigó paradigm of forbidden/missing patterns: a detailed analysis. *The European Physical Journal B*, 85(12):1–12.
- Rosso, O., Martin, M., Figliola, A., Keller, K., and Plastino, A. (2006). Eeg analysis using wavelet-based information tools. *Journal of neuroscience methods*, 153(2):163–182.
- Rosso, O., Olivares, F., and Plastino, A. (2015). Noise versus chaos in a causal fisher-shannon plane. *Papers in Physics*, 7(0).
- Rosso, O., Zunino, L., Pérez, D., Figliola, A., Larrondo, H., Garavaglia, M., Martín, M., and Plastino, A. (2007a). Extracting features of gaussian self-similar stochastic processes via the bandt-pompe approach. *Physical Review E*, 76(6):061114.
- Rosso, O. A., Carpi, L. C., Saco, P. M., Ravetti, M. G., Plastino, A., and Larrondo, H. A. (2012b). Causality and the entropy–complexity plane: Robustness and missing ordinal patterns. *Physica A: Statistical Mechanics and its Applications*, 391(1):42–55.
- Rosso, O. A., De Micco, L., Larrondo, H. A., Martín, M. T., and Plastino, A. (2010a). Generalized statistical complexity measure. *International Journal of Bifurcation and Chaos*, 20(03):775–785.
- Rosso, O. A., Larrondo, H., Martin, M., Plastino, A., and Fuentes, M. (2007b). Distinguishing noise from chaos. *Physical review letters*, 99(15):154102.
- Rosso, O. A., Micco, L. D., Plastino, A., and Larrondo, H. A. (2010b). Info-quantifiers’ map-characterization revisited. *Physica A: Statistical Mechanics and its Applications*, 389(21):4604 – 4612.
- Rosso, O. A., Olivares, F., Zunino, L., De Micco, L., Aquino, A. L., Plastino, A., and Larrondo, H. A. (2013). Characterization of chaotic maps using the permutation bandt-pompe probability distribution. *The European Physical Journal B*, 86(4):1–13.
- Saco, P. M., Carpi, L. C., Figliola, A., Serrano, E., and Rosso, O. A. (2010). Entropy analysis of the dynamics of el niño/southern oscillation during the holocene. *Physica A: Statistical Mechanics and its Applications*, 389(21):5022 – 5027.
- Schindler, K., Gast, H., Stieglitz, L., Stibal, A., Hauf, M., Wiest, R., Mariani, L., and Rummel, C. (2011). periictal intracranial eeg indicate deterministic dynamics in human epileptic seizures. *epilepsia*. 52 (10): 1771–80.[doi. *Epilepsia*, 53(1):225.
- Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.
- Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Champaign, USA.
- Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

- Sheng, S., Zhang, L., and Gao, R. (2006). A systematic sensor placement strategy for enhanced defect detection in rolling bearings. *IEEE Sensors Journal*, 6:1346–1354.
- Sinai, I. (1959a). On the concept of entropy for a dynamic system. *Doklady Akademii Nauk SSSR*, 124(4):768–771.
- Sinai, Y. (1959b). Flows with finite entropy. In *Doklady Akad. Nauk SSSR*, volume 125. Lomonosov Moscow State Univ.
- Sinn, M. and Keller, K. (2011). Estimation of ordinal pattern probabilities in gaussian processes with stationary increments. *Computational Statistics & Data Analysis*, 55(4):1781–1790.
- Subasi, A. (2007). Eeg signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, 32(4):1084–1093.
- Timmer, J. and Koenig, M. (1995). On generating power law noise. *Astronomy and Astrophysics*, 300:707.
- Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487.
- Tukey, J. W. (1977). *Exploratory data analysis*, volume 2. Reading, Mass.
- Veisi, I., Pariz, N., and Karimpour, A. (2007). Fast and robust detection of epilepsy in noisy eeg signals using permutation entropy. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 200–203. IEEE.
- Vignat, C. and Bercher, J.-F. (2003). Analysis of signals in the fisher–shannon information plane. *Physics Letters A*, 312(1-2):27–33.
- Voss, R. F. and Clarke, J. (1975). $1/f$ noise in music and speech. *Nature*, 258:317–318.
- Vuong, P. L., Malik, A. S., and Bornot, J. (2014). Weighted-permutation entropy as complexity measure for electroencephalographic time series of different physiological states. *Phys. Rev. E*, pages 979–984.
- Wold, H. (1938). *A study in the analysis of stationary time series*. PhD thesis, Almqvist & Wiksell.
- Yan, R., Liu, Y., and Gao, R. X. (2012). Permutation entropy: a nonlinear statistical measure for status characterization of rotary machines. *Mechanical Systems and Signal Processing*, 29:474–484.
- Yohai, V. and Boente, G. (2004). Introducción a la inferencia estadística. *Notas no publicadas. ICFCE y N.(UBA)-Buenos Aires*.
- Zanin, M., Zunino, L., Rosso, O. A., and Papo, D. (2012). Permutation entropy and its main biomedical and econophysics applications: a review. *Entropy*, 14(8):1553–1577.

- Zelechower, J., Astudillo, J., Traversaro, F., Redelico, F., Luna, D., Quiros, F., San, E. R., and Risk, M. (2017). Infrastructure for big data in the intensive care unit. *Studies in health technology and informatics*, 245:1346–1346.
- Ziehmann, C., Smith, L. A., and Kurths, J. (1999). The bootstrap and lyapunov exponents in deterministic chaos. *Physica D: Nonlinear Phenomena*, 126(1):49–59.
- Zunino, L., Olivares, F., and Rosso, O. A. (2015). Permutation min-entropy: An improved quantifier for unveiling subtle temporal correlations. *EPL (Europhysics Letters)*, 109(1):10005.
- Zunino, L., Olivares, F., Scholkmann, F., and Rosso, O. A. (2017). Permutation entropy based time series analysis: Equalities in the input signal can lead to false conclusions. *Physics Letters A*, 381(22):1883–1892.
- Zunino, L., Pérez, D., Martín, M., Garavaglia, M., Plastino, A., and Rosso, O. (2008a). Permutation entropy of fractional brownian motion and fractional gaussian noise. *Physics Letters A*, 372(27):4768–4774.
- Zunino, L., Pérez, D., Kowalski, A., Martín, M., Garavaglia, M., Plastino, A., and Rosso, O. (2008b). Fractional brownian motion, fractional gaussian noise, and tsallis permutation entropy. *Physica A: Statistical Mechanics and its Applications*, 387(24):6057 – 6068.

Apéndice A

Algoritmos bootstrap

A.1. Algoritmo 1

Algorithm 1 Algoritmo del bootstrap paramétrico para la Entropía de Permutación

- 1: $T \leftarrow$ largo de la serie de tiempo
 - 2: **set** m
 - 3: **set** τ
 - 4: **computar** $\hat{P}_T(\pi_i)$ (Ecuación 6.1) de la serie de tiempo
 - 5: **computar** $\hat{\mathcal{H}}_T$ (Ecuación 6.5) de la serie de tiempo
 - 6: **computar** \hat{P}_T^{ij} (Ecuación 6.3) de la serie de tiempo
 - 7: $b \leftarrow 1$
 - 8: **while** $b \leq B$ **do**
 - 9: $i \leftarrow 1$
 - 10: $s_i^*(b) \leftarrow \pi_k$ c.p. $\hat{P}_T(\boldsymbol{\pi})$ {i. e. el estado inicial de la b-ésima replica bootstrap}
 - 11: **while** $i \leq T - m + 1$ **do**
 - 12: $s_{(i+1)}^*(b) \leftarrow \pi_k$ w.p. $\hat{P}_T^{ik}(\boldsymbol{\pi})$ {i. e. el i-ésimo estado para la b-ésima replica bootstrap}
 - 13: $i \leftarrow i + 1$
 - 14: **end while**
 - 15: **estimar** $\hat{P}^*(\boldsymbol{\pi})$ usando $\mathbf{S}^*(b)$ Ecuación 6.1
 - 16: **estimar** $\hat{\mathcal{H}}_T^*(b)$ usando $\hat{P}^*(\boldsymbol{\pi})$ y la Ecuación 6.5 {i. e. la b-ésima estimación bootstrap}
 - 17: **end while**
-

A.2. Algoritmo 2

Algorithm 2 Algoritmo para construir un intervalo de confianza de nivel $1 - \alpha$ para la Entropía de Permutación

- 1: **while** $b \leq B$ **do**
 - 2: **generar** $\hat{\mathcal{H}}_T^*(b)$
 - 3: **end while**
 - 4: **computar** $\hat{\mathcal{H}}_T^*(\bullet) = \frac{1}{B} \sum_{i=1}^B \hat{\mathcal{H}}_T^*(i)$
 - 5: **ordenar** $\delta^*(b) = \hat{\mathcal{H}}_T^*(b) - \hat{\mathcal{H}}_T^*(\bullet)$ en orden creciente
 - 6: **set** nivel de confianza $1 - \alpha$
 - 7: **computar** $\delta_{\frac{\alpha}{2}}^* \leftarrow \left\{ \delta_{\frac{\alpha}{2}}^* / \frac{\#(\delta^* < \delta_{\frac{\alpha}{2}}^*)}{B} \leq \frac{\alpha}{2} \right\}$
 {i. e. Si $B = 1000$ y $\alpha = 0,1$ elegir el elemento 50 del vector ordenado δ^* }
 - 8: **computar**
 $\delta_{(1-\frac{\alpha}{2})}^* \leftarrow \left\{ \delta_{(1-\frac{\alpha}{2})}^* / \frac{\#(\delta^* < \delta_{(1-\frac{\alpha}{2})}^*)}{B} \leq 1 - \frac{\alpha}{2} \right\}$
 {i. e. Si $B = 1000$ y $\alpha = 0,1$ elegir el elemento 950 del vector ordenado δ^* }
 - 9: El límite inferior del intervalo de confianza es
 $\text{máx}(2\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_T^*(\bullet) + \delta_{\frac{\alpha}{2}}^*, 0)$
 - 10: El límite superior del intervalo de confianza es
 $\text{mín}(2\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_T^*(\bullet) + \delta_{(1-\frac{\alpha}{2})}^*, 1)$
-

A.3. Algoritmo 3

Algorithm 3 Algoritmo para el test de hipótesis de la diferencia de Entropías de Permutación

- 1: **computar** $\hat{\mathcal{H}}_{1T}$ la Entropía de Permutación de la 1ra serie de tiempo
 - 2: **computar** $\hat{\mathcal{H}}_{2T}$ la Entropía de Permutación de la 2da serie de tiempo
 - 3: **computar** $\hat{\Delta}_T = \hat{\mathcal{H}}_{1T} - \hat{\mathcal{H}}_{2T}$
 - 4: **while** $b \leq B$ **do**
 - 5: **generar** $\hat{\mathcal{H}}_{1T}^*(b)$ la réplica bootstrap de la 1ra serie de tiempo
 - 6: **generar** $\hat{\mathcal{H}}_{2T}^*(b)$ la réplica bootstrap de la 2da serie de tiempo
 - 7: **end while**
 - 8: **for** i in 1 to B **do**
 - 9: **for** k in 1 to B **do**
 - 10: **computar** $\Delta_T^*(n) = \hat{\mathcal{H}}_{1T}^*(i) - \hat{\mathcal{H}}_{2T}^*(k)$
 - 11: **end for**
 - 12: **end for**
 - 13: **computar** $\hat{\Delta}_T^*(\bullet) = \frac{1}{B^2} \sum_{i=1}^{B^2} \Delta_T^*(n)$
 - 14: **ordenar** $\delta^*(n) = \Delta_T^*(n) - \hat{\Delta}_T^*(\bullet)$ en orden creciente
 - 15: **set** nivel de confianza $1 - \alpha$
 - 16: **computar** $\delta_{\frac{\alpha}{2}}^* \leftarrow \left\{ \delta_{\frac{\alpha}{2}}^* / \frac{\#\left(\delta^* < \delta_{\frac{\alpha}{2}}^*\right)}{B} \leq \frac{\alpha}{2} \right\}$
 {i. e. Si $B = 1000$ y $\alpha = 0,1$ elegir el elemento 50 del vector ordenado δ^* }
 - 17: **compute**
 $\delta_{(1-\frac{\alpha}{2})}^* \leftarrow \left\{ \delta_{(1-\frac{\alpha}{2})}^* / \frac{\#\left(\delta^* < \delta_{(1-\frac{\alpha}{2})}^*\right)}{B} \leq 1 - \frac{\alpha}{2} \right\}$
 {i. e. Si $B = 1000$ y $\alpha = 0,1$ elegir el elemento 950 del vector ordenado δ^* }
 - 18: El límite inferior del intervalo de confianza es
 $\hat{\Delta}_T + \delta_{\frac{\alpha}{2}}^*$
 - 19: El límite superior del intervalo de confianza es
 $\hat{\Delta}_T + \delta_{(1-\frac{\alpha}{2})}^*$
 - 20: Si 0 no pertenece al intervalo
 Entonces $\mathcal{H}_1 \neq \mathcal{H}_2$ con una probabilidad de cometer un error α
-

Apéndice B

Regla de Scott para la cantidad de intervalos de un histograma

La estimación de la Entropía de Amplitud presentada en el Capítulo 4 se basa en la estimación de la función de probabilidades de las amplitudes de los valores de la serie de tiempo mediante la función histograma. En esta Tesis una heurística simple es utilizada para definir la cantidad de intervalos de dicha función histograma: la regla de Scott. Tiene la propiedad de que es cercana al límite superior universal del ancho máximo óptimo y en el caso de ser una distribución Gaussiana minimiza el Error Medio Cuadrático Integrado (Scott [1979]).

En la Figura B.1, \mathcal{H}_a es calculada para distribuciones no correlacionadas Gaussianas, Uniformes y Gaussianas en función de la varianza de la variable aleatoria simulada σ^2 usando la regla de Scott. Se muestra que mediante esta regla es posible distinguir entre las distintas distribuciones al calcular \mathcal{H}_a independientemente de la varianza de la variable aleatoria. Por lo tanto esta regla es robusta frente a cambios en la varianza y muestra ser muy útil para la construcción del plano $\mathcal{H} \times \mathcal{H}_a$.

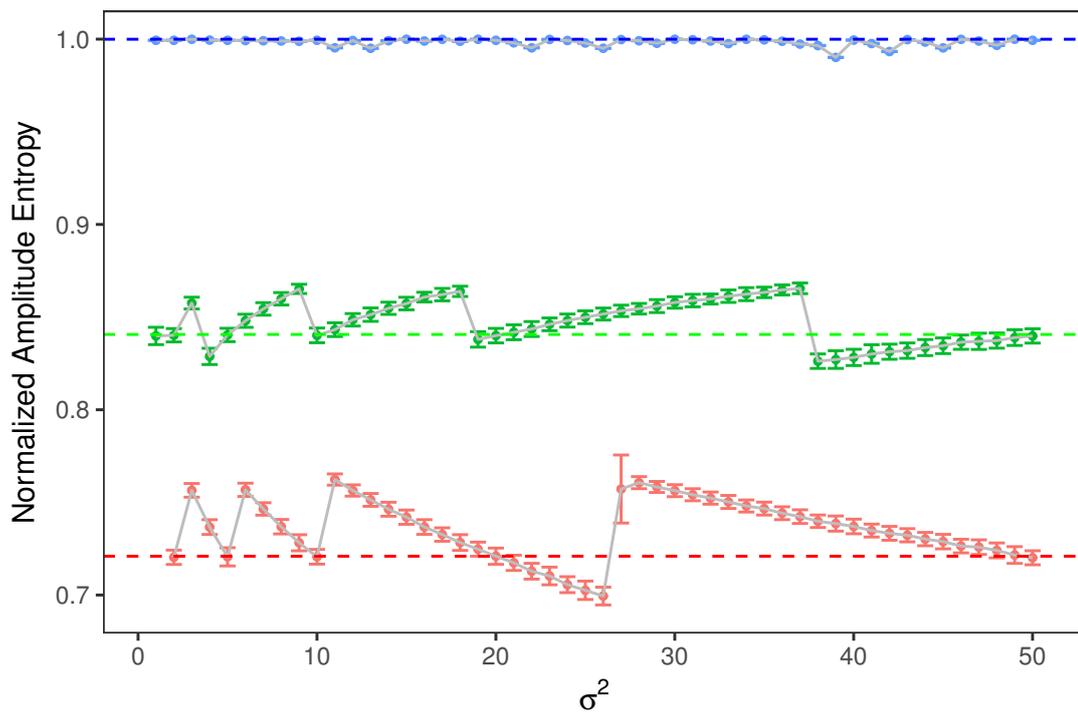


Figura B.1 **La Entropía de Amplitud para los tres procesos estocásticos descorrelacionados.** En azul: la distribución Uniforme, verde: distribución Gaussiana y rojo: distribución Exponencial. Cada una de éstas fue simulada usando distintas varianzas (eje x). La regla de Scott fue utilizada para definir la cantidad de intervalos. La Entropía de Amplitud calculada de esta manera es capaz de distinguir entre las distintas distribuciones de probabilidad para un amplio rango de varianzas.

Apéndice C

Códigos en R

```
1
2 ###FUNCION DE DISTRIBUCION DE BANDT Y POMPE PARA m=3,4,5,6
3 require(combinat)
4
5 # x es la serie a analizar
6 # d es el retardo de tiempo
7 # m es la dimension de embedding
8
9 bandt.pdf<-function(x,d=1,m=3)
10
11 {
12   T<-length(x)
13   dmax<-1
14   proba<-vector("numeric")
15
16   if (m==3)
17   {
18     y0<-x[1:(T-2*d)]
19     y1<-x[(d+1):(T-d)]
20     y2<-x[(2*d+1):(T)]
21
22     s<-2*((y0>y1)+(y0>y2))+(y1>y2)
23
24     h<-is.na(y0) | is.na(y1) | is.na(y2) | (y0==y1) | (y0==y2) | (y1==y2)
```

```

25     dupli<-sum((y0==y1)|(y0==y2)|(y1==y2))
26     s<-s+6*h
27
28     p<-hist(s,breaks=c(-1:10),plot=FALSE)$counts
29     # cuenta la cantidad de veces que aparece cada patron
30     if (sum(p[1:6])==0) {
31         proba<-rep(0,6)
32     } else
33     {
34         proba<-p[1:6] / sum(p[1:6])
35     }
36
37     freq<-p[1:6]
38 }
39
40 if (m==4)
41 {
42     y0<-x[1:(T-3*d)]
43     y1<-x[(d+1):(T-2*d)]
44     y2<-x[(2*d+1):(T-d)]
45     y3<-x[(3*d+1):(T)]
46
47     s<-6*((y0>y1)+(y0>y2)+(y0>y3))+2*((y1>y2)+(y1>y3))+(y2>y3)
48
49
50
51     h<-is.na(y0)|is.na(y1)|is.na(y2)|is.na(y3)|(y0==y1)|(y0==y2)
52         |(y1==y2)|(y0==y3)|(y1==y3)|(y2==y3)
53     s<-s+24*h
54     dupli<-sum((y0==y1)|(y0==y2)|(y1==y2)|(y0==y3)|(y1==y3)|(y2==
55         y3))
56
57     p<-hist(s,breaks=c(-1:47),plot=FALSE)$counts
58
59     if (sum(p[1:24])==0) {
60         proba<-rep(0,24)

```

```

60     } else
61     {
62         proba<-p[1:24] / sum(p[1:24])
63     }
64     freq<-p[1:24]
65 }
66
67 if (m==5)
68 {
69     y0<-x[1:(T-4*d)]
70     y1<-x[(d+1):(T-3*d)]
71     y2<-x[(2*d+1):(T-2*d)]
72     y3<-x[(3*d+1):(T-d)]
73     y4<-x[(4*d+1):(T)]
74
75
76     s<-24*((y0>y1)+(y0>y2)+(y0>y3)+(y0>y4))+
77         6*((y1>y2)+(y1>y3)+(y1>y4))+2*((y2>y3)+(y2>y4))+(y3>y4)
78
79     h<-is.na(y0) | is.na(y1) | is.na(y2) | is.na(y3) | is.na(y4) |
80         (y0==y1) |(y0==y2) |(y1==y2) |(y0==y3) |(y1==y3) |(y2==y3) |
81         (y0==y4) |(y1==y4) |(y2==y4) |(y3==y4)
82
83     s<-s+120*h
84     dupli<-sum( (y0==y1) |(y0==y2) |(y1==y2) |(y0==y3) |(y1==y3) |(y2
85         ==y3) |
86         (y0==y4) |(y1==y4) |(y2==y4) |(y3==y4))
87
88     p<-hist(s, breaks=c(-1:239), plot=FALSE)$counts
89
90     if (sum(p[1:120])==0) {
91         proba<-rep(0,120)
92     } else
93     {
94         proba<-p[1:120] / sum(p[1:120])
95     }
96     freq<-p[1:120]

```

```

96   }
97
98   if (m==6)
99   {
100     y0<-x[1:(T-5*d)]
101     y1<-x[(d+1):(T-4*d)]
102     y2<-x[(2*d+1):(T-3*d)]
103     y3<-x[(3*d+1):(T-2*d)]
104     y4<-x[(4*d+1):(T-d)]
105     y5<-x[(5*d+1):T]
106
107     s<-120*((y0>y1)+(y0>y2)+(y0>y3)+(y0>y4)+(y0>y5))+24*((y1>y2)
108       +(y1>y3)+(y1>y4)+(y1>y5))+
109       6*((y2>y3)+(y2>y4)+(y2>y5))+2*((y3>y4)+(y3>y5))+(y4>y5)
110
111     h<-is.na(y0) | is.na(y1) | is.na(y2) | is.na(y3) | is.na(y4) | is.na(y5
112       ) |
113       (y0==y1) | (y0==y2) | (y0==y3) | (y0==y4) | (y0==y5) |
114       (y1==y2) | (y1==y3) | (y1==y4) | (y1==y5) |
115       (y2==y3) | (y2==y4) | (y2==y5) |
116       (y3==y4) | (y3==y5) |
117       (y4==y5)
118
119     s<-s+720*h
120     dupli<-sum((y0==y1) | (y0==y2) | (y0==y3) | (y0==y4) | (y0==
121       y5) |
122       (y1==y2) | (y1==y3) | (y1==y4) | (y1==y5) |
123       (y2==y3) | (y2==y4) | (y2==y5) |
124       (y3==y4) | (y3==y5) |
125       (y4==y5))
126
127     p<-hist(s, breaks=c(-1:1439), plot=FALSE)$counts
128
129     if (sum(p[1:720])==0) {
130       proba<-rep(0,720)
131     } else
132     {

```

```

130     proba<-p[1:720] / sum(p[1:720])
131   }
132   freq<-p[1:720]
133 }
134
135 if (m==7)
136 {
137   y0<-x[1:(T-6*d)]
138   y1<-x[(d+1):(T-5*d)]
139   y2<-x[(2*d+1):(T-4*d)]
140   y3<-x[(3*d+1):(T-3*d)]
141   y4<-x[(4*d+1):(T-2*d)]
142   y5<-x[(5*d+1):(T-d)]
143   y6<-x[(6*d+1):T]
144
145   s<-720*((y0>y1)+(y0>y2)+(y0>y3)+(y0>y4)+(y0>y5)+(y0>y6))+
146     120*((y1>y2)+(y1>y3)+(y1>y4)+(y1>y5)+(y1>y6))+24*((y2>y3)
147     +(y2>y4)+(y2>y5)+(y2>y6))+
148     6*((y3>y4)+(y3>y5)+(y3>y6))+2*((y4>y5)+(y4>y6))+(y5>y6)
149
150   h<-is.na(y0) | is.na(y1) | is.na(y2) | is.na(y3) | is.na(y4) | is.na(y5
151     ) | is.na(y6)
152   (y0==y1) | (y0==y2) | (y0==y3) | (y0==y4) | (y0==y5) | (y0==y6)
153   (y1==y2) | (y1==y3) | (y1==y4) | (y1==y5) | (y1==y6) |
154   (y2==y3) | (y2==y4) | (y2==y5) | (y2==y6) |
155   (y3==y4) | (y3==y5) | (y3==y6) |
156   (y4==y5) | (y4==y6) |
157   (y5==y6)
158
159   s<-s+5040*h
160   dupli<-sum((y0==y1) | (y0==y2) | (y0==y3) | (y0==y4) | (y0==y5) | (y0==
161     y6) |
162     (y1==y2) | (y1==y3) | (y1==y4) | (y1==y5) | (y1==y6) |
163     (y2==y3) | (y2==y4) | (y2==y5) | (y2==y6) |
164     (y3==y4) | (y3==y5) | (y3==y6) |
165     (y4==y5) | (y4==y6) |
166     (y5==y6) )

```

```
164
165     p<-hist(s,breaks=c(-1:10079),plot=FALSE)$counts
166
167     if (sum(p[1:5040])==0) {
168         proba<-rep(0,5040)
169     } else
170     {
171         proba<-p[1:5040] / sum(p[1:5040])
172     }
173     freq<-p[1:5040]
174 }
175 palabras<-sort( unlist( permn( as.character(1:m),fun=paste,collapse
176     =" ")))
177 names(freq)<-palabras
178 names(proba)<-palabras
179 salida<-list(freq,proba,dupli)
180 names(salida)<-c("Frequency","Pi","Repeated")
181 salida
181 }
```

```

1 # SERIES AUTORREGRESIVAS EXPONENCIALES CON EL ALGORITMO DE LEWIS AND
  LAWRENCE (1980)
2 rho<-c(0.75,0.5,0.25,0.125)
3 alpha<-2*rho/(1+rho)
4
5 largo<-250000
6 en<-vector(length=largo)
7 epsilon<-function(largo,alpha,beta)
8 {
9   En<-rexp(largo,rate=1)
10  proba<-(1-beta)/(1-beta+alpha*beta)
11  bn<-rbinom(largo,size=1,proba)
12  en[bn==1]<-En[bn==1]
13  en[bn==0]<-(1-alpha)*beta*En[bn==0]
14  en
15 }
16
17 for(i in alpha)
18 {
19   for(j in 1:10)
20   {
21     x<-vector(length=largo)
22     beta<-1/(2-i)
23     rho<-i*beta
24     en<-epsilon(largo,i,beta)
25     bn<-rbinom(largo,1,i)
26
27     x[1]<-rexp(1,rate=1)
28     for(l in 1:(largo-1))
29     {
30       x[l+1]<-en[l+1]+bn[l+1]*beta*x[l]
31
32     }
33
34     write.csv(x,paste0("exp-",rho,"run",j,".csv"))
35   }

```

```

36 }
37
38
39 #NEARA(1) EN CASO DE CORRELACIONES NEGATIVAS
40
41
42 antiVn<-function(largo , alpha)
43 {
44   Un<-runif(largo)
45   Vn<-as.numeric(Un<=alpha)
46   Vnp<-as.numeric((Un>=(1-alpha)))
47   cbind(Vn,Vnp)
48 }
49
50 antiKn<-function(largo , alpha , beta)
51 {
52   p<-(1-beta)/(1-(1-alpha)*beta)
53   Un<-runif(largo)
54   Kn<-ifelse(Un<=p,1,(1-alpha)*beta)
55   Knp<-ifelse(Un>=(1-p),1,(1-alpha)*beta)
56   cbind(Kn,Knp)
57 }
58
59 antiEn<-function(largo)
60 {
61   c=0.025
62   d=0.9
63   lambdap=1
64   lambdaq=1
65   a<-c*lambdap
66   b<-(1-d)*lambdaq
67   Pn<-rexp(largo , rate=lambdap)
68   Qn<-rexp(largo , rate = lambdaq)
69   Un<-runif(largo)
70   valorV<-function(U) {-log((1-U)/(1-c))}
71   valorW<-function(U) {-log(U/d) }
72   Vn<-valorV(Un)*as.numeric(Un>c)

```

```

73 Wn<-valorW(Un)*as.numeric(Un<d)
74 En<-a*Pn+Vn
75 Enp<-b*Qn+Wn
76 cbind(En,Enp)
77 }
78
79 ### rho 0.05 si alpha=0.3 y beta =0.3
80 ### rho 0.10 si alpha=0.6 y beta =0.4
81 ### rho 0.15 si alpha=0.6 y beta =0.7
82 ### rho 0.05 si alpha=0.5 y beta =1
83
84 coeficientes<-matrix(c
      (0.3,0.3,0.05,0.6,0.4,0.1,0.6,0.7,0.15,0.5,1,0.2),byrow=TRUE,ncol
      =3)
85 for(j in 1:10)
86 {
87   for(i in 1:4)
88   {
89     print(alpha<-coeficientes[i,1])
90     print(beta<-coeficientes[i,2])
91     print(rho<-coeficientes[i,3])
92     en<-NULL
93     enp<-NULL
94     En<-antiEn(largo)
95     Kn<-antiKn(largo,alpha,beta)
96     en<-En[,1]*Kn[,1]
97     enp<-En[,2]*Kn[,2]
98     Vn<-antiVn(largo,alpha)
99     Xn<-NULL
100    Xnp<-NULL
101    Xn[1]<-rexp(1)
102    Xnp[1]<-rexp(1)
103    for(i in 2:largo)
104    {
105      Xn[i]<-en[i]+beta*Vn[i,1]*Xnp[i-1]
106      Xnp[i]<-enp[i]+beta*Vn[i,2]*Xn[i-1]
107    }

```

```
108
109     write.csv(Xn, paste0("exp-N", rho, "run", j, ".csv"))
110 }
111 }
112
113
114 ## Ruido exponencial descorrelacionado
115
116 for (j in 1:10)
117 {
118     x<-rexp(250000)
119     write.csv(x, paste0("exp-", 0, "run", j, ".csv"))
120
121 }
```

```
1 #SERIES AUTOREGRESIVAS UNIFORMES CON EL ALGORITMO DE CHERNICK
2 #LLAMADO UAR (1)
3
4 rdunif1<-function(k,n,pos=TRUE)
5 {
6     valoresposibles<-seq(0,(k-1)/k,1/k)
7     if (!pos) valoresposibles<-(-valoresposibles)
8     muestra<-sample(valoresposibles,n,replace=TRUE)
9     muestra
10 }
11 ##### valores de rho solo pueden ser iguales a 1/k siendo enteros
12 k<-c(1,2,3,4,5,6,7,8,9)
13 largo<-1000000
14
15 for (i in k)
16 {
17     x<-vector(length=largo)
18     rho<-1/i
19     for (j in 1:10)
20     {
21         x[1]<-rdunif1(i,1)
22         for (l in 1:(largo-1))
23         {
24             x[l+1]<-rho*x[l]+rdunif1(i,1)
25
26         }
27         write.csv(x,paste0("uni-",round(rho,2),"run",j,".csv"))
28     }
29 }
30
31 k<-c(1,2,3,4,5,6,7,8,9)
32
33 for (i in -k)
34 {
35     x<-vector(length=largo)
36     rho<-1/i
```

```
37     for (j in 1:10)
38     {
39         print(j)
40         x[1]<-rdunif1(-i,1, pos = FALSE)
41         for (l in 1:(largo-1))
42         {
43             x[l+1]<-rho*x[l]+rdunif1(-i,1, pos = FALSE)
44
45         }
46         write.csv(x, paste0("uni-N", round(-rho,2), "run", j, ".
47             csv"))
48     }
49
50 ### Ruido uniforme descorrelacionado
51
52 for (j in 1:10)
53 {
54     x<-runif(1000000)
55     write.csv(x, paste0("uni-",0, "run", j, ".csv"))
56 }
```

```
1 ##CODIGO PARA GENERAR REPLICAS BOOTSTRAP
2
3 library(gtools)
4 bootBP<-function(x,m=3,d=1,repet=100)
5 {
6   ### Funciones auxiliares
7   simula<-function(Mtrans,pi,n)
8   #simula los estados de una serie de tiempo dados las #probabilidades
9     de B&P
10  {
11    x<-vector()
12    states<-colnames(Mtrans)
13    x[1]<-sample(states,1,prob=pi)
14    for(i in 1:(n-1))
15    {
16      x[i+1]<-sample(states,size=1,prob=Mtrans[,x[i]])
17    }
18  }
19
20  entropia<-function(proba)
21  {
22    N<-length(proba)
23    c<-proba[proba!=0]
24    (-1)*sum(c*log(c))/(log(N))
25  }
26
27  fisher.info<-function(pi)
28  {
29    if(sum(pi)==0) {
30      return(NA)}
31    N<-length(pi)
32    if(pi[1]==1|pi[N]==1) {fo<-1}
33    else {fo<-0.5}
34    p0<-pi[1:N-1]
35    p1<-pi[2:N]
```

```

36 f<-( ( sqrt(p1)-sqrt(p0) )^2)
37 fisher.info<-fo*sum (f)
38 fisher.info
39 }
40
41 Complexity<-function(pi)
42 {
43 N<-length(pi)
44 s<-function(pi) {
45 N<-length(pi)
46 c<-pi[pi!=0]
47 (-1)*sum(c*log(c))/(log(N))
48 } ##calcula la entropia no normalizada
49
50 pe<-rep((1/N),N) ## la distribucion uniforme
51
52 q0<-(-2)/(((N+1)/N)*log(N+1))-2*log(2*N)+log(N) ## cte de
53 normalizacion
54
55 j<-s((pi+pe)/2)-(s(pi)+s(pe))/2 ## distancia de Jensen - Shannon
56
57 q<-q0*j ##desequilibrio
58
59 Complexity<-s(pi)*q
60 Complexity
61 }
62 Mapping<-function(x,m,d)
63 {
64 T=length(x)
65 n<-factorial(m)
66 if (m==3)
67 {
68 y0<-x[1:(T-2*d)]
69 y1<-x[(d+1):(T-d)]
70 y2<-x[(2*d+1):(T)]
71 s<-2*((y0>y1)+(y0>y2))+(y1>y2)

```

```

72
73 # h sirve para eliminar los elementos iguales y los datos faltantes
74
75 h<-is.na(y0) | is.na(y1) | is.na(y2) | (y0==y1) | (y0==y2) | (y1==y2)
76 s<-s+6*h
77 repeated<-sum((y0==y1) | (y0==y2) | (y1==y2))
78 brek<-10
79 }
80 if (m==4)
81 {
82   y0<-x[1:(T-3*d)]
83   y1<-x[(d+1):(T-2*d)]
84   y2<-x[(2*d+1):(T-d)]
85   y3<-x[(3*d+1):(T)]
86
87   s<-6*((y0>y1)+(y0>y2)+(y0>y3))+2*((y1>y2)+(y1>y3))+(y2>y3)
88   h<-is.na(y0) | is.na(y1) | is.na(y2) | is.na(y3) | (y0==y1) | (y0==y2) | (y1
      ==y2) | (y0==y3) | (y1==y3) | (y2==y3)
89   s<-s+24*h
90   repeated<-sum((y0==y1) | (y0==y2) | (y1==y2) | (y0==y3) | (y1==y3) | (y2==
      y3))
91   # h sirve para eliminar los elementos iguales y los datos
      faltantes
92   brek<-47
93 }
94
95 if (m==5)
96 {
97   y0<-x[1:(T-4*d)]
98   y1<-x[(d+1):(T-3*d)]
99   y2<-x[(2*d+1):(T-2*d)]
100  y3<-x[(3*d+1):(T-d)]
101  y4<-x[(4*d+1):(T)]
102
103
104  s<-24*((y0>y1)+(y0>y2)+(y0>y3)+(y0>y4))+
105    6*((y1>y2)+(y1>y3)+(y1>y4))+2*((y2>y3)+(y2>y4))+(y3>y4)

```

```

1106 h<-is.na(y0) | is.na(y1) | is.na(y2) | is.na(y3) | is.na(y4) |
1107     (y0==y1) |(y0==y2) |(y1==y2) |(y0==y3) |(y1==y3) |(y2==y3) |
1108     (y0==y4) |(y1==y4) |(y2==y4) |(y3==y4)
1109 s<-s+120*h
1110 dupli<-sum( (y0==y1) |(y0==y2) |(y1==y2) |(y0==y3) |(y1==y3) |(y2==y3) |
1111             (y0==y4) |(y1==y4) |(y2==y4) |(y3==y4) )
1112 }
1113
1114 if (m==6)
1115 {
1116   y0<-x[1:(T-5*d)]
1117   y1<-x[(d+1):(T-4*d)]
1118   y2<-x[(2*d+1):(T-3*d)]
1119   y3<-x[(3*d+1):(T-2*d)]
1120   y4<-x[(4*d+1):(T-d)]
1121   y5<-x[(5*d+1):T]
1122
1123   s<-120*((y0>y1)+(y0>y2)+(y0>y3)+(y0>y4)+(y0>y5))+24*((y1>y2)+(y1>y3
1124     )+(y1>y4)+(y1>y5))+
1125     6*((y2>y3)+(y2>y4)+(y2>y5))+2*((y3>y4)+(y3>y5))+(y4>y5)
1126   h<-is.na(y0) | is.na(y1) | is.na(y2) | is.na(y3) | is.na(y4) | is.na(y5) |
1127     (y0==y1) |(y0==y2) |(y0==y3) |(y0==y4) |(y0==y5) |
1128     (y1==y2) |(y1==y3) |(y1==y4) |(y1==y5) |
1129     (y2==y3) |(y2==y4) |(y2==y5) |
1130     (y3==y4) |(y3==y5) |
1131     (y4==y5)
1132
1133   s<-s+720*h
1134   dupli<-sum((y0==y1) |(y0==y2) |(y0==y3) |(y0==y4) |(y0==y5) |
1135             (y1==y2) |(y1==y3) |(y1==y4) |(y1==y5) |
1136             (y2==y3) |(y2==y4) |(y2==y5) |
1137             (y3==y4) |(y3==y5) |
1138             (y4==y5) )
1139 }
1140 s[s<n]
1141 }

```

```
142 n<-factorial(m)
143 simb<-c("0",LETTERS)
144 allstates<-apply(permutations(length(simb),2,simb, repeats.allowed =
    TRUE),1,paste0,collapse=" ")[1:n]
145 trans_states<-apply(permutations(length(allstates),2,allstates,
    repeats.allowed = TRUE),1,paste0,collapse=" ")
146
147 s<-Mapping(x,m=m,d=d)
148 states<-allstates[s+1]
149
150
151 ns=length(states)
152 while (sum(states==states[ns])==1) {
153   states<-states[-ns]
154   ns<-ns-1}
155
156 pi<-vector("numeric",length = n)
157 names(pi)<-allstates
158 pix<-as.numeric(table(states))/sum(as.numeric(table(states)))
159 names(pix)<-names(table(states))
160 pi[names(pix)]<-pix
161 hatentropy<-entropia(pi)
162 hatfisher<-fisher.info(pi)
163 hatnmp<-sum(pi==0)
164 hatcomplex<-Complexity(pi)
165
166 ss<-vector("character",length=ns-1)
167 ss<-paste0(states[1:(ns-1)],states[2:ns])
168 ss1<-as.numeric(table(ss))
169 names(ss1)<-names(table(ss))
170
171 Mtrans<-matrix(nrow=n,ncol=n)
172 names(Mtrans)<-trans_states
173 Mtrans[names(ss1)]<-ss1
174 Mtrans[is.na(Mtrans)]<-0
175 Mtrans<-matrix(Mtrans,nrow=n,ncol=n)
176 normalize<-function(x)
```

```

177 {
178 if (sum(x)==0) {y<-rep(0,length(x))} else
179 {
180   y<- (x/sum(x))
181 }
182 }
183 Mtrans<-as.matrix(apply(Mtrans,2,normalize))
184 colnames(Mtrans)<-allstates
185 rownames(Mtrans)<-allstates
186 entropy<-vector()
187 missing<-vector()
188 fisher<-vector()
189 complex<-vector()
190 cantmuestra<-length(x)
191 for (b in 1:repet)
192 {
193   s<-simula(Mtrans,pi,n=ns)
194   print(paste0("rep=",b,"m=",m,"n",cantmuestra))
195   pboot<-table(s)/sum(table(s))
196   pix<-vector("numeric",length = n)
197   names(pix)<-names(pi)
198   pix[names(pboot)]<-pboot
199   entropy[b]<-entropia(pix)
200   missing[b]<-sum(pix==0)
201   fisher[b]<-fisher.info(pix)
202   complex[b]<-Complexity(pix)
203 }
204 salida<-data.frame(entropy,missing,complex,fisher,rep(hatentropy,
    repet),rep(hatnmp,repet),rep(hatcomplex,repet),rep(hatfisher,repet)
    ))
205 names(salida)<-c("Entropy","NMP","Complexity","Fisher","hatE","hatC",
    "hatNMP","hatF")
206 salida
207 }

```