

Propiedades estadísticas de la correlación integral.



Departamento de Matemática
Universidad Técnica Federico Santa María

Autor

Eduardo Ignacio Rubio Marín

Profesor guía

Ronny Vallejos

Memoria presentada como requisito parcial para optar por el título de
Ingeniería Civil Matemática

Propiedades estadísticas de la correlación integral.

Abstract

This is a detailed study on correlation integral, a relevant tool in the field of statistics, which is interesting to study due to its theoretical importance and its possible applications in various areas. The main objective of this work is to provide a comprehensive review of the fundamentals and theoretical advancements on correlation integral, compiling and analyzing key results from the literature.

To achieve this, a bibliographic review of works on correlation integral was carried out, focusing on bringing together in one place, and with a unified notation, both the fundamental requirements for its study and its main properties. Subsequently, numerical simulations were implemented to study the addressed properties, using the R software.

The obtained results match the studies and represent them graphically. This not only helps provide a better theoretical understanding of correlation integral and its properties but also offers new perspectives for future research.

In conclusion, this work makes a significant contribution to the literature on correlation integral by clearly synthesizing existing results and providing experiments that complement theoretical predictions.

Propiedades estadísticas de la correlación integral.

Resumen

En esta memoria se realiza un estudio detallado sobre la correlación integral, una herramienta relevante en el campo de la estadística, que es interesante de estudiar debido a su importancia teórica y sus posibles aplicaciones en diversas áreas. El objetivo principal de este trabajo es proporcionar una revisión exhaustiva de los fundamentos y avances teóricos sobre la correlación integral, recopilando y analizando resultados clave de la literatura.

Para ello, se llevó a cabo una revisión bibliográfica de los trabajos sobre la correlación integral, centrándose en acunar en el mismo lugar y con una notación unificada tanto los requerimientos fundamentales para su estudio como sus propiedades principales. Posteriormente, se implementaron simulaciones numéricas para estudiar las propiedades abordadas, utilizando el software y lenguaje de programación R.

Los resultados obtenidos concuerdan con los resultados estudiados y los presentan de forma gráfica. Esto además de ayudar a proporcionar una mejor comprensión teórica de la correlación integral y sus propiedades, también ofrece nuevas perspectivas para futuras investigaciones.

En conclusión, este trabajo ofrece una contribución significativa a la literatura sobre la correlación integral, al sintetizar de manera clara los resultados existentes y proporcionar experimentos que complementan las predicciones teóricas.

Agradecimientos

Me tomo esta página para intentar expresar mis agradecimientos a quienes estuvieron conmigo durante esta larga etapa académica.

Primero a mi madre y a mi padre que siempre lo dieron todo para apoyarme en mis estudios y en mi vida en general, el que yo pudiera estudiar lo que quería y que haya llegado hasta aquí es gracias a ellos.

A mi hermano al que siempre he tenido como rol a seguir y como compañero de la vida, de la misma forma a mi hermana con su amor inmenso que siempre me da.

A mi abuelita Silvia que siempre tuvo un lugar para mí en su hogar, en mis muchos y constantes viajes entre Viña del Mar y Punta Arenas.

A las amistades que hice en este recorrido, Martín y Carlos como los primeros en mis primeros años de estudio, Ricardo que tanto me ayudó así como amigo como también de manera académica durante los muchos ramos que cursamos juntos. A mi grupo cercano que me acobijó durante la mayor parte de mis estudios, no los nombraré a todos pero saben bien quienes son, algunos que tengo presentes mientras escribo esto son Bastián, Bruno, Gustavo e Ike, que me estuvieron ayudando y aconsejando en esta última instancia. Y tantas más personas con quienes compartí durante este largo periodo de mi vida.

A mis amigos de Punta Arenas, en particular aquellos con los que tuve la suerte de coincidir por mucho tiempo al vivir en la misma ciudad, mis amigos más antiguos y cercanos como Benjamin, Ivo y Javier.

Finalmente a mi Profesor guía Ronny Vallejos, quien tuvo la mejor de las disposiciones posibles a la hora de apoyarme y guiarme durante todas las etapas de este trabajo.

Gracias.

Índice general

1. Introducción	7
1.1. Introducción	7
1.2. Aplicaciones de la Correlación Integral	9
2. Preliminares y Notación	11
2.1. Preliminares Series de Tiempo	11
2.2. Notación y definiciones Series de Tiempo	11
2.2.1. Procesos Estocásticos	11
2.2.2. Estacionariedad	12
2.2.3. Procesos Gaussianos	13
2.3. Correlación Integral	15
2.3.1. Correlación Integral en el Análisis de un proceso no lineal . . .	19
3. Propiedades Estadísticas	22
3.1. Contexto	22
3.2. Estadístico BDS	22
4. Experimentos	27
4.1. Simulaciones de Datos	27
5. Conclusiones	32
6. Apéndice	34
6.1. Demostraciones	34
6.2. Gráficos	36

Índice de figuras

4.1. Gráfico Q-Q de datos IID con $d = 2$	29
4.2. Gráfico Q-Q de datos IID con $d = 3$	29
4.3. Gráfico Q-Q de datos Autorregresivos con $d = 2$	30
4.4. Gráfico Q-Q de datos Autorregresivos con $d = 3$	31
6.1. Gráfico Q-Q de datos Autorregresivos con $d = 1$	36
6.2. Gráfico Q-Q de datos Autorregresivos con $d = 4$	36
6.3. Gráfico Q-Q de datos Autorregresivos con $d = 5$	36
6.4. Gráfico Q-Q de datos IID con $d = 1$	37
6.5. Gráfico Q-Q de datos IID con $d = 4$	37
6.6. Gráfico Q-Q de datos IID con $d = 5$	37
6.7. Gráficos box-plot IID valores de d	38
6.8. Gráficos box-plot Autorregresivos valores de d	38

Capítulo 1

Introducción

1.1. Introducción

La correlación integral es una herramienta estadística muy versátil que se puede aplicar a diversos campos. Además de ser una herramienta crucial en el análisis de datos, donde juega un papel fundamental en diversas áreas de estudio como en el análisis del caos y la teoría de sistemas dinámicos, también se puede emplear a otros campos mas aplicados como la física, la biología, la economía y las ciencias sociales.

Su principal uso radica en la capacidad de cuantificar la similitud entre diferentes conjuntos de datos, lo que es especialmente útil en el análisis de series de tiempo. Esta técnica se centra en la evaluación de la estructura de los datos, permitiendo identificar patrones que no son evidentes a simple vista, especialmente en datos que evolucionan en el tiempo. Se puede definir para series de tiempo como un coeficiente normalizado que representa el número de pares de puntos cercanos de esta dentro de su espacio de fase ([Acosta et al., 2024](#)).

En el campo de la física, la correlación integral se utiliza para estudiar la dinámica de sistemas complejos, como los sistemas climáticos o los fenómenos astrofísicos. Permite a los científicos comprender mejor cómo evolucionan estos sistemas con el tiempo y predecir su comportamiento futuro.

En biología, esta herramienta se aplica en el estudio de sistemas ecológicos y biológicos, ayudando a entender las interacciones complejas entre diferentes especies y su entorno. Por ejemplo, puede ser usada para analizar patrones de migración de animales o la propagación de enfermedades en poblaciones.

En el ámbito económico y social, la correlación integral se emplea para analizar series de tiempo relacionadas con datos financieros, económicos o demográficos. Esto permite a los analistas y tomadores de decisiones entender mejor las tendencias del mercado, los ciclos económicos y los patrones de comportamiento social.

Uno de los aspectos más importantes de la correlación integral es su capacidad para manejar datos de alta dimensionalidad. En muchas aplicaciones modernas, como el aprendizaje automático y la minería de datos, se trabaja con conjuntos de datos grandes y complejos. La correlación integral es particularmente útil en estos escenarios, ya que puede revelar estructuras ocultas y correlaciones en los datos que otros métodos podrían pasar por alto.

Usar la correlación integral puede ser una técnica versátil y poderosa que proporciona perspectivas valiosas en una amplia gama de campos científicos y aplicaciones prácticas. Su capacidad para analizar la similitud y la estructura en diversos datos la hace indispensable en el mundo moderno, donde sigue creciendo la cantidad y complejidad de la información disponible. Esto es posible debido a que la correlación integral de un conjunto de puntos en un espacio euclidiano es definida como la probabilidad de que dos puntos seleccionados aleatoriamente tengan una distancia menor o igual a una distancia r . Definiendo correctamente esta distancia para el espacio adecuado de dimensión n en el que se esté trabajando. Dado que estas definiciones se pueden aplicar de manera generalizada y flexible dentro de cada contexto sin mayor dificultad es que la aplicación de la correlación integral en diversos formatos se hace llamativa.

Un valor alto de la correlación integral sugiere que los datos son muy similares entre sí, mientras que un valor bajo sugiere que los datos son muy diferentes entre sí. Así, la correlación integral proporciona una medida cuantitativa de la similitud entre los datos. Esto es aplicable tanto dentro de un mismo grupo de datos a través del tiempo como diferentes datos que podrían estar relacionados o de los cuales se espera lo estén o se parezcan de alguna manera.

La correlación integral está estrechamente relacionada con el concepto de dimensión fractal, una medida de la complejidad de un objeto. La correlación integral es uno de los métodos más utilizados para estimar la dimensión fractal de los atractores caóticos a partir de series de tiempo, con un gran rendimiento en datos experimentales. En particular, la pendiente de la línea que mejor se ajusta a la correlación integral en un gráfico log-log proporciona una estimación de la dimensión fractal.

1.2. Aplicaciones de la Correlación Integral

La correlación integral es una técnica ampliamente aplicable, extendiéndose a campos tan diversos como la física, la ingeniería y las ciencias sociales. Un área destacada de su uso es el análisis de series de tiempo, donde se emplea para describir cómo las propiedades de las series evolucionan a lo largo del tiempo ([Acosta et al., 2024](#)). Esto es crucial para entender fenómenos que varían temporalmente, proporcionando una visión más profunda de las tendencias y patrones subyacentes.

Además, la correlación integral se ha mostrado eficaz en la identificación de patrones en datos de alta dimensionalidad y en el estudio de la dependencia espacial en datos geográficos. Esto amplía su aplicabilidad, facilitando el análisis en campos que manejan grandes volúmenes de información compleja.

En la práctica, su uso se extiende al análisis tanto de datos temporales como espaciales en múltiples disciplinas. Algunos ejemplos notables incluyen:

Análisis de Imágenes: Utilizando la correlación integral para medir texturas o rugosidades en imágenes, se obtiene información valiosa sobre su estructura espacial. Esto es particularmente útil en la clasificación de imágenes o en la detección de características específicas. [Sarkar and Chaudhuri \(1994\)](#) destacan en este campo.

Estudio de Sistemas Dinámicos: Aquí, la correlación integral ayuda a analizar la estructura del espacio de fases de un sistema, proporcionando perspectivas sobre la dinámica subyacente y la identificación de comportamientos como el caos o complejidades. [Grassberger and Procaccia \(1983\)](#) son referentes en este ámbito.

Estudios Biomédicos: En este sector, se utiliza para analizar la estructura de señales biomédicas, como las series temporales de electroencefalogramas (EEG). Esta aplicación es crucial para comprender el estado de salud de los pacientes y resulta instrumental en el diagnóstico de condiciones médicas. [Mormann et al. \(2005\)](#) han contribuido significativamente a esta área.

Estudio de Sistemas Climáticos: Investigadores como [Di et al. \(2018\)](#) han aplicado métodos computacionales y algoritmos para identificar y aproximarse a la dimensión de correlación asociada a modelos de cambio climático en China, demostrando la versatilidad de esta técnica en el estudio de fenómenos ambientales. Este último caso es remarcable por el uso de la correlación integral de series de tiempo correspondientes a datos climáticos dentro de un modelo computacional para poder medir el parecido entre datos de precipitaciones y datos la temperatura del aire dentro de una región específica. El uso de la correlación integral en este caso se da para evaluar si los datos son estocásticos o no, usando gráficos log-log con respecto a esta y a la misma distancia r con un exponente al que llaman exponente de correlación, el cual tiene sus raíces en una de las propiedades fundamentales de la correlación integral y sus tendencias, la cual será abordada en su respectivo capítulo.

En este caso esta propiedad se aplicó a un ya existente criterio computacional para evaluar la relación entre lluvia y temperatura, el cual fue fortalecido gracias a la inclusión de los cálculos aplicados de la correlación integral. Es así como en este y otros ámbitos la correlación integral puede aportar o fortalecer diversos enfoques dentro de un mismo campo o investigación. Ésta no solo es una herramienta valiosa para los científicos y analistas de datos, sino que también se ha convertido en un componente integral en una variedad de campos de investigación, ayudando a descifrar y entender la complejidad inherente a muchos sistemas y fenómenos.

Capítulo 2

Preliminares y Notación

2.1. Preliminares Series de Tiempo

Una serie de tiempo es una secuencia de observaciones recogidas o registradas en puntos sucesivos en el tiempo. Es una forma de analizar el conjunto de datos en un intervalo de tiempo usando técnicas estadísticas para poder estudiar y modelar el comportamiento de un fenómeno evolucionando a través del tiempo. Si se está observando una única variable, se suele tener como objetivo el generar un modelo descriptivo y predictivo de la variable. Si el caso incluye un grupo de variables se suele estudiar las posibles relaciones entre estas y su evolución en conjunto.

Las series de tiempo se utilizan en una variedad de campos, como economía, meteorología, ciencias de la salud y muchas otras disciplinas donde los cambios en los datos a lo largo del tiempo son de interés.

Cada observación en la serie de tiempo está asociada con un sello de tiempo único. Las series de tiempo pueden tener diferentes frecuencias de observación, por ejemplo, diaria, mensual, trimestral o anual. Algunos aspectos importantes de las series de tiempo incluyen la tendencia (la dirección general de los datos), la estacionariedad (patrones repetitivos dentro de períodos fijos de tiempo), y la autocorrelación (la correlación de la serie con versiones rezagadas de sí misma).

2.2. Notación y definiciones Series de Tiempo

2.2.1. Procesos Estocásticos

Sea $(\Omega, \mathcal{F}, P)$ un espacio de Probabilidad, sea T un conjunto arbitrario de índices y $(E, \mathcal{B})$ un espacio medible.

Se define un proceso estocástico ([Capasso and Bakstein, 2021](#)) como una familia $\{X_t, t \in T\}$ de variables aleatorias $X_t : (\Omega, \mathcal{F}) \rightarrow (E, \mathcal{B})$ ordenadas con respecto al conjunto de índices T .

Se llaman espacio de probabilidad subyacente a $(\Omega, \mathcal{F}, P)$ y espacio de fase a $(E, \mathcal{B})$ del proceso $\{X_t, t \in T\}$, con X_t la variable aleatoria referente al estado del proceso en el tiempo $t \in T$.

Se tiene que para cada $w \in \Omega$, la función trayectoria del proceso w se escribe como:

$$X(\cdot, w) : t \in T \rightarrow X_t(w) \in E.$$

2.2.2. Estacionariedad

Se define un proceso estocástico estacionario como aquel en que su distribución permanece igual con respecto al avance del tiempo, por lo que se requiere que ciertas propiedades de este sean invariantes en el tiempo.

Se suele referir a dos tipos comunes de estacionariedad ([Percival and Walden, 2020](#)): Estacionariedad Completa y Estacionariedad de segundo orden.

La estacionariedad completa, o dicho también como proceso estrictamente estacionario, consiste en que la estructura probabilística del proceso sea invariante bajo un desplazamiento en el tiempo. Es decir, para una colección de instantes $(t_1, \dots, t_n) \in T$, la función de distribución de probabilidad conjunta de $(X_{t_1}, \dots, X_{t_n})$, $\forall n \in \mathbb{N}$ será la misma para una colección donde se desplaza cada elemento de la colección la misma cantidad de veces.

Dado que es difícil trabajar procesos estrictamente estacionarios con modelos de series de tiempo de interés, se suele usar la opción simplificada de estos, la estacionariedad de segundo orden.

Se dice que el proceso $\{X_t, t \in T\}$ es estacionario de segundo orden o débilmente estacionario si para cada $n \in \mathbb{Z}$ y cualquier conjunto $(t_1, \dots, t_n) \in T$, existen todos los momentos hasta de orden 2 de $(X_{t_1}, \dots, X_{t_n})$, son finitos y son iguales a los momentos $(X_{t_1+h}, \dots, X_{t_n+h})$ para algún $h \in T$.

Notando $E[X_{t_j}, X_{t_k}]$ como momento de segundo orden para $t_j, t_k \in T$ y de manera más general $E[X_{t_j}^l, X_{t_k}^m]$ como momento de orden $l + m$.

De lo anterior se pueden desprender las funciones de media $m : T \rightarrow \mathbb{C}$ y covarianza $\gamma : T \times T \rightarrow \mathbb{C}$ respectivamente tal que:

$$\begin{aligned} m(t) &:= E[X(t)] , \\ \gamma(t, s) &:= Cov(X_t, X_s) = E[X_t, X_s] - m(t)m(s). \end{aligned}$$

Se dice un proceso estocástico centrado si se cumple $m(t) = 0$. Para el caso particular $t = s$ se tiene la función varianza $\sigma^2(t) = Cov(X_t, X_t) = Var[X_t]$.

Así entonces para un proceso estocástico estacionario de segundo orden se cumple para $m \in \mathbb{C}$ y una función $C : T \rightarrow \mathbb{C}$ que: $m(t) = m$ y $\gamma(t, s) = C(t - s)$. A C se le denomina función de covarianza del proceso.

El tamaño de la covarianza dependerá directamente de las unidades en las que se mide X_t , por lo que para poder interpretar mejor estos valores se suele normalizar las covarianzas, produciendo así la función de correlación, dada por : $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$. Midiendo así la correlación entre X_t y X_{t+h} .

2.2.3. Procesos Gaussianos

Un proceso gaussiano es un tipo de proceso estocástico donde cada conjunto de puntos en un rango dado tiene una distribución normal multivariada. Esto implica que para cualquier conjunto de puntos temporales, la colección de valores que el proceso toma en esos puntos sigue una distribución normal.

Se dice que un proceso estocástico X_t es un proceso gaussiano si para cualquier elección de tiempos $\{t_1, \dots, t_n\}$, el vector de variables aleatorias $\{X_{t_1}, \dots, X_{t_n}\}$ tiene una distribución normal multivariada. Como esta distribución está caracterizada por un vector de medias μ y una matriz de covarianzas Σ , se pueden establecer condiciones sobre estas para encontrar distintos ejemplos de procesos estacionarios.

Un proceso gaussiano es estrictamente estacionario si es un proceso débilmente estacionario ([Percival and Walden, 2020](#)).

Un proceso $\{X_t, t \in T\}$ se dice **ruido blanco** o proceso puramente aleatorio si la secuencia de variables aleatorias $\{X_t\}$ son independientes e idénticamente distribuidas. De lo anterior se tiene que las funciones de media y varianza serán constantes y la covarianza nula para un ruido blanco ([Chatfield, 1996](#)).

El ruido blanco al ser un proceso puramente aleatorio suele usarse como base de construcciones de procesos mas complicados.

Modelos autorregresivos de media móvil (ARMA)

Los modelos autorregresivos de media móvil son una clase de modelos ampliamente utilizada para analizar las series de tiempo. Los modelos ARMA son la combinación de dos modelos más simples: los modelos autorregresivos (AR) y los modelos de media móvil (MA).

En los **modelos Autorregresivos AR**, el valor actual de la serie de tiempo se supone depende linealmente de sus propios valores previos. Un modelo AR de orden p , denotado como $AR(p)$, utiliza los p valores previos de la serie de tiempo.

Sea $\{Z_t, t \in \mathbb{Z}\}$ un ruido blanco con media cero y varianza σ_z^2 . Se define a $\{X_t, t \in \mathbb{Z}\}$ como un modelo $AR(p)$ o modelo autorregresivo de orden p si :

$$X_t = \alpha_1 X_{t-1} + \dots + \alpha_p X_{t-p} + Z_t, \quad (2.1)$$

para α_i constantes ([Chatfield, 1996](#)).

En los **modelos de Media Móvil MA**, el valor actual de la serie de tiempo se supone dependiente linealmente de los términos de error previos. Un modelo MA de orden q , denotado como $MA(q)$, utiliza los q errores anteriores en la serie de tiempo.

Sea $\{Z_t, t \in \mathbb{Z}\}$ un ruido blanco con media cero y varianza σ_z^2 . Se define a $\{X_t, t \in \mathbb{Z}\}$ como un modelo $MA(q)$ o modelo de media móvil de orden q si :

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \dots + \beta_q Z_{t-q}, \quad (2.2)$$

para β_i constantes (Chatfield, 1996). Si el ruido blanco asociado $\{Z_t, t \in \mathbb{Z}\}$ distribuye normal, entonces $\{X_t, t \in \mathbb{Z}\}$ será un proceso normal fuertemente estacionario.

Un **modelo ARMA(p, q)** combina estos dos conceptos. Se asume que el valor actual de la serie de tiempo depende tanto de los valores anteriores de la serie como de los errores anteriores.

Sea $\{Z_t, t \in \mathbb{Z}\}$ un ruido blanco con media cero y varianza σ_z^2 constante. Se define a $\{X_t, t \in \mathbb{Z}\}$ como un modelo $ARMA(p, q)$ o modelo autorregresivo de media móvil si :

$$X_t = \alpha_1 X_{t-1} + \dots + \alpha_p X_{t-p} + \beta_0 Z_t + \beta_1 Z_{t-1} + \dots + \beta_q Z_{t-q}, \quad (2.3)$$

para α_i y β_i constantes. (Chatfield, 1996). Un modelo $ARMA(p, q)$ tiene la ventaja de que al usarse para describir una serie de tiempo suele requerir menos parámetros que al usar modelos $AR(p)$ o $MA(q)$ puros.

Estimación de los Modelos ARMA

La estimación de los parámetros en un modelo ARMA es un paso crítico para modelar series temporales. Generalmente, este proceso se realiza mediante el método de máxima verosimilitud. Este método busca encontrar los valores de los parámetros del modelo que hacen que los datos observados sean lo más probables posible, dados estos parámetros.

La maximización de la función de verosimilitud implica utilizar técnicas de optimización numérica, ya que la solución analítica rara vez es factible debido a la complejidad del modelo. Durante este proceso, se calcula la probabilidad de los datos observados bajo diferentes configuraciones de parámetros y se seleccionan aquellos que maximizan esta probabilidad.

Además de la máxima verosimilitud, existen otros métodos para estimar los parámetros de un modelo ARMA, como el método de los momentos y el método de mínimos cuadrados. Sin embargo, la máxima verosimilitud es preferida debido a tener propiedades estadísticas deseables, como lo es la consistencia y la eficiencia asintótica, especialmente para muestras de grandes tamaños.

Es importante señalar que la correcta especificación del orden de los componentes es crucial antes de la estimación de los parámetros. Esto generalmente se realiza mediante criterios de información, que equilibran la complejidad del modelo con el ajuste a los datos.

Una vez estimados los parámetros, el modelo ARMA puede ser utilizado para hacer pronósticos de la serie temporal, analizar su estructura y entender las relaciones dinámicas que gobiernan los datos observados.

2.3. Correlación Integral

Sea $\{X_t, t \in \mathbb{Z}\}$ un proceso estocástico y $\{X_1, \dots, X_N\}$ una serie de tiempo de longitud N y sean los vectores $V_i = (X_i, \dots, X_{i+d-1})^T$, con $i \in \{1, \dots, N-d+1\}$ y $d \in \mathbb{N}$ dimensión del espacio fase (Wolff, 1990).

Para un $r > 0$ fijo, se define la Correlación Integral del proceso como:

$$C_N(r, d) = \frac{2}{N^2} \sum_{i=1}^{N-d} \sum_{j=i+1}^{N-d+1} \delta_{i,j}, \quad (2.4)$$

donde

$$\delta_{i,j} = \begin{cases} 1 & \text{si } \|V_i - V_j\| < r \\ 0 & \text{en otro caso} \end{cases}, \quad (2.5)$$

con $\frac{2}{N^2}$ una constante de normalización.

Aquí entonces, la Correlación Integral representa un coeficiente para el número de pares de vectores diferentes en una serie de Tiempo, con respecto a la distancia r .

La Correlación Integral en Series de Tiempo

La correlación integral es una técnica útil en el análisis de series de tiempo, particularmente para series de tiempo no lineales o caóticas. Permite la cuantificación de la similitud entre diferentes series de tiempo y puede usarse para identificar patrones y estructuras en los datos.

La correlación integral es una técnica estadística usada para cuantificar la similitud entre diferentes series de tiempo. Esta técnica es especialmente útil para analizar series de tiempo no lineales o caóticas.

En el contexto de las series de tiempo, la correlación integral se utiliza para evaluar la autocorrelación en los datos. La autocorrelación es el grado en que los valores en una serie de tiempo están correlacionados con sus propios valores pasados. Para las series de tiempo, la correlación integral a menudo se utiliza en el análisis de la dependencia temporal.

La correlación integral se basa en la idea de contar el número de pares de puntos en la serie de tiempo que están dentro de una cierta distancia r el uno del otro. Si la serie de tiempo es densa y los puntos están cerca el uno del otro, entonces la correlación integral será alta. Por otro lado, si la serie de tiempo es dispersa y los puntos están lejos el uno del otro, entonces la correlación integral será baja.

En Particular, la Correlación Integral cumple con

$$\lim_{N \rightarrow \infty} C_N(r, d) \propto r^d. \quad (2.6)$$

La correlación integral se usa en el análisis de series de tiempo para estudiar la estructura de los datos. Por ejemplo, se puede usar para identificar si los datos presentan algún tipo de patrón o estructura, o si parecen ser aleatorios.

Una aplicación importante de la correlación integral en series de tiempo es en el análisis de la dimensión fractal. En el caso de las series de tiempo, se puede usar la correlación integral para estimar la dimensión fractal de los datos.

Esto se hace trazando la correlación integral en función del radio r en un gráfico log-log. Si los datos siguen un patrón fractal, entonces este trazado debe ser aproximadamente una línea recta, y la pendiente de esta línea es una estimación de la dimensión fractal de los datos.

Correlación Integral en ruido blanco

Sea $\{\epsilon_t, t \in \mathbb{N}\}$ un ruido blanco gaussiano con varianza σ^2 . Para la serie de tiempo $\{\epsilon_1, \dots, \epsilon_N\}$ se asume que es observación del proceso aleatorio, tal que $\epsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0; \sigma^2)$ para $i \in \{1, \dots, N\}$.

Tomando $j > i$, se tiene $\sqrt{2\sigma^2}z_{i,j} = \epsilon_i - \epsilon_j \sim \mathcal{N}(0; 1)$, por lo que para la esperanza de $C_N(r, d)$:

$$\begin{aligned}
E[C_N(r, d)] &= \frac{2}{N^2} \sum_{j=1}^{N-d} (N-d+1-j) E[\delta_{1,1+j}] \\
&= \frac{2}{N^2} \sum_{j=1}^{N-d} (N-d+1-j) P(\|z\| < \frac{r}{\sqrt{2\sigma^2}}) \\
&= \frac{2}{N^2} \sum_{j=1}^{N-d} (N-d+1-j) P(z \in B(0, \frac{r}{\sqrt{2\sigma^2}})) \\
&= \frac{2}{N^2} \int_{B(0, \frac{r}{\sqrt{2\sigma^2}})} f_Z(z) dz \sum_{j=1}^{N-d} (N-d+1-j) \\
&= \left(1 + \frac{1-2d}{N} + \frac{d(d-1)}{N^2}\right) c_d f_Z(\xi) \frac{r^d}{(2\sigma^2)^{\frac{d}{2}}}, \quad (2.7)
\end{aligned}$$

con $c_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)}$ volumen de la esfera unitaria y $f_Z(\xi)$ la función de densidad de probabilidad gaussiana. Se tiene por el Teorema del valor medio para integrales múltiples que existe ξ . Aplicando limite $N \rightarrow \infty$ con la ecuación anterior se tiene :

$$\lim_{N \rightarrow \infty} E[C_N(r, d)] \propto r^d.$$

Correlación Integral en modelos ARMA

Sea $\{X_t, t \in \mathbb{Z}\}$ un proceso estacionario $ARMA(1, 1)$ y $\{X_1, \dots, X_N\}$ una serie de tiempo de largo $N \in \mathbb{N}$ que representa el tramo observado del proceso subyacente, definido como:

$$X_i = \alpha X_{i-1} + \epsilon_i + \beta \epsilon_{i-1}; \quad i \in \{1, \dots, N\} \quad (2.8)$$

con $\{\epsilon_i\}$ ruido blanco gaussiano de varianza σ^2 y $|\alpha| < 1$ constante. Para este proceso la función de correlación es:

$$\rho(h) = \alpha^{h-1} \rho(1), \quad \rho(1) = \frac{(1 + \alpha\beta)(\alpha + \beta)}{1 + 2\alpha\beta + \beta^2}.$$

Tomando $j > i$, se tiene $(X_i - X_j) \sim \mathcal{N}(0; 2\gamma(0) - 2\gamma(j-i))$, por lo que para $Z \sim \mathcal{N}(0; 1)$, $d = 1$, $\delta_{i,j}$ y $r_{j-i} = \frac{1}{(2\gamma(0) - 2\gamma(j-i))^{\frac{1}{2}}}$ se tiene para la esperanza de $C_N(r, 1)$:

$$\begin{aligned}
E[C_N(r, 1)] &= \frac{2}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \delta_{1,1+j} \\
&= \frac{2}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N P(|Z| < r_{j-i}) + \frac{2}{N^2} (N-1) P(|Z| < r_1) \\
&= \frac{2}{N^2} \sum_{j=1}^{N-2} (N-1-j) P(|Z| < r_{j+1}) + \left(\frac{2}{N} - \frac{2}{N^2}\right) P(|Z| < r_1) \\
&= \frac{4}{N^2} \sum_{j=1}^{N-2} (N-1-j) \phi(\xi_{j+1}) r_{j+1} + \left(\frac{2}{N} - \frac{2}{N^2}\right) P(|Z| < r_1).
\end{aligned}$$

(2.9)

Obteniendo la última igualdad por el Teorema del valor medio para integrales, donde $\{\xi_j\} \in (0, r_j)$ es una secuencia positiva.

Sea $r' = \frac{r}{(2\gamma(0))^{\frac{1}{2}}}$, entonces:

$$\begin{aligned}
E[C_N(r, 1)] &\approx \frac{4r'}{N^2} \sum_{j=1}^{N-2} (N-1-j) \phi(\xi_{j+1}) + \frac{1}{N} \left(4r' \sum_{j=1}^{\infty} \phi(\xi_j) \left(\frac{1}{(1 - \rho(1)\psi^j)^{\frac{1}{2}}} - 1 \right) \right. \\
&+ \left. \frac{2}{N} P(|Z| < r_1) - \frac{1}{N^2} \left(2r' \sum_{j=1}^{\infty} \phi(\xi_j) \left(\frac{1}{(1 - \rho(1)\psi^j)^{\frac{1}{2}}} - 1 \right) + 2P(|Z| < r_1) \right) \right).
\end{aligned}$$

(2.10)

Teniendo $|\alpha| < 1$ se cumple que ambas series con respecto a $\phi(\xi_j)$ en la suma anterior convergen y se puede probar también convergencia para el inicio de la suma a $2r'\phi(\xi)$ ([Furdui, 2013](#)). Luego como $r' \propto r$ se tiene:

$$\lim_{N \rightarrow \infty} E[C_N(r, 1)] \propto r. \quad (2.11)$$

Lo anterior visto para $ARMA(1, 1)$ puede ser generalizado a través del siguiente Teorema:

Teorema 2.1

Sea $\{X_i, i \in \mathbb{Z}\}$ un proceso estacionario gaussiano con función de correlación $\rho(h)$, $h \in \mathbb{Z}$. Para una serie de tiempo $\{X_1, \dots, X_N\}$ se satisface que :

$$\lim_{N \rightarrow \infty} E[C_N(r, d)] \propto r^d. \quad (2.12)$$

2.3.1. Correlación Integral en el Análisis de un proceso no lineal

Existen dos categorías de series de tiempo no lineales: Modelos lineales dinámicos multiproceso o MDLM y los modelos autorregresivos no lineales o NLAR, ambos capaces de generar varias características no lineales para series de tiempo (Glasbey, 2001).

Se define $\{X_i, i \in \mathbb{Z}\}$ como un modelo autorregresivo no lineal de primer orden o NLAR(1) tal que, para una función específica f y $\{\epsilon_i\}$ errores independientes, entonces (Tong, 1990):

$$X_{i+1} = f(X_i, \epsilon_i). \quad (2.13)$$

Jones (1978) consideró :

$$X_{i+1} = \lambda(X_i) + \epsilon_{i+1}, \quad (2.14)$$

con $\epsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0; 1)$, independiente de X_j para $j < i$ y $\lambda(\cdot)$ función fija real.

El modelo anterior pertenece a la familia de modelos $\{X_{i+1}(\beta)\}$ generada por:

$$X_{i+1}(\beta) = a + bX_i(\beta)\beta(\lambda X_i(\beta) - bX_i(\beta) - a) + \epsilon_{i+1}, \quad (2.15)$$

con a y b constantes fijas. De esta manera se puede descomponer $X_{i+1}(\beta)$ es una parte lineal y una parte residual.

Entonces se puede suponer que $X_i(\beta)$ puede ser representada por el proceso estacionario:

$$X_i(\beta) = X_i^{(0)} + \beta X_i^{(1)} + \frac{1}{2!} \beta^2 X_i^{(2)} + \dots \quad (2.16)$$

siendo cada $\{X_i^{(r)}; i \in \{1, \dots, N\}; r \in \mathbb{N}\}$ series estacionarias. Si para esta expansión se utiliza lo escrito en (2.15) se pueden escribir los componentes de los procesos $\{X^{(r)}\}$ de manera que se obtienen:

$$X_{i+1}^{(0)} = a + bX_i^{(0)} + \epsilon_{i+1},$$

$$X_{i+1}^{(1)} = bX_i^{(1)} + \Delta (X_i^{(0)}),$$

$$X_{i+1}^{(2)} = bX_i^{(2)} + 2 \Delta^{(1)} (X_i^{(0)})X_i^{(1)},$$

con la notación correspondiente a los diferenciales siendo:

$$\Delta (x) = \lambda(x) - bx - a,$$

$$\Delta^{(k)} (x) = \frac{d^k}{dx^k} \Delta (x),$$

Con λ constante. Para el caso particular donde $a = 0$ y $b = 0$ se tiene la función de densidad del proceso estacionario $\{X_{i+1}(\beta)\}$ generada por:

$$X_{i+1}(\beta) = \beta\lambda X_i(\beta) + \epsilon_{i+1}. \quad (2.17)$$

Por lo que se puede escribir $X_{i+1}(\beta)$ tal que:

$$X_{i+1}(\beta) = b\lambda X_i(\beta) + \epsilon_{i+1} = b\lambda(X_i^{(0)} + \beta X_i^{(1)} + \frac{1}{2!}\beta^2 X_i^{(2)}) + \epsilon_{i+1}$$

$$= b\lambda(\epsilon_i + \beta \Delta (X_{i-1}^{(0)}) + \frac{1}{2}\beta^2 2 \Delta^{(1)} (X_{i-1}^{(0)})X_{i-1}^{(1)}) + \epsilon_{i+1}$$

$$= b\lambda(\epsilon_i + \beta\lambda\epsilon_{i-1} + \beta^2 \frac{d}{dx}(\lambda\epsilon_{i-1})) + \epsilon_{i+1}$$

$$= b\lambda(\epsilon_i + \beta\lambda\epsilon_{i-1} + \beta^2\lambda^2\epsilon_{i-1}) + \epsilon_{i+1}.$$

Así truncando la expansión original de (2.16) se pueden calcular las aproximaciones de los momentos estacionarios $\{X_i\beta\}$, por ejemplo las esperanzas, varianzas y covarianzas tal que, para $\{\tilde{X}_i\beta\}$ truncado se tiene:

$$E[\tilde{X}_i(\beta)] = 0$$

$$V(\tilde{X}_i) = 1 + \lambda^2 \beta^2 + \lambda^4 \beta^4,$$

$$Cov(\tilde{X}_{i+1}; \tilde{X}_i) = \lambda \beta (1 + \lambda^2 \beta^2),$$

$$Cov(\tilde{X}_{i+2}; \tilde{X}_i) = \lambda^2 \beta^2,$$

$$Cov(\tilde{X}_{i+h}; \tilde{X}_i) = 0; \quad h \geq 3.$$

Entonces para este proceso se tendrá que la correlación integral cumple con:

$$\lim_{N \rightarrow \infty} E[C_N(r, d)] \propto r^d; \quad d = 1, 2. \quad (2.18)$$

Capítulo 3

Propiedades Estadísticas

3.1. Contexto

En este capítulo se presentan algunas de las principales propiedades estadísticas de la correlación integral. A través del siguiente desarrollo, se establecen bases necesarias para distintos Teoremas que se abarcaran en este capítulo. Aplicando esta vez la correlación integral a casos más concretos con hipótesis estrictas o bien definidas las cuales apuntan a poder aplicarse en contextos de datos reales. Los tres Teoremas presentados en este capítulo se emplean para construir estadísticos que permiten realizar pruebas sobre datos y evaluar su distribución y origen correspondiente.

3.2. Estadístico BDS

Sea $\{u_t, t \in \mathbb{Z}\}$ un proceso estocástico estrictamente estacionario de distribución conocida F . Sean los vectores $u_t^d = \{u_t, u_{t+1}, \dots, u_{t+d-1}\}$ con $d \in \mathbb{N}$ dimensión del espacio fase y con distribuciones F_d .

Para un $r > 0$ fijo, se define la Correlación Integral del proceso como:

$$C_N(r, d) = \frac{2}{N^2} \sum_{i=1}^{N-d} \sum_{j=i+1}^{N-d+1} \delta_{i,j}, \quad (3.1)$$

donde

$$\delta_{i,j} = \begin{cases} 1 & \text{si } \|u_i^d - u_j^d\| < r \\ 0 & \text{en otro caso} \end{cases}, \quad (3.2)$$

Denotando

$$C(r, d) = \lim_{N \rightarrow \infty} C_N(r, d), \quad (3.3)$$

Con esta correlación integral se puede diferenciar datos idénticamente distribuidos obtenidos de manera aleatoria a datos provenientes de series de tiempo caóticas (Lai, 1999).

Para datos compuestos de observaciones independientes se puede escribir:

$$C(r, d) = \int \int \delta_{u,v}^* dF_d(u) dF_d(v) = C_l(r, d), \quad (3.4)$$

con

$$\delta_{u,v}^* = \begin{cases} 1 & \text{si } \|u - v\| < r \\ 0 & \text{en otro caso} \end{cases}, \quad (3.5)$$

Se puede medir la desviación de lo observado con respecto a la hipótesis nula de independencia usando $C_N(r, d) - C_l(r, d)$

Sea $K := \int (F(u+r) - F(u-r))^2 dF_d(u)$

y por simplicidad notemos $C = C_l(r, \cdot)$ y $C^d = C_l(r, d)$.

Teorema 3.1

Si $\{u_t, t \in \mathbb{Z}\}$ son independientes idénticamente distribuidos y se tiene $K > C^2$, entonces:

$$\mathcal{L} \left(\frac{\sqrt{N}(C_N(r, d) - C_l(r, d))}{\sigma_{r,d}} \right) \rightarrow \mathcal{N}(0, 1), \quad (3.6)$$

donde:

$$\frac{\sigma_{r,d}^2}{4} = K(r, d) - C(r, 2d) + 2 \sum_{i=1}^{d-1} [K(r, d-i)C(r, 2i) - C(r, 2d)]. \quad (3.7)$$

Cuando no se conoce la distribución F no se pueden calcular C y K , por lo que se estudian las propiedades de:

$$T_N(r, d) = C_N(r, d) - C_l(r, d), \quad (3.8)$$

Teorema 3.2

Si $\{u_t, t \in \mathbb{Z}\}$ son independientes idénticamente distribuidos y $d \geq 2$, entonces:

$$\mathcal{L} \left(\frac{\sqrt{N}T_N(r, d)}{V_N(r, d)} \right) \longrightarrow \mathcal{N}(0, 1), \quad (3.9)$$

donde:

$$\begin{aligned} \frac{V_N^2(r, d)}{4} &= d(d-1)\hat{C}_N(r, 2d-2)(\hat{K}_N - \hat{C}_N(r, 2)) + \hat{K}_N^d - \hat{C}_N(r, 2d) \\ &+ 2 \sum_{j=1}^{d-1} [\hat{C}_N(r, 2j)(\hat{K}_N^{d-j} - \hat{C}_N(r, 2d-j)) - d\hat{C}_N(r, 2d-2)(\hat{K}_N - \hat{C}_N(r, 2))], \end{aligned}$$

con:

$$\hat{C}_N = \frac{2}{N^2} \sum_{i=1}^{N-d} \sum_{j=i+1}^{N-d+1} \delta_{i,j}, \quad (3.10)$$

y

$$\hat{K}_N = \frac{2}{N^3} \sum_{i=1}^{N-p} \sum_{j=i+1}^{N-p+1} \sum_{k=j+1}^{N-p+2} \delta_{i,j} \delta_{j,k}, \quad (3.11)$$

Se denomina a $\frac{\sqrt{N}T_N(r, d)}{V_N(r, d)}$ como el estadístico BDS ([Chen and Tsay, 1993](#)) ([Wolff, 1994](#)).

El estadístico BDS se puede utilizar en tests con respecto a la hipótesis nula de que los datos sean independientes e idénticamente distribuidos. Esto con respecto a si el estadístico BDS es diferente o igual a cero. En caso de ser significativamente diferente a cero se rechaza la hipótesis nula, sugiriendo la presencia de dinámicas no lineales o caóticas.

Para datos caóticos de gran dimensionalidad como lo son los creados de manera aleatoria por distintos algoritmos computacionales, el test del estadístico BDS falla a la hora de diferenciar estos de datos que vengan de procesos independientes e idénticamente distribuidos.

El siguiente estadístico se puede utilizar para tests de la misma índole y funciona en donde el BDS puede fallar ([Savit and Green, 1991](#)).

Sea

$$X_N(r, d) = \frac{C_N^2(r, d)}{C_N(r, d-1)C_N(r, d+1)}, \quad (3.12)$$

con

$$X_N(r, 1) = \frac{C_N^2(r, 1)}{C_N(r, 2)}. \quad (3.13)$$

Teorema 3.3

Si $\{u_t, t \in \mathbb{Z}\}$ son independientes idénticamente distribuidos, $d \geq 2$ y $K > C^2$, entonces:

$$\mathcal{L} \left(\frac{\sqrt{N} \ln(X_N(r, d))}{\sigma_{r,d}^*} \right) \longrightarrow \mathcal{N}(0, 1), \quad (3.14)$$

donde:

$$\sigma_{r,d}^{*2} = \frac{\sigma_{r,d-1}^2}{C_l(r, 2d-2)} + \frac{4\sigma_{r,d}^2}{C_l(r, 2d)} + \frac{\sigma_{r,d+1}^2}{C_l(r, 2d+2)} + \rho, \quad (3.15)$$

con ρ la covarianza entre $(C_N(r, d-1) - C_l(r, d-1))$, $(C_N(r, d) - C_l(r, d-1))$ y $(C_N(r, d+1) - C_l(r, d-1))$

En resumen, estos 3 Teoremas se utilizan para obtener estadísticos que permiten realizar pruebas sobre datos y evaluar su distribución y origen correspondiente. A continuación, se detalla cómo cada uno de estos teoremas contribuye al desarrollo de dichos estadísticos y sus respectivas aplicaciones, destacando las limitaciones y requisitos para su implementación práctica.

El primer Teorema aborda el caso ideal, en el cual se asume que se conocen las funciones de distribución involucradas. Aunque este escenario proporciona una base teórica sólida, resulta poco realista en la práctica debido a las dificultades para obtener dichas funciones en contextos aplicados. En particular, la implementación de este teorema requiere calcular expresiones como $C_N(r, d) - C_l(r, d)$, así como disponer de los valores de las varianzas $\sigma_{r,d}$. Estos cálculos, a su vez, dependen de los valores exactos de las funciones de correlación C y K , lo cual puede no ser factible en la mayoría de las aplicaciones empíricas. Por tanto, el primer teorema se presenta como una referencia teórica más que como una herramienta directamente aplicable en entornos experimentales.

El segundo Teorema, en cambio, se define de manera que no requiere el conocimiento explícito de las funciones de distribución, facilitando así su aplicación práctica. Bajo esta formulación, se puede emplear el teorema con base en las dimensiones d seleccionadas, aunque aún se enfrenta la necesidad de calcular $T_N(r, d)$. Adicionalmente, el teorema introduce la variable $V_N(r, d)$, cuyo cálculo no es trivial y cuya complejidad varía significativamente según el caso de estudio. Esto sugiere que, aunque más accesible que el primer teorema, su implementación práctica sigue siendo desafiante y su aplicabilidad dependerá de la capacidad para estimar $V_N(r, d)$ con precisión.

El tercer Teorema se presenta como una alternativa más sencilla, ya que utiliza la expresión $X_N(r, d)$, que en ciertos contextos puede ser menos compleja de calcular y aplicar. En el caso particular de series independientes e idénticamente distribuidas (IID), el tercer teorema permite una implementación directa, siempre que se puedan calcular las correlaciones integrales para un valor específico de d . En este contexto, la obtención de $X_N(r, d)$ se simplifica, dado que las correlaciones integrales pueden ser estimadas con relativa facilidad. Esto convierte al tercer teorema en una opción más viable para analizar series temporales simples o casos donde la estructura de dependencia no es demasiado compleja.

Capítulo 4

Experimentos

4.1. Simulaciones de Datos

Con el objetivo de estudiar en detalle el comportamiento de los estadísticos planteados anteriormente se eligió la expresión $\sqrt{N} \ln(X_N(r, d))$ para distintos valores del parámetro d , donde se ha decidido realizar un análisis bajo la hipótesis nula de que los datos sean independientes e idénticamente distribuidos (IID). Específicamente, bajo esta hipótesis, se asume que los datos provienen de una distribución normal, lo que garantiza que no haya correlación alguna entre las observaciones y que su comportamiento estadístico sea consistente a lo largo del tiempo.

Para llevar a cabo este análisis, se procedió a generar un conjunto de series de tiempo simuladas a partir de esta distribución normal estándar. Se consideraron tres configuraciones distintas con el fin de evaluar la robustez de los resultados: en el primer caso, se generaron 100 series con una longitud de $N = 1000$ observaciones cada una; en el segundo, se simularon 500 series, manteniendo el mismo tamaño muestral de $N = 1000$; y finalmente, en el tercer caso, se crearon 1000 series de igual longitud. La elección de estas configuraciones tiene como propósito analizar cómo varía el comportamiento de la estadística en cuestión a medida que se incrementa el número de series, y, por ende, se explora un mayor rango de variabilidad posible.

La simulación de estos diferentes escenarios permite observar con mayor precisión la estabilidad y la sensibilidad de $\sqrt{N} \ln(X_N(r, d))$ frente a cambios en la cantidad de datos disponibles, proporcionando un marco robusto para validar la hipótesis nula y asegurando que los resultados obtenidos no se vean afectados por un número limitado de observaciones o por fluctuaciones aleatorias en los datos generados. En consecuencia, el análisis empírico se ve fortalecido y se puede confiar en la validez de los resultados obtenidos bajo estas condiciones experimentales.

Para cada una de estas series generadas se puede evaluar el comportamiento de las expresiones $\sqrt{N} \ln(X_N(r, 1))$, $\sqrt{N} \ln(X_N(r, 2))$, $\sqrt{N} \ln(X_N(r, 3))$ y $\sqrt{N} \ln(X_N(r, 4))$, respectivamente. A partir de estos resultados, se procede a realizar pruebas de bondad de ajuste con el fin de evaluar la adecuación de la distribución subyacente y determinar si existe evidencia en contra de la hipótesis nula.

Las series IID se generaron siguiendo la estructura previamente establecida y, a continuación, se les aplicaron las fórmulas de correlación integral correspondientes, así

como las de $X_N(r, d)$ respectivamente. El análisis reveló que, dependiendo del valor de d y r , es posible obtener resultados donde las correlaciones integrales se aproximan a cero. Esto ocurre particularmente cuando el valor de d es alto, incluso en el caso de datos generados bajo la hipótesis de independencia. Esta situación sugiere la necesidad de ser especialmente cautelosos al evaluar la validez de las correlaciones para valores grandes de d , dado que podría llevar a resultados no informativos o interpretaciones erróneas. En efecto, se observó que para d mayores a 8, la correlación integral se reduce a cero en datos IID, lo cual compromete la estabilidad de las métricas calculadas y podría invalidar cualquier conclusión derivada de estos resultados.

En contraste, el parámetro r presenta una menor problemática en cuanto a su impacto sobre el valor de la correlación. Si bien su influencia es notable, especialmente en la medida en que afecta el rango de valores considerados en la sumatoria de δ , este parámetro no tiende a generar resultados nulos en la correlación integral. Esto se debe a que, por definición, un incremento en r aumenta la probabilidad de que los valores de δ cercanos contribuyan a la sumatoria de la correlación, mitigando el riesgo de valores cero en las correlaciones integrales.

Sin embargo, a pesar de estas observaciones, el cálculo del estadístico $\sqrt{N} \ln(X_N(r, d))$ no presenta mayores inconvenientes siempre y cuando no se seleccionen valores de r demasiado pequeños con respecto a los valores de las series. Por otro lado, para valores de d muy grandes, el cálculo de este estadístico puede enfrentar dificultades considerables, ya que la expresión $\sqrt{N} \ln(X_N(r, d))$ involucra la división por correlaciones integrales que, bajo estas condiciones, tienden a ser cero. Este fenómeno subraya la importancia de elegir con cuidado los valores de d y r utilizados en el análisis para evitar problemas numéricos que puedan sesgar o distorsionar las conclusiones obtenidas. Por consiguiente, se recomienda mantener un rango razonable para estos parámetros durante la evaluación empírica para garantizar resultados consistentes y significativos. Esto se condice con experimentos como en [Lai \(1999\)](#) donde también se optó para el apartado de experimentos numéricos por utilizar los casos con valores pequeños para d .

Durante las pruebas realizadas con series generadas bajo la hipótesis de independencia e idéntica distribución (IID), se pudo constatar que, a medida que se incrementaba el número de series simuladas, el estadístico $\sqrt{N} \ln(X_N(r, d))$ exhibía un comportamiento que confirmaba su convergencia hacia la normalidad asintótica. Específicamente, en experimentos iniciales con 100 series, ya se podía observar un comportamiento aproximado a la normalidad para este estadístico, mientras que, de manera menos pronunciada, el mismo patrón se comenzaba a reflejar también en los valores de $X_N(r, d)$.

Al aumentar la cantidad de series a 500, tanto los gráficos resultantes como las pruebas de normalidad aplicadas corroboraron de manera clara y robusta que tanto $\sqrt{N} \ln(X_N(r, d))$ como $X_N(r, d)$ tendían a una distribución normal. Esto sugiere que el comportamiento de normalidad asintótica se establece con mayor evidencia a medida que se incrementa el tamaño del conjunto de datos, lo que se traduce en una mayor estabilidad y precisión en la evaluación de estos estadísticos. Dicho comportamiento se torna aún más evidente conforme se incrementa el número de series evaluadas.

Estos resultados indican que la aproximación a la normalidad asintótica se manifiesta no solo en el estadístico $\sqrt{N} \ln(X_N(r, d))$, sino también, aunque en menor medida, en la propia distribución de $X_N(r, d)$. La rápida convergencia observada incluso con 100 series sugiere que el tamaño muestral requerido para alcanzar un comportamiento aproximadamente normal es relativamente bajo, haciendo factible la aplicación de pruebas basadas en esta normalidad en contextos experimentales donde no se dispone de un número muy elevado de realizaciones. En consecuencia, para un análisis más riguroso y detallado, la utilización de al menos 500 series se recomienda como un estándar para asegurar la validez de los resultados obtenidos y garantizar que el comportamiento asintótico de los estadísticos analizados refleje de manera adecuada la estructura teórica esperada bajo la hipótesis nula.

Se puede observar por ejemplo en las Figuras 4.1 y 4.2 los gráficos tipo $Q - Q$ de $\sqrt{N} \ln(X_N(r, 2))$ y $\sqrt{N} \ln(X_N(r, 3))$ respectivamente donde se evidencia lo anteriormente descrito.

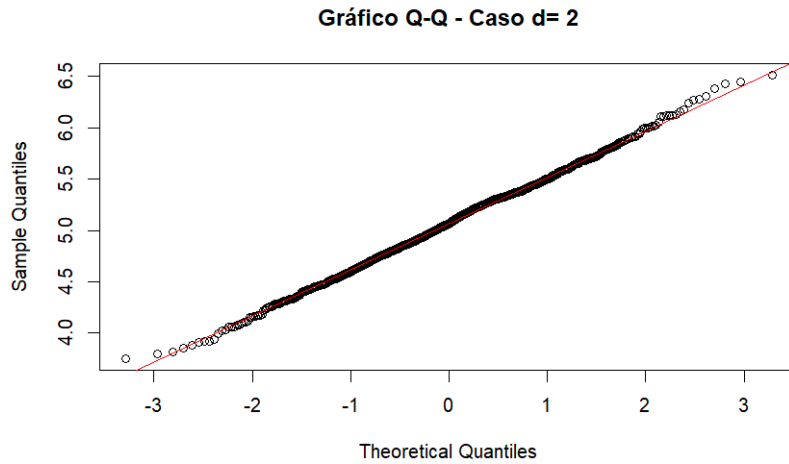


Figura 4.1: Gráfico Q-Q de datos IID con $d = 2$

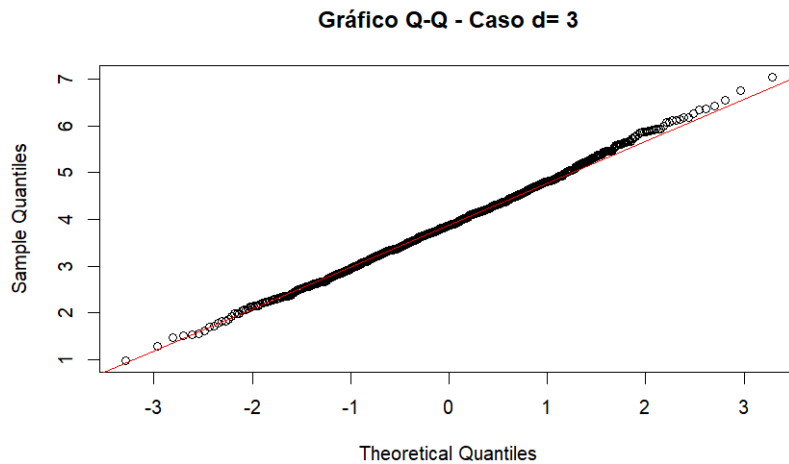


Figura 4.2: Gráfico Q-Q de datos IID con $d = 3$

En el caso de datos generados a partir de un proceso autorregresivo de primer orden (AR(1)), se observó que la expresión $\sqrt{N} \ln(X_N(r, d))$ mantenía una tendencia a la normalidad asintótica, aunque de manera menos evidente en comparación con las series independientes e idénticamente distribuidas (IID). En este contexto, se detectaron dificultades con los valores elevados de d en una etapa más temprana que en el caso IID. Sin embargo, al incrementar el número de series simuladas de 100 a 500 y luego a 1000, se hizo aún más visible el aumento de esta tendencia hacia la normalidad asintótica.

Los resultados obtenidos a través de los gráficos y las pruebas de bondad de ajuste corroboraron este comportamiento, confirmando que, a pesar de que los procesos AR(1) no cumplen con los supuestos de independencia requeridos por los teoremas que establecen la normalidad asintótica, la aplicación numérica del estadístico $\sqrt{N} \ln(X_N(r, d))$ mostró ser viable y estable bajo ciertas condiciones. Este hallazgo resulta particularmente interesante y abre nuevas líneas de investigación, ya que sugiere que la estructura de dependencia temporal inherente a los procesos AR(1) no impide completamente la convergencia a la normalidad de los estadísticos considerados.

Además, se puede observar que, al analizar los gráficos para $\sqrt{N} \ln(X_N(r, 2))$ y $\sqrt{N} \ln(X_N(r, 3))$ correspondientes a series AR(1) en las Figuras 4.3 y 4.4, la normalidad asintótica se mantiene en términos generales, aunque con ciertas variaciones que no se presentan en las series IID. Este fenómeno podría deberse a la presencia de correlación serial que caracteriza a los procesos autorregresivos, la cual afecta las propiedades de las correlaciones integrales y, en consecuencia, la distribución del estadístico.

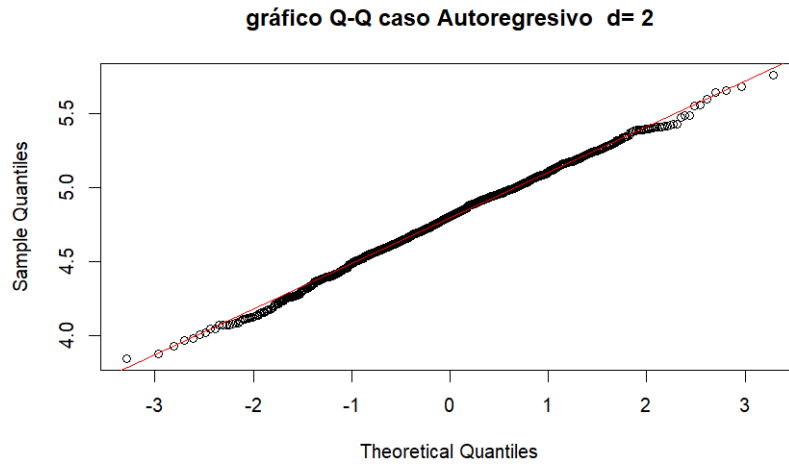


Figura 4.3: Gráfico Q-Q de datos Autorregresivos con $d = 2$

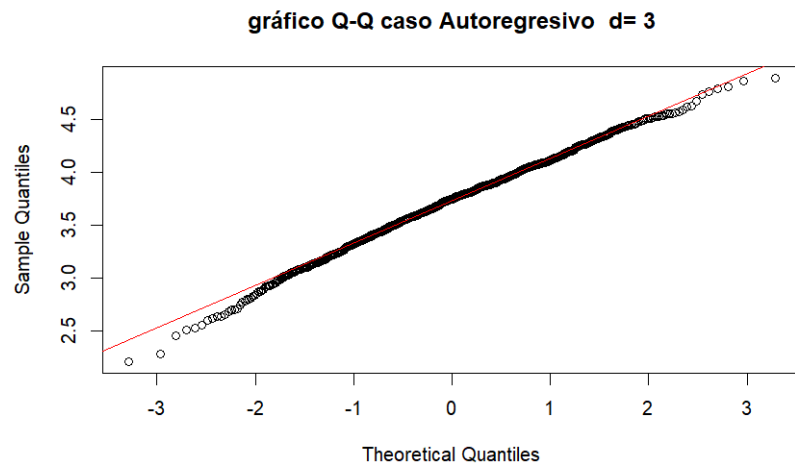


Figura 4.4: Gráfico Q-Q de datos Autorregresivos con $d = 3$

Capítulo 5

Conclusiones

En este capítulo se presentan las conclusiones derivadas del estudio sobre las propiedades estadísticas de la correlación integral. A partir de las definiciones y notaciones dadas y junto con los experimentos realizados y el análisis gráfico de los resultados, se exponen las principales observaciones y aportes de este trabajo, junto con posibles direcciones futuras para la investigación en esta área.

Este estudio representa una contribución significativa a la literatura sobre la correlación integral, ya que sintetiza y organiza de manera coherente diversas propiedades fundamentales de esta herramienta estadística bajo una notación unificada. Al reunir estos resultados y presentarlos de manera estructurada, se facilita su análisis y comprensión, proporcionando un recurso útil tanto para investigadores como para profesionales que deseen profundizar en el tema.

Además, este trabajo no solo compila las propiedades más relevantes, sino que también ofrece una introducción exhaustiva al concepto de correlación integral, abordando los conocimientos y definiciones básicas necesarias para entender su funcionamiento y aplicación. Esta introducción, permite a los lectores adquirir una comprensión sólida de los principios subyacentes antes de adentrarse en aspectos más complejos, proporcionando una base sólida para la exploración de sus propiedades y usos en contextos científicos.

Al ofrecer una perspectiva integrada y didáctica sobre la correlación integral, este estudio se convierte en una referencia accesible y valiosa para quienes deseen emplear esta herramienta en el análisis de datos o para aquellos interesados en su desarrollo teórico, contribuyendo así al avance del conocimiento en este campo y promoviendo su aplicación en una amplia variedad de investigaciones científicas y académicas.

Los resultados obtenidos a partir de los experimentos numéricos realizados respaldan de manera robusta la validez de los teoremas y propiedades teóricas analizadas en este estudio. En particular, se pudo observar que el comportamiento gráfico del estadístico $\sqrt{N} \ln(X_N(r, d))$ se ajusta a las predicciones teóricas en diferentes escenarios y tamaños muestrales. Al aplicar distintos tamaños de muestra para cada experimento, se evidenció que, a medida que el tamaño muestral aumentaba, la tendencia hacia la normalidad asintótica se volvía más evidente y clara, consolidando así la convergencia del estadístico hacia su distribución límite.

Este comportamiento es consistente con lo esperado teóricamente, ya que los incrementos en el tamaño muestral reducen la influencia de las fluctuaciones aleatorias y permiten una mayor estabilidad en la estimación de los parámetros del modelo. De hecho, a partir de muestras de mayor tamaño, la aproximación a la normalidad asintótica se vuelve notablemente pronunciada, tanto en la visualización gráfica como en los resultados de las pruebas de bondad de ajuste aplicadas. Esta concordancia entre los resultados empíricos y las predicciones teóricas demuestra la fiabilidad del estadístico $\sqrt{N} \ln(X_N(r, d))$ como una herramienta adecuada para el análisis de este tipo de datos bajo las condiciones establecidas en los teoremas.

A pesar de los resultados obtenidos, este estudio presenta algunas limitaciones, como en el caso de los ejemplos numéricos de donde se obtuvieron los gráficos donde sólo se estudió una parte del estadístico del Teorema 3.3 y se dejaron de lado los posibles estudios hacia las varianzas de este. Esto en un principio se intentó modelar y estimar pero se terminó descartando de este trabajo en concreto por lo que podría ser materia de un trabajo futuro.

En definitiva, la capacidad de $\sqrt{N} \ln(X_N(r, d))$ para aproximarse a la normalidad asintótica en procesos autorregresivos, incluso cuando los supuestos teóricos no se cumplen plenamente, representa un resultado prometedor y digno de un análisis más profundo. A medida que se aumenta el número de series, se mitigan los efectos adversos de la dependencia temporal, logrando así una tendencia más clara hacia la normalidad. Esto sugiere que, en escenarios prácticos, el uso de este estadístico podría extenderse a datos no necesariamente IID, siempre que se consideren con cautela las limitaciones impuestas por la estructura de dependencia en los datos.

En resumen, en este trabajo se pudo englobar lo que son los principales resultados con respecto al estudio y comprensión de la correlación integral, así como importantes propiedades que hacen que esta sea relevante de estudiar. Las conclusiones aquí presentadas demuestran la importancia de la correlación integral y su potencial para seguir desarrollándose en futuras investigaciones, con aplicaciones que podrían extenderse a diversos campos tanto teóricos como aplicados.

Capítulo 6

Apéndice

6.1. Demostraciones

Teorema 3.3

Si $\{u_t, t \in \mathbb{Z}\}$ son independientes idénticamente distribuidos , $d \geq 2$ y $K > C^2$, entonces:

$$\mathcal{L} \left(\frac{\sqrt{N} \ln(X_N(r, d))}{\sigma_{r,d}^*} \right) \longrightarrow \mathcal{N}(0, 1), \quad (6.1)$$

donde:

$$\sigma_{r,d}^{*2} = \frac{\sigma_{r,d-1}^2}{C_l^{2d-2}} + \frac{4\sigma_{r,d}^2}{C_l^{2d}} + \frac{\sigma_{r,d+1}^2}{C_l^{2d+2}} + \rho, \quad (6.2)$$

con ρ la covarianza entre $(C_N(r, d-1) - C_l^{d-1})$, $(C_N(r, d) - C_l^{d-1})$ y $(C_N(r, d+1) - C_l^{d-1})$.

Demostración

Se puede notar que

$$\begin{aligned} \ln X_N(r, d) &= 2 \ln C_N(r, d) - \ln C_N(r, d-1) - \ln C_N(r, d+1) \\ &= (1/\sqrt{n}) \left[2\sqrt{n} (\ln C_N(r, d) - \ln C_1(r, d)) \right. \\ &\quad - \sqrt{n} (\ln C_N(r, d-1) - \ln C_l(r, d-1)) \\ &\quad \left. - \sqrt{n} (\ln C_N(r, d+1) - \ln C_l(r, d+1)) \right] \end{aligned}$$

Por expansión de Taylor, se tiene:

$$\begin{aligned}
\sqrt{n} (\ln C_N(r, d) - \ln C_1(r, d)) &= (\sqrt{n}/C_1(r, d)) (C_N(r, d) - C_1(r, d)) \\
&\quad + (C_1^*/\sqrt{n}) [\sqrt{n} (C_N(r, d) - C_1(r, d))]^2 \\
\sqrt{n} (\ln C_N(r, d-1) - \ln C_1(r, d-1)) &= (\sqrt{n}/C_1(r, d-1)) (C_N(r, d-1) - C_1(r, d-1)) \\
&\quad + (C_2^*/\sqrt{n}) [\sqrt{n} (C_N(r, d-1) - C_1(r, d-1))]^2, \\
\sqrt{n} (\ln C_N(r, d+1) - \ln C_1(r, d+1)) &= (\sqrt{n}/C_1(r, d+1)) (C_N(r, d+1) - C_1(r, d+1)) \\
&\quad + (C_3^*/\sqrt{n}) [\sqrt{n} (C_N(r, d+1) - C_1(r, d+1))]^2,
\end{aligned}$$

donde C_1^*, C_2^* y C_3^* son constantes obtenidas de la expansión de Taylor.

Sea

$$\begin{aligned}
\xi_n &= (2C_1^*/\sqrt{n}) [\sqrt{n} (\delta_{i,j}(d) - C_1(r, d))]^2 \\
&\quad - (C_2^*/\sqrt{n}) [\sqrt{n} (\delta_{i,j}(d-1) - C_1(r, d-1))]^2 \\
&\quad - (C_3^*/\sqrt{n}) [\sqrt{n} (\delta_{i,j}(d+1) - C_1(r, d+1))]^2,
\end{aligned}$$

entonces

$$\begin{aligned}
\sqrt{n} \ln X_N(r, d) &= (2\sqrt{n}/C_1(r, d)) (C_N(r, d) - C_1(r, d)) \\
&\quad + (\sqrt{n}/C_1(r, d-1)) (C_N(r, d-1) - C_1(r, d-1)) \\
&\quad + (\sqrt{n}/C_1(r, d+1)) (C_N(r, d+1) - C_1(r, d+1)) + \xi_n.
\end{aligned}$$

Queda claro entonces que $\xi_n \rightarrow 0$ cuando $n \rightarrow \infty$.

Por el Teorema 3.1,

$(\sqrt{n}/C_1(r, d)) (C_N(r, d) - C_1(r, d))$, $(\sqrt{n}/C_1(r, d-1)) (C_N(r, d-1) - C_1(r, d-1))$, y $(\sqrt{n}/C_1(r, d+1)) (C_N(r, d+1) - C_1(r, d+1))$ convergen a variables aleatorias distribuidas normalmente. Por lo tanto,

$$\mathcal{L} \left(\sqrt{n} \ln X_N(r, d) / \sigma_{r,d}^* \right) \rightarrow \mathcal{N}(0, 1) \quad ,$$

donde $\sigma_{r,d}^{*2} = 4\sigma_{r,d}^2/C_1(r, 2d) + \sigma_{r,d-1}^2/C_1^{2d-2} + \sigma_{r,d+1}^2/C_1^{2d+2} + \rho$, y, calculando las covarianzas se tiene,

$$\begin{aligned}
\rho &= -4 \sum_{j=1}^{d-1} [\Sigma_j K^j C_1(r, 2d-1-2j) - C_1(r, 2d-1)] / C_1(r, 2d-1) \\
&\quad - 4 \sum_{j=1}^d [\Sigma_j K^j C_1(r, 2d+1-2j) - C_1(r, 2d+1)] / C_1(r, 2d+1) \\
&\quad + 2 \sum_{j=1}^{d-1} [\Sigma_j K^j C_1(r, 2d-2j) - C_1(r, 2d-1)] / C_1(r, 2d)
\end{aligned}$$

donde Σ_j denota la suma de todos los posibles términos tal que hay exactamente j pares del mismo ítem entre los productos.

6.2. Gráficos

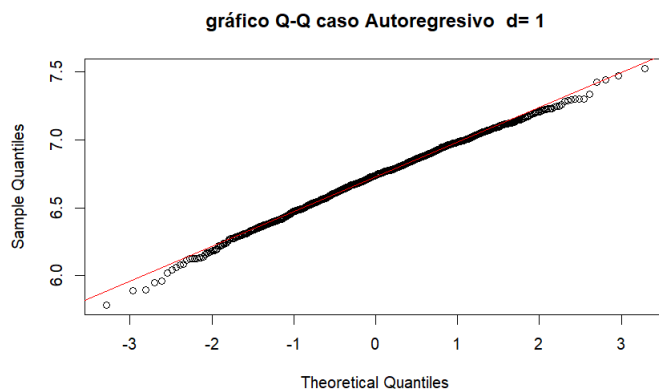


Figura 6.1: Gráfico Q-Q de datos Autorregresivos con $d = 1$

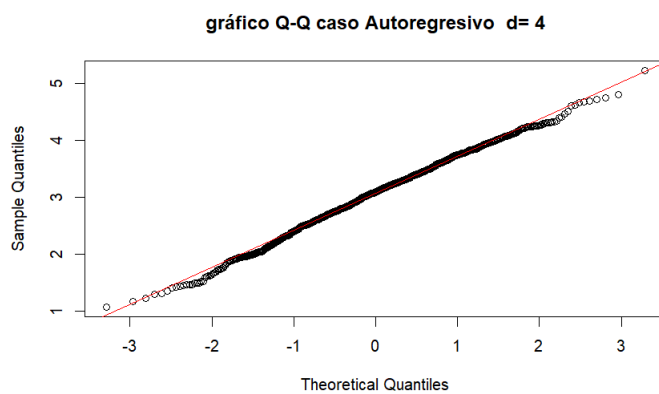


Figura 6.2: Gráfico Q-Q de datos Autorregresivos con $d = 4$

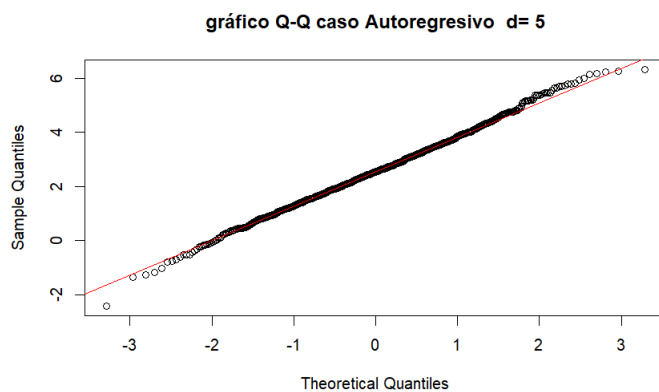
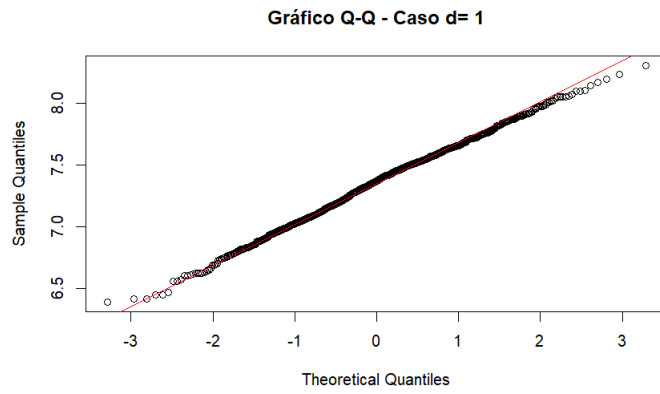
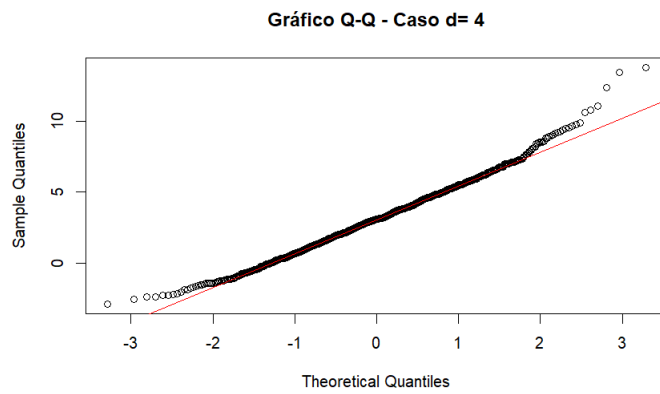
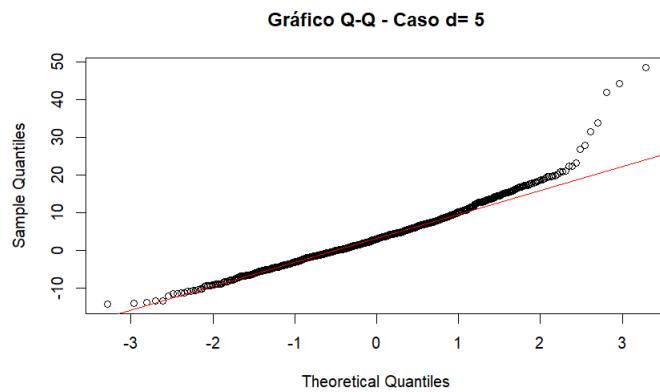


Figura 6.3: Gráfico Q-Q de datos Autorregresivos con $d = 5$

Figura 6.4: Gráfico Q-Q de datos IID con $d = 1$ Figura 6.5: Gráfico Q-Q de datos IID con $d = 4$ Figura 6.6: Gráfico Q-Q de datos IID con $d = 5$

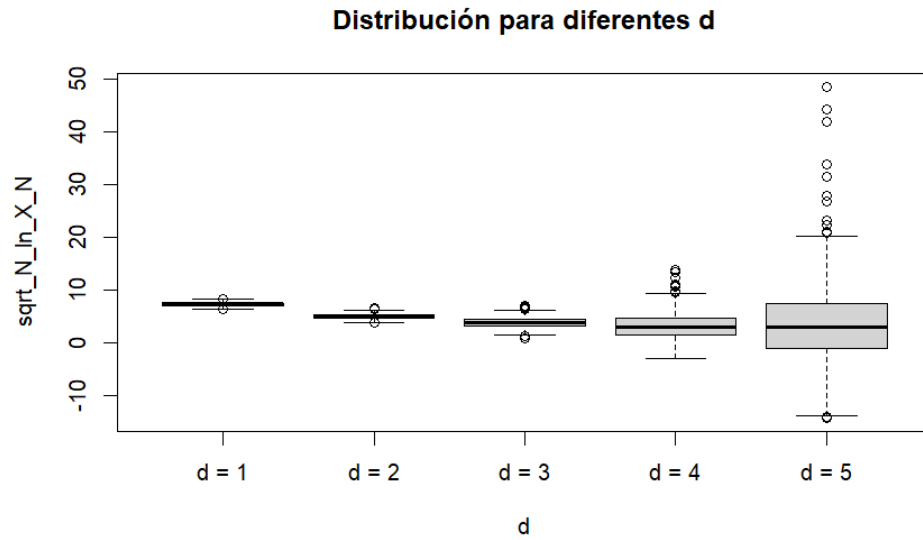


Figura 6.7: Gráficos box-plot IID valores de d

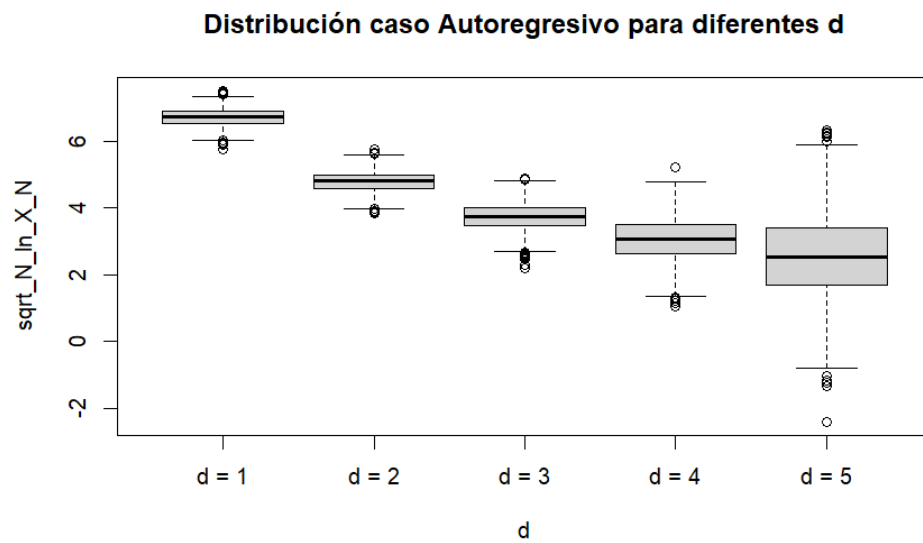


Figura 6.8: Gráficos box-plot Autorregresivos valores de d

Gráficos extraídos de las simulaciones obtenidas en R, trabajo completo en [Github](#)

Bibliografía

- Acosta, J., Vallejos, R., and Gómez, J. (2024). Correlation integral for stationary gaussian time series. *Sankhya A*, 86(1):191–214.
- Capasso, V. and Bakstein, D. (2021). Stochastic processes. In *An Introduction to Continuous-Time Stochastic Processes: Theory, Models, and Applications to Finance, Biology, and Medicine*, pages 81–201. Springer.
- Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting*, 15(7):495–508.
- Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308.
- Di, C., Wang, T., Yang, X., and Li, S. (2018). An improved grassberger–procaccia algorithm for analysis of climate system complexity. *Hydrology and Earth System Sciences*, 22(10):5069–5079.
- Furdui, O. (2013). Limits, series, and fractional part integrals. *Problems in Mathematical Analysis*, New York.
- Glasbey, C. (2001). Non-linear autoregressive time series with multivariate gaussian mixtures as marginal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):143–154.
- Grassberger, P. and Procaccia, I. (1983). Characterization of strange attractors. *Physical review letters*, 50(5):346.
- Jones, R. H. (1978). Multivariate autoregression estimation using residuals. In *Applied Time Series Analysis I*, pages 139–162. Elsevier.
- Lai, D. (1999). Asymptotic distributions of the correlation integral based statistics. *Journal of Nonparametric Statistics*, 10(2):127–135.
- Mormann, F., Kreuz, T., Rieke, C., Andrzejak, R. G., Kraskov, A., David, P., Elger, C. E., and Lehnertz, K. (2005). On the predictability of epileptic seizures. *Clinical neurophysiology*, 116(3):569–587.
- Percival, D. B. and Walden, A. T. (2020). *Spectral analysis for univariate time series*, volume 51. Cambridge University Press.
- Sarkar, N. and Chaudhuri, B. B. (1994). An efficient differential box-counting approach to compute fractal dimension of image. *IEEE Transactions on systems, man, and cybernetics*, 24(1):115–120.

- Savit, R. and Green, M. (1991). Time series and dependent variables. *Physica D: Nonlinear Phenomena*, 50(1):95–116.
- Tong, H. (1990). *Non-linear time series: a dynamical system approach*. Oxford university press.
- Wolff, R. C. (1990). A note on the behaviour of the correlation integral in the presence of a time series. *Biometrika*, 77(4):689–697.
- Wolff, R. C. (1994). Independence in time series: another look at the bds test. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 348(1688):383–395.