



**CIMAT**

---

---

Centro de Investigación en Matemáticas, A.C.

**Métodos para Series de Tiempo: Detección  
de Memoria Larga y Modelos Funcionales  
Robustos**

TESIS

Que para obtener el Grado de:

**Doctor en Ciencias  
con Orientación en Probabilidad y Estadística**

P R E S E N T A:

Israel Martínez Hernández

Director de tesis:

Dra. Graciela González Farías

Guanajuato, Gto, México

Diciembre de 2017



## **Integrantes del Jurado.**

**Presidente: Dr. Daniel Hernández Hernández**  
CIMAT

**Secretario: Dr. Rolando José Biscay Lirio**  
CIMAT

**Vocal: Dr. Joaquín Ortega Sánchez**  
CIMAT

**Vocal: Dr. Ramsés Humberto Mena Chávez**  
IIMAS-UNAM

**Vocal y director de tesis: Graciela González Farías**  
CIMAT

**Asesor:**

---

Dra. Graciela González Farías.

**Sustentante:**

---

M.C. Israel Martínez Hernández.



*A mis padres, Andrés Martínez Robles y Magdalena Hernández Reyes.*

*A Carolina Euán Campos.*



# Agradecimientos

Agradezco a la Dra. Graciela González Farías por todo su apoyo. Por haber escuchado mis ideas sin sentido al inicio de mi doctorado. Por haber tenido la paciencia en la última etapa del mismo. Por sus sabias sugerencias cuando la requerí. Por haber sido no solo una asesora de tesis si no también como a alguien parte de la familia.

Agradezco a los profesores que me recibieron en mis estancias académicas. Al Dr Jesús Gonzalo de la Universidad Carlos III de Madrid, por haberme permitido trabajar y discutir con él acerca de los problemas relevantes en la tesis. Al profesor Marc Genton, de King Abdullah University of Science and Technology, por haberme brindado la oportunidad de formar parte de su grupo y desarrollar la segunda parte de la tesis.

Agradezco a todos los miembros del comité evaluador de esta tesis; Dr Daniel Hernández, Dr Joaquín Ortega, Dr Rolando Biscay y al Dr Ramsés Mena, por sus comentarios y sugerencias a la tesis.

Agradezco a todos los profesores que contribuyeron en mi formación académica. Especialmente a los profesores de CIMAT del área de probabilidad y estadística.

Agradezco a CIMAT y CONACYT por la beca que se me otorgó. En particular los apoyos del proyecto CONACYT, CB-2015-01-252996.





# Resumen

La tesis consiste en el estudio de dos problemas, ambos en un contexto de series temporales; prueba de persistencia basada en la entropía y estimación robusta del proceso autoregresivo funcional.

En la primera parte se estudia el concepto de memoria larga en series de tiempo no lineales,  $\{X_t\}$ , usando la información mutua como una medida de dependencia. Esta medida es usada similarmente a la medida de covarianza en el caso lineal. Se considera la información mutua a pares, es decir, la información mutua de  $(X_t, X_{t-h})$ , con  $h = 0, 1, \dots$ . Se supone que dicha información mutua es una función continua en  $h$ . A partir de esto, se propone una prueba de hipótesis basada en la entropía del proceso para discriminar si el proceso tiene memoria corta o memoria larga. Esta prueba es con un enfoque funcional y prueba envolvente. Los valores críticos son obtenidos de simulaciones bajo la hipótesis nula usando una noción de orden en las curvas.

En la segunda parte, se propone un estimador robusto en el modelo autoregresivo funcional de orden uno, ordenando las curvas observadas de acuerdo a la similaridad de ciertas propiedades que dichas curvas deben satisfacer. Se asume que el operador de coeficiente es un operador integral. La estimación es obtenida usando el método de mínimos cuadrados ponderados. Se propone una función de ponderación basada en la noción de profundidad. Se muestra que la función de influencia correspondiente es acotada y que el estimador preserva las propiedades deseables.



# Índice general

<b>Agradecimientos</b>	<b>III</b>
<b>Resumen</b>	<b>V</b>
<b>1. Introducción</b>	<b>1</b>
<b>2. Conceptos Preliminares</b>	<b>9</b>
2.1. Series de tiempo y memoria larga . . . . .	9
2.2. Variables aleatorias en espacio de funciones . . . . .	14
2.3. Ordenamiento de curvas basado en profundidades . . . . .	18
2.3.1. Profundidad de Banda Modificada . . . . .	19
2.3.2. Medida de cercanía de Stahel-Donoho . . . . .	21
2.4. Ajuste de curvas . . . . .	23
<b>3. Detección de Persistencia en Series de Tiempo No Lineales</b>	<b>27</b>
3.1. Pruebas de persistencia existentes . . . . .	28
3.2. Persistencia y el concepto mixing . . . . .	32
3.3. Información mutua y cópulas . . . . .	36
3.3.1. Estimación de la información mutua . . . . .	40
3.4. Prueba de persistencia basada en la entropía . . . . .	42
3.4.1. Prueba envolvente . . . . .	43
3.4.2. Prueba de hipótesis de memoria larga . . . . .	46
3.5. Estudio de simulación Monte Carlo . . . . .	50
3.5.1. Comportamiento de $T(h)$ . . . . .	50
3.5.2. Nivel de significancia y potencia . . . . .	53
<b>4. Estimación Robusta de FAR(1) Basada en Profundidades</b>	<b>59</b>
4.1. Métodos de estimación para el modelo FAR(1) . . . . .	59
4.2. Efecto de datos atípicos en la estimación del modelo FAR(1) . . . . .	63

---

4.3. Estimación vía mínimos cuadrados . . . . .	65
4.3.1. Construcción del estimador robusto basado en profundidades . . . . .	68
4.4. Función de influencia del estimador DLS . . . . .	71
4.4.1. Función de influencia . . . . .	71
4.4.2. Función de influencia empírica . . . . .	75
4.5. Estudio de simulación Monte Carlo . . . . .	76
4.5.1. Esquema de la simulación . . . . .	77
4.5.2. Resultados de Simulación . . . . .	79
4.5.3. Función de influencia empírica . . . . .	87
<b>Bibliografía</b>	<b>91</b>

# Capítulo 1

## Introducción

En esta tesis se estudian dos problemas diferentes pero ambos en el contexto de series temporales. El primero es el estudio de memoria larga o dependencia a largo plazo en series de tiempo no lineales, y el segundo es el estudio de un estimador robusto en el modelo autoregresivo funcional. En ambos se usan las herramientas de Análisis de Datos Funcionales (FDA) y otros conceptos relacionados como profundidad funcional.

El estudio de memoria larga en series de tiempo, comenzó con los trabajos de Hurst ([Hurst, 1951, 1956](#)), cuando estudiaba los datos del río Nilo con el propósito de construir una presa. Sean  $X_1, \dots, X_n$   $n$  observaciones recolectadas en el tiempo y sea  $S_m = X_1 + \dots + X_m$  las sumas parciales donde  $m = 0, 1, \dots$ , con  $S_0 = 0$ , y consideremos el estadístico de Hurst definido como,

$$\frac{R}{S}(X_1, \dots, X_n) = \frac{\max_{0 \leq i \leq n} (S_i - \frac{i}{n} S_n) - \min_{0 \leq i \leq n} (S_i - \frac{i}{n} S_n)}{\left( \frac{1}{n} \sum_{i=1}^n (X_i - \frac{1}{n} S_n)^2 \right)^{1/2}}. \quad (1.1)$$

Si los datos son independientes (o “cerca” de serlos), el estadístico  $\frac{R}{S}(X_1, \dots, X_n)$  crece en el orden de  $n^{0.5}$ , es decir  $\frac{R}{S}(X_1, \dots, X_n) \sim n^{0.5}$ . Cuando Hurst calculó el estadístico con los datos del río Nilo, se percató que el orden de crecimiento era aproximadamente  $n^{0.74}$ , lo cual atrajo atención y se volvió un fenómeno interesante, llamado el *fenómeno de Hurst*. También, este fenómeno conllevó al estudio de procesos estocásticos que explicaran tal fenómeno.

En los trabajos de Mandelbrot ([Mandelbrot and Wallis, 1968](#)) se propuso modelar el fenómeno de Hurst usando incrementos del Movimiento Browniano Fraccionario (FBM). Un proceso FBM es definido como un proceso Gaussiano  $\{B_H(t), t \geq 0\}$  tal que:

- $\mathbb{E}(B_H(t)) = 0$
- $B_H(0) = 0$
- $\mathbb{E}((B_H(t) - B_H(s))^2) = \sigma^2 |t - s|^{2H}$  con  $0 < H < 1$  y  $\sigma > 0$ .

Este proceso tiene una propiedad de auto-similitud,  $B_H(ct) \stackrel{L}{=} c^H B_H(t)$  para  $c > 0$ . Observar que, cuando  $H = 1/2$ ,  $B_H(t)$  es el movimiento Browniano.

Un ruido Gaussiano Fraccionario (FGN)  $X = \{X_k, k = 0, 1, \dots\}$  es definido como los incrementos del proceso FBM, esto es

$$X_k = B_H(k+1) - B_H(k).$$

Observemos que, por definición, FGN es estacionario de segundo orden (ver Definición 2.4) con función de covarianza

$$\gamma(k) = \text{Cov}(X_{n+k}, X_n) = \frac{\sigma^2}{2} [(k-1)^{2H} + (k+1)^{2H} + k^{2H}],$$

luego usando una expansión de Taylor se puede ver que

$$\rho(k) = \text{Corr}(X_{n+k}, X_n) \sim H(2H-1)k^{-2(1-H)}.$$

Si  $1/2 < H < 1$ , se tiene que

$$\sum_{k \geq 0} \rho(k) = \infty. \tag{1.2}$$

Además, la varianza de las sumas parciales es tal que  $\text{Var}(X_1 + \dots + X_n) = \text{Var}(B_H(n)) = \sigma^2 n^{2H}$ . Entonces, si  $\{X_i, i = 1, \dots\}$  es un FGN, el estadístico  $R/S$  crece con un orden de  $n^H$ , lo cual explica el fenómeno de Hurst. El punto clave en esta idea es la velocidad con la que decrece la correlación entre  $X_t$  y  $X_{t-k}$  cuando  $k \rightarrow \infty$ . La tasa de decrecimiento es lento. Así, el proceso es aún estacionario pero con una estructura de dependencia más fuerte que un proceso estacionario usual. En general, cuando la estructura de dependencia es fuerte, en el sentido de la ecuación (1.2) se dice que la serie de tiempo tiene memoria larga. Esta no es la única manera de definir memoria larga, también se ha usado la varianza de las sumas parciales, ver Sección 2.1.

A partir de la idea de explicar el fenómeno de Hurst usando FGN, la memoria larga también se ha estudiado desde una perspectiva de procesos autosimilares (ver Samorodnitsky, 2016, Cap. 8).

En la práctica, identificar procesos con memoria larga puede ser una tarea difícil. En Bhattacharya et al. (1983) se muestra el problema de distinguir entre procesos

estacionarios con memoria larga y procesos no estacionarios con memoria corta. Sea  $\beta$  tal que  $0 < \beta < 1/2$  y sea  $a \geq 0$ , consideremos el modelo

$$X_i = Y_i + (a + i)^{-\beta}, \quad i = 1, 2, \dots, \quad (1.3)$$

donde  $Y_i$  son i.i.d. con varianza finita. Observemos que el modelo tiene una tendencia determinista pero esta desaparece cuando  $i$  es grande, es decir, asintóticamente es estacionario. Dado que la sucesión  $Y_i$  es independiente, el proceso no tiene memoria larga. En el artículo mencionado se muestra que si los datos satisfacen el modelo (1.3), entonces el estadístico  $\frac{R}{S}(X_1, \dots, X_n)$  crece con un orden de  $n^{1-\beta}$ , similar al caso del FGN con  $H = 1 - \beta$ . Por lo tanto, con el estadístico de Hurst, los procesos

$$X_n = B_H(n+1) - B_H(n) \text{ y } X_n = Y_n + (a+n)^{H-1},$$

son indistinguibles, siendo de memoria larga el primer proceso y de memoria corta el segundo. Con esto se concluye que el estadístico R/S no es adecuado para procesos estocásticos no lineales en general.

Por otro lado, la teoría de los modelos Autoregresivos de Promedio Móvil (ARMA) ofrece una manera de clasificar a los procesos relacionándolos con la no estacionariedad. Un proceso  $\{X_t\}$  que satisface el modelo ARMA( $p, q$ ), es definido como

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}, \quad (1.4)$$

con  $\phi_p \neq 0$ ,  $\theta_q \neq 0$  y  $\{w_t\}$  es ruido blanco con media cero y varianza  $\sigma_w^2 > 0$ . En este contexto la clasificación de los procesos es bien conocida. Un proceso ARMA es estacionario si los ceros del polinomio  $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$  están fuera del círculo unitario, (ver [Brockwell and Davis, 2006](#)). En este caso el proceso tiene memoria corta y se dicen que son procesos integrados de orden cero,  $I(0)$ . Si el polinomio  $\phi(z)$  tiene ceros sobre el círculo unitario entonces se tiene que el proceso es no estacionario y contiene una componente de caminata aleatoria. Esto hace que el proceso tenga una dependencia fuerte con su pasado, este problema es conocido como *raíz unitaria*, (ver [Maddala and Kim, 1998](#)). Un ejemplo de raíz unitaria es la caminata aleatoria,  $X_t = X_{t-1} + w_t$ , en este caso si se toma la primera diferencia,  $\Delta X_t := X_t - X_{t-1}$ , se obtiene que  $\Delta X_t = w_t$  es estacionario. Estos procesos también se conocen como procesos integrados de orden uno,  $I(1)$ . En general un proceso es integrado de orden  $d$ ,  $I(d)$ , si  $\Delta^d X_t$  es estacionario con  $d \in \mathbb{N}$  el entero más pequeño que hace que el proceso sea estacionario. El parámetro de integración  $d$  también puede ser fraccionario, cuando ese es el caso, se dice que es un proceso integrado fraccionario, ARFIMA( $p, d, q$ ). Esto permite construir procesos con memoria larga a partir de

procesos estacionarios, definiendo  $X_t = \Delta^{-d}Y_t$ , con  $Y_t$  un proceso estacionario,  $\Delta^{-d} := (1 - L)^{-d}$ , donde  $L$  es el operador de retraso y

$$(1 - z)^{-d} = \sum_{j=0}^{\infty} (-1)^d \binom{d}{j} z^j, \quad |z| < 1.$$

En consecuencia,

$$X_t = \Delta^{-d}Y_t = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} Y_{t-j}. \quad (1.5)$$

Si  $0 < d < 1/2$  se obtiene un proceso estacionario con memoria larga, en el sentido de que la covarianza de  $X_t$  decrece lentamente satisfaciendo (1.2). Y claramente, también podemos obtener un proceso de memoria larga en el modelo ARMA sin la integración (1.5), reemplazando el ruido blanco  $w_t$  por un FGN, esto es  $X_t = \phi_1 + \dots + \phi_p X_{t-p} + \varepsilon^H$ , donde  $\{\varepsilon^H\}$  es FGN. Existe una relación entre los procesos integrados fraccionarios y FGN, la relación es  $d = H - 1/2$ . La clasificación de los procesos usando el concepto de procesos integrados  $I(d)$ , es válida en un contexto de procesos lineales o dependencias lineales, ya que está basada en la medida de covarianza y (1.5) lo que no resulta adecuado para procesos no lineales.

En la actualidad es de gran importancia el estudio de series de tiempo no lineales. Se ha evidenciado que muchos fenómenos tienen una dinámica no lineal y cada vez es más común hablar de procesos no lineales. Por ejemplo, en econometría la dinámica en el mercado es no lineal, ver [Dufrénot and Mignon \(2002\)](#), [Teräsvirta et al. \(2010\)](#), [Tong \(2015\)](#). Por lo tanto, es necesario tener herramientas para procesos no lineales, en particular para el estudio de memoria larga, ya que esto también da pie a conceptos como, cointegración no lineal.

Algunos ejemplos de series de tiempo no lineales son: modelos GARCH, modelos Bilineales, modelos SETAR (ver Sección 3.5) y en general modelos de la forma  $X_t = h(X_{-1}, \dots, X_{t-p}) + w_t$  con  $h$  no lineal. Los modelos no lineales son capaces de describir dinámicas complejas, pueden reproducir asimetrías y fenómenos de saltos, algo que los modelos lineales no pueden hacer.

Para el estudio de memoria larga en procesos no lineales se requieren medidas de dependencia que no estén basadas de la linealidad del proceso. En el contexto de series de tiempo no lineales se ha definido la memoria larga usando conceptos de  $\alpha$ -mixing y ergodicidad. Sea

$$\alpha_X(n) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \mathcal{F}_{-\infty}^k, B \in \mathcal{F}_{k+n}^{\infty}\},$$

donde  $\mathcal{F}_a^b$  es el  $\sigma$ -álgebra generado por  $(X_a, X_{a+1}, \dots, X_b)$ . El proceso  $X = \{X_n\}$  es llamado  $\alpha$ -mixing si  $\alpha_X(n) \rightarrow 0$  cuando  $n \rightarrow \infty$ . Recordemos que  $\alpha$ -mixing implica



---

ergodicidad. Entonces, un proceso que no cumple con la condición de  $\alpha$ -mixing y que es ergódico, se dice que tiene memoria larga. En la práctica es difícil hacer una prueba sobre si el proceso es  $\alpha$ -mixing. Para esto, muchas veces se usa el estadístico  $R/S$ , que como ya hemos comentado, no necesariamente tiene un buen desempeño en procesos estocásticos en general.

También se ha usado el enfoque de la evolución de la varianza que se mencionó en el caso de los modelos ARMA. En [Berenguer-Rico and Gonzalo \(2014\)](#) se propone el concepto del *índice de sumabilidad*, con el objeto de clasificar a los procesos no lineales usando la misma idea que clasificación  $I(d)$  en el caso lineal y con ello se define cointegración no lineal, ver Sección 2.1. Otra medida que se ha usado en el contexto no lineal es la información mutua (definida en Sección 2.1). La información mutua es una medida de dependencia que no está basada en la linealidad del proceso pero requiere el conocimiento de las distribuciones finito dimensionales, y como es de imaginarse, la estimación de estas distribuciones no es simple cuando el proceso es no lineal.

En la primera parte de esta tesis estudiaremos el concepto de memoria larga en series de tiempo no lineales, consideraremos la información mutua como una medida de dependencia, y la estimación estará basada en la cópula para evadir el problema de la estimación de las distribuciones. Esta medida es usada similarmente a la medida de covarianza en el caso lineal. Se estudia el comportamiento de la información mutua a pares, es decir, la información mutua de  $(X_t, X_{t+h})$ , con  $h = 0, 1, \dots$ . A partir de esto se propone una prueba de hipótesis usando un enfoque funcional y la prueba envolvente. Los resultados serán comparados con los valores correspondientes del estadístico  $R/S$ .

Por otro lado, el Análisis de Datos Funcionales (FDA) se ha convertido en una herramienta usual, una parte es debido a que en la actualidad se cuenta con la facilidad de recolectar datos, en forma casi continua. Por ejemplo, datos financieros (observaciones cada hora o minuto), datos relacionados con estudios del cerebro (las observaciones pueden ser por cada milisegundos), datos del clima (observaciones cada hora o minuto), las mismas densidades espectrales obtenidas de una serie de tiempo, etc..

FDA fue popularizado por los trabajos de [Ramsay \(1982\)](#) y [Ramsay and Dalzell \(1991\)](#). En [Ramsay and Silverman \(2005\)](#) se puede encontrar un estudio detallado de FDA, presentando la mayoría de las herramientas para el análisis. FDA supone que los datos son funciones continuas  $\{X_i(t), t \in D \subseteq \mathbb{R}\}$ , así, intrínsecamente son datos de dimensión infinita que contiene vasta información para el análisis de datos.

La estructura de dimensión infinita hace FDA sea interesante desde el punto de vista matemático, estadístico y computacional.

La teoría de FDA no se limita solamente sobre conjuntos de funciones definidos en  $\mathbb{R}$ , puede ser aplicado sobre espacios de Hilbert en general, por ejemplo las observaciones pueden tener una componente espacial, pueden ser superficies, etc..

Existen dos enfoques principales para el análisis de datos funcionales. El primero es asumir que los datos son observados en un subconjunto denso de  $D$  sin errores de medición, y usar el método de interpolación para aproximar la función. La desventaja es el costo computacional debido a la alta dimensionalidad que se involucra. El segundo enfoque es asumir que los datos observados están contaminados por un ruido aleatorio, que podría ser errores de medición (lo que tendría mucho sentido) y que la curva subyacente es una curva diferenciable, es decir  $X_{\text{obs}}(t) = X(t) + e(t)$ , con  $X(t)$  diferenciable. Así, la inferencia estadística está basada en la curva subyacente  $X(t)$ , asumiendo alguna base de funciones finita  $\{\phi_i(t)\}_1^J$  tal que

$$X(t) \approx \sum_{k=1}^J c_{n,k} \phi_k(t).$$

La ventaja de este enfoque es la reducción de dimensionalidad y adicionalmente es más fácil hacer cumplir la propiedad de diferenciability en  $X(t)$ , considerando algún tipo de penalización, ver Sección 2.4 para ajuste de curvas.

Muchos de los datos funcionales pueden considerarse como observaciones independientes, por ejemplo las curvas que representan el crecimiento de las personas. Pero también puede ser que el supuesto de independencia no sea razonable, por ejemplo, en fenómenos relacionados con el clima se tienen periodos largos de observación, donde las observaciones son cada minuto (o por hora) durante varios años, en este tipo de datos es natural observar una periodicidad por día, lo cuál hace natural pensar en datos funcionales, donde cada dato funcional es la curva que representa los datos diarios. En este caso es difícil pensar que las curvas sean independientes. Esta idea nos conduce a una serie de tiempo de curvas y se conoce como *series de tiempo funcional*.

En el ejemplo anterior, la idea es obtener series de tiempo funcionales a través de procesos estocásticos  $\{\zeta_t\}$ , dividiendo el proceso por un periodo  $d$  ( $d$  puede ser día, semana, mes) específico, obteniendo así los datos funcionales  $X_n(t) = \zeta_t \mathbf{1}_{\{t \in [n\delta, (n+1)\delta)\}}$ . En general, no tienen por que provenir de un proceso estocástico, como los espectros que pueden ser curvas por separado pero con dependencia en el tiempo.

Para el análisis de series de tiempo funcional, no se requiere que las observaciones en cada periodo de tiempo tengan una estructura de dependencia específica, es más,

$X_i(t_1), \dots, X_i(t_k)$  puede ser no estacionario o de memoria larga. El supuesto de estacionariedad es sobre el periodo donde está definido el dato funcional y no dentro del periodo.

Uno de los modelos más estudiados en series de tiempo funcionales son los modelos autoregresivos funcionales de orden  $p$ , FAR( $p$ ), definidos como

$$X_n(t) = \sum_{j=1}^p \Phi_j(X_{n-j})(t) + W_n(t), \quad (1.6)$$

donde  $\Phi_i$  son operadores lineales y  $\{W_n\}$  ruido blanco funcional, ver Sección 4.1. En la práctica  $p$  toma valores pequeños tales como  $p = 1$  ó  $2$  debido a la complejidad de la estructura del modelo, (ver Kokoszka and Reimherr, 2013). Así, muchos estudios se han centrado en el caso cuando  $p = 1$ . En esta tesis asumimos  $p = 1$  y también que la serie funcional sea estacionaria, ver Sección 4.1. En Bosq (2000) se pueden encontrar las condiciones para la estacionariedad del FAR(1), así como el estimador del operador de coeficiente  $\Phi$  y sus propiedades, y en general la teoría de procesos lineales funcionales tanto para espacios de Hilbert como espacios de Banach.

Otro problema común con datos reales, es la presencia de datos atípicos. La presencia de este tipo de datos afecta seriamente el análisis estadístico y es por eso que ha recibido una gran atención en la literatura. Existen varios artículos de investigación que tratan sobre detección de datos atípicos en datos funcionales, como Hyndman and Shang (2010); Sun and Genton (2011); Dai and Genton (2017); Hubert et al. (2015) por mencionar algunos.

Por otro lado, ha habido algunas propuestas de estimadores robustos en el contexto de series de tiempo funcional. En particular para las autofunciones, ver Gervini (2008); Bali et al. (2011); Lee et al. (2013) entre otros. Sin embargo, cuando se trata de estimadores robustos para el operador de coeficiente en el modelo autoregresivo funcional, no encontramos muchas referencias. Una manera es usar las autofunciones robustas para aproximar el operador de covarianza y obtener el estimador de la ecuación normal, similar al caso real. En esta tesis se propone un estimador robusto para el modelo autoregresivo funcional usando el método de mínimos cuadrados ponderados y asumiendo una base finita para los datos funcionales.

Esta tesis está organizada de la siguiente manera. En el Capítulo 2 se introducen conceptos generales que serán usados en los capítulos posteriores, tales como memoria larga en procesos, variables aleatorias funcionales y profundidades funcionales. En el Capítulo 3 se desarrolla un método para detección de persistencia usando la

información mutua como una medida de dependencia. Se detalla la estimación de la información mutua combinando el concepto de cópula y enfoque funcional. La contribución principal que se presenta en este capítulo es la propuesta de una prueba de hipótesis para detección de memoria larga en las series de tiempo lineales y no lineales, considerando la entropía del proceso y enfoque funcional. Se presenta estudios de simulación bajo diferentes esquemas con resultados satisfactorios tanto para procesos lineales y no lineales. Estos resultados también son comparados con el estadístico  $R/S$ . En el Capítulo 4 se detalla la propuesta de un estimador robusto para el modelo autoregresivo funcional de orden uno. Usando la noción de profundidad se define las ponderaciones en las observaciones. El estimador propuesto resulta ser un estimador de mínimos cuadrados ponderados, mostrando un excelente desempeño, relativo a lo que se describe en la literatura actual. También, se muestra que la función de influencia correspondiente es acotada. La contribución principal de este capítulo es el uso de la profundidad en la ponderación de los datos funcionales para la obtención del estimador robusto, con función de influencia acotada. En los estudios de simulación se comparan los resultados con el estimador obtenido de autofunciones robustas propuesta en [Hyndman and Ullah \(2007\)](#).

# Capítulo 2

## Conceptos Preliminares

En este capítulo se describen conceptos generales que serán de utilidad a lo largo de esta tesis. En la Sección 2.1 se introducen conceptos de memoria larga en un contexto de procesos lineales y no lineales. En la Sección 2.2, se describe el concepto de variables aleatorias en un espacio de funciones, así como la relación de estas variables aleatorias con procesos estocásticos continuos. En la Sección 2.3, se presenta la noción de profundidad con datos funcionales, como una forma de ordenar curvas. Finalmente en la Sección 2.4 se hace una revisión breve de ajuste de curvas, ya que este concepto es esencial en el análisis de datos funcionales, pues en la práctica, sólo se puede observar una muestra finita de la curva, por lo que es necesario estimar la curva subyacente para el estudio estadístico.

### 2.1. Series de tiempo y memoria larga

El concepto de memoria larga (o persistencia) en procesos no lineales desarrollada en esta tesis, surge a partir del problema de cointegración no lineal. Para desarrollar la teoría de cointegración no lineal, es necesario saber de la estructura de dependencia de una serie de tiempo en un contexto de procesos estocásticos no necesariamente lineales (ver Aparicio and Escribano (1998), Dufrénot and Mignon (2002) y Berenguer-Rico and Gonzalo (2014)), por lo que es necesario una herramienta que proporcione información sobre la estructura de dependencia en un contexto más general. Recordemos la definición de un proceso estocástico.

**Definición 2.1** *Sea  $T \subseteq \mathbb{R}^+$ , un proceso estocástico  $\zeta = \{\zeta_t, t \in T\}$  es una familia de variables aleatorias que toma valores en  $\mathbb{R}$  definido en un espacio de probabilidad  $(\Omega, \mathcal{F}, \mathbb{P})$ .*

El conjunto  $T$  es interpretado como el tiempo. En la práctica no es común observar el proceso estocástico bajo diferentes escenarios,  $w \in \Omega$ , en muchas ocasiones se observa para un solo evento  $w \in \Omega$ , conocido como trayectoria. A continuación escribimos la definición formal.

**Definición 2.2** *La trayectoria de un proceso  $\zeta = \{\zeta_t, t \in T\}$  se define como la función  $t \mapsto \zeta_t(\omega)$  para cada  $\omega \in \Omega$ .*

Dado que solo se puede observar una trayectoria, es necesario hacer supuestos sobre el proceso para poder hacer inferencia estadística, estos supuestos son la estacionariedad y la ergodicidad.

**Definición 2.3** *Un proceso estocástico  $\zeta = \{\zeta_t, t \in T\}$  se dice fuertemente estacionario si sus distribuciones finito-dimensionales son invariantes bajo traslaciones. Esto es, sea  $t_1, \dots, t_n$  un conjunto de índices y  $h > 0$ , entonces*

$$(\zeta_{t_1}, \dots, \zeta_{t_n}) \stackrel{L}{=} (\zeta_{t_1+h}, \dots, \zeta_{t_n+h}),$$

donde " $\stackrel{L}{=}$ " significa igualdad en distribución.

Otra definición de estacionariedad es usando los primeros dos momentos del proceso.

**Definición 2.4** *Un proceso estocástico  $\zeta = \{\zeta_t, t \in T\}$  se dice débilmente estacionario o estacionario de segundo orden si*

1.  $\mathbb{E}(\zeta_t) = \mu$ , esto es, la media del proceso no depende del tiempo.
2.  $\text{Cov}(\zeta_t, \zeta_s) = \gamma(s - t)$ , esto es, la función de covarianza sólo depende de la diferencia entre los tiempos  $t$  y  $s$ .

Para el concepto de ergodicidad ver [Samorodnitsky \(2016\)](#).

Observemos que fuertemente estacionario implica débilmente estacionario si existe el segundo momento del proceso. En general no se tiene que estacionariedad débil implica estacionariedad fuerte. Si el proceso es un proceso Gaussiano, entonces se tiene esta implicación, ya que los procesos Gaussianos están caracterizados por los dos primeros momentos.

Para procesos con segundo momento finito, la función de correlación (ACF) de  $\zeta$  es definida como

$$\rho(s, t) = \frac{\text{Cov}(\zeta_t, \zeta_s)}{\sqrt{\text{Var}(\zeta_s)}\sqrt{\text{Var}(\zeta_t)}}. \quad (2.1)$$

En consecuencia, si el proceso es débilmente estacionario,  $\rho(s, t) = \rho(s - t)$ . Un caso particular de procesos estocásticos es cuando el tiempo  $T$  es discreto.

**Definición 2.5** Sea  $\zeta = \{\zeta_t, t \in T\}$  un proceso estocástico tal que  $T = \mathbb{Z}$ . Una serie de tiempo se define como una trayectoria parcial de  $\zeta$ , i.e.,

$$\{\zeta_1, \zeta_2, \dots, \zeta_N\}.$$

Se usará la notación de  $\{X_1, X_2, \dots, X_N\}$  para referirnos a una serie de tiempo.

Una serie de tiempo (proceso) se dice que es lineal si se puede expresar como una combinación lineal de variables aleatorias independientes  $\{\varepsilon_t\}$

$$X_t = \mu + \sum_{i=-\infty}^{\infty} \phi_i \varepsilon_{t-i}, \quad \{\varepsilon_t\} \sim \text{i.i.d.}(0, 1), \quad t \in T.$$

Se dice que una serie de tiempo lineal es causal si  $\phi_i = 0$  para toda  $i < 0$ .

En la literatura se pueden encontrar diferentes definiciones de procesos de memoria larga (memoria corta). En general, tiene que ver con el grado de dependencia del proceso con su pasado. Consideremos por un instante que  $\rho(h)$  es la medida de dependencia, entonces se puede estudiar que tan rápido la función de correlación

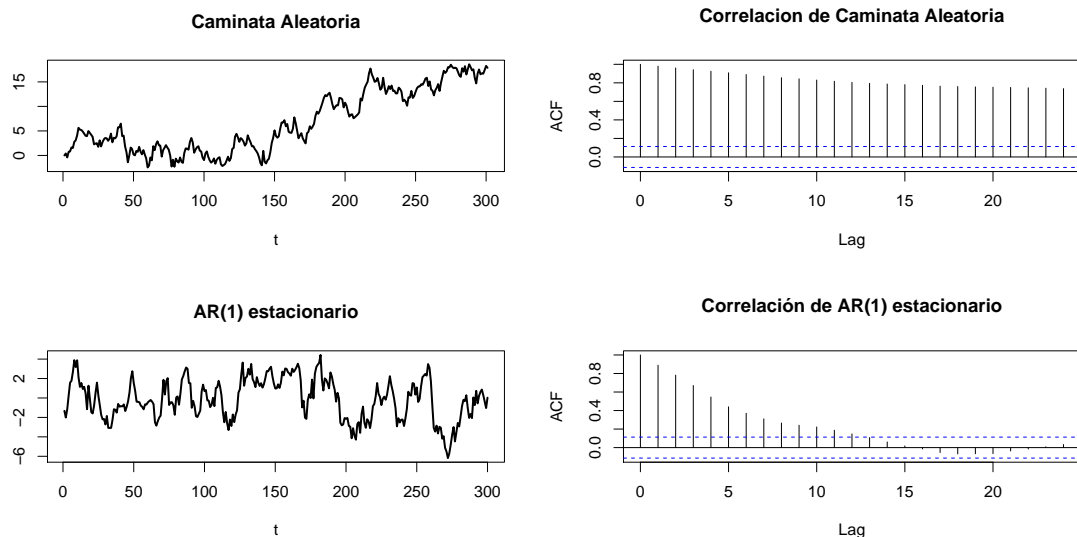


Figura 2.1: Simulación de una serie de tiempo no estacionaria (caminata aleatoria) y una serie de tiempo estacionaria (AR(1)) con sus correspondientes funciones de autocorrelación.

$\rho(h)$  decrece cuando el parámetro de retraso  $h$  tiende a infinito. Si la función de correlación decrece “rápido” el proceso correspondiente se dice que tiene memoria corta y si esto decrece “lento” entonces se dice que el proceso correspondiente es de memoria larga.

Por ejemplo, consideremos un proceso autorregresivo de orden uno (AR(1)) definido por  $X_t = \theta X_{t-1} + \varepsilon_t$ . Si  $\theta = 1$  se tiene que  $X_t$  una caminata aleatoria y no es estacionario. En este caso  $X_t$  es un proceso de memoria larga ya que depende fuertemente de su pasado y su función de correlación decrece “lentamente”, de hecho  $\rho(t, t-h) = \sqrt{(t-h)/t}$ . Si  $|\theta| < 1$ , entonces  $X_t$  es un proceso estacionario y es de memoria corta con  $\rho(h) = \theta^h$ . La figura 2.1 muestra las series de tiempo simuladas de estos dos procesos, caminata aleatoria y AR(1) estacionario con  $\theta = 0.9$ , con sus respectivas funciones de correlación. En ella se puede observar los diferentes comportamientos de la estructura de correlación.

A continuación escribimos la definición formal de memoria larga que aparece con más frecuencia en la literatura para procesos lineales.

**Definición 2.6** *Sea  $\{X_t\}$  una serie de tiempo. Se dice que  $\{X_t\}$  tiene memoria larga si cumple algunas de las siguientes condiciones:*

1. *La función de correlación de la serie de tiempo, definida en 2.1, no es absolutamente sumable,*

$$\sum_{h=0}^{\infty} |\rho(h)| = \infty.$$

2. *La función de correlación de la serie de tiempo es tal que*

$$\rho(n) = L(n)n^{-d},$$

*con  $0 < d < 1$  y  $L$  es una función de variación lenta en el infinito.*

3. *La serie de tiempo satisface*

$$\text{Var}(X_1 + \dots + X_n) = L(n)n^{2-d},$$

*con  $0 < d < 1$  y  $L$  es una función de variación lenta en el infinito.*

El ruido Gaussiano fraccionario con índice  $H$ , definido en el Capítulo 1, satisface las propiedades de la definición anterior con  $1/2 < H < 1$ . Cabe mencionar que en general las condiciones anteriores no son equivalentes. Recientemente [Berenguer-Rico and Gonzalo \(2014\)](#) propusieron el concepto de índice de sumabilidad que está en la misma línea que el punto 3 de la definición anterior y es válido en el



contexto de procesos no lineales, en la Sección 3.1 se presentan más detalles sobre este idea. También podemos encontrar definiciones en el espacio de frecuencias, ver [Samorodnitsky \(2016\)](#).

De igual manera, en el contexto de procesos estocásticos no lineales, se han propuesto varias definiciones de memoria larga. A continuación mencionamos algunas de ellas. Denotemos por  $\mathcal{F}_t$  el  $\sigma$ -álgebra generado por  $\{X_{t-h}, h \geq 0\}$ .

**Definición 2.7** *Sea  $\{X_t\}$  una serie de tiempo.*

1.  $\{X_t\}$  tiene memoria larga en media si la esperanza condicional de la variable aleatoria al tiempo  $t$  dada la información al tiempo  $t - h$  no aproxima a la esperanza sin el condicionamiento, cuando  $h \rightarrow \infty$

$$\lim_{h \rightarrow \infty} \|\mathbb{E}(X_t | \mathcal{F}_{t-h}) - \mathbb{E}(X_t)\| \neq 0,$$

de lo contrario, si el límite es cero, se dice que la serie de tiempo tiene memoria corta.

2. Si  $\{X_t\}$  tiene momentos de orden  $k$  finito, se dice que tiene memoria larga si

$$\sum_{\tau_1=-\infty}^{\infty} \dots \sum_{\tau_k=-\infty}^{\infty} |C_k(\tau_1, \dots, \tau_k)| = \infty, \quad k = 2, 3, \dots$$

donde  $C_k(\tau_1, \dots, \tau_k)$  es el cumulante de orden  $k$  de la serie de tiempo. Si la suma es finita, se dice que tiene memoria corta.

Por otro lado, existen varias formas de definir procesos no lineales, por ejemplo los procesos TAR o SETAR que son procesos definidos en varios regímenes, y en cada régimen es un proceso lineal. También se pueden obtener proceso no lineales a partir de una transformación no lineal de un proceso lineal. Otro ejemplo usual en la literatura de procesos no lineales son los modelos bilineales. Un modelo bilineal,  $B(1, 0, 1, 1)$  es definido como,

$$X_t = \phi_1 X_{t-1} + \phi_{11} X_{t-1} \varepsilon_{t-1} + \varepsilon_t.$$

Una forma de obtener memoria larga en este modelo es usar la misma idea que en el caso lineal, esto es a través del ruido blanco  $\varepsilon_t$ , es decir, se considera ruido blanco fraccionario  $w^H$  o integrado  $\varepsilon_t = (1 - L)^{-d} w_t$ , con  $|d| < 1/2$  y  $w_t \sim N(0, \sigma_w^2)$ , donde  $L$  es el operador de retraso.

Para leer más de estas definiciones, ver [Dufrénot and Mignon \(2002\)](#), [Samorodnitsky \(2016\)](#) y [Granger and Hallman \(1991\)](#). En esta tesis usaremos la siguiente definición de un proceso con memoria larga.

**Definición 2.8** Sea  $X_t$  una serie de tiempo y sea  $\delta(h, t) := \delta(X_t, X_{t-h})$  una medida no negativa de dependencia, que no está basada en la linealidad del proceso, tal que  $\delta(X_t, X_t) = 1$  y  $\delta(X_t, X_{t-h})$  decrece de acuerdo a  $h \rightarrow \infty$ . La serie de tiempo  $X_t$  se dice que tiene memoria larga si para toda  $t$ ,

$$\lim_{T \rightarrow \infty} \sum_{h=1}^T \delta(h, t) = \infty. \quad (2.2)$$

Y diremos que es de memoria corta si la suma es finita.

La definición anterior es similar al que se propone en [Aparicio and Escribano \(1998\)](#). La medida  $\delta$  se detallará en el capítulo 3 y también se propone una prueba de hipótesis basada en la entropía.

## 2.2. Variables aleatorias en espacio de funciones

Otra clase de variables aleatorias que son de interés en esta tesis son las variables aleatorias funcionales. A continuación presentamos los conceptos básicos para el análisis de este tipo de variables. Se denotará por  $\mathbb{B}(E)$  la  $\sigma$ -álgebra de Borel de  $E$ . Entonces, la definición general de una variable aleatoria  $E$ -valuada definida sobre  $(\Omega, \mathcal{F})$  es la siguiente.

**Definición 2.9** Sean  $(\Omega, \mathcal{F})$  y  $(E, \mathcal{E})$  espacios medibles. Una función  $\mathcal{F}/\mathcal{E}$ -medible es toda función

$$X : \Omega \rightarrow E,$$

tal que  $X^{-1}(A) \in \mathcal{F}$  para toda  $A \in \mathcal{E}$ .

Una variable aleatoria funcional  $X$  es una función  $\mathcal{F}/\mathcal{E}$ -medible donde  $\mathcal{E}$  es el  $\sigma$ -álgebra de algún espacio de funciones  $E$ . En este escrito  $E$  será un espacio de Hilbert de las funciones reales medibles de cuadrado integrable definidas en un subconjunto compacto  $D \subset \mathbb{R}$ , siendo más específico, consideraremos el espacio de Hilbert

$$\mathbb{H} = L^2(D, \mathcal{H}, \lambda),$$

con producto interno,

$$\langle f, g \rangle = \int_D f(t)g(t)dt,$$

donde  $\mathcal{H} = \mathbb{B}(D)$  y  $\lambda$  la medida de Lebesgue. Se denotará por  $\|\cdot\|_{\mathbb{H}}$  la norma correspondiente en  $\mathbb{H}$ . Así, diremos que  $X$  es una variable aleatoria funcional si

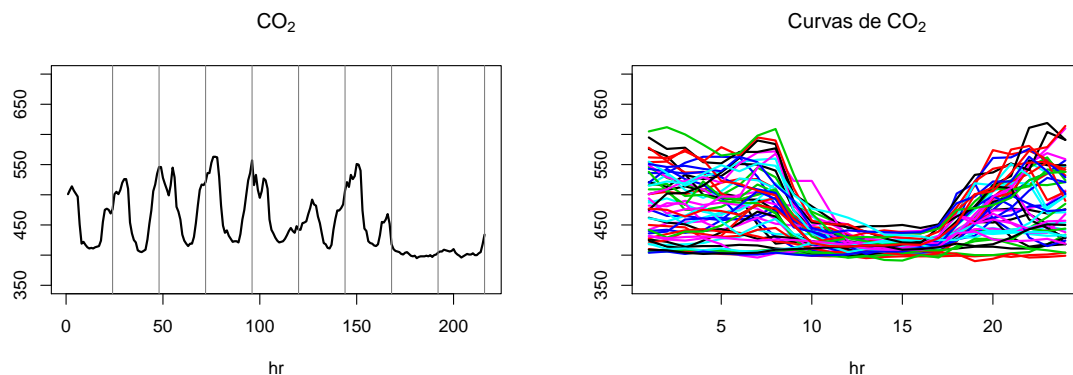


Figura 2.2: Ejemplo de una variable aleatoria funcional. La figura del lado izquierdo es una serie de tiempo que representa las mediciones por cada hora de  $\text{CO}_2$  y las líneas verticales indican el final del día correspondiente. La figura del lado derecho representa las curvas diarias de las series de tiempo de  $\text{CO}_2$ .

es una función  $\mathcal{F}/\mathcal{H}$ -medible. Análogamente al caso de variables aleatorias en  $\mathbb{R}$ , si existe dependencia en estas variables a través del tiempo de observación, entonces se habla de una serie de tiempo funcional.

Una forma de obtener series de tiempo funcionales es a través de procesos estocásticos continuos. El siguiente lema nos dice cuándo esto es posible.

**Lema 2.1** *Si  $B$  es un espacio de Banach, entonces  $X : (\Omega, \mathcal{F}) \rightarrow (B, \mathbb{B}(B))$  es una variable aleatoria que toma valores en  $B$  si y solo si  $x^*(X)$  es una variable aleatoria que toma valores en los reales para toda  $x^*$  que esté en el espacio dual de  $B$ .*

Por lo tanto, una serie de tiempo funcional puede ser construida de la siguiente manera. Sea  $Y = (Y_t, 0 \leq t \leq 1)$  un proceso estocástico que toma valores en  $\mathbb{R}$  con trayectoria continua. Entonces  $Y$  define una variable aleatoria funcional que toma valores en  $\mathbb{H}$  con  $D = [0, 1]$ , ya que  $\int_0^1 Y_t x^*(t) dt$  es una variable aleatoria para toda  $x^* \in \mathbb{H}$ , pues el espacio dual de  $\mathbb{H}$  es  $\mathbb{H}$ , (ver [Bosq, 2000](#), Cap. 1). La Figura 2.2 muestra un ejemplo de una variable aleatoria funcional obtenida de un proceso estocástico, del lado izquierdo se muestra la serie de tiempo de 9 días consecutivos con mediciones en cada hora, las líneas verticales representan el día completo (24 horas). La figura del lado derecha representa los datos funcionales obtenidos de los datos diarios (24 observaciones). Observemos que en este caso, los datos funcionales no son independientes, pues la curva del día actual depende del día anterior. Así, se

tiene que los datos funciones tienen dependencia en el tiempo (días), es decir, son una serie de tiempo funcional.

Denotaremos por  $\mathcal{B}_{\mathbb{H}}$  el conjunto de todos los operadores lineales de  $\mathbb{H}$  en  $\mathbb{H}$ , esto es,

$$\mathcal{B}_{\mathbb{H}} = \{\Phi : \mathbb{H} \rightarrow \mathbb{H}; \Phi \text{ es acotado y lineal}\}.$$

y llamaremos  $\mathbb{H}$ -variable aleatoria a la variable aleatoria que toma valores en  $\mathbb{H}$ , es decir,  $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{H}, \mathcal{H})$ . También, denotaremos por  $L_{\mathbb{H}}^p(\Omega, \mathcal{F}, \mathbb{P})$  el conjunto de  $\mathbb{H}$ -variables aleatorias con momento finito de orden  $p$ ,

$$L_{\mathbb{H}}^p(\Omega, \mathcal{F}, \mathbb{P}) = \{X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{H}, \mathcal{H}, \lambda); \mathbb{E}(\|X\|^p) < \infty\}.$$

Si  $X$  es una  $\mathbb{H}$ -variable aleatoria y  $\mathbb{E}(\|X\|) < \infty$ , entonces existe un elemento único  $\mu$  de  $\mathbb{H}$ , tal que

$$\mathbb{E}(\langle X, z \rangle) = \langle \mu, z \rangle, \quad (2.3)$$

para toda  $z \in \mathbb{H}$ . Este elemento es llamado el valor esperado de  $X$  y se denotará como  $\mathbb{E}(X)$ .

Ahora, si  $X \in L_{\mathbb{H}}^2(\Omega, \mathcal{F}, \mathbb{P})$  entonces el operador de covarianza  $\Gamma_X$  de  $X$  se define como un elemento de  $\mathcal{B}_{\mathbb{H}}$ ,

$$\Gamma_X(z) = \mathbb{E}(\langle X - \mu, z \rangle (X - \mu)), \quad (2.4)$$

donde  $z \in \mathbb{H}$ . Si  $Y$  es otra  $\mathbb{H}$ -variable aleatoria definida sobre el mismo espacio de probabilidad que  $X$ , entonces el operador de covarianza cruzada es definido como

$$\Gamma_{X,Y}(z) = \mathbb{E}(\langle X - \mathbb{E}(X), z \rangle (Y - \mathbb{E}(Y))), \quad z \in \mathbb{H}. \quad (2.5)$$

Sin pérdida de generalidad, supongamos  $\mathbb{E}(X) = 0$ , entonces

$$\begin{aligned} \Gamma_{X,Y}(z)(t) &= \mathbb{E}(\langle X, z \rangle Y(t)) \\ &= \int_D \mathbb{E}(X(s)Y(t))z(s)ds \\ &= \int_D \gamma_{X,Y}(s,t)z(s)ds, \end{aligned}$$

donde  $\gamma_{X,Y}(s,t) = \mathbb{E}(X(s)Y(t))$  es la función de covarianza entre  $X(s)$  y  $Y(t)$ .

Similar al análisis de datos multivariados, un objeto de interés son los autovalores y autofunciones del operador de covarianza. Las autofunciones  $v_j(t)$ ,  $j \geq 1$ , son definidas como funciones sobre  $D$  tales que,

$$\Gamma_X(v_j)(t) = \lambda_j v_j(t),$$

donde  $\lambda_j$  son los correspondientes autovalores. El conjunto de los autovalores  $\{\lambda_j\}$  es un conjunto numerable ya que el operador  $\Gamma_X$  es compacto.

Supongamos que se observan  $N$  funciones aleatorias,  $X_1(t), \dots, X_N(t)$ , independientes de la variable aleatoria funcional  $X(t)$  definidas en el espacio de Hilbert  $\mathbb{H}$ , la versión muestral de  $\mu$  en (2.3) es definida como,

$$\hat{\mu}_X(t) = \frac{1}{N} \sum_{i=1}^N X_i(t), \quad t \in D,$$

y el operador de covarianza muestral en (2.4) es definido como,

$$\hat{\Gamma}_X(z) = \frac{1}{N} \sum_{i=1}^N \langle X_i, z \rangle X_i, \quad z \in \mathbb{H},$$

o equivalentemente la función de covarianza muestral,

$$\hat{\gamma}_X(s, t) = \frac{1}{N} \sum_{i=1}^N X_i(s)X_i(t), \quad s, t \in D.$$

Se puede demostrar que  $\Gamma_X$  (al igual que  $\hat{\Gamma}_X$ ) es un operador de Hilbert-Schmidt, esto es, existe una base ortonormal  $\{v_i\}$  de  $\mathbb{H}$  tal que  $\Gamma_X(z) = \sum_{i=1}^{\infty} \lambda_i \langle v_i, z \rangle v_i$ . Tomando esto en cuenta, la norma del operador  $\Gamma_X$  se define como

$$\|\Gamma_X\|_S^2 := \sum_{i=1}^{\infty} \lambda_i^2.$$

El siguiente teorema garantiza la consistencia de los estimadores definidos anteriormente.

**Teorema 2.1** *Sea  $X$  una  $\mathbb{H}$ -variable aleatoria con  $\mathbb{E}(X) = \mu$  y operador de covarianza  $\Gamma$ .*

1. Si  $X \in L_{\mathbb{H}}^4$ , entonces  $\mathbb{E}(\|\hat{\Gamma} - \Gamma\|_S^2) = O(1/N)$ .
2. Si  $X \in L_{\mathbb{H}}^2$ , entonces  $\mathbb{E}(\|\hat{\mu} - \mu\|_{\mathbb{H}}^2) = O(1/N)$ .

Para leer más sobre variables aleatoria funcionales y operadores correspondientes, consultar [Bosq \(2000\)](#).

## 2.3. Ordenamiento de curvas basado en profundidades

Para hacer inferencia estadística, muchas veces se requiere ordenar los datos, por ejemplo en la estimación de la distribución acumulada o simplemente para calcular cierto cuantil. Lo anterior es fácil si tenemos observaciones en  $\mathbb{R}$ , caso contrario si las observaciones están en  $\mathbb{R}^d$ ,  $d \geq 2$ . En la literatura se han propuesto varios enfoques para ordenar datos multivariados o datos funcionales, uno de estos enfoques es el concepto de profundidad, ver [Zuo and Serfling \(2000\)](#). El concepto de profundidad es una medida de centralidad, donde las observaciones son ordenadas del “dato central” hacia el dato menos central, es decir, observaciones que están cerca del dato central se les asigna un mayor índice de profundidad, el dato que tiene un índice de profundidad cerca de cero se considera que es un dato extremo o atípico. Algunas aplicaciones de este concepto son las siguientes: definir  $p\%$ -región central, estimar la mediana o detectar datos atípicos.

**Definición 2.10** *La profundidad de  $x \in \mathbb{H}$  asociada a la ley  $\mathbb{P}$  es definida como*

$$\begin{aligned} \mathcal{D} : \mathbb{H} &\rightarrow [0, 1] \\ x &\mapsto D(x, \mathbb{P}) \end{aligned}$$

tal que  $\mathcal{D}$  cumple las siguientes condiciones:

1. *Invarianza afín (Affine invariance):  $\mathcal{D}(Ax + b, \mathbb{P}_{X+b}) = \mathcal{D}(x, \mathbb{P}_X)$ , para todo operador  $A$  lineal y biyectivo y para toda  $b \in H$ .*
2. *Alcanza el máximo en el dato más central: Si  $X$  es una variable aleatoria con distribución  $\mathbb{P}$  simétrica y centrada en  $z$ , entonces*

$$\mathcal{D}(z, \mathbb{P}) = \sup_{x \in H} D(x, \mathbb{P}).$$

3. *Se anula en el infinito:  $\mathcal{D}(x, \mathbb{P}) \rightarrow 0$  cuando  $\|x\| \rightarrow \infty$ .*
4. *Monótona con respecto al punto más profundo: Si  $z$  es el más profundo con respecto a  $\mathbb{P}$ , entonces*

$$\mathcal{D}(x, \mathbb{P}) \leq \mathcal{D}(z + \alpha(x - z), \mathbb{P}),$$

para toda  $\alpha \in [0, 1]$ .

Se denotará por  $d_n(x) = \mathcal{D}(x, \hat{\mathbb{P}}_n)$  la versión muestral de la profundidad  $\mathcal{D}(x, \mathbb{P})$ .

Existen discusiones si las propiedades enlistadas en la definición anterior son necesarias, lo que si se está de acuerdo es que son deseables. Por ejemplo, la propiedad 1 podría ser suplida por invarianza bajo transformaciones ortogonales. La propiedad 3 puede ser omitida mientras se cumpla la propiedad 4. Las propiedades más importante son la 2 (aunque no tanto si la distribución correspondiente es multimodal), la 4 y adicionalmente la propiedad de consistencia,  $\mathcal{D}(x, \hat{\mathbb{P}}_n) \rightarrow \mathcal{D}(x, \mathbb{P})$ . En general dependerá del contexto del problema y de lo que se quiere resolver. Para leer acerca de definición y discusión del concepto de profundidad se pueden consultar Nieto-Reyes and Battey (2016), Long and Huang (2016), Serfling and Wijesuriya (2017).

En el caso de variables aleatorias funcionales existen varias definiciones de profundidad. A continuación se presentan dos de ellas: profundidad de banda modificada y profundidad de proyección. Estas serán de utilidad para la metodología presentada en los Capitulo 3 y 4.

### 2.3.1. Profundidad de Banda Modificada

En López-Pintado and Romo (2009) se propone el concepto de profundidad de banda y profundidad de banda modificada (MB y MBD, por sus siglas en inglés), usando un enfoque basado en gráficos. A continuación detallamos la idea.

**Definición 2.11** Sea  $X(s), s \in D$  una  $\mathbb{H}$ -variable aleatoria y  $\{X_1(s), \dots, X_N(s)\}$  copias i.i.d. de  $X(s)$ . La profundidad MBD de  $x(s) \in \mathbb{H}$  relativa a  $X(s)$  es definida como

$$\text{MBD}(x, X) = \sum_{j=2}^N \mathbb{E} \left( \lambda \left( s \in D : \min_{i=1, \dots, j} X_i(s) \leq x(s) \leq \max_{i=1, \dots, j} X_i(s) \right) \right),$$

donde  $\lambda$  es la medida de Lebesgue en  $D$ .

Consideremos una muestra de curvas  $\mathbf{X} = \{x_1(s), \dots, x_n(s), s \in D\}$  en  $\mathbb{H}$ . La profundidad de banda modificada consiste en formar bandas en  $\mathbb{R}^2$  determinadas por subconjuntos de curvas  $\{x_{i_1}, \dots, x_{i_j}\}$ , para alguna  $j \in \{2, \dots, n\}$  fija, esto es

$$B(x_{i_1}, \dots, x_{i_j}) = \left\{ (t, y) : t \in D, \min_{i=1, \dots, j} x_i(t) \leq y \leq \max_{i=1, \dots, j} x_i(t) \right\}.$$

Para  $x(s) \in \mathbf{X}$ , el conjunto  $t \in D$  con  $x(t)$  en la banda  $B(x_{i_1}, \dots, x_{i_j})$  está dado por

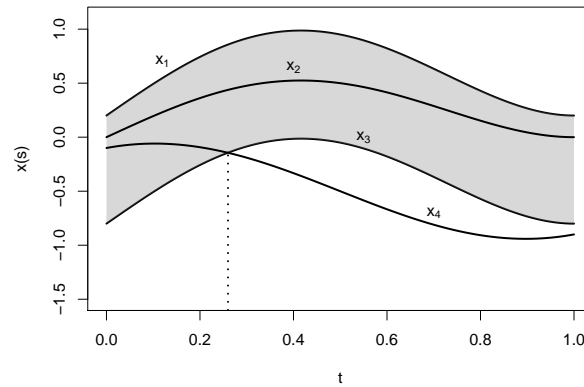


Figura 2.3: Ejemplo del concepto de profundidad de banda modificada. La curva  $x_2$  está dentro de la banda y la curva  $x_4$  solo pasa un tiempo menor a .3 en dicha banda.

$$A(x; x_{i_1}, \dots, x_{i_j}) = \left\{ t \in D : \min_{r=i_1, \dots, i_j} x_r(t) \leq x(t) \leq \max_{r=i_1, \dots, i_j} x_r(t) \right\}. \quad (2.6)$$

Luego, usando la medida de Lebesgue  $\lambda$  en  $D$ , se tiene que la proporción del tiempo que  $x$  pertenece a la banda es  $\lambda_r(A(x; x_{i_1}, \dots, x_{i_j})) = \lambda(A(x; x_{i_1}, \dots, x_{i_j}))/\lambda(A)$ . Y en consecuencia, la profundidad de la curva  $x$  en el conjunto  $\mathbf{X}$  se define como el promedio de las proporciones sobre todas los posibles subconjuntos de  $x_{i_1}, \dots, x_{i_j}$ ,

$$\text{MBD}_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < \dots < i_j \leq n} \lambda_r\{A(x; x_{i_1}, \dots, x_{i_j})\}.$$

La figura 2.3 muestra la idea general de este concepto, usando cuatro curvas y  $j = 2$ . En ella se observa que la banda está delimitada por las curvas  $x_1$  y  $x_3$ , también que la curva  $x_2$  queda completamente contenida dentro de la banda a diferencia de la curva  $x_4$ .

La profundidad MBD cumple las propiedades de ser invariable bajo transformación, invariable bajo traslaciones, se anula en el infinito, y es consistente, ver [López-Pintado and Romo \(2009\)](#).



### 2.3.2. Medida de cercanía de Stahel-Donoho

Otro concepto que está relacionado con la profundidad es el concepto de la medida de cercanía (outlyingness measure),  $o(x, \mathbb{P}) : \mathbb{H} \rightarrow [0, \infty)$ . Esta medida se define similarmente a la función de profundidad. La relación entre la profundidad y la medida de cercanía es la siguiente ecuación,

$$D(x, \mathbb{P}) = \frac{1}{1 + o(x, \mathbb{P})}, \quad (2.7)$$

consultar [Zuo and Serfling \(2000\)](#) para ver más detalles. Observar que el orden para los datos obtenido del concepto de cercanía es equivalente al de profundidad, lo que difiere es el índice asignado a cada observación, las observaciones que están más alejadas del centro se les asigna un valor grande de  $o(x, \mathbb{P})$ , y decreciendo cuando las observaciones se acercan al centro, asignando el valor más pequeño a la observación más central.

Un ejemplo es la medida de cercanía de Stahel-Donoho. Sea  $X$  una variable aleatoria que toma valores en los reales con ley  $\mathbb{P}$ , entonces, la medida de cercanía de Stahel-Donoho  $SD(z, \mathbb{P})$  se define como

$$SD(z, \mathbb{P}) = \frac{|z - \text{med}(X)|}{\text{mad}(X)},$$

donde  $\text{mad}(X)$  es la desviación absoluta en la media<sup>1</sup> de  $X$  y  $\text{med}(X)$  es la mediana de  $X$ . Se puede extender esta definición al caso de variables aleatorias funcionales tomando la integral sobre  $D$ ,

$$\text{FSD}(z(t), \mathbb{P}) = \int_D SD(z(t), \mathbb{P}_t)^2 dt.$$

Luego, usando la ecuación (2.7) se obtiene una definición de profundidad conocida como la profundidad de proyección, la cual usaremos en el [Capítulo 4](#). Esta definición cumple las propiedades de ser invariante bajo transformación, se maximiza en el centro, monotonicidad con respecto al punto más profundo y también consistencia, ver [Dai and Genton \(2017\)](#).

La [Figura 2.4](#) muestra la idea de este concepto. Si suponemos que la curva  $x_3$  es la mediana y si  $\text{mad}(X)$  es una constante, entonces la cercanía de cada curva

---

<sup>1</sup> $\text{mad}(X)$  es una medida robusta de la variabilidad de la variable aleatoria  $X$  definida como  $\text{mad}(X) = 1.4826\text{med}(|X_i - \text{med}(X)|)$ .

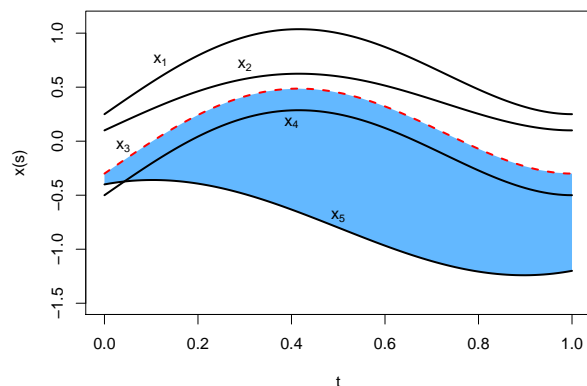


Figura 2.4: Ejemplo del concepto de medida de cercanía de Stahel-Donoho. Si la curva  $x_3$  representa la mediana y si el valor de  $\text{mad}(X)$  es constante, entonces el valor de cercanía de la curva  $x_5$  es el área al cuadrado de la región coloreada.

$x_j(t)$  es el valor del área al cuadrado de la región determinada entre la curva  $x_j(t)$  y la curva  $x_3(t)$ . Por ejemplo, la cercanía de la curva  $x_5$ , es el área al cuadrado de la región coloreada. De esta manera, si la curva está cerca de la curva mediana, su correspondiente cercanía toma un valor pequeño y si ésta está lejos de la curva mediana, entonces la cercanía toma un valor grande. Notemos que, el valor de la cercanía de una curva  $x(s)$  aumenta si  $x(s)$  se aleja cada vez más de la curva mediana. Esto no sucede con la versión empírica del concepto de MBD, pues si  $x(s)$  se aleja cada vez más de la mediana, el valor de la profundidad se mantiene constante.

Recientemente, acorde a la importancia que los datos funcionales han adquirido, se han propuesto varios conceptos de profundidad en este contexto, conceptos que toman en cuenta la forma de la curva y/o ciertas propiedades específicas de los datos funcionales. Algunos de estos trabajos se puede leer en, [Arribas-Gil and Romo \(2015\)](#), [Hubert et al. \(2015\)](#), [Huang and Sun \(2016\)](#) y [Dai and Genton \(2017\)](#). Obviamente no existe una única forma de ordenar estas curvas por lo que diferentes conceptos de profundidad ordenan las curvas de diferentes formas.

## 2.4. Ajuste de curvas

En fenómenos reales donde se asume que los datos son funciones o curvas,  $x(t)$ ,  $t \in D \subset \mathbb{R}$ , es imposible observar dicha curva para toda  $t \in D$ , lo que se observa es una muestra finita, a lo más numerable. Denotemos por  $y_i$  la observación de  $x(t)$  en  $t_i \in D$ . Otro punto a considerar en los datos reales, es que usualmente las observaciones están contaminadas, esto es, lo que realmente se observa en cada  $t_i$  es  $x(t_i)$  más un error.

Supongamos que se observa una muestra de  $x(t)$  en  $n$  sitios,  $(t_1, y_1), \dots, (t_n, y_n)$ , y supongamos también que  $x(t)$  es elemento de  $\mathbb{H}$ , donde  $\mathbb{H}$  es el espacio de Hilbert de funciones definidas en un dominio  $D$ . Entonces, el objetivo es encontrar una estimación  $\hat{x} \in \mathbb{H}$  tal que  $\hat{x}(t)$  aproxime “bien” a  $x(t)$ . Existen dos tipos de aproximaciones: interpolación y regresión. En ambos se pueden usar bases globales (como bases de Fourier o polinomios), o bases locales (como bases B-splines, bases radiales, núcleos reproductores, etc.). Se usa el método de interpolación cuando se supone que las observaciones son sin error, es decir  $y_i = x(t_i)$ .

En este trabajo se usa la aproximación vía regresión, ya que reduce el ruido en la estimación y es más simple manejar el concepto de suavidad. En consecuencia, supondremos que los datos observados,  $(t_1, y_1), \dots, (t_n, y_n)$ , son tales que

$$y_i = x(t_i) + \varepsilon_i,$$

donde  $\{\varepsilon_i\}$  son variables aleatoria no correlacionadas. Para encontrar  $\hat{x}(t)$ , supondremos también que  $x(t)$  es de la forma

$$x(t) = \sum_{j=1}^p \beta_j b_j(t) = \boldsymbol{\beta}^\top \mathbf{b}(t),$$

donde  $\mathbf{b}(t) = (b_1(t), \dots, b_p(t))^\top$  es una base finita y el número de bases  $p$  es un parámetro a determinar.

Lo anterior es equivalente a pensar que el espacio  $\mathbb{H}$  es el espacio de Hilbert que está generado por la base  $\{b_1(t), \dots, b_p(t)\}$ .

Así, dada una base  $\mathbf{b}(t)$  el problema consiste en estimar los coeficientes de la base,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ , y estos serán determinados minimizando el error cuadrático medio

$$\psi(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - x(t_i))^2,$$

o en notación matricial,

$$\psi(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|^2,$$

donde  $\mathbf{y} = (y_1, \dots, y_n)^\top$  y  $\mathbf{B} = (b_k(t_j))_{1 \leq k \leq p, 1 \leq j \leq n}$ . La solución a este problema es,

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{y},$$

y consecuentemente,

$$\hat{x}(t) = \mathbf{b}^\top(t) \hat{\boldsymbol{\beta}}.$$

Adicionalmente, podría requerirse propiedades específicas para  $\hat{x}(t)$ , como “suavidad”, para ello se puede ajustar la curva usando mínimos cuadrados con penalización, es decir, minimizando la función

$$\psi(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - x(t_i))^2 + \lambda \int (x^{(m)}(t))^2 dt$$

o en términos matriciales,

$$\psi(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\top \mathbf{A} \boldsymbol{\beta}$$

donde

$$\mathbf{A} = \int \mathbf{b}^{(m)}(t) \mathbf{b}^{(m)\top}(t) dt,$$

y  $\lambda$  es un parámetro de ajuste. En este caso la solución es

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{A})^{-1} \mathbf{B}^\top \mathbf{y}. \quad (2.8)$$

El parámetro  $\lambda$  usualmente se escoge usando validación cruzada. Para más detalles se puede consultar [Ramsay and Silverman \(2005\)](#), para B-spline ver [Wahba \(1990\)](#).

Para aclarar la idea del ajuste de curva, consideremos un ejemplo. Se evalúa la función  $x(t) = 10 + \sin(x)$  en 50 puntos igualmente espaciados en  $[0, 6]$ . Luego esta muestra es contaminada con ruidos independientes  $\varepsilon_i \sim N(0, 1)$  en cada punto. Supongamos que se observa  $y_i = x(t_i) + \varepsilon_i$ ,  $i = 1, \dots, 50$ . En este ejemplo, se consideran dos bases; base de Fourier y base Spline cúbico, ambos con 5 componentes, es decir,  $p = 5$ . Se ajusta la curva usando la muestra  $(y_i, t_i)$  para obtener  $\hat{\boldsymbol{\beta}}$  y consecuentemente  $\hat{x}(t)$ . La Figura 2.5 muestra los resultados, la curva continua es la curva sin contaminación y es la que se quiere aproximar, las curvas discontinuas representan la aproximación usando las bases correspondientes. Como se puede observar en este caso, la aproximación es bastante buena, considerando el tamaño de muestra y número de base.

La teoría de ajuste de curvas, tiene aparte sus fundamentos teóricos, principalmente tiene que ver con espacios de Hilbert con núcleo reproductor y no es objetivo

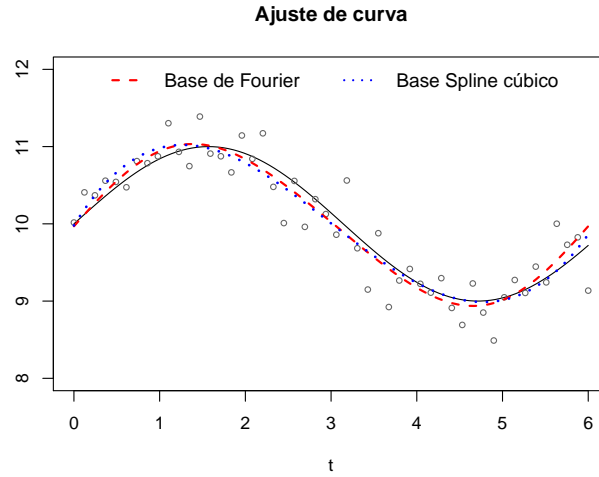


Figura 2.5: La curva con línea continua representa la función sin contaminación  $x(t)$ , los puntos representan  $y_i = x(t_i) + \varepsilon_i$ , y las curvas discontinuas representan el ajuste.

de esta tesis. En esta tesis supondremos que se ha observado la curva para toda  $t \in D$  y esta curva será  $\hat{x}(t)$  obtenida del ajuste.

Cabe mencionar que el estudio de funciones aleatorias y procesos estocásticos están naturalmente ligados cuando dicho proceso estocástico de interés es en sí mismo una función aleatoria que varía en el tiempo. Esto da origen al estudio de series de tiempo funcionales. Dentro de este enfoque el modelo más estudiado, es el modelo FAR(1),

$$X_n(t) = \Phi(X_{n-1})(t) + W_n(t),$$

donde  $\Phi \in B_{\mathbb{H}}$  y  $W_n(t)$  es un ruido blanco funcional. Se hablará más a detalle de este modelo en el Capítulo 4.



## Capítulo 3

# Detección de Persistencia en Series de Tiempo No Lineales

En este capítulo se detalla el método propuesto en [Martínez-Hernández and González-Farías \(2017\)](#). Las series de tiempo con memoria larga aparecen en muchos contextos de la ciencia, por ejemplo en climatología, econometría, física, biología, en secuencias de ADN, tráfico de datos en las redes. El conocimiento del tipo de estructura de dependencia en las series de tiempo es importante ya que a partir de esto se pueden proponer ciertos modelos específicos, por ejemplo los modelos ARFIMA, o en el caso de más de una serie, considerar una posible cointegración entre ellas. El concepto de memoria larga es muy general, en el sentido que podemos hablar de memoria larga en procesos estacionarios y no estacionarios, por ejemplo procesos estacionarios con ruido Gaussiano fraccionario, de igual manera procesos con una tendencia, procesos con componentes una caminata aleatoria. Lo relevante del concepto de memoria larga es que brinda una manera de clasificar a los procesos estocásticos (ver [Aparicio and Escribano, 1998](#); [Dufrénot and Mignon, 2002](#); [Berenguer-Rico and Gonzalo, 2014](#)).

Por otro lado, es sabido que en muchos fenómenos, sobre todo en econometría, con frecuencia las series de tiempo son no lineales, consultar [Sercu et al. \(1995\)](#); [Malliaropulos \(2000\)](#); [Escribano \(2004\)](#); [Stokes and Hinich \(2011\)](#); [Berenguer-Rico and Gonzalo \(2014\)](#). Esto conlleva a que resulte de interés tener medidas de dependencia que tengan un buen desempeño tanto en el contexto lineal y no lineal.

La información mutua (Definición [3.2](#)) se ha usado como una medida de dependencia en muchas áreas de la ciencia, y en particular para medir persistencia en series de tiempo no lineales, ver [Granger and Lin \(1994\)](#), [Dufrénot and Mignon \(2002\)](#) y [Giannerini et al. \(2015\)](#), por mencionar algunos.

En la literatura cuando se usa alguna medida de dependencia entre las series de tiempo  $X_t$  y  $X_{t-h}$ , es común estudiar el decaimiento de la medida cuando  $h \rightarrow \infty$ . También es común estudiar la evolución de la varianza de la suma acumulada de la serie de tiempo, como discutiremos más adelante. A partir de estos enfoques se define una prueba de hipótesis para la memoria larga en la serie de tiempo. Otra alternativa de la prueba de hipótesis es definirla sobre un conjunto de valores de  $h \in \mathbb{N}$ , esto es, para cada  $h$  se infiere sobre la dependencia del proceso. En este enfoque puede presentarse el problema de pruebas múltiples.

En esta tesis se propone usar la información mutua como una medida de dependencia, con esto se puede calcular la dependencia de una serie de tiempo al tiempo  $t$  con su pasado al tiempo  $t - h$ ,  $h \in \mathbb{N}$ . Esta medida es independiente de la linealidad del proceso, como se explica más adelante, además de que tiene las propiedades deseadas; está bien definida tanto para procesos lineales como no lineales, toma el valor cero en el caso independiente y es comparable con la correlación cuando el proceso es lineal. Una desventaja es que requiere del conocimiento de las leyes distribucionales, lo cual no es una tarea fácil en el contexto de procesos no lineales, pero no es una limitación del todo ya que existen varios métodos para estimar densidades en casos generales.

Proponemos una prueba de persistencia sobre un conjunto de valores de  $h$ , usando un enfoque funcional. Se supondrá que la información mutua es una función continua sobre el parámetro de retraso  $h$ , de esta manera, en vez de hacer la prueba sobre  $h$  objetos, se hace una prueba sobre un solo objeto, la curva obtenida de los  $h$  objetos.

Usaremos “ $\xrightarrow{L}$ ” para denotar convergencia en distribución.

### 3.1. Pruebas de persistencia existentes

Antes de desarrollar el método propuesto, se hace una revisión de las pruebas existentes sobre memoria larga en el contexto de series de tiempo no lineales. Dado que el estudio de memoria larga está fuertemente ligado con el estudio de la no estacionariedad del proceso, varias pruebas están motivadas desde esta perspectiva y más generales que la noción basada en las series integradas.

Como hemos mencionado anteriormente, una prueba popular es la prueba de Hurst, conocida también como el estadístico  $R/S$ . Sea  $X_1, \dots, X_n$  una secuencia de  $n$  observaciones, definamos  $S_m = X_1 + \dots + X_m$ , las sumas parciales donde



$m = 0, 1, \dots$ , con  $S_0 = 0$ . El estadístico de Hurst es el siguiente

$$\frac{R}{S}(X_1, \dots, X_n) = \frac{\max_{0 \leq i \leq n} (S_i - \frac{i}{n}S_n) - \min_{0 \leq i \leq n} (S_i - \frac{i}{n}S_n)}{\left(\frac{1}{n} \sum_{i=1}^n (X_i - \frac{1}{n}S_n)^2\right)^{1/2}}. \quad (3.1)$$

Observemos que  $S_n/n$  es la media muestral, de aquí  $\max_{0 \leq i \leq n} (S_i - \frac{i}{n}S_n)$  mide hasta qué punto las sumas parciales quedan por encima de la línea recta que se produciría si todas las observaciones fueran iguales (a la media de la muestra). La diferencia entre el máximo y el mínimo del numerador es la diferencia entre las posiciones más altas y más bajas de las sumas parciales con respecto a la línea recta de crecimiento uniforme.

Si las observaciones  $\{X_i\}$  son i.i.d.(0, 1) entonces

$$\frac{1}{\sqrt{n}} \frac{R}{S}(X_1, \dots, X_n) \xrightarrow{L} \left( \sup_{0 \leq t \leq 1} B_0(t) - \inf_{0 \leq t \leq 1} B_0(t) \right), \quad (3.2)$$

donde  $B_0$  es un puente Browniano en  $[0, 1]$ ;  $B_0(t) := B(t) - tB(1)$  con  $B(t)$  un Movimiento Browniano.

Si  $\{X_i\}$  son ruidos Gaussianos fraccionarios, con índice de autosimilitud  $H$ , entonces el estadístico de Hurst satisface (ver [Mandelbrot, 1965](#)),

$$\frac{1}{n^H} \frac{R}{S}(X_1, \dots, X_n) \xrightarrow{L} \left( \sup_{0 \leq t \leq 1} (B_H(t) - tB_H(1)) - \inf_{0 \leq t \leq 1} (B_H(t) - tB_H(1)) \right), \quad (3.3)$$

donde  $B_H(t)$  es un movimiento Browniano fraccionario con índice de autosimilitud  $H$ . Entonces  $\mathbb{E}(R/S) \sim C_H n^H$ , cuando  $n \rightarrow \infty$ , donde  $C_H$  es una constante que no depende de  $n$ .

El estadístico  $R/S$ , es una prueba no paramétrica y en un principio no requiere de la hipótesis de linealidad del proceso. Pero tiene serios problemas con procesos no lineales. En [Bhattacharya et al. \(1983\)](#) se muestra que si la serie de tiempo es de la forma

$$X_i = Y_i + (a + i)^{-\beta},$$

con  $\{Y_i\}$  i.i.d.,  $0 < \beta < 1/2$ , entonces

$$\frac{1}{n^{1-\beta}} \frac{R}{S}(X_1, \dots, X_n) \rightarrow \frac{C_\beta}{\sigma}. \quad (3.4)$$

Por lo tanto, el estadístico  $R/S$  tiene el mismo comportamiento en procesos lineales con memoria larga y procesos no estacionarios con una tendencia. Como

consecuencia, estos procesos son indistinguibles bajo el estadístico  $R/S$ . Tomando en cuenta el problema anterior [Lo \(1991\)](#) propone una modificación a la prueba para intentar resolver la sensibilidad a la dependencia a corto plazo, considerando lo siguiente,

$$\frac{R}{S}(X_1, \dots, X_n) = \frac{\max_{0 \leq i \leq n} (S_i - \frac{i}{n} S_n) - \min_{0 \leq i \leq n} (S_i - \frac{i}{n} S_n)}{\hat{\sigma}^2 + 2 \sum_{j=1}^q w_j(q) \gamma_j}, \quad (3.5)$$

donde  $w_j(q) = 1 - \frac{j}{q+1}$ ,  $q < n$  y  $\gamma_j$  denota la autocovarianza. Un inconveniente con esta modificación es la elección del parámetro  $q$  ya que los resultados dependen mucho de este parámetro. En general, estas pruebas no tienen un buen desempeño cuando el proceso es no lineal (ver [Dufrénot and Mignon, 2002](#), Cap 3).

[Higuchi \(1988\)](#) propuso un método que permite estimar la dimensión fractal  $D$  de la serie de tiempo. Sea  $S_m = X_1 + \dots + X_m$ , y sea

$$L(m) = \frac{n-1}{m^3} \sum_{i=1}^m \left( \frac{n-i}{m} \right)^{-1 \lfloor (n-i)/m \rfloor} \sum_{k=1}^{\lfloor (n-i)/m \rfloor} |S_{i+km} - S_{i+(k-1)m}|,$$

donde  $n$  es el tamaño de la serie de tiempo. Entonces se tiene que  $\mathbb{E}(L(m)) \sim C_H m^{-D}$ , donde  $D = 2 - H$  y  $m$  es el tamaño del bloque. Tomando logaritmo se obtiene  $\log(\mathbb{E}(L(m))) = \log(C_H) - D \log(k) + \nu_k$ , de esta manera, proyectando  $\log(\mathbb{E}(L(m)))$  sobre  $\log(k)$  se obtiene un estimador de la dimensión fractal,  $\hat{D}$ , y en consecuencia el de la memoria larga  $\hat{H} = 2 - \hat{D}$ . En un contexto no lineal, la relación entre dimensión fractal y el índice de Hurst puede no ser lineal, leer [Gneiting and Schlather \(2004\)](#). Esto hace que esta prueba no sea adecuada en series de tiempo no lineales.

Una forma de describir la dependencia en los procesos no lineales es a través de los momentos de orden más grandes que dos, (ver [Brillinger, 1975](#)). Y obviamente, se requiere la existencia de estos momentos en el proceso. [Hinich \(1996\)](#) propuso una prueba no paramétrica basada en este enfoque, (ver también [Brooks and Hinich, 2001](#)). Definamos los cumulantes empíricos de orden dos y tres,

$$\hat{C}(\tau_1) = \frac{1}{T - \tau_1} \sum_{t=1}^{T-\tau_1} x_t x_{t+\tau_1}$$

$$\hat{C}(\tau_1, \tau_2) = \frac{1}{T - m} \sum_{t=1}^{T-m} x_t x_{t+\tau_1} x_{t+\tau_2},$$

donde  $m = \max\{\tau_1, \tau_2\}$ .

Bajo el supuesto de independencia  $\hat{C}(\tau_1) = 0 = \hat{C}(\tau_1, \tau_2)$ . Si existe dependencia, al menos uno de los cumulantes debería ser diferente de cero. Para dicha prueba de dependencia [Hinich \(1996\)](#) propone los siguientes estadísticos,  $H_1$  y  $H_2$ ,

$$H_1 = \sum_{s=1}^L \sqrt{T-s} \left| \hat{C}(s) \right|^2$$

$$H_2 = \sum_{s=2}^L \sum_{r=1}^{s-1} \sqrt{T-s} \left| \hat{C}(r, s) \right|^2$$

donde  $L = T^c$ , con  $0 < c < 1/2$ . Observemos que  $H_1$  captura la dependencia lineal mientras que  $H_2$  captura las posibles dependencias no lineales. Esta prueba tiene una potencia muy baja, (ver [Dufrénot and Mignon, 2002](#), Cap. 3).

[Berenguer-Rico and Gonzalo \(2014\)](#) propusieron el concepto del índice de sumabilidad con el propósito de generalizar la cointegración lineal, al mismo tiempo, este concepto es utilizado para medir el grado de persistencia usando la evolución de la varianza del proceso. Una serie de tiempo  $\{X_t\}$  se dice que es sumable con índice  $\alpha$ , si existe una función de variación lenta en el infinito,  $L(n)$ , y una sucesión determinista,  $m_t$ , tal que

$$S_n = \frac{1}{n^{\frac{1}{2}+\alpha}} L(n) \sum_{t=1}^n (X_t - m_t) = O_p(1), \quad (3.6)$$

donde  $\alpha$  es el número real más pequeño que hace que  $S_n$  esté acotado en probabilidad. El índice de sumabilidad  $\alpha$  indica el grado de persistencia. Por ejemplo, si  $\{X_t\}$  es independiente, entonces  $\alpha = 0$  y si es una caminata aleatoria, entonces  $\alpha = 1$ .

La estimación de  $\alpha$  es obtenida a través de un estimador de mínimos cuadrados del modelo de regresión que resulta al tomar el logaritmo de  $S_n^2$  en la ecuación (3.6), esto es

$$Y_k = \beta \log k + U_k, \quad k = 1, \dots, n,$$

donde  $\beta = 1 + 2\alpha$ ,  $Y_k = \log \left( \sum_{t=1}^k (y_t - m_t) \right)$  y  $U_k = O_p(1)$ . Así,  $\hat{\alpha} = (\hat{\beta} - 1)/2$ . El problema es que se desconoce la distribución (finita o asintótica) de  $\hat{\alpha}$  por lo cual es difícil cuantificar la incertidumbre de este estimador. Para conocer más detalles de este concepto ver [Berenguer-Rico and Gonzalo \(2014\)](#).

En general, se puede adaptar la siguiente prueba de hipótesis. Sea  $I(h)$  alguna medida de dependencia de  $(X_t, X_{t-h})$ , y definamos  $\{X_t\}$  de memoria corta si  $I(h) = o(h^{-d})$ , para  $0 \leq d \leq 1$ , lo cual implica que  $I(h) < bh^{-1}$  para algún  $b$  y

$h > h_0$ , de lo contrario, la serie de tiempo tiene memoria larga, es decir, existen  $c$  y  $h_1$  tal que  $I(h) \approx ch^{-d}$ , para toda  $h > h_1$ . Aplicando logaritmo ambos lados se obtiene

$$\log I(h) \approx \log c - d \log h + \epsilon_h, \quad h \gg h_1,$$

donde  $\epsilon_h$  es una sucesión de errores. Por lo tanto, se puede definir la prueba de hipótesis de memoria corta,  $H_0 : d > 1$  contra la alternativa  $H_1 : d \leq 1$ . Para este tipo de pruebas se requiere una serie de tiempo con longitud muy larga y no tiene buen desempeño (ver [Aparicio and Escribano, 1998](#)).

Otros enfoques que se han usado son: atractores, memoria larga (corta) en media, memoria corta (larga) en distribución, ver [Granger and Hallman \(1991\)](#), [Granger \(1995\)](#) y [Escanciano and Hualde \(2009\)](#).

## 3.2. Persistencia y el concepto mixing

La definición de un proceso con memoria corta se relaciona con un proceso  $I(0)$ . Un proceso  $I(0)$  se ha definido de diferentes maneras. Una de ellas está relacionada con el teorema de límite central funcional, ver [Müller \(2008\)](#), [Davidson \(2009\)](#) y [Berenguer-Rico and Gonzalo \(2014\)](#). Por otro lado, en series de tiempo, a menudo se usa la medida  $R(h) := \text{Corr}(X_n, X_{n+h})$  para definir memoria corta (memoria larga). Específicamente, una serie de tiempo  $\{X_n, n \in \mathbb{Z}\}$  se dice que tiene memoria corta si

$$\sum_n R(n) < \infty \tag{3.7}$$

y memoria larga si  $\sum_n R(n) = \infty$ . Para entender las razones detrás de el por qué la definición anterior y la definición que usaremos, introducimos los conceptos básicos de un proceso mixing.

Sea  $(\Omega, \mathcal{F}, \mathbb{P})$  un espacio de probabilidad. Para cualquier  $\sigma$ -álgebra  $\mathcal{A} \subset \mathcal{F}$  sea  $L^2(\mathcal{A})$  el espacio de variables aleatorias sobre  $\mathcal{A}$  real valuadas y segundo momento finito. Para dos  $\sigma$ -álgebras,  $\mathcal{A}$  y  $\mathcal{B} \subset \mathcal{F}$ , se definen las siguientes medidas de dependencia:

$$\begin{aligned} \rho(\mathcal{A}, \mathcal{B}) &:= \sup |\text{Corr}(f, g)|, f \in L^2(\mathcal{A}), g \in L^2(\mathcal{B}); \\ \alpha(\mathcal{A}, \mathcal{B}) &:= \sup |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \mathcal{A}, B \in \mathcal{B}; \\ I(\mathcal{A}, \mathcal{B}) &:= \sup \sum_{i=1}^I \sum_{j=1}^J \mathbb{P}(A_i \cap B_j) \log \left( \frac{\mathbb{P}(A_i \cap B_j)}{\mathbb{P}(A_i)\mathbb{P}(B_j)} \right). \end{aligned}$$

El supremo en la definición de  $I(\mathcal{A}, \mathcal{B})$  es tomado sobre particiones  $\{A_1, \dots, A_I\}$  y  $\{B_1, \dots, B_J\}$  de  $\Omega$  tal que  $A_i \in \mathcal{A}$  para cada  $i$  y  $B_j \in \mathcal{B}$  para cada  $j$ . También en la definición de  $I(\mathcal{A}, \mathcal{B})$ , se toma por conveniencia  $0/0 := 0$  y  $0 \log(0) := 0$ .

Sea  $\mathbf{X} = (X_n, n \in \mathbb{Z})$  una sucesión de variables aleatorias. Para  $-\infty \leq J \leq L \leq \infty$ , definamos el  $\sigma$ -álgebra

$$\mathcal{F}_J^L := \sigma(X_k, J \leq k \leq L, k \in \mathbb{Z}).$$

A continuación se definen los coeficientes de dependencia:

$$\rho(n) := \sup_{j \in \mathbb{Z}} \rho(\mathcal{F}_{-\infty}^j, \mathcal{F}_{j+n}^\infty);$$

$$\alpha(n) := \sup_{j \in \mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^j, \mathcal{F}_{j+n}^\infty);$$

$$I(n) := \sup_{j \in \mathbb{Z}} I(\mathcal{F}_{-\infty}^j, \mathcal{F}_{j+n}^\infty).$$

Con las definiciones anteriores, se dice que  $\mathbf{X}$  es  $\alpha$ -mixing si  $\alpha(n) \rightarrow 0$  cuando  $n \rightarrow \infty$ . Análogamente con el resto de las medidas;  $\rho$ -mixing y  $I$ -mixing. También se usa el término “regular en información” en vez de  $I$ -mixing. Se tienen las siguientes implicaciones,  $I$ -mixing  $\Rightarrow$   $\alpha$ -mixing,  $\rho$ -mixing  $\Rightarrow$   $\alpha$ -mixing, y no hay otra relación.

Observemos que, si  $\mathbf{X}$  es estrictamente estacionaria, entonces la condición  $\alpha$ -mixing (igual con las demás) se simplifica a

$$\alpha(n) = \alpha(\mathcal{F}_{-\infty}^0, \mathcal{F}_n^\infty).$$

Para leer más a detalle sobre estos conceptos ver [Bradley \(2005\)](#).

Ahora, sea  $S_n = X_1 + \dots + X_n$  y sea  $S^{(n)}$  definido como

$$S^{(n)}(t) = S_{[nt]} - [nt]\mu, \quad 0 \leq t \leq 1,$$

con  $\mu$  la media del proceso. El siguiente teorema establece las condiciones para que se cumpla el Teorema del Límite Central Funcional para procesos  $\alpha$ -mixing.

**Teorema 3.1** *Sea  $\mathbf{X} = \{X_n, n \in \mathbb{N}\}$  un proceso con media cero y  $\alpha$ -mixing tal que, para algún  $\delta > 0$ ,  $\mathbb{E}|X_0|^{2+\delta} < \infty$ . Supongamos que  $\text{Var}(S_n) \rightarrow \infty$  cuando  $n \rightarrow \infty$  y para alguna  $K < \infty$*

$$\mathbb{E}|S_n|^{2+\delta} \leq K(\text{Var}(S_n))^{1+\delta/2}, \quad \forall n.$$

*Entonces, las sumas parciales convergen débilmente en  $D[0, 1]$*

$$\frac{1}{\text{Var}(S_n)^{1/2}} S^{(n)} \xrightarrow{L} B, \quad (3.8)$$

donde  $B$  es un movimiento Browniano. Más aún, si

$$\sum_{n=1}^{\infty} |R(n)| < \infty,$$

entonces, el siguiente límite existe,

$$\lim_{n \rightarrow \infty} \frac{\text{Var}(S_n)}{n} := \sigma_*^2,$$

y se cumple el teorema de límite central funcional,

$$\frac{1}{\sqrt{n}} S_n \xrightarrow{L} \sigma_* B.$$

Si la serie de tiempo es estacionaria, se tiene

$$\text{Var}(S_n) = \sigma^2 \left( n + 2 \sum_{i=1}^{n-1} (n-i) R(i) \right).$$

Esta expresión está relacionada con la tasa de decrecimiento de la correlación del proceso. Si  $\sum_{n=1}^{\infty} |R(n)| < \infty$ , claramente el límite  $\lim_{n \rightarrow \infty} \frac{\text{Var}(S_n)}{n} = \sigma^2 (1 + 2 \sum_{i=1}^{\infty} R(i))$  existe. Esto es, la varianza de  $S_n$  crece linealmente con  $n$ . Cuando esto sucede se puede pensar que la serie de tiempo no está tan lejos del caso i.i.d.

Si la varianza de  $S_n$  crece más rápido que una función de variación regular con exponente más grande que uno ( $\text{Var}(S_n) \sim n^H$ , con  $H > 1$ ), no es posible la convergencia al movimiento Browniano  $B$ , sin importar que normalización se use en (3.8) y en consecuencia no se cumple el teorema de límite central funcional. El siguiente teorema respalda esta afirmación.

**Teorema 3.2 (Lamperti)** *Sea  $\{U(t), t \geq 0\}$ , un proceso estocástico y sea  $a_n \uparrow \infty$  una sucesión de números positivos. Supongamos que existe un proceso  $\{Y(t), t \geq 0\}$  tal que*

$$\frac{1}{a_n} U(nt) \xrightarrow{L} Y(t), \quad (3.9)$$

*cuando  $n \rightarrow \infty$ . Supongamos que  $\mathbb{P}(Y(1) \neq 0) > 0$  y continua en probabilidad. Entonces para alguna  $H > 0$ ,  $\{Y(t), t \geq 0\}$  es un proceso autosimilar con índice  $H$  y la sucesión  $\{a_n\}$  es de variación regular con exponente  $H$  ( $a_n \sim n^H$ ).*

Se sabe que si  $\{X_t\}$  es un proceso Gaussiano estacionario, entonces  $\{X_t\}$  es  $\alpha$ -mixing si y sólo si  $R(n) \rightarrow 0$ , cuando  $n \rightarrow \infty$ . Tomando en cuenta el resultado anterior podemos ver el porqué la definición de memoria corta a través de (3.7).

La medida  $R(n)$  es una medida de dependencia lineal y su uso no es apropiado en un contexto de series de tiempo no lineales. En vez de usar  $R(k)$  consideraremos la medida de información mutua de la Definición 3.2. Para ello, hacemos una breve revisión del concepto mixing relacionado a  $I(n)$ .

Supongamos que  $\mathbf{X} = (X_k, k \in \mathbb{Z})$  es una cadena de Markov, entonces la condición  $\alpha$ -mixing (similarmente con las otras condiciones) se simplifica a

$$\alpha(n) = \sup_{j \in \mathbb{Z}} \alpha(\sigma(X_j), \sigma(X_{j+n})), \quad (3.10)$$

donde  $\sigma(X)$  denota la  $\sigma$ -álgebra generada por la variable aleatoria  $X$ .

Si  $\mathbf{X}$  es una cadena de Markov estrictamente estacionaria, la condición anterior se reduce a

$$\alpha(n) = \alpha(\sigma(X_0), \sigma(X_n)). \quad (3.11)$$

El siguiente Lema se puede encontrar en Bradley (1983).

**Lema 3.1** *Sea  $X = (X_k, k \in \mathbb{Z})$  una cadena de Markov estrictamente estacionaria con espacio de estado  $\mathcal{E}$  contable. Entonces para cada  $n \geq 1$ ,*

$$I(n) = I(\sigma(X_0), \sigma(X_n)).$$

El lema anterior es de gran relevancia en la práctica ya que permite inferir sobre  $I$ -mixing del proceso a través del estudio de los vectores a pares  $(X_t, X_{t-n})$ . Así, podemos usar la Definición 3.2 para definir memoria corta (memoria larga), cubriendo una gran clase de procesos no lineales, y como hemos mencionado anteriormente, si el proceso es  $I$ -mixing entonces es  $\alpha$ -mixing. Entonces, si reemplazamos  $\alpha$ -mixing por  $I$ -mixing, adicionalmente a los supuestos en el teorema de límite central funcional, se tiene que el resultado aún es válido, con la ventaja de que  $I(n)$  es más fácil de estimar comparado con  $\alpha(n)$ .

Con lo anterior, tenemos la siguiente definición de memoria corta (larga) para series de tiempo.

**Definición 3.1** *Sea  $X_t$  una serie de tiempo y sea  $T(h)$  el coeficiente de información mutua definida en (3.16). Decimos que la serie de tiempo tiene memoria larga si*

$$\lim_{m \rightarrow \infty} \sum_{h=0}^m T(h) = \infty,$$

*y diremos que es de memoria corta si la suma es finita.*

Se han estudiado propiedades  $\alpha$ -mixing en algunos modelos no lineales, por ejemplo en los modelos bilineales, ver [Tuan \(1986\)](#), usando representación de Markov. Para modelos GARCH ver [Masry and Tjøstheim \(1995\)](#). Para modelos TAR( $p$ ) ver [Dufrénot and Mignon \(2002\)](#).

### 3.3. Información mutua y cópulas

A continuación definimos la información mutua que es usada como medida de dependencia.

**Definición 3.2** *Sea  $\{X_t\}$  y  $\{X_{t-h}\}$  dos procesos estocásticos con densidad conjunta  $f(X_t, X_{t-h})$  y marginales  $f_t(X_t)$  y  $f_{t-h}(X_{t-h})$ . La información mutua entre  $X_t$  y  $X_{t-h}$  es definido como*

$$\delta(X_t, X_{t-h}) = \mathbb{E}_f \left[ \log \left( \frac{f(X_t, X_{t-h})}{f_t(X_t)f_{t-h}(X_{t-h})} \right) \right]. \quad (3.12)$$

La información mutua mide la distancia basada en la entropía entre la distribución conjunta  $f(x, y)$  y el producto de las marginales,  $f_t(x)f_{t-h}(y)$ , que representa la independencia entre  $\{X_t\}$  y  $\{X_{t-h}\}$ .

La información mutua tiene la propiedad de que no depende de la linealidad del proceso, pero depende de la función de densidad conjunta y las marginales correspondientes. En [Blumentritt and Schmid \(2012\)](#) se reescribe la información mutua en términos de la cópula de la función de distribución conjunta  $(X_t, X_{t-h})$ . Con esto se puede estimar  $\delta(X_t, X_{t-h})$  estimando la cópula de  $(X_t, X_{t-h})$ .

Las cópulas son una herramienta para describir estructuras de dependencia. A continuación se da la definición formal.

**Definición 3.3** *Una cópula bidimensional es una función de distribución conjunta en  $[0, 1] \times [0, 1]$  con marginales uniformes, esto es*

$$C(u, v) = \mathbb{P}(U \leq u, V \leq v),$$

donde  $U$  y  $V$  son variables aleatorias uniformes en  $[0, 1]$ .

La definición anterior es válida para cualquier dimensión  $d \geq 2$ , es decir, definido sobre  $[0, 1]^d$ . En esta tesis solo usamos el caso bidimensional.

El siguiente teorema es la base para las aplicaciones ya que establece la relación entre la cópula y la función de distribución conjunta. Denotamos por  $\text{Im}(F)$  la imagen de la función  $F$ .



**Teorema 3.3 (Teorema de Sklar)** *Sea  $F$  una función de distribución conjunta con marginales  $F_1$  y  $F_2$ . Entonces existe una cópula  $C$  tal que para toda  $(x, y) \in \mathbb{R}^2$ ,*

$$F(x, y) = C(F_1(x), F_2(y)). \quad (3.13)$$

*Si las marginales son continuas, entonces  $C$  es única. En otro caso  $C$  es única en  $\text{Im}(F_1) \times \text{Im}(F_2)$ . Inversamente, si  $C$  es una cópula y  $F_1, F_2$  son funciones de distribución, entonces la función  $F$  definida por (3.13) es una distribución conjunta con marginales  $F_1$  y  $F_2$ .*

Como podemos observar, la cópula  $C(u, v)$  describe la dependencia entre  $X_1$  y  $X_2$  y no depende de las marginales,  $F_1$  y  $F_2$ , ya que  $F_i(X_i)$  siempre es una variable aleatoria uniforme. Si la cópula  $C(u, v)$  es absolutamente continua con respecto a la medida de Lebesgue, entonces su función de densidad es

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v},$$

y en consecuencia,

$$c(F_1(x_1), F_2(x_2)) = \frac{f(x_1, x_2)}{f_1(x_1)f_2(x_2)}. \quad (3.14)$$

Algunas medidas de dependencia que están relacionadas con la cópula son: La  $\tau$  de Kendall y la  $\rho$  de Spearman. Recordemos la definición de estas dos medidas. Sea  $(X_1, X_2)$  un vector aleatorio, entonces la  $\rho$  de Spearman de las variables aleatorias  $X_1$  y  $X_2$  es definida como,

$$\rho_S(X_1, X_2) = 3(\mathbb{P}((X_1 - Y_1)(X_2 - Z_2) > 0)) - \mathbb{P}((X_1 - Y_1)(X_2 - Z_2) < 0),$$

donde  $(X_1, X_2)$ ,  $(Y_1, Y_2)$  y  $(Z_1, Z_2)$  son copias independientes. La  $\tau$  de Kendall de las variables aleatorias  $X_1$  y  $X_2$  es definida como,

$$\tau(X_1, X_2) = \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0) - \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) < 0),$$

donde  $(X'_1, X'_2)$  es independiente de  $(X_1, X_2)$  y con la misma ley. Entonces, si  $C$  es la cópula de  $(X_1, X_2)$ ,

$$\tau_S(X_1, X_2) = 4 \int \int C(u, v) dC(u, v) - 1,$$

y

$$\rho_S(X_1, X_2) = 12 \int \int (C(u, v) - uv) dudv.$$

**Observación.** La función de correlación no es función de la cópula.

Las cópulas también pueden describir otros tipos de dependencias, como dependencias en los valores extremos (ver [Nelsen, 2006](#)).

Otra propiedad interesante de las cópulas es que se pueden acotar superiormente e inferiormente, estas cotas son conocidas como las cotas de Fréchet-Hoeffding. El siguiente teorema muestra este resultado.

**Teorema 3.4** *La cópula  $C(u, v)$  tiene las cotas*

$$W^2 = \max\{u + v - 1, 0\} \leq C(u, v) \leq \min\{u, v\} = M^2.$$

Se puede ver que  $W^2(u, v) = \mathbb{P}(U \leq u, 1 - U \leq v)$  y  $M^2(u, v) = \mathbb{P}(U \leq u, U \leq v)$ , son cópulas. Esta cota se puede definir para dimensiones más grandes que dos, sin embargo  $W^d$  no será una cópula.

Un ejemplo de cópulas que se debe tener en cuenta es la correspondiente de dos variables aleatorias independientes. Sea  $(X_1, X_2)$  dos variables aleatorias independientes, así la distribución conjunta es el producto de las marginales, entonces la cópula es

$$C(x, y) = xy.$$

En la literatura, esta cópula es denotada por  $\Pi$ .

Existen varios enfoques para estimar una cópula, una forma es vía máxima verosimilitud, esto siempre y cuando se suponga un modelo para dicha cópula. Otra manera es vía no paramétrica, en este caso se puede proponer estimadores usando algún tipo de núcleo. En esta tesis se usa la estimación vía no paramétrica y es detallada más adelante. Para leer más sobre cópulas, se puede consultar [Nelsen \(2006\)](#) y [Joe \(2015\)](#).

Como hemos explicado, las cópulas son herramientas que ayudan a describir la estructura de dependencia subyacente por si sola. En el método que se propone, para la detección de persistencia, las cópulas son usadas para describir la información mutua y a partir de este último se hace la inferencia para la estructura de dependencia.

Regresando a la información mutua, y considerando la ecuación [\(3.14\)](#), podemos

reescribir  $\delta(X_t, X_{t-h})$  en términos de la cópula de la siguiente manera,

$$\begin{aligned}
\delta_C(X_t, X_{t-h}) &= \int \int \log \left( \frac{f(x, y)}{f_t(x)f_h(y)} \right) f(x, y) dx dy \\
&= \int \int \log (c(F_t(x), F_h(y))) c(F_t(x), F_h(y)) f_t(x) f_h(y) dx dy \\
&= \int_{[0,1]^2} \log (c(u, v)) c(u, v) du dv \\
&= \mathbb{E}_C [\log(c(U, V))], \tag{3.15}
\end{aligned}$$

donde  $U = F_t(X_t)$  y  $V = F_h(X_{t-h})$ . Con esta nueva expresión, la información mutua basada en la cópula,  $\delta_C$ , no depende de las distribuciones marginales, solamente de la estructura de dependencia conjunta. La siguiente proposición muestra las propiedades de  $\delta_C$ .

**Proposición 3.1** (*Blumentritt and Schmid 2012*). *Sea  $\mathbf{X} = (X_1, X_2)$  un vector aleatorio. Sea  $C(u, v)$  la cópula de  $\mathbf{X}$  y  $c(u, v)$  la correspondiente densidad de la cópula. Entonces,*

- $\delta_C(\mathbf{X}) \geq 0$  y  $\delta_C(\mathbf{X}) = 0$  si y sólo si  $X_1$  y  $X_2$  son independientes o en término de la cópula, esto es  $\delta_C(\mathbf{X}) = 0$  si y sólo si  $C = \Pi$ .
- $\delta_C(X_1, X_2)$  es invariante bajo transformaciones monótonas crecientes (decrecientes) por componente.
- $\delta_C(X_1, X_2) = \delta_C(X_2, X_1)$
- Sea  $c_m$  una sucesión de densidades de cópulas que converge uniformemente a una densidad de cópula  $c$  cuando  $m \rightarrow \infty$ . Entonces,  $\delta_{c_m} \rightarrow \delta_c$  cuando  $m \rightarrow \infty$ .

El último punto indica que  $\delta_C$  puede ser estimada consistentemente si la estimación de la cópula es consistente. Observemos que el rango de  $\delta_c$  es  $[0, \infty)$ . Con el objeto de poder comparar con otras medidas de dependencia, como la correlación, se prefiere usar el coeficiente de información mutua  $T(X_t, X_{t-h})$ , definido por una transformación,

$$T(X_t, X_{t-h}) = [1 - \exp\{-2\delta_C(X_t, X_{t-h})\}]^{1/2}. \tag{3.16}$$

En este caso se tiene que  $T(X_t, X_{t-h}) = 0$  si y solamente si  $X_t$  y  $X_{t-h}$  son independientes, y  $T(X_t, X_{t-h}) = 1$  si  $X_t$  es una transformación monótona creciente de  $X_{t-h}$ .

El coeficiente de información mutua se puede pensar como una generalización de la función de correlación definida en (2.1). Sea  $X_t$  un proceso Gaussiano, y sea  $\gamma(t, t-h) = \text{Cov}(X_t, X_{t-h})$ . Entonces,

$$\delta(X_t, X_{t-h}) = -\frac{1}{2} \log \left( 1 - \frac{\gamma(t, t-h)^2}{\text{Var}(X_t)\text{Var}(X_{t-h})} \right) = -\frac{1}{2} \log (1 - \rho(t, t-h)^2),$$

y por lo tanto,

$$T(X_t, X_{t-h}) = |\rho(t, t-h)|.$$

En esta tesis proponemos hacer una prueba de persistencia en series de tiempo considerando  $T(h)$  y suponiendo que es una función continua en  $h$ . Así, la prueba de hipótesis es sobre un sólo objeto  $T(h)$  en vez de una hipótesis múltiple sobre los diferentes retrasos  $h = 1, 2, \dots$ .

### 3.3.1. Estimación de la información mutua

Estamos interesados en estimar  $T(h)$  como una función continua en  $h$ . Esta estimación consiste en dos etapas. Primero se estiman  $\hat{T}(h_j)$ , para  $h_j = 0, 1, \dots, m$ , obteniendo una muestra de tamaño  $m+1$  para  $T(h)$ , con  $m \in \mathbb{N}$ . En la segunda etapa se asume una base finita de funciones y se ajusta una curva usando mínimos cuadrados para finalmente obtener  $\hat{T}(h)$  para toda  $h \in [0, m]$ .

Antes de detallar la estimación  $\hat{T}(h)$ , primero describimos la estimación para  $\delta_C(X_t, X_{t-h_j})$ . En la literatura se han propuesto varios métodos para la estimación de  $\delta$ . [Granger and Lin \(1994\)](#) usa la relación de la información mutua con la entropía de Shannon;

$$\delta(X, Y) = H(X) + H(Y) - H(X, Y),$$

donde  $H(X, Y) = -\int \log(f(x, y)) f(x, y) dx dy$  y  $H(X) = -\int \log(f_X(x)) f_X(x) dx$ . Usando un estimador  $\hat{f}$  vía núcleo, se obtienen los estimadores para  $H(X, Y)$  y  $H(X)$  y en consecuencia  $\hat{\delta}$ . Este tipo de estimador no tiene buenas propiedades cuando el proceso es no lineal, al menos en nuestro caso. Otros estimadores se pueden leer en [Paninski \(2003\)](#) o [Walters-Williams and Li \(2009\)](#).

En este trabajo se considera la estimación propuesta en [Blumentritt and Schmid \(2012\)](#), que está basado en la ecuación (3.15).

**Suposición 3.1**  $\delta(X_t, X_{t-h})$  depende solamente de  $h$  y escribiremos  $\delta(h)$  y  $T(h)$ .

De ahora en adelante se asume la Suposición 3.1. Si denotamos por  $\hat{c}(u, v)$  la estimación de la cópula  $c(u, v)$ , entonces el estimador de la información mutua  $\delta$

es definida como,

$$\hat{\delta}(h) = \frac{1}{n} \sum_{j=1}^n \log(\hat{c}(u_j, v_j)), \quad (3.17)$$

donde  $n$  es la longitud de la muestra y  $u_j$  y  $v_j$  son las *pseudobservaciones* de la cópula que se definen más adelante.

Por lo tanto, la estimación de la cópula  $c(u, v)$  juega un papel importante en la estimación de  $\delta$ . La estimación no paramétrica tiene la ventaja de que no es necesario especificar un modelo para la cópula, o equivalentemente, no requiere de un supuesto de la forma específica de estructura de dependencia, de lo contrario se podría restringir la estimación de la cópula para una estructura de dependencia específica, (ver [Ibragimov and Lentzas, 2008](#); [Chen et al., 2009](#); [Lagerås, 2010](#)).

Recordar que al usar un estimador tipo núcleo, se presentan problemas de estimación en la frontera; sesgo y varianza. En procesos con memoria larga, muchas veces las dependencias están en los valores extremos por lo cual es importante reducir la varianza de la estimación en la frontera.

Tomando en cuenta que el dominio es  $[0, 1] \times [0, 1]$  y el problema de la frontera, se prefiere usar en la estimación de la cópula un núcleo Beta (ver [Chen, 1999](#); [Bouezmarni and Rolin, 2003](#)).

La definición del núcleo Beta,  $K(x, \alpha, \beta)$ , es

$$K(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad x \in [0, 1] \text{ y } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Luego la definición del estimador de la densidad de la cópula,  $c(u, v)$ , del correspondiente vector aleatorio  $(X_t, X_{t-h_j})$  usando un núcleo Beta, es definido como,

$$\hat{c}_j(u, v) = \frac{1}{n - h_j} \sum_{l=1}^{n-h_j} K(\hat{U}_{l+h_j}, \frac{u}{b} + 1, \frac{1-u}{b} + 1) K(\hat{V}_l, \frac{v}{b} + 1, \frac{1-v}{b} + 1), \quad (3.18)$$

donde  $\hat{U}_l = \hat{F}_t(X_l)$ ,  $\hat{V}_{l-h_j} = \hat{F}_{t-h_j}(X_{l-h_j})$  y

$$\hat{F}_{t-h}(x) = \frac{1}{n+1} \sum_{j=1}^n \mathbb{1}_{\{X_{t-h,j} \leq x\}}.$$

con  $\{X_{t,j}\}_{j \geq 1}$  una muestra de  $\{X_t\}$ . Las observaciones  $\{\hat{U}_l\}$  y  $\{\hat{V}_i\}$  son conocidas como *pseudobservaciones*.

Finalmente se obtiene que, para cada  $h_j$ , el estimador de la información mutua basada en cópulas es

$$\hat{\delta}_C(h_j) = \frac{1}{n - h_j} \sum_{l=1}^{n-h_j} \log \left[ \hat{c}(\hat{U}_{l+h_j}, \hat{V}_l) \right]. \quad (3.19)$$

Consecuentemente, obtenemos una muestra de la curva  $T(h)$ , aplicando la transformación definida en (3.16). Esto es, tenemos la muestra  $t_j = \left(1 - \exp \{-2\hat{\delta}_C(h_j)\}\right)^{1/2}$  para  $h_j = 0, 1, 2, \dots, m$ .

Ahora, supongamos que tenemos una base finita de funciones,  $\mathbf{b}(t) = (b_1(t), \dots, b_p(t))^\top$ , y que

$$T(h) = \sum_{k=1}^p a_k b_k(h).$$

Los coeficientes  $\{a_i\}$  son estimados minimizando la función

$$\sum_{i=1}^{m+1} (t_i - T(h_i))^2 + \lambda \int (T^{(2)}(h))^2 dh. \quad (3.20)$$

Con esto, se obtiene la estimación de la curva

$$\hat{T}(h) = \sum_{k=1}^p \hat{a}_k b_k(h), \quad h \in [0, m],$$

con  $\hat{\mathbf{a}} = (\hat{a}_1, \dots, \hat{a}_p)$  como en (2.8). Ver Capítulo 2.3 para los detalles de ajuste de curva.

En el resto de este capítulo, supondremos que  $\hat{T}(h)$  es una función conocida y que corresponde a la serie de tiempo  $\{X_t\}$ . La denotaremos simplemente como  $T(h)$ .

### 3.4. Prueba de persistencia basada en la entropía

En esta sección se describe la prueba de hipótesis usando el estadístico  $T(h)$  con enfoque funcional. Para hacer una prueba de hipótesis, se requiere los valores críticos. Usualmente, estos valores críticos son determinados ya sea usando la distribución asintótica del estadístico o a través de simulaciones bajo la hipótesis nula. Nuestra propuesta es con el segundo enfoque usando la prueba envolvente.

### 3.4.1. Prueba envolvente

Sea  $\{X_t\}$  la serie de tiempo sobre la que se quiere hacer la prueba de hipótesis. Sea  $T_{\text{obs}}(h)$  la función empírica observada correspondiente al coeficiente de información mutua de la serie de tiempo. Denotaremos por  $D = [0, m]$  el dominio de la función  $T_{\text{obs}}(h)$  y a la vez usaremos la notación  $T_1(h) \equiv T_{\text{obs}}(h)$ . Observemos que  $m$  es el máximo de los retrasos en donde  $T_1(h)$  es evaluada.

El interés es probar si  $T_1(h)$  es una observación de la variable aleatoria funcional  $T_0$  bajo la nula, es decir, la hipótesis nula es de la forma  $H_0 : T_1 \sim T_0$ . Dado que la distribución de  $T_0$  es desconocida, se recurre al método Monte Carlo, esto es, se simula  $M$  curvas  $T_i(h)$ ,  $i = 2, \dots, M + 1$ , del modelo bajo la hipótesis nula (suponiendo que esto es posible). Entonces, la prueba de hipótesis puede ser reemplazada por  $H_0 : T_1(h), T_2(h), \dots, T_{M+1}(h)$  son i.i.d.. Si  $H_0$  no es correcta, entonces  $T_1(h)$  sería una observación “extrema”. Para determinar si la curva  $T_1(h)$  es extrema dentro de las simulaciones, usamos la prueba envolvente. La prueba envolvente define una banda de no rechazo  $[T_{\text{low}}(h), T_{\text{upp}}(h)]$ , que constituye una región para las observaciones más verosímiles bajo  $H_0$  a través de un ordenamiento de curvas. Así, si  $T_1(h)$  no pertenece a dicha banda es indicio de que posiblemente proviene de otra ley diferente de  $T_0$ .

Como podemos ver, la prueba envolvente compara gráficamente la función empírica  $T_1(h)$  con su contraparte de la hipótesis nula. La ventaja de la prueba envolvente es que es un método que proporciona una herramienta para interpretar los resultados de manera gráfica y que es de gran ayuda en la práctica.

En general, este tipo de prueba, donde se ordenan los datos obtenidos de las simulaciones Monte Carlo es conocida como la prueba de Monte Carlo de Barnards, (ver [Dufour, 2006](#)). A continuación escribimos la definición formal.

**Definición 3.4** *Una banda de no rechazo es una banda que está determinada por las funciones  $T_{\text{low}}(h)$  y  $T_{\text{upp}}(h)$ . Se llama prueba envolvente al estadístico de prueba que rechaza la hipótesis nula si la función observada  $T_1(h)$  no está completamente dentro de la envolvente, esto es,*

$$\phi_{\text{env}}(T_1) = \mathbb{1}_{\{\exists h \in D : T_1(h) \notin [T_{\text{low}}, T_{\text{upp}}]\}}. \quad (3.21)$$

Para determinar la banda se requiere ordenar las curvas, y para ello se usa el concepto de profundidad que a continuación describimos.

Usamos la profundidad de banda modificada (MBD) definida en el Capítulo 2.3, tomando en cuenta las curvas  $T_i(h)$  así también como su derivada  $T'_i(h)$ , ya que la derivada proporciona información sobre que tan rápido decrece la curva. Esto es, calculamos la proporción del tiempo que la curva  $T_i(h)$  pasa en una región definida y

también se calcula la proporción del tiempo que la derivada  $T'_i(h)$  pasa en otra región definida por las curvas correspondientes a las derivadas. Siendo más específicos, sea  $\mathbf{T} = \{T_1(h), \dots, T_{M+1}(h)\}$  una muestra de curvas y sea  $T \in \mathbf{X}$ , definimos  $A^*$  como

$$A^*(T; T_{i_1}, \dots, T_{i_j}) = w_1 A(T; T_{i_1}, \dots, T_{i_j}) + w_2 A(T'; T'_{i_1}, \dots, T'_{i_j}),$$

donde  $w_i$  representan los pesos,  $w_1 + w_2 = 1$  y  $A$  es como en (2.6). En este caso, tenemos que

$$\text{MBD}_{M+1}^{*(j)}(x) = \binom{M+1}{j}^{-1} \sum_{1 \leq i_1 < \dots < i_j \leq M+1} \lambda_r \{A^*(T; T_{i_1}, \dots, T_{i_j})\}. \quad (3.22)$$

**Observación:** En la práctica el valor de  $j$  que se usa es  $j = 2$ , ya que el orden de las curvas es estable con los valores de  $j = 2, \dots, J$  (ver Sun and Genton, 2011).

Denotemos por  $d_i$  el valor de la profundidad de la curva  $T_i$  según la ecuación (3.22), esto es,

$$d_i = \text{MBD}_{M+1}^{*(2)}(T_i). \quad (3.23)$$

Con esto tenemos un método para ordenar las curvas de acuerdo al orden decreciente de los valores  $d_i$ . Denotemos por  $T_{[i]}(h)$  la curva asociada con el  $i$ -ésimo valor más grande de los valores de la profundidad. Como hemos explicado en la sección 2.3,  $T_{[1]}(h)$  será la curva más profunda o la curva mediana, y la curva  $T_{[M+1]}(h)$  es la curva más extrema.

Usando el orden obtenido de las profundidades, definimos la banda de no rechazo como sigue. Sea  $\alpha \in (0, 1)$  el nivel de significancia, entonces  $T_{\text{low}}(h)$  y  $T_{\text{upp}}(h)$  son definidos como,

$$T_{\text{low}}^\alpha(h) = \min_{i=1, \dots, k_\alpha} T_{[i]}(h) \text{ y } T_{\text{upp}}^\alpha(h) = \max_{i=1, \dots, k_\alpha} T_{[i]}(h),$$

donde  $k_\alpha = \lfloor (1 - \alpha)(M + 1) \rfloor$ . En consecuencia, la banda que no toma en cuenta la porción  $\alpha$  de las curvas más extremas es definida como,

$$C_\alpha = \{(h, T(h)) : t \in D, T_{\text{low}}^\alpha(h) \leq T(h) \leq T_{\text{upp}}^\alpha(h)\}, \quad (3.24)$$

Por lo tanto, la prueba envolvente es definida como,

$$\phi_{\text{env}}(T_1) = \mathbb{1}_{\{\exists h \in D : T_1(h) \notin C_\alpha\}}. \quad (3.25)$$

Esto es, se rechaza  $H_0$  si  $T_1(h)$  no pertenece completamente a la banda  $C_\alpha$ .



Para el estudio de la potencia de esta prueba, consideremos la prueba Monte Carlo, donde el  $p$ -valor correspondiente es definido como,

$$p = \frac{1}{M+1} \left( 1 + \sum_{i=2}^{M+1} \mathbb{1}_{\{T_i < T_1\}} \right), \quad (3.26)$$

con  $<$  un orden estricto,  $T_i < T_j$  o  $T_j < T_i$ . así, la prueba Montecarlo es

$$\phi_\alpha(T_1) = \mathbb{1}_{(p \leq \alpha)},$$

esto es, rechazar  $H_0$  si  $p \leq \alpha$ . Para el siguiente resultado definimos el concepto de intercambiabilidad, esto es

$$\mathbb{P}((T_1, \dots, T_{M+1}) \in A) = \mathbb{P}((T_{\sigma(1)}, \dots, T_{\sigma(M+1)}) \in A),$$

para cualquier conjunto medible  $A$ .

**Proposición 3.2** *Sea  $T_1, \dots, T_{M+1}$  funcionales estadísticos intercambiables y sea “ $<$ ” un orden tal que*

$$\mathbb{P}(T_i < T_j \text{ o } T_j < T_i) = 1, \quad \forall i \neq j.$$

*Entonces, si  $M$  es tal que  $\alpha(M+1)$  es un entero, la prueba  $\phi(T_1) = \mathbb{1}_{(p \leq \alpha)}$ , es tal que*

$$\mathbb{P}(\phi(T_1)) = \alpha,$$

*donde  $p$  es como en (3.26).*

La proposición anterior nos dice que la hipótesis nula es rechazada con nivel de significancia  $\alpha$ , siempre que el orden sea estricto, ver [Myllymäki et al. \(2017\)](#).

Ahora, veamos la relación entre la prueba envolvente definida en (3.25) y la prueba Montecarlo. Para ello definamos los siguientes  $p$ -valores,

$$p_- = \frac{1}{M} \sum_{i=1}^{M+1} \mathbb{1}_{\{d_i < d_1\}} \quad \text{y} \quad p_+ = \frac{1}{M} \sum_{i=1}^{M+1} \mathbb{1}_{\{d_i \leq d_1\}}. \quad (3.27)$$

El siguiente resultado establece la relación entre estos dos métodos.

**Proposición 3.3** *Sean  $p_-$  y  $p_+$  definidos como en (3.27). Supongamos que las profundidades son tales que*

$$\mathbb{P}(d_i \neq d_j) = 1, \quad \forall i \neq j.$$

*Entonces se cumplen las siguientes propiedades,*

1.  $T_1$  pertenece completamente a la banda  $C_\alpha$ , esto es,  $T_{\text{low}}^\alpha(h) < T_1(h) < T_{\text{upp}}^\alpha(h)$  para toda  $h \in I$ , si y solamente si  $p_- > \alpha$ .
2.  $T_1$  está fuera de la banda  $C_\alpha$ , esto es,  $T_1(h) < T_{\text{low}}^\alpha(h)$  o  $T_1(h) > T_{\text{upp}}^\alpha(h)$  para algún  $h \in I$ , si y sólo si  $p_+ \leq \alpha$ .

**Demostación.** Probemos la primera propiedad, tenemos que  $T_1(h)$  cumple  $T_{\text{low}}^\alpha(h) < T_1(h) < T_{\text{upp}}^\alpha(h)$  para toda  $h \in D$  si y solamente si  $d_1 > d_{[i]}$ , con  $i = k_\alpha, k_\alpha + 1, \dots, M + 1$ . Observemos que,

$$\begin{aligned} p_- &= \frac{1}{M+1} \sum_{i=1}^{M+1} \mathbb{1}_{\{d_{[i]} < d_1\}} \\ &= \frac{1}{M+1} \sum_{i=1}^{k_\alpha} \mathbb{1}_{\{d_{[i]} < d_1\}} + \alpha, \end{aligned}$$

y como  $d_{[k_\alpha]} < d_1$ , tenemos que  $\frac{1}{M+1} \sum_{i=1}^{k_\alpha} \mathbb{1}_{\{d_{[i]} < d_1\}} > 0$ . Por lo tanto se tiene el resultado.

Equivalentemente para la propiedad dos,  $T_1(h) < T_{\text{low}}^\alpha(h)$  o  $T_1(h) > T_{\text{upp}}^\alpha(h)$  para algún  $h \in I$ , si y sólo si  $d_1 < d_{[k_\alpha]}$ , esto es  $d_1 < d_{[i]}$  para  $i = 1, \dots, k_\alpha$ . Entonces

$$\begin{aligned} p_+ &= \frac{1}{M+1} \sum_{i=1}^{k_\alpha} \mathbb{1}_{\{d_{[i]} \leq d_1\}} + \frac{1}{M+1} \sum_{i=k_\alpha+1}^{N+1} \mathbb{1}_{\{d_{[i]} \leq d_1\}} \\ &= \frac{1}{M+1} \sum_{i=k_\alpha+1}^{M+1} \mathbb{1}_{\{d_{[i]} \leq d_1\}} \\ &\leq \alpha. \end{aligned}$$

Por lo tanto se tiene el resultado.  $\square$

El resultado anterior muestra que la prueba envolvente, además de ser una prueba de hipótesis, es una herramienta de visualización que ayuda a entender el por qué la decisión es rechazada o no rechazada en la prueba.

### 3.4.2. Prueba de hipótesis de memoria larga

Sea  $\mathcal{P}$  el conjunto de todas las series de tiempo; lineales, no lineales, con memoria corta y con memoria larga. Sea  $\mathcal{P}_0 \subset \mathcal{P}$  el subconjunto que contiene a los procesos lineales AR estables, con memoria corta.

$$\mathcal{P}_0 = \{\text{Procesos lineales AR}(p) \text{ estables de memoria corta}\}, \quad (3.28)$$

y sea

$$\mathcal{P}_1 = \mathcal{P}_0^c,$$

el complemento de  $\mathcal{P}_0$ . Denotemos por  $\mathbb{P}_0$  el conjunto de las leyes de los procesos que pertenecen a la clase  $\mathcal{P}_0$  y  $\mathbb{P}_1$  al conjunto de las leyes correspondientes a los procesos en  $\mathcal{P}_1$ . Dada una serie de tiempo  $\mathbf{X} = \{X_t\}$ , la prueba de hipótesis sobre memoria larga es definida como

$$H_0 : \mathbf{X} \text{ pertenece a la clase } \mathcal{P}_0,$$

vs

$$H_1 : \mathbf{X} \text{ pertenece a la clase } \mathcal{P}_1,$$

usando el estadístico  $T(h)$ .

Esta prueba de hipótesis se puede pensar de la siguiente manera. Sea  $F_n(\cdot; \mathbb{P})$  la distribución del estadístico  $T(h)$  bajo la ley  $\mathbb{P}$  de la serie de tiempo  $\mathbf{X}$ . Dado que  $\mathbb{P}$  es desconocida  $F_n(\cdot; \mathbb{P})$  es desconocida. Se estima  $\mathbb{P}$  con  $\hat{Q}$  bajo  $H_0$  ajustando un modelo  $\text{AR}(p)$ , estable y con memoria corta. Con esto,  $\hat{Q}$  satisface la restricción de  $H_0$  y obviamente si  $\mathbb{P} \in \mathbb{P}_0$ ,  $\hat{Q} \approx \mathbb{P}$  y en consecuencia  $F_n(\cdot; \hat{Q}) \approx F_n(\cdot; \mathbb{P})$ . Por otro lado, si  $\mathbb{P} \in \mathcal{P}_1$ ,  $\hat{Q}$  aproxima una ley diferente que pertenece a  $\mathbb{P}_0$ . Sin embargo, si  $\mathbb{P}$  representa un proceso no lineal con memoria corta, aún así puede que  $\hat{Q}$  aproxime a  $\mathbb{P}$  en cierto sentido, ya que los procesos  $\text{AR}(p)$  pueden aproximar a procesos estacionarios no lineales. En este caso las entropías serán similares y en consecuencia  $F_n(\cdot; \hat{Q}) \approx F_n(\cdot; \mathbb{P})$ . Por ejemplo los procesos bilineales estacionarios se puede aproximar con un proceso  $\text{AR}(p)$  en el sentido de la estructura de covarianza. En cambio un proceso GARCH no es aproximable con un proceso  $\text{AR}(p)$  ya que este proceso es un ruido blanco desde el punto de vista de un proceso lineal, es decir, la covarianza no es capaz de detectar la estructura de dependencia.

Si  $\mathbb{P} \in \mathbb{P}_1$  representa un proceso con memoria larga, entonces  $\hat{Q}$  no aproximará a  $\mathbb{P}$  y las entropías correspondientes a los procesos con ley  $\hat{Q}$  y  $\mathbb{P}$  serán totalmente diferentes por lo que  $F_n(\cdot; \hat{Q})$  no aproximará a  $F_n(\cdot; \mathbb{P})$  y consecuencia  $T_1(h)$  no será parte de la muestra de  $F_n(\cdot; \hat{Q})$ .

Con lo anterior podemos reformular la prueba de hipótesis de la siguiente manera. Sea  $\mathcal{P}'_1$  definido como la clase

$$\mathcal{P}'_1 = \{ \text{Procesos con memoria larga} \},$$

y definir la prueba de hipótesis como

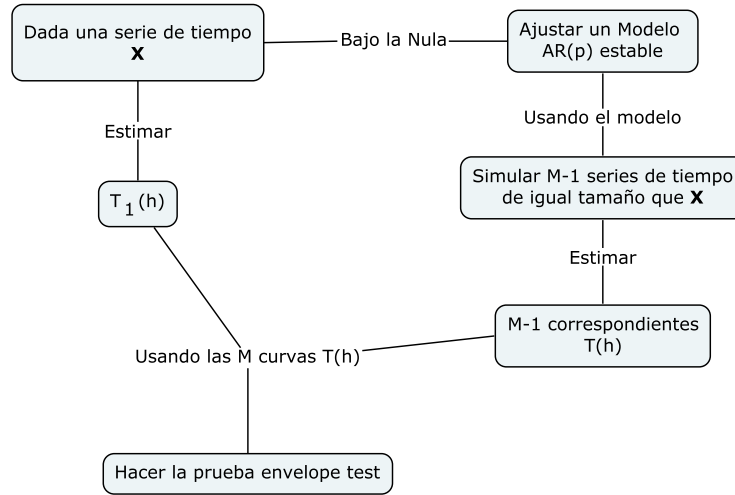


Figura 3.1: Diagrama de la prueba de hipótesis.

$$H_0 : \mathbf{X} \text{ pertenece a la clase } \mathcal{P}_0$$

vs

$$H'_1 : \mathbf{X} \text{ pertenece a la clase } \mathcal{P}'_1,$$

usando el estadístico  $T(h)$ .

Dado que se desconoce la ley de  $T(h)$  bajo  $H_0$ , usamos la prueba envolvente para determinar los valores críticos, ajustando un modelo  $AR(p)$  a  $\mathbf{X}$  tal que el modelo es estable y de memoria corta. Luego de este modelo se simulan  $\mathbf{X}_2^*, \dots, \mathbf{X}_M^*$  series de tiempo y se estiman  $T_2(h), \dots, T_M(h)$ . Dado un nivel de significancia  $\alpha$  se construye la banda de no rechazo definida en (3.24). En la Figura 3.1 se ilustra el diagrama del procedimiento de la prueba de hipótesis.

**Observación 1.** La prueba de hipótesis definida no es una prueba de procesos autoregresivos estables contra procesos autoregresivos no estables. Es una prueba de procesos con entropía idéntica a un proceso autoregresivo estable contra un proceso con entropía diferente a un proceso autoregresivo estable. Es decir, es una prueba sobre la memoria del proceso en términos de la entropía.

**Observación 2.** Si una serie de tiempo  $\{X_t\}$  tiene memoria larga,  $T(h)$  decrecerá más lento a cero a diferencia si es de memoria corta. Más aún, si  $\{X_t\}$  tiene memoria

corta, entonces  $T(h)$  debe ser similar a la de un proceso AR estable satisfaciendo  $\sum_{h \geq 0} T(h) < \infty$ .

La prueba anterior, se puede usar en un contexto más general. Esto es, considerando

$$\mathcal{P}_0 = \{\text{Procesos lineales estables con memoria corta}\},$$

donde un proceso lineal estable es definido como un proceso,

$$X_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i}, \quad \psi_0 = 1, \quad \{Z_i\} \sim \text{i.i.d.}, \quad \sum_{i=-\infty}^{\infty} |\psi_i| < \infty,$$

utilizando el mismo procedimiento anterior; ajustar un modelo  $\text{AR}(p)$  bajo la hipótesis nula. Esto debido a que los procesos  $\text{AR}(p)$  son densos en el espacio de procesos lineales estables, en el sentido de que se puede describir cualquier estructura de correlación a través de la estructura de correlación de un proceso  $\text{AR}(p)$ . La justificación se lleva a cabo utilizando las densidades espectrales.

**Definición 3.5** Sea  $\{X_t\}$  una serie de tiempo estacionaria con media cero y función de covarianza  $R(\cdot)$  tal que satisface  $\sum_{h=-\infty}^{\infty} |R(h)| < \infty$ . Entonces la densidad espectral de  $\{X_t\}$  es definida como la función  $f$  tal que

- $f$  es simétrica, es decir,  $f(\lambda) = f(-\lambda)$ ,
- $f(\lambda) \geq 0$  para toda  $\lambda \in (-\pi, \pi]$ ,
- $R(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda$ .

**Teorema 3.5** Sea  $R(k)$  una función definida en los enteros tal que  $\sum_{h=-\infty}^{\infty} |R(h)| < \infty$ . Entonces  $R(k)$  es función de autocovarianza de un proceso estacionario  $\{X_t\}$  si y sólo si

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} R(h) \geq 0, \quad \forall \lambda \in [-\pi, \pi].$$

**Teorema 3.6** Sea  $f$  una densidad espectral continua simétrica y  $\epsilon > 0$ . Entonces existe un proceso  $\text{AR}(p)$  causal

$$X_t = \sum_{i=1}^p \phi_i X_{t-i} + Z_t,$$

tal que su densidad espectral  $f_X$ ,

$$|f_X(\lambda) - f(\lambda)| < \epsilon, \quad \forall \lambda \in [-\pi, \pi].$$

Con los resultados anteriores, se concluye que la estructura de covarianza de un proceso lineal puede ser aproximado por la estructura de covarianza de un proceso  $AR(p)$  (ver [Brockwell and Davis, 2006](#), Cap. 4).

## 3.5. Estudio de simulación Monte Carlo

En esta sección se presentan estudios de simulación con el propósito de evaluar el desempeño de la prueba de persistencia usando la prueba envolvente y la entropía del proceso. En el esquema de simulación se consideran procesos lineales y no lineales con distintos grados de dependencia. Para la manipulación de las funciones  $T(h)$  se usa el paquete *fda* de [R Core Team \(2017\)](#). Para la generación de procesos de memoria larga se usa el paquete *fracdiff*. En cada esquema de simulación, se consideran series de tiempo  $\{X_t\}$  con tamaño de muestra  $n = 1000$ .

### 3.5.1. Comportamiento de $T(h)$

En esta parte mostramos el comportamiento de las curvas  $T(h)$  bajo diferentes modelos con memoria corta y memoria larga. Específicamente, consideramos modelos

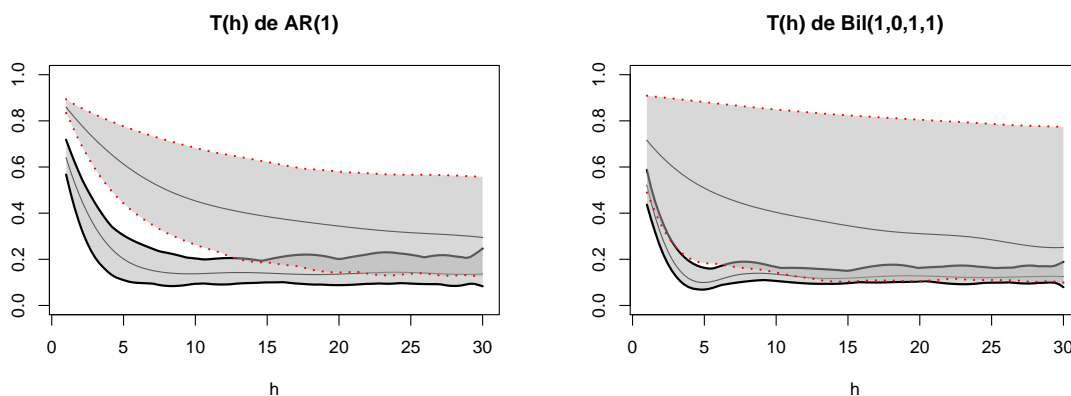


Figura 3.2: La región coloreada representa el 95% de las curvas más profundas de  $T(h)$  (región cuantil) en cada escenario. Lado izquierdo corresponde al modelo  $AR(1)$  con memoria corta y memoria larga. Lado derecho corresponde al modelo Bilineal con memoria corta y memoria larga.

AR(1);  $x_t = 0.7x_{t-1} + \varepsilon_t^d$ , y Bilineal(1, 1, 0, 1);  $x_t = 0.4x_{t-1} + 0.5x_{t-1}\varepsilon_{t-1} + \varepsilon_t^d$  donde  $\{\varepsilon_t^d\}$  es tal que  $\Delta^d \varepsilon_t$  es ruido blanco (ver ecuación (1.5)) con  $d = 0$  y  $d = 0.4$ .

Para cada modelo se simulan series de tiempo  $\mathbf{X}_j = \{X_{t,j}\}$  y se estiman  $\{T_j(h)\}$  con  $j = 1, \dots, 2500$  y  $t = 1, \dots, 1000$ . La Figura 3.2 muestra la región que representa el 95% de las curvas más profundas usando (3.22), esta región se representa gráficamente como el área en color gris y se puede pensar como la región del cuantil 95% de las curvas  $\{T_j(h)\}$ . La curva continua contenida en cada región representa el comportamiento más común (curva más profunda) entre las curvas del correspondiente modelo. La primera gráfica (lado izquierdo) corresponde al modelo AR(1). La región delimitada por las curvas continuas corresponden al error con  $d = 0$ , es decir, un proceso AR con memoria corta. La región delimitada por las curvas discontinuas corresponden a  $d = 0.4$ , es decir, un proceso AR de memoria larga. En este caso podemos observar que la intersección de las regiones es mínima, sugiriendo que la prueba de hipótesis puede tener buen desempeño. La segunda gráfica (lado derecho) corresponde al modelo Bilineal, de igual manera la región delimitada por curvas continuas representa al proceso con memoria corta ( $d = 0$ ) y la otra región corresponde al proceso de memoria larga ( $d = 0.4$ ). La región correspondiente al proceso de memoria corta es similar al caso lineal con memoria corta. Sin embargo, la región que representa la memoria larga tiene mayor variabilidad comparada con el caso lineal. Aún así, la intersección es mínima. Cabe mencionar que el parámetro de memoria en los modelos simulados es alto,  $d = 0.4$ , cercano a un proceso con raíz

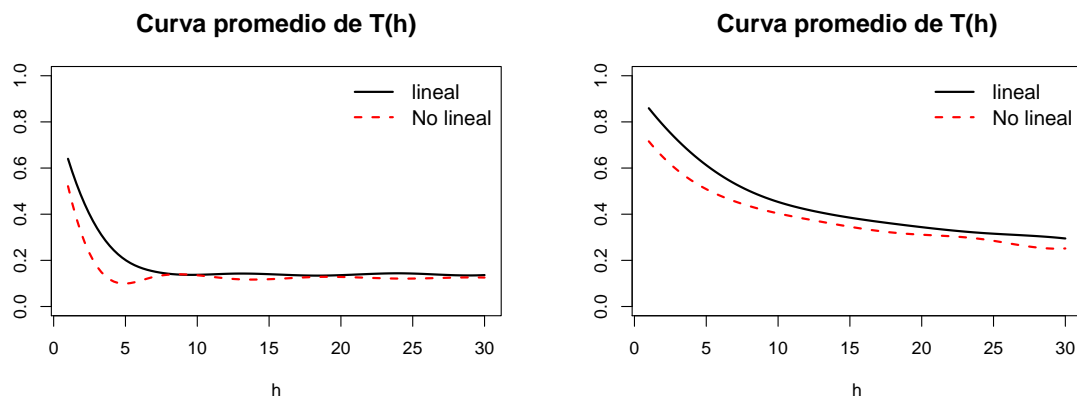


Figura 3.3: Curvas promedio de  $T(h)$ . Lado izquierdo corresponde a los modelos con memoria corta. Lado derecho corresponde a los modelos con memoria larga.

unitaria, así, la variabilidad esperada en las curvas es grande.

En la figura 3.3 se comparan conjuntamente las curvas promedio de cada modelo simulado. En ella podemos observar que las curvas en promedio decrecen a una velocidad similar sin importar si proviene del modelo lineal o no lineal (con el mismo parámetro  $d$  de memoria larga). Para el caso de memoria larga el decaimiento es pronunciado en el intervalo  $[0, 5]$ , mientras que en el caso de memoria larga, el decrecimiento es más lento. Estos comportamientos se comparan con lo que se observa en un correlograma, a diferencia que este método no depende de la linealidad del proceso.

La Figura 3.4 presenta las curvas promedio de  $T(h)$  del modelo AR(1);  $x_t = \phi x_{t-1} + \varepsilon_t$ , con  $\{\varepsilon_t\}$  ruido blanco y  $\phi = 0.1, 0.2, 0.3, 0.5, 0.7, 0.8, 0.9, 1$ . Todos ellos corresponden a procesos con memoria corta excepto el caso  $\phi = 1$ , que es una caminata aleatoria. En esta gráfica observamos que de acuerdo el parámetro  $\phi$  tiende a uno,  $T(h)$  decrece cada vez más lento pero aún con una velocidad pronunciada que no se observa en procesos de memoria larga.

En general, parece que  $T(h)$  describe correctamente la estructura de dependencia.

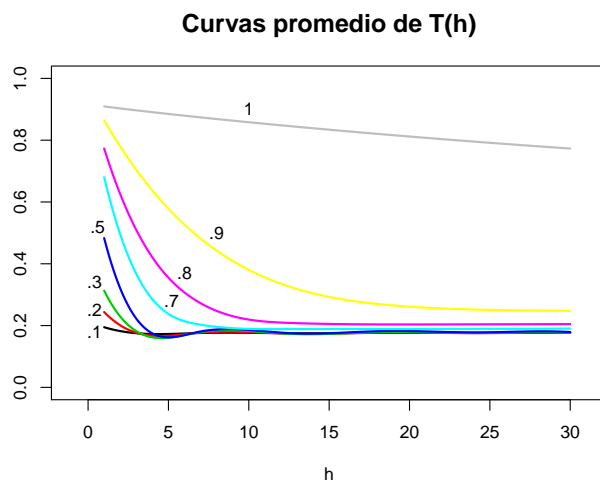


Figura 3.4: Curvas promedio de procesos AR(1) con diferentes parámetros. Los números en las curvas indican el valor del parámetro correspondiente al modelo.



### 3.5.2. Nivel de significancia y potencia

Simulamos series de tiempo  $\{X_t, t = 1, \dots, n\}$  con tamaño de muestra  $n = 1000$ , usando diferentes modelos que se enlistan a continuación:

1.  $x_t = 0.5x_{t-1} + \varepsilon_t$
2.  $x_t = 0.5x_{t-1} + \varepsilon_t^d$
3.  $x_t = x_{t-1} + \varepsilon_t$
4.  $x_t = y_t + (0.5 + t)^{-0.2}$
5.  $x_t = 0.3x_{t-1} + 0.2x_{t-1}\varepsilon_{t-1} + \varepsilon_t$
6.  $x_t = 0.3x_{t-1} + 0.2x_{t-1}\varepsilon_{t-1} + \varepsilon_t^d$
7.  $x_t = x_{t-1} + 0.2x_{t-1}\varepsilon_{t-1} + \varepsilon_t$
8.  $x_t = (0.4x_{t-1} + \varepsilon_t)\mathbb{1}_{(x_{t-1} \leq 0)} + (0.6x_{t-1} + \varepsilon_t^d)\mathbb{1}_{(x_{t-1} > 0)}$
9.  $x_t = (0.4x_{t-1} + \varepsilon_t^d)\mathbb{1}_{(x_{t-1} \leq 0)} + (0.6x_{t-1} + \varepsilon_t^d)\mathbb{1}_{(x_{t-1} > 0)}$
10.  $x_t = \varepsilon_t\sqrt{\sigma_t}, \sigma = 0.1 + 0.4x_{t-1}^2 + 0.2\sigma_{t-1}$
11.  $x_t = \varepsilon_t\sqrt{\sigma_t}, \sigma = 0.1 + 0.8x_{t-1}^2 + 0.2\sigma_{t-1}$

donde  $\varepsilon_t \sim \text{i.i.d. } N(0, 1)$ ,  $\{\varepsilon_t^d\}$  es tal que  $\Delta^d \varepsilon_t$  es ruido blanco (ver ecuación (1.5)) y  $\{y_t\}$  es una sucesión i.i.d.. El valor de  $d \in (0, 1/2)$  corresponde al parámetro de memoria larga y consideramos;  $d = 0.2, 0.3, 0.4$ .

Observemos que los modelos 1, 2, 3 y 4 son modelos lineales y el resto son no lineales. Dada la simulación  $\mathbf{X} = \{X_1, \dots, X_n\}$ , se ajusta un modelo AR( $p$ ) bajo la hipótesis nula. A partir del modelo ajustado se simulan series de tiempo de igual tamaño que  $\mathbf{X}$ . Para cada simulación bajo  $H_0$ ,

$$\mathbf{X}_i^* = \{X_{1,i}, \dots, X_{n,i}\}, i = 2, \dots, M,$$

se estima  $\delta(h)$  usando (3.19) y (3.18) con  $h = 1, \dots, h_{max}$ , donde  $h_{max} = 30$  y  $M = 1000$ . Luego hacemos la transformación para obtener una muestra,  $\{t_{1,i}, \dots, t_{m,i}\}$ , de cada curva de coeficiente de información mutua,

$$\{T_i(h), i = 2, \dots, M\}.$$

Finalmente, para ajustar la curva, usamos una base de splines cúbicos,  $\mathbf{b}(h) = (b_1(t), \dots, b_p)^T$ , con número de base  $p = 8$ . Obteniendo así el ajuste de curva,

$$T_i(h) = \sum_{j=1}^8 \hat{a}_{ij} b_j(h),$$

donde  $\hat{\mathbf{a}}_i = (\hat{a}_{i1}, \dots, \hat{a}_{ip})^T$  es solución del problema de optimización (3.20).

Como antes, denotemos por  $T_1(h)$  la curva de coeficiente de información mutua correspondiente a  $\mathbf{X}$ . A partir de

$$\{T_i(h), i = 1, 2, \dots, M\},$$

se determina la banda de no rechazo,  $C_{0.05}$ , definido en (3.24). Una vez definida la región, se cuantifica si  $T_1(h)$  pertenece completamente a dicha región. Este procedimiento se repite 1000 veces. También, para cada simulación se calcula el estadístico R/S definido en (3.5), usando la función *RoverS* en el paquete *fractal* de R. Recordemos que estos valores son estimaciones del parámetro  $H = 1/2 - d$ . En teoría, si  $H \approx 1/2$  indica que el proceso es de memoria corta, y si  $H \in (1/2, 1)$  indica que el proceso es de memoria larga. En la práctica esto no es el caso ya que la estimación de R/S depende mucho del método a usar, y es por eso que simulamos procesos ruido blanco y calculamos los valores de R/S tomando en cuenta los cuantiles 0.05, 0.5 y 0.95. Así, entre más grande sean los valores de R/S comparados con los del proceso ruido blanco más alta es la probabilidad de memoria larga. Análogamente al caso del ruido blanco, presentamos los cuantiles de los valores de R/S para cada modelo a simular. Por lo tanto, la comparación con la prueba envolvente no es directa.

Primero mostramos los resultados de los modelos que corresponden a procesos de memoria corta; modelo 1, 4, 5 y 10. La Tabla 3.1 presenta dichos resultados.

Tabla 3.1: Resultados de las proporciones de rechazo y valores del estadístico R/S para procesos de memoria corta.

	Env test	$\overline{R/S}$	$(R/S_{.05}, R/S_{.95})$
WN		0.65	(0.64, 0.67)
Modelo 1	0.06	.76	(0.75, 0.78)
Modelo 4	0.05	0.65	(0.64, 0.67)
Modelo 5	0.06	0.73	(0.71, 0.74)
Modelo 10	0.66	0.66	(0.64, 0.68)

Tabla 3.2: Resultados de las proporciones de rechazo y valores del estadístico R/S para procesos no estacionarios.

	Env test	$\overline{R/S}$	$(R/S_{.05}, R/S_{.95})$
WN		0.65	(0.64, 0.67)
Modelo 3	1	0.88	(0.86, 0.89)
Modelo 7	0.7	0.87	(0.86, 0.88)
Modelo 11	0.94	0.66	(0.64, 0.68)

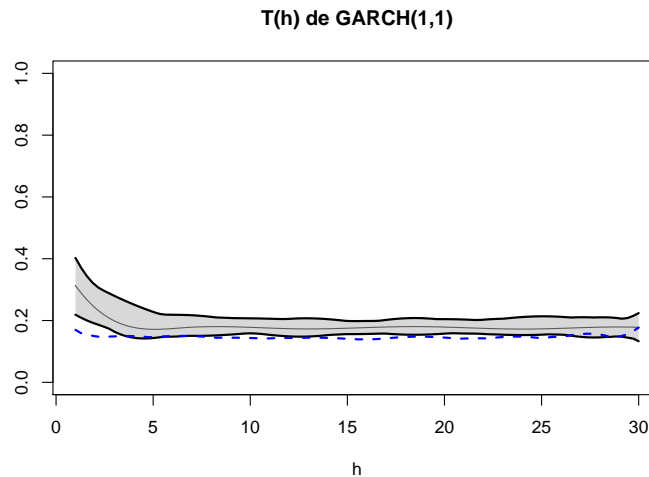


Figura 3.5: La región coloreada representa el 95 % de las curvas más profundas de  $\{T_i(h)\}$  correspondiente al modelo 10. La línea discontinua representa una cota superior de las curvas  $T(h)$  correspondientes al ruido blanco.

Podemos observar que la proporción de rechazo obtenida con la prueba envolvente para las simulaciones con el modelo 1, 4 y 5 es satisfactoria, y en estos casos podemos concluir que se mantiene el nivel de significancia  $\alpha = 0.05$ . Con respecto al modelo 10, se obtiene una proporción de rechazo del 0.66. Esto es debido a lo que se discutió anteriormente; la estructura de dependencia del proceso GARCH es completamente no lineal y la medida de covarianza no detecta tal dependencia por lo cual el modelo ajustado bajo la nula, en teoría debería ser un ruido blanco. A pesar de que el modelo tiene memoria corta, tengamos en cuenta que la hipótesis nula es proceso lineal con memoria corta, obteniendo así una proporción de rechazo de 66 %. Por otro lado, si se observa la curva  $T(h)$  correspondiente a este modelo se puede apreciar que efectivamente corresponde a un proceso de memoria corta por el tipo decaimiento. En la Figura 3.5 se muestra el comportamiento de  $T(h)$  para este modelo; se simularon 2500 series y se estimaron los correspondientes  $T_i(h)$ . Como referencia, la curva discontinua corresponde a la cota superior de la región que representa el 95 % de las curvas más profundas de un ruido blanco. Con esto concluimos que la prueba no tiene potencia alta bajo la alternativa de modelos no lineales con memoria corta en general. Sin embargo, dado que la prueba envolvente proporciona una herramienta visual y ayuda a entender la toma de decisión, se puede concluir de aquí que para el

caso del modelo 10 se tiene un proceso de memoria corta.

Por otro lado, los resultados del estadístico R/S indican que estos procesos son de memoria larga excepto para los modelos 4 y 10, aunque es contradictorio para el modelo 4 por (3.4). Cabe señalar que estos resultados se deben interpretar con precaución.

Ahora, veamos los resultados correspondientes a los modelos no estacionarios; modelos 3, 7 y 11. Estos modelos pertenecen a la hipótesis alternativa. Entonces, la proporción de rechazo representa la potencia de la prueba. La Tabla 3.2 presenta dichos resultados. La proporción de rechazo para los modelos 3 y 11 son satisfactorios. Observemos que para el modelo 3 se obtiene una proporción de rechazo de uno, esto es razonable pues la entropía de una caminata aleatoria decrece “muy lento” y es significativamente diferente a la entropía de procesos con memoria corta, de igual manera para el modelo 11, concluyendo que la prueba tiene buena potencia en estos casos. Para el modelo 7 se obtiene una proporción de rechazo del 70 %. En este caso la prueba no tiene buena potencia. Esto puede ser debido a la gran variabilidad en  $T(h)$  para este modelo. Con respecto a los valores del estadístico R/S, tenemos que para el modelo 11, los valores obtenidos corresponden a un proceso de memoria corta lo cual es errónea. Para los modelos 3 y 7 podemos concluir que acierta en la detección de persistencia.

Finalmente, la Tabla 3.3 presenta los resultados correspondientes a los modelos estacionarios con memoria larga; modelos 2, 6, 8 y 9. Para el modelo lineal 2, el resultado no es lo que uno esperaría (debería estar cerca de uno), obteniendo una proporción de rechazo del 76 % para el caso  $d = 0.2$  y del 93 % para el caso  $d = 0.4$ . Concluimos que en este caso la prueba de hipótesis tiene menos potencia cuando  $d = 0.2$ , que corresponde a un proceso de memoria larga pero cerca del caso de memoria corta.

El resto de los modelos; 6, 8, 9 representan procesos no lineales. Observemos que para el modelo 6, obtenemos una proporción de rechazo del 80 % cuando  $d = 0.2$  y del 98 % cuando  $d = 0.4$ . Podemos concluir que en este caso la prueba tiene buen desempeño. Con respecto a los modelos 8 y 9, tenemos que la proporción de rechazo es menor comparado con todos los casos anteriores bajo la hipótesis alternativa. Obtenemos una proporción de rechazo aproximadamente del 50 % cuando  $d = 0.2$ . Este porcentaje mejora de acuerdo al incremento del valor de  $d$ , obteniendo un rechazo alrededor del 85 % cuando  $d = 0.4$ . La baja potencia en este caso podría ser debido a que la estructura de memoria larga es interrumpida cada vez que  $X_t \leq 0$ , y esto hace más difícil la cuantificación de la memoria larga.

Los resultados del estadístico R/S indican que todos estos modelos son de memoria larga. Como hemos dicho anteriormente, estos resultados se deben tratar con

Tabla 3.3: Resultados de las proporciones de rechazo y valores del estadístico R/S para procesos de memoria larga.

	Env test	$\overline{R/S}$	$(R/S_{.05}, R/S_{.95})$
WN		0.65	(0.64, 0.67)
$d = 0.2$			
Modelo 2	0.76	0.8	(0.79, 0.82)
Modelo 6	0.8	0.77	(0.75, 0.79)
Modelo 8	0.49	0.78	(0.76, 0.81)
Modelo 9	0.55	0.79	(0.77, 0.82)
$d = 0.3$			
Modelo 2	0.87	0.82	(0.81, 0.83)
Modelo 6	0.93	0.79	(0.76, 0.82)
Modelo 8	0.83	0.79	(0.74, 0.83)
Modelo 9	0.72	0.8	(0.76, 0.84)
$d = 0.4$			
Modelo 2	0.93	0.84	(0.83, 0.85)
Modelo 6	0.98	0.81	(0.73, 0.86)
Modelo 8	0.87	0.8	(0.7, 0.85)
Modelo 9	0.84	0.81	(0.71, 0.85)

cierta moderación. Otra manera de interpretar estos resultados es tomar como referencia los valores correspondientes del modelo 1 para procesos de memoria corta; (0.75, 0.78). Con este enfoque todos los modelos bajo la nula son bien detectados. No así para los modelos bajo la alternativa, por ejemplo los modelos 6, 8, 8 y 11 incluyendo los casos cuando  $d = 0.4$ .

La Figura 3.6 muestra un ejemplo de como se visualiza la prueba envolvente. Del lado izquierdo se simuló un modelo con memoria corta, podemos observar que el correspondiente  $T_1(h)$  decae rápido y pertenece dentro de la banda bajo  $H_0$ . Del lado derecho corresponde a un modelo de memoria larga, en este caso observamos que  $T_1(h)$  decrece lento y no pertenece a la banda obtenida bajo  $H_0$ .

En general, podemos concluir que la prueba envolvente es más certera comparada con el estadístico R/S. Tal como hemos mencionado anteriormente, la prueba envolvente también proporciona una herramienta visual que ayuda en la toma de decisiones, viendo el comportamiento del decaimiento de  $T(h)$  adicionalmente a la

información si la curva  $T(h)$  pertenece o no la región en cuestión.

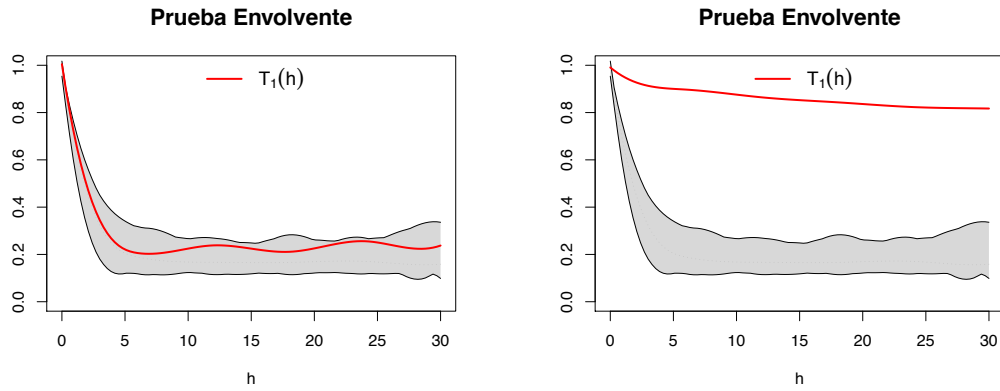


Figura 3.6: Lado izquierdo:  $T_1(h)$  corresponde a un modelo con memoria corta. Lado derecho:  $T_1(h)$  corresponde a un modelo con memoria larga.

# Capítulo 4

## Estimación Robusta de FAR(1) Basada en Profundidades

En este capítulo se propone un método de estimación robusta para el operador de coeficiente en el modelo FAR(1). Primero, hacemos una breve revisión de los métodos comunes para la estimación del operador de coeficiente y luego presentamos en detalle el método propuesto en [Martínez-Hernández et al. \(2017a\)](#).

### 4.1. Métodos de estimación para el modelo FAR(1)

Sea  $\mathbf{X} = \{X_n(t); t \in D, n \in \mathbb{Z}\}$  una sucesión de variables aleatorias funcionales, se dice que  $\mathbf{X}$  es estacionaria si  $\mathbb{E}(\|X_n\|^2) < \infty$ ,  $n \in \mathbb{Z}$ , la esperanza no depende de  $n$ ,  $\mathbb{E}(X_n) = \mu$  y el operador de covarianza satisface  $\langle \Gamma_{X_{n+h}, X_{m+h}}(z), y \rangle = \langle \Gamma_{X_n, X_m}(z), y \rangle$  para  $m, n \in \mathbb{Z}$ ,  $z, y \in \mathbb{H}$ .

En particular si  $\mu = 0$ , el operador de covarianza no depende de  $n$ ,  $\Gamma_X = \Gamma_{X_n}$  y el operador de covarianza cruzada entre  $X_n$  y  $X_m$  es equivalente a cero para  $n \neq m$ ,  $\Gamma_{X_m, X_n} = 0$ , decimos que  $X_n(t)$  es un ruido blanco funcional y lo denotaremos como  $\{W_n(t), n \in \mathbb{Z}\}$

Recordemos que una serie de tiempo funcional FAR(1) es definida como una sucesión de variables aleatorias funcionales  $\{X_n(t); t \in D, n \in \mathbb{Z}\}$  en  $\mathbb{H}$  tal que

$$(X_n - \mu)(t) = \Phi(X_{n-1} - \mu)(t) + W_n(t), \quad (4.1)$$

donde  $\Phi$  es un operador en  $\mathcal{B}_{\mathbb{H}}$ ,  $\mathbb{E}(X_n) = \mu$ , y  $\{W_n(t)\}$  es un ruido blanco funcional ([Bosq, 2000](#)). Sin pérdida de generalidad, supondremos que  $\mu = 0$ .

Un problema crucial es la estimación de  $\Phi(z)$ . Si el modelo FAR(1) se satisface y considerando que  $\Gamma_{X_n, W_n} \equiv 0$ , entonces para toda  $z, y \in \mathbb{H}$  se tiene que

$$\begin{aligned} \langle \Gamma_{X_n, X_{n+1}}(z), y \rangle &= \langle \mathbb{E}(\langle X_n, z \rangle X_{n+1}), y \rangle \\ &= \mathbb{E}(\langle X_n, z \rangle \langle X_{n+1}, y \rangle) \\ &= \mathbb{E}(\langle X_n, z \rangle \langle \Phi(X_n) + W_n, y \rangle) \\ &= \mathbb{E}(\langle X_n, z \rangle \langle X_n, \Phi^*(y) \rangle) + \mathbb{E}(\langle X_n, z \rangle \langle W_n, y \rangle) \\ &= \langle \Gamma_{X_n}(z), \Phi^*(y) \rangle + \langle \Gamma_{X_n, W_n}(z), y \rangle \\ &= \langle \Phi \Gamma_{X_n}(z), y \rangle, \end{aligned}$$

donde  $\Phi^*$  denota el operador adjunto de  $\Phi$ . Luego, si la serie de tiempo funcional es estacionaria, tenemos que

$$\Gamma_1 = \Phi \Gamma_0, \quad (4.2)$$

y si  $\Gamma_0$  es invertible, entonces  $\Phi = \Gamma_1 \Gamma_0^{-1}$ . Sin embargo,  $\Gamma_0^{-1}$  no es acotado, notemos que por el teorema de Mercer,  $\Gamma_0^{-1} = \sum_{j=1}^{\infty} \lambda_j^{-1} v_j \otimes v_j$ , donde  $v_j$  son las autofunciones de  $\Gamma_0$  y  $\lambda_j$  los autovalores correspondientes. Dado que  $\Gamma_0$  es un operador nuclear (Bosq, 2000), implica que los autovalores tienden a cero, en consecuencia  $\lambda_j^{-1} \rightarrow \infty$ . Existen varios trabajos de investigación enfocados en obtener una versión de  $\Gamma_0^{-1}$  acotada y así obtener un estimador de  $\Phi$ . El método más usado es el de análisis de componentes principales (PCA), que consiste en considerar solamente los primeros  $k$  autofunciones correspondientes a los autovalores más grandes y en consecuencia definir una aproximación de la inversa de  $\Gamma_0$  como

$$\Gamma_0^{-1} \approx \sum_{j=1}^k \lambda_j^{-1} v_j \otimes v_j. \quad (4.3)$$

Una crítica a esta aproximación es que no se puede ignorar el resto de los autovalores. Otra propuesta se puede encontrar en Cardot et al. (1999) la idea es proyectar los datos en el espacio generado por un número finito de autofunciones de  $\Gamma_0$  y así obtener una inversa de  $\Gamma_0$  acotada. En Bosq (2000) se usa la misma idea anterior para predecir el modelo FAR(1), a saber, el estimador propuesto para  $\Phi$  es

$$\hat{\Phi} = \pi^k \hat{\Gamma}_1 \hat{\Gamma}_0^{-1} \pi^k,$$

donde  $\pi^k$  es el operador de proyección sobre el espacio generado por  $\{\hat{v}_1, \dots, \hat{v}_k\}$  y  $\hat{\Gamma}_0^{-1}$  es como en (4.3) con las autofunciones empíricas. En general, se puede aplicar cualquier método de regularización para calcular la inversa, i.e., calcular  $(\Gamma_0 + \alpha_n)^{-1}$  en vez de  $\Gamma_0^{-1}$  (ver Crambes and Mas, 2013).



Una clase importante de operadores son los operadores integrales. Sea  $\beta(t, s)$  una función medible en  $\mathbb{R}^2$ . El operador integral  $\Phi$  con kernel  $\beta(t, s)$  es definido como

$$\Phi(z)(s) = \int \beta(t, s)z(t) dt, \quad z \in \mathbb{H}.$$

Los operadores integrales representan una gran clase de operadores acotados y pueden aproximar a otros operadores. Como podemos observar los operadores integrales toman en cuenta la estructura funcional de las covariables y la estructura interna de cada covariable funcional. También, se puede pensar como una generalización del modelo de regresión en el caso multivariado, ya que la integral se puede ver como una suma infinita.

Asumiendo que el operador  $\Phi$  es un operador integral, el modelo FAR(1) en 4.1 se reescribe como

$$X_n(s) = \Phi(X_{n-1})(s) + W_n(s) = \int_D \beta(t, s)X_{n-1}(t) dt + W_n(s), \quad (4.4)$$

y se puede ver que el estimador de  $\beta(t, s)$  usando la versión empírica de  $\Gamma_1$  en (4.2) y el enfoque PCA para aproximar la inversa (4.3), es

$$\hat{\beta}(t, s) = \frac{1}{N-1} \sum_{l=1}^{N-1} \sum_{j=1}^k \sum_{i=1}^k \hat{\lambda}_j^{-1} \langle X_l, \hat{v}_j \rangle \langle X_{l+1}, \hat{v}_i \rangle \hat{v}_j(s) \hat{v}_i(t), \quad (4.5)$$

donde  $\hat{v}_i(t)$  es autofunción de  $\hat{\Gamma}_0$ . La Figura 4.1 ilustra los resultados obtenidos usando (4.5) con diferentes valores de  $k$  (número de componentes principales) y considerando un tamaño de muestra  $N = 1000$ . Podemos observar que  $\hat{\beta}(t, s)$  no es un buen estimador en el sentido de que no aproxima bien el kernel  $\beta(t, s)$  original. El comportamiento de los estimadores no mejora cuando aumenta  $k$ , al contrario empeora, debido a que valores pequeños de  $\hat{\lambda}_j$  se convierten en valores grandes en la estimación. Los estimadores tampoco mejoran si aumenta el tamaño de muestra  $N$ . En consecuencia, este enfoque no es de utilidad cuando se quiere explicar la dinámica del fenómeno a través de  $\beta(t, s)$ , y en muchas aplicaciones es de interés saber de que manera las observaciones anteriores afectan a las actuales.

Bajo el supuesto del operador integral, otro enfoque para la estimación de  $\Phi$  es a través de  $\beta(t, s)$ , usando el estimador de mínimos cuadrados (LS). En este enfoque también aparece el problema de la inversión, por lo cual es necesario algún tipo de regularización. En Ramsay and Dalzell (1991) se propone un estimador de mínimos cuadrados penalizados (PeLS). En Ramsay and Silverman (2005) se discuten varios enfoques, incluyendo el supuesto de bases de funciones finitas y regularización. En

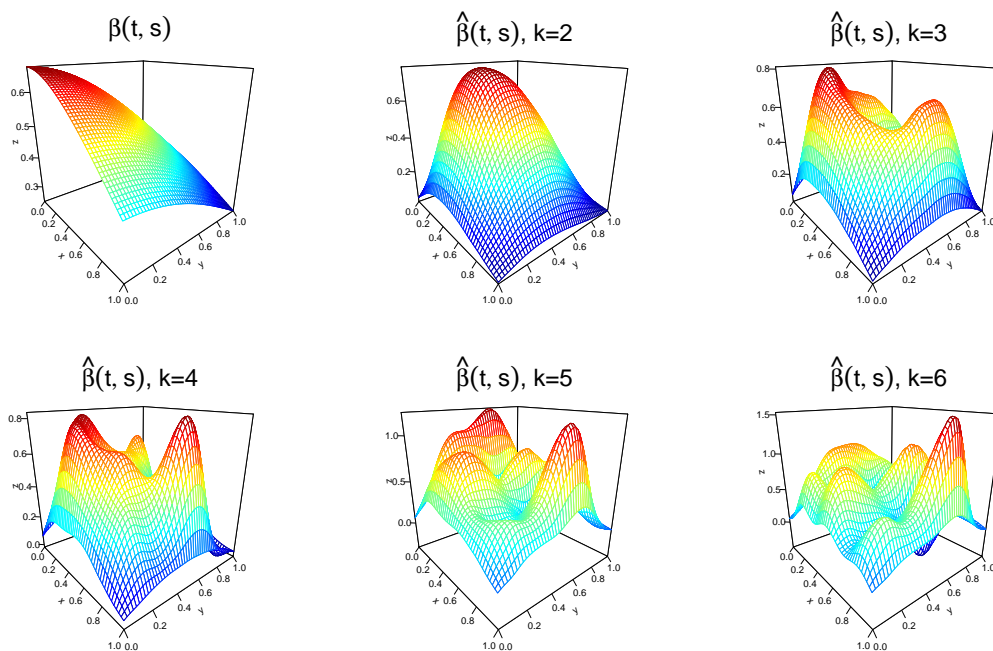


Figura 4.1: Estimadores de  $\beta(t, s)$  usando (4.2) y con aproximación de  $\Gamma_0^{-1}$  como en (4.3), tomando en cuenta diferentes valores de  $k$ .

Wu and Müller (2011) se calcula el estimador usando mínimos cuadrados ponderados. Ivanescu et al. (2015) extiende la idea de regresión funcional penalizado usando bases P-splines para la regularización y en Scheipl and Greven (2016) se puede encontrar sobre la identificabilidad de este método.

La Figura 4.2 ilustra los estimadores obtenidos usando el método PeLS y LS definidos en (4.7) con  $w_{i,i-1} = 1$  para toda  $i$ . El método LS corresponde a  $P(\beta(t, s)) \equiv 0$ . Se considera nuevamente  $N = 1000$ . Observemos que el estimador LS es sin penalización y entonces se tiene el problema de la inversión, obteniendo un estimador muy ruidoso y no satisfactorio<sup>1</sup>. En cambio, en el estimador obtenido del método PeLS,  $\hat{\beta}_{\text{PeLS}}(t, s)$ , se considera una regularización en la inversión a través de la penalización, presentando resultados satisfactorios y acertados para la interpretación en las aplicaciones.

<sup>1</sup>Esto es equivalente a tomar todas las componentes principales en el método PCA.

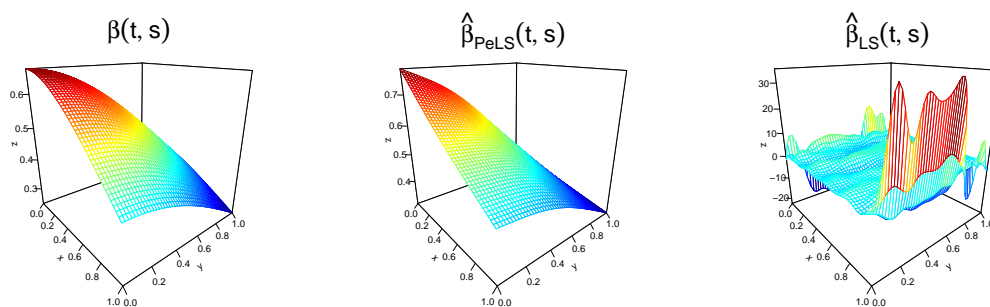


Figura 4.2: Estimadores de  $\beta(t, s)$  usando el método PeLS y LS.

## 4.2. Efecto de datos atípicos en la estimación del modelo FAR(1)

Adicionalmente al problema de la inversión de  $\Gamma_0$ , los datos reales pueden estar contaminados por diferentes tipos de datos atípicos y afectan seriamente la inferencia estadística. Los datos atípicos en series de tiempo son diferentes al caso de observaciones independientes, esto es debido a la dependencia temporal. Existen varios tipos de datos atípicos; en la media, en la varianza, aislados, etc.. Por otro lado, existen varios modelos de contaminación en series de tiempo, aquí asumiremos el modelo por reemplazo. Sea  $X_n$  el proceso estocástico del cual se observa la trayectoria sin contaminación, sea  $Z_n$  un proceso Bernoulli tal que  $\mathbb{P}(Z_n = 0) = 1 - \varepsilon$  con  $\varepsilon \geq 0$  y  $V_n$  el proceso generador de datos atípicos. Se supone que se observa

$$Y_n(t) = X_n(t) + Z_n V_n(t),$$

con  $V_n$  independiente de  $X_n$  y  $X_n$  satisface el modelo FAR(1). El efecto de este tipo de datos solo es en el tiempo  $n$  tal que  $Z_n = 1$  y no contribuye en la observación  $X_{n+1}$ <sup>2</sup> (ver Maronna et al., 2006, Chapter 8).

Para ilustrar el efecto de este tipo de contaminación en los datos, simulamos  $\{X_i(s); s \in [0, 1], i = 1, \dots, 500\}$ , del modelo FAR(1) con kernel  $\beta(t, s) = c \exp\{-(t^2 + s^2)/2\}$ , donde  $c$  es tal que  $\|\Phi\|_{\mathcal{S}} = \{\int \int \beta^2(t, s) dt ds\}^{1/2} = 0.5$ . Fijamos 5 observaciones a contaminar de las 500 simuladas, sumando una constante  $\zeta$ ; ver Modelo 2 en sección 4.5.1. Luego, estimamos  $\beta(t, s)$  para cada muestra simulada

<sup>2</sup>No son como los datos atípicos en la innovación que tiene un efecto no solo en el tiempo que se observa si no también en las observaciones consecuentes.

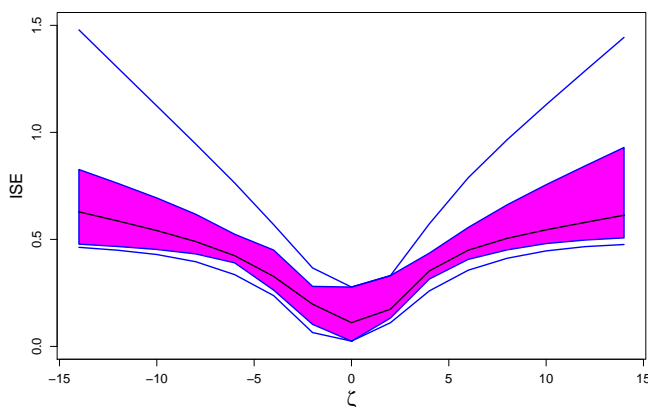


Figura 4.3: Boxplot funcional de los valores de ISE como función de la magnitud de contaminación  $\zeta$  usando el modelo de contaminación Modelo 2, en la estimación de  $\beta(t, s)$ .

usando el estimador PeLS,  $\hat{\beta}_{\text{PeLS}}$  (definido en sección 4.5), y calculamos el error cuadrático integrado (ISE) (definido en (4.18)) como una función de la magnitud de perturbación  $\zeta$ . La Figura 4.3 describe el comportamiento del error de estimación, podemos observar que este incrementa cuando la magnitud de la perturbación  $\zeta$  crece. Otra propiedad que se ve afectada es la norma del operador  $\Phi$  correspondiente, bajo datos contaminados la norma resulta ser más grande que la verdadera, .5 en este caso, y más aún puede llegar a ser mayor que 1 indicando erróneamente que el proceso es no estacionario, en la sección 4.5 se muestran estos resultados.

Una manera de obtener un estimador robusto de  $\Phi$  en el modelo FAR(1) es usando autofunciones robustas en el enfoque de PCA. Algunos trabajos de investigación en autofunciones robustas son Locantore et al. (1999), Gervini (2008), Bali et al. (2011), Lee et al. (2013), y Dürre et al. (2016). Sin embargo, las ideas anteriores consisten en realizar alguna transformación de los datos que afecta los valores y/o los órdenes de los autovalores y que puede cambiar la aproximación de  $\Gamma_0$  al truncarla (Dürre et al., 2016).

Otro enfoque es cambiar la norma en la función a optimizar, considerando por ejemplo usar la norma  $L_1$  o alguna modificación de la norma  $L_2$ . Más general, se puede obtener un estimador robusto usando métodos como M-estimador y S-estimador. En Maronna and Yohai (2013) se propone un estimador robusto en regresión funcional con respuesta escalar usando bases splines y MM-estimador. En Gervini (2012) se

propone estimar  $\beta(t, s)$  para el modelo de regresión funcional, expresando los datos en términos de autofunciones robustas y usando M-estimador.

En la práctica es difícil identificar datos atípicos ya que estos pueden manifestarse de diferentes maneras, es así que mucho métodos han sido desarrollados con el propósito de detectarlos. Un enfoque común es usar el concepto de profundidad, [Sun and Genton \(2011\)](#) extendieron la idea del boxplot a las variables aleatorias funcionales, *boxplot funcional*, esto está basado en la profundidad de banda modificada ([López-Pintado and Romo, 2009](#)) y puede también estar basado en otras definiciones de profundidades. El boxplot funcional detecta eficientemente datos atípicos provenientes de un desplazamiento en la media, inclusive si los datos son dependientes a través del tiempo ([Sun and Genton, 2012](#)). Otras ideas para detección de datos atípicos en datos funcionales se pueden encontrar en [Hyndman and Shang \(2010\)](#), [Hubert et al. \(2015\)](#), y [Arribas-Gil and Romo \(2015\)](#). En [Dai and Genton \(2017\)](#) se propone una noción de medida de cercanía direccional como una medida de centralidad para datos funcionales multivariados, esto permite ordenar las curvas de acuerdo a las magnitudes y las formas de cada curva. Nuestra propuesta está basada en detectar observaciones atípicas usando boxplot funcional y ordenando las curvas con las mismas ideas que en [Dai and Genton \(2017\)](#). Al mismo tiempo usamos las profundidades para ponderar las curvas y en consecuencia obtener un estimador robusto de  $\beta(t, s)$  usando mínimos cuadrados, con la ventaja de que es computacionalmente fácil de calcular comparado con otros métodos.

En esta tesis se propone un estimador robusto para  $\beta(t, s)$  bajo la presencia de datos atípicos por reemplazo. El método propuesto considera el concepto de profundidad como una medida de centralidad para definir pesos en el método PeLS. Ya que el uso de la profundidad en este contexto resulta natural, es decir, las curvas que están alejadas del dato más central son naturalmente ponderadas por su respectiva profundidad, el cual resulta ser un valor más pequeño que uno. Para la identificación de las observaciones atípicas, usamos el boxplot funcional basado en la profundidad correspondiente a la medida de cercanía con el propósito de tomar en cuenta varios aspectos de la curva al momento de ordenarlas.

### 4.3. Estimación vía mínimos cuadrados

Sea  $\{X_n(t)\}$  una serie de tiempo funcional que satisface el modelo FAR(1). El siguiente supuesto hace que el estimador del operador de coeficiente  $\Phi$  sea identificable.

**Suposición 4.1** *El operador de covarianza de la serie de tiempo funcional,  $\Gamma_0 =$*

$\mathbb{E}(X_1 \otimes X_1)$  satisfacen  $\text{Ker}\Gamma_0 = \{0\}$ .

También, supondremos que el operador de coeficiente es un operador integral.

**Suposición 4.2** *El operador de coeficiente  $\Phi$  es un operador integral en  $B_{\mathbb{H}}$  (acotado y lineal) con kernel  $\beta(t, s)$ , esto es*

$$\Phi(z)(s) = \int_D \beta(t, s)z(t)dt, \quad z \in \mathbb{H}.$$

En este caso se considera la norma de operadores de Hilbert-Schmidt. La siguiente condición sobre el operador integral hace que la serie de tiempo sea estacionaria (ver [Bosq, 2000](#))).

**Suposición 4.3** *El operador integral es tal que*

$$\|\Phi\|_S = \left\{ \int_D \int_D \beta^2(t, s) dt ds \right\}^{1/2} < 1.$$

Denotemos por  $\boldsymbol{\psi}(s) = \{\psi_1(s), \dots, \psi_{l_1}(s)\}^T$  y  $\boldsymbol{\phi}(s) = \{\phi_1(s), \dots, \phi_{l_2}(s)\}^T$  dos conjuntos de bases que están definidos sobre el dominio  $D$ . Se supondrá que la observación  $X_n(s)$  puede ser expresada en términos de la base  $\boldsymbol{\psi}(s)$ , es decir,  $X_n(s) = \mathbf{a}_n^T \boldsymbol{\psi}(s) = \sum_{i=1}^{l_1} a_{n,i} \psi_i(s)$  y la observación  $X_{n-1}(s)$  puede ser expresada en términos de la base  $\boldsymbol{\phi}(s)$ , esto es,  $X_{n-1}(s) = \mathbf{b}_{n-1}^T \boldsymbol{\phi}(s) = \sum_{i=1}^{l_2} b_{n-1,i} \phi_i(s)$ , para toda  $n = 2, \dots, N$ .

Con el supuesto de que el operador  $\Phi$  es un operador integral, la estimación se reduce a estimar  $\beta(t, s)$ , ya que  $\Phi$  está determinado por el kernel  $\beta(t, s)$ . Para ello, vamos a aproximarla por una superficie que resulta de un producto tensorial de bases B-splines.

Sean  $\mathbb{S}_1 = \text{span}\{\nu_1, \dots, \nu_{k_1}\}$ , y  $\mathbb{S}_2 = \text{span}\{\eta_1, \dots, \eta_{k_2}\}$ , donde  $\boldsymbol{\nu} = \{\nu_1, \dots, \nu_{k_1}\}^T$  y  $\boldsymbol{\eta} = \{\eta_1, \dots, \eta_{k_2}\}^T$  son bases de funciones B-splines. El espacio lineal generado por  $\{\nu_i(t)\eta_j(s)\}$  se denotará por  $\mathbb{S}_1 \otimes \mathbb{S}_2$ . Entonces  $\beta(t, s)$  es aproximado con una superficie que pertenece al espacio  $\mathbb{S}_1 \otimes \mathbb{S}_2$ .

**Suposición 4.4** *El kernel  $\beta(t, s)$  es una superficie tensorial definida como*

$$\beta(t, s) = \boldsymbol{\nu}^T(t) \boldsymbol{\Theta} \boldsymbol{\eta}(s), \quad (4.6)$$

donde  $\boldsymbol{\Theta}$  es la matriz de coeficientes de tamaño  $k_1 \times k_2$ .

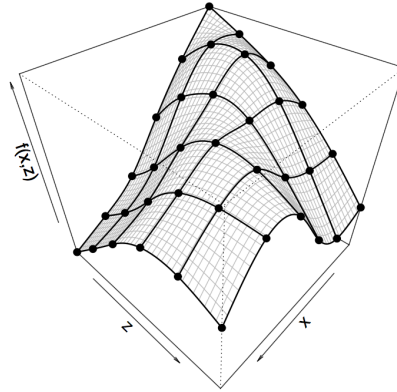


Figura 4.4: Idea de aproximación de una superficie usando producto tensorial B-splines, los puntos representan los nodos. Imagen recuperada de [Wood \(2006\)](#).

La Figura 4.4 ilustra la idea de una superficie de producto tensorial, en cada rectángulo definido por los nodos en las coordenadas, se aproxima la superficie usando B-splines. Si fijamos la coordenada  $z$ , tenemos una curva  $\beta(z_0, x) = \sum_{j=1}^{k_2} \theta_j(z_0)\eta_j(x)$ , en consecuencia, si variamos  $z_0$  se obtiene una superficie, pero si los coeficientes  $\theta_j(z)$  son a la vez curvas en  $\mathbb{S}_1$ , es decir,  $\theta_i(z) = \sum_{j=1}^{k_1} \theta_{i,j}\nu_j(z)$ , entonces tenemos que  $\beta(z, x) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \theta_{i,j}\nu_j(z)\eta_j(x)$ .

Para obtener una estimación de la superficie que sea diferenciable o que tenga cierta propiedad específica, es necesario aplicar un método de regularización o penalización, esto también será de gran ayuda para el problema de inversión que aparece en la estimación de dicha superficie. Por lo tanto, se desarrolla un método de mínimos cuadrados penalizados usando la representación tensorial de  $\beta(t, s)$  para obtener un estimador.

Adicionalmente, supondremos que la serie de tiempo funcional observada está generada por el modelo  $Y_n(t) = X_n(t) + Z_n v(t, \zeta)$ , donde  $Z_n$  es un proceso Bernoulli tal que  $\mathbb{P}(Z_n = 1) = 1 - \varepsilon$ ,  $\varepsilon \geq 0$ ,  $X_n$  es un proceso FAR(1),  $v(\zeta, t)$  es el generador de datos atípicos con tamaño de contaminación  $\zeta$  tal que  $v(\zeta, t) = O(\zeta)$  y es independiente de  $X_n$  y  $Z_n$ .

**Notación:** Denotemos por  $\mathbf{J}_{\nu\eta} = \int \nu(s)\eta^T(s) ds$  la matriz del producto interno de  $\nu(s)$  y  $\eta(s)$ .

### 4.3.1. Construcción del estimador robusto basado en profundidades

Proponemos un estimador de  $\beta(t, s)$  usando mínimos cuadrados ponderados y penalizados, con el objeto de reducir el efecto de los datos que son considerados como atípicos. El uso de la noción de profundidad de curvas resulta natural en este contexto, así la ponderación será función de ella. Como hemos explicado en capítulos anteriores, la profundidad de una variable aleatoria alcanza su máximo valor en el centro de los datos (mediana), decreciendo cuando estos datos se alejan del centro. De esta manera, curvas que están alejadas de la curva más central son naturalmente ponderados por su respectiva profundidad con respecto a la muestra completa.

Sea  $d(X_j(t))$  la profundidad de  $X_j(t)$  y denotemos por  $w_{i-1,i} := W(d(X_{i-1}(t)), d(X_i(t)))$  la función de ponderación basada en la profundidad. Entonces,  $\hat{\beta}(t, s)$  es obtenido de la solución del siguiente problema de minimización,

$$\hat{\Theta} = \min_{\Theta} \left[ \int_D \sum_{i=2}^N w_{i-1,i} \left\{ X_i(t) - \int_D X_{i-1}(s) \boldsymbol{\nu}^T(s) \Theta \boldsymbol{\eta}(t) ds \right\}^2 dt + P(\beta(t, s)) \right], \quad (4.7)$$

donde  $P(\beta(t, s))$  representa una penalización cuadrática marginal que puede ser expresada como

$$P(\beta(t, s)) = \lambda_t \text{vec}(\Theta)^T \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_t \text{vec}(\Theta) + \lambda_s \text{vec}(\Theta)^T \mathbf{P}_s \otimes \mathbf{J}_{\nu\nu} \text{vec}(\Theta), \quad (4.8)$$

con  $\lambda_t$  y  $\lambda_s$  parámetros de ajuste que controlan el grado de la penalización,  $\mathbf{P}_t$  y  $\mathbf{P}_s$  son matrices que representan las penalizaciones marginales en las coordenadas  $t$  y  $s$  respectivamente, en el dominio de la superficie  $\beta(t, s)$ . Para esta representación de  $P(\beta(t, s))$  se puede consultar [Wood \(2006\)](#).

Sean  $\mathbf{A}^T$  y  $\mathbf{B}^T$  matrices de coeficientes con filas  $\mathbf{a}_i^T$  y  $\mathbf{b}_j^T$ , respectivamente, y sea  $\mathbf{W}$  una matriz diagonal de  $(N-1) \times (N-1)$  con elementos  $w_{i-1,i}$ ,  $i = 2, \dots, N$ .

Por lo anterior, observemos que la estimación se reduce a estimar la matriz de coeficientes  $\Theta$ . Usando la ecuación (4.2) y la versión empírica del operador de covarianza, la matriz  $\hat{\Theta}$  debe satisfacer

$$(\mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{B} \mathbf{W} \mathbf{B}^T \mathbf{J}_{\phi\nu} + \lambda_t \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_t + \lambda_s \mathbf{P}_s \otimes \mathbf{J}_{\nu\nu}) \text{vec}(\hat{\Theta}) = \text{vec}(\mathbf{J}_{\phi\nu}^T \mathbf{B} \mathbf{W} \mathbf{A}^T \mathbf{J}_{\psi\eta}^T). \quad (4.9)$$

De esta manera, podemos definir un estimador robusto de  $\beta(t, s)$  como  $\hat{\beta}(t, s) = \boldsymbol{\nu}^T(t) \hat{\Theta} \boldsymbol{\eta}(s)$ , donde  $\hat{\Theta}$  es la solución de la ecuación (4.9). Observemos que esta estimación depende de la ponderación a definir que a la vez depende de la definición de profundidad. A continuación detallamos dicha ponderación. La definición de la función de ponderación consiste en dos etapas. Primero identificamos curvas que son



posiblemente atípicas, usando boxplot funcional basado en la profundidad obtenida por la cercanía direccional, definida en (2.7). Luego, las curvas catalogadas como atípicas son ponderadas, usando los valores de dicha profundidad.

Sea  $\mathbf{X} = \{X_i(t); t \in D, i = 1, \dots, N\}$  una muestra de curvas y sea  $X_{[i]}(t)$  la curva con el  $i$ -ésimo valor de profundidad más grande, así  $X_{[1]}(t)$  representa la curva más profunda. Recordemos que el boxplot funcional se define tomando la banda que define el 50 % de las curvas más profundas.

$$C_{0.5} = \left\{ (t, y(t)) : \min_{r=1, \dots, \lceil n/2 \rceil} X_{[r]}(t) \leq y(t) \leq \max_{r=1, \dots, \lceil n/2 \rceil} X_{[r]}(t) \right\}.$$

Luego, esta banda es expandida por un factor de 1.5 (el factor puede ser modificado), obteniendo así una banda que define la frontera para las curvas atípicas. Cualquier curva que no esté completamente contenida en la banda es catalogada como potencialmente atípica (ver Sun and Genton, 2011). Denotemos por  $I_B \subseteq \{1, \dots, N\}$  el subconjunto de índices tal que  $X_k(t)$ , con  $k \in I_B$ , no es una curva atípica de acuerdo al boxplot funcional.

Proponemos utilizar la profundidad basada en la medida de cercanía. Para cada  $t \in D$ , sea  $d\text{SDO}(X_0(t), F_{X(t)})$  definido como

$$d\text{SDO}(X_0(t), F_{X(t)}) = v(t) \frac{|X_0(t) - \text{median}\{X(t)\}|}{\text{MAD}\{X(t)\}} \quad (4.10)$$

esto es un caso particular de la medida de cercanía direccional de Stahel-Donoho (Dai and Genton, 2017), donde

$$v(t) = \frac{X_0(t) - \text{median}(X(t))}{|X_0(t) - \text{median}\{X(t)\}|},$$

y  $\text{MAD}(X)$  es la desviación absoluta en la media de  $X$ . Entonces, la versión funcional se obtiene integrando sobre todo  $t \in D$ ,

$$\text{FO}(X, F_X) = \int_D d\text{SDO}(X(t), F_{X(t)})^2 dt.$$

Finalmente, usando la relación entre cercanía y profundidad, definimos la siguiente profundidad relacionado con FO,

$$d\text{O}(X, F_X) = \frac{1}{1 + \text{FO}(X, F_X)}. \quad (4.11)$$

Regresando a la estimación de  $\beta(t, s)$ , para la construcción de los pesos en (4.7) hacemos uso de la profundidad  $dO$  y la herramienta boxplot funcional. Denotemos por  $dO_i$  la profundidad correspondiente a la curva  $X_i(t)$  y sea

$$c = \min\{dO_k; k \in I_B\},$$

el mínimo de los valores de las profundidades de la muestra de curvas que no son atípicas.

**Definición 4.1** Sea  $\{X_i(s); s \in D, i = 1, \dots, N\}$  una serie de tiempo funcional FAR(1). Sea  $dO_i$  la profundidad correspondiente a la observación  $X_i(s)$ . Definimos los pesos  $w_{i-1,i}$  como

$$w_{i-1,i} = W(dO_i, dO_{i-1}) = \mathbb{1}_{[c,1]}(d_i) + dO_i dO_{i-1} \mathbb{1}_{[0,c)}(d_i), \quad (4.12)$$

donde  $d_i = \min\{dO_i, dO_{i-1}\}$ ,  $\mathbb{1}_A(x)$  es la función indicadora y  $c$  es el mínimo de los valores  $dO_j$  tal que  $X_j(s)$  no es atípica de acuerdo al boxplot funcional basado en la profundidad  $dO$ .

Observemos que la correspondiente medida de cercanía de  $X_k(t) + v(\zeta, t)$  es  $FO(X_k + v, F_X) = FO(X_k, F_X) + \int v^2(\zeta, t)/MAD\{X(t)\}dt + \int v(\zeta, t)[X_k(t) - \text{median}\{X(t)\}]/MAD\{X(t)\}dt$ . Así,  $FO(X_k + v, F_X)$  tiene mismo orden que  $v^2(\zeta, t)$ . Entonces, el efecto de las curvas atípicas en el problema de optimización (4.7) es removido por sus correspondientes valores de profundidad  $dO(X_k + v, F_X)$ .

**Definición 4.2** Sea  $\{X_i(s); s \in D, i = 1, \dots, N\}$  una serie de tiempo funcional autoregresiva de orden uno, FAR(1). Definamos el DLS estimador como

$$\hat{\beta}_{DLS}(t, s) = \boldsymbol{\nu}^T(t) \hat{\boldsymbol{\Theta}}_{DLS} \boldsymbol{\eta}(s),$$

donde  $\hat{\boldsymbol{\Theta}}_{DLS}$  es la solución al problema de optimización (4.9) con ponderaciones  $w_{i-1,i}$  definidas en (4.12).

El algoritmo 1 presenta los pasos a seguir para obtener el DLS estimador para la serie de tiempo funcional.

### Algoritmo 1:

---

1. Para  $i = 1, \dots, N$ , calcular  $dO_i$  de (4.11).

2. Usar el boxplot funcional basado en  $\{dO_i\}_{i=1}^N$  y fijar  $c = \min\{dO_k; X_k(t) \text{ no es atípica}\}$ .
  3. Calcular  $w_i$  y obtener  $\hat{\Theta}$  como solución a la ecuación (4.9).
- 

Como podemos observar, el estimador propuesto, DLS, reduce el efecto de los posibles datos atípicos en la estimación del kernel  $\beta(t, s)$ , y esto se hace ponderando las curvas con sus respectivos profundidades. Observemos que  $FO(X, F_X)$  puede ser descompuesto como sigue

$$\begin{aligned}
 FO(X, F_X) &= \int_D dSDO(X(t), F_{X(t)})^2 dt \\
 &= \int_D (dSDO(X(t), F_{X(t)}) - MO(X, F_X) + MO(X, F_X))^2 dt \\
 &= \int_D (dSDO(X(t), F_{X(t)}) - MO(X, F_X))^2 dt + MO(X, F_X)^2 \\
 &= MO(X, F_X)^2 + VO(X, F_X),
 \end{aligned}$$

donde

$$MO(X, F_X) = \int_D dSDO(X(t), F_{X(t)}) dt,$$

y

$$VO(X, F_X) = \int_D (dSDO(X(t), F_{X(t)}) - MO(X, F_X))^2 dt.$$

La componente  $MO(X, F_X)$  describe la posición de  $X$  con respecto a la curva mediana y representa datos atípicos en magnitud. Por otro lado  $VO(X, F_X)$  mide el cambio de las curvas una vez extraída la curva mediana, es decir, representa la forma de la curva (como suavidad, oscilaciones, etc.) que es ajena a la estructura general de los datos. Para ver más detalles leer [Dai and Genton \(2017\)](#).

## 4.4. Función de influencia del estimador DLS

### 4.4.1. Función de influencia

Para medir la robustez del estimador propuesto, usamos la Función de Influencia (IF) ([Hampel et al., 1986](#)). El IF es un caso particular de la derivada de Gâteaux. Consideremos una variable aleatoria con distribución  $F$  en  $\mathbb{R}$ . Sea  $x_0 \in \mathbb{R}$  y  $T$  un

funcional definido en el espacio de las distribuciones,  $\mathcal{F} = \{F + \varepsilon(\delta_x - F)\}$  con  $\varepsilon \in [0, 1]$ . IF es definido como

$$\text{IF}(x_0; T, F) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} [T\{(1 - \varepsilon)F + \varepsilon\delta_{x_0}\} - T(F)],$$

donde  $\delta_x$  es la distribución que asigna masa uno a  $x$ . Esto representa la tasa de cambio en el funcional  $T$  como resultado de una pequeña contaminación por otra distribución  $G = \delta_x$ .

Sin embargo, usar la definición  $(1 - \varepsilon)F + \varepsilon\delta_{x_0}$  como contaminación de un proceso estocástico no es adecuado por la dependencia en el tiempo. [Martin and Yohai \(1986\)](#) introdujeron otra definición de IF en el contexto de series de tiempo, una serie de tiempo contaminada es definida como

$$Y_t^\varepsilon = (1 - Z_t^\varepsilon)X_t + Z_t^\varepsilon V_t,$$

donde  $X_t$  es un proceso estacionario,  $Z_t$  es un proceso Bernoulli tal que  $\mathbb{P}(Z_t^\varepsilon = 1) = \varepsilon$  y  $V_t$  es un proceso generador de de datos atípicos. Entonces, el IF para series de tiempo del funcional  $T$  es definido como

$$\text{IF}(F_{X,Z,V}; T) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \{T(F_Y^\varepsilon) - T(F_X)\}, \quad (4.13)$$

donde  $F_Y^\varepsilon$  es la distribución conjunta de  $Y_t$ ,  $F_X$  es la distribución de  $X_t$ ,  $Z_t^\varepsilon$  y  $V_t$ . Observemos que la definición (4.13) es válida para cualquier tipo de variable aleatoria, incluyendo variables aleatorias funcionales, es decir, cuando  $X_t$  y  $V_t$  son procesos estocásticos funcionales.

La siguiente proposición muestra que la correspondiente IF del estimador DLS es acotado, esta propiedad es deseable en estimadores robustos. Sea  $T_{\text{DLS}} : \mathcal{F} \rightarrow \mathbb{S}_1 \otimes \mathbb{S}_2$  el funcional  $T_{\text{DLS}}(F_X) = \beta_{\text{DLS}}$  donde  $\beta_{\text{DLS}}(t, s) = \boldsymbol{\nu}^T(t)\boldsymbol{\Theta}_{\text{DLS}}\boldsymbol{\eta}(s)$ ,  $\boldsymbol{\Theta}_{\text{DLS}}$  satisface

$$\begin{aligned} \boldsymbol{\Theta}_{\text{DLS}} = \arg \min_{\boldsymbol{\Theta}} \left[ \int_D \mathbb{E}_{F_X} \left\{ W(X_{n-1}, X_n) \left( X_n(t) - \int_D X_{n-1}(s) \boldsymbol{\nu}^T(s) \boldsymbol{\Theta} \boldsymbol{\eta}(t) ds \right)^2 \right\} dt \right. \\ \left. + P(\beta(t, s)) \right], \end{aligned}$$

y  $P(\beta(t, s))$  es como en (4.8). Notemos que  $\boldsymbol{\Theta}_{\text{DLS}}$  determina los valores de  $T_{\text{DLS}}$ , y sin lugar a confusión, denotaremos por  $\text{vec}(T_{\text{DLS}})$  la correspondiente vectorización de  $\boldsymbol{\Theta}_{\text{DLS}}$ , esto es  $\text{vec}(\boldsymbol{\Theta}_{\text{DLS}})$ .

**Proposición 4.1** Sea  $X_t(s) = \mathbf{b}_t^T \boldsymbol{\phi}(s)$  una serie de tiempo FAR(1) y sea  $V_t(s) = \mathbf{v}_t^T \boldsymbol{\phi}(s)$  el proceso generador de datos atípicos. La función de influencia de  $T_{\text{DLS}}$  está dada por

$$\begin{aligned} & \text{vec}\{\text{IF}_t(F_{X,Z,V}; T_{\text{DLS}})\} \\ &= \left\{ \mathbb{E}_{F_X} \left( \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{b}_t \mathbf{b}_t^T \mathbf{J}_{\phi\nu} + \lambda_t \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_t + \lambda_s \mathbf{P}_s \otimes \mathbf{J}_{\nu\nu} \right)^{-1} \right\} \left\{ \mathbb{E}_{F_{V_t, V_{t+1}}} (w_{t,t+1} \mathbf{v}_t \mathbf{v}_{t+1}^T) \right. \\ & \quad \left. - \mathbb{E}_{F_{V_t}} (w_{t,t+1} \mathbf{v}_t \mathbf{v}_t^T) \text{vec}(\boldsymbol{\Theta}_{\text{DLS}}) \right\}, \end{aligned} \quad (4.14)$$

y es acotada.

Notemos que  $\text{vec}\{\text{IF}_t(F_{X,Z,V}; T_{\text{DLS}})\} = \text{IF}_t(F_{X,Z,V}; \text{vec}(T_{\text{DLS}}))$ , de esta manera, tenemos que (4.14) define la función de influencia de  $T_{\text{DLS}}$ .

Para probar la proposición anterior, primero introducimos el siguiente lema.

**Lema 4.1** Sea  $V_t(\zeta, s) = X_t(s) + v(\zeta, s)$  el proceso generador de datos atípicos tal que  $v(\zeta, s) = O(\zeta)$ ,  $X_t(s)$  es un proceso estacionario y  $\zeta \in \mathbb{R}$ . Entonces, para cada  $t$ ,

$$\text{FO}(V_t, F_X) = O(\zeta^2), \text{ cuando } \zeta \rightarrow \infty.$$

**Demostración.** Para cada  $s$ , sea

$$g(\zeta, s) = \frac{1}{\text{MAD}\{X_1(s)\}} [X_t(s) + v(\zeta, s) - \text{median}\{X_1(s)\}].$$

Entonces,

$$g(\zeta) = \frac{X_t(s) - \text{median}}{\text{MAD}(X_1(s))} + \frac{v(\zeta, s)}{\text{MAD}(X_1(s))} = O(1) + O(\zeta).$$

Así,

$$\text{FO}(V_t, F_X) = \int_D \text{dSDO}(V_t(s), F_X)^2 \text{d}s = \int_D g(\zeta, s)^2 \text{d}s = O(\zeta^2).$$

□

**Demostración. (Proposición 4.1)** Supongamos que  $X_t(s)$  y  $V_t(\zeta, s)$  tienen una representación en bases finita;  $X_t(s) = \mathbf{x}_t^T \boldsymbol{\phi}(s)$  y para cada  $\zeta$ ,  $V_t(\zeta, s) = \mathbf{v}_t^T \boldsymbol{\phi}(s)$ . Sea

$F_X$  y  $F_Y$  la distribución conjunta de  $X_t$  y  $Y_t$ , respectivamente. Sea  $\text{vec}(\Theta_{\text{DLS}})$  tal que  $\beta_{\text{DLS}}(t_1, t_2) = \boldsymbol{\nu}^T(t_1) \Theta_{\text{DLS}} \boldsymbol{\eta}(t_2)$ . Definamos  $\mathcal{X}_{t,t}$ ,  $\mathcal{V}_{t,t}$ ,  $\mathcal{X}_{t,t+1}$ ,  $\mathcal{V}_{t,t+1}$ , y  $\mathcal{P}$  como

$$\begin{aligned}\mathcal{X}_{t,t} &:= \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{x}_t \mathbf{x}_t^T \mathbf{J}_{\phi\nu}, \\ \mathcal{V}_{t,t} &:= \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{v}_t \mathbf{v}_t^T \mathbf{J}_{\phi\nu}, \\ \mathcal{X}_{t,t+1} &:= \text{vec} \left( \mathbf{J}_{\phi\nu}^T \mathbf{x}_t \mathbf{x}_{t+1}^T \mathbf{J}_{\psi\eta}^T \right), \\ \mathcal{V}_{t,t+1} &:= \text{vec} \left( \mathbf{J}_{\phi\nu}^T \mathbf{v}_t \mathbf{v}_{t+1}^T \mathbf{J}_{\psi\eta}^T \right), \\ \mathcal{P} &:= \lambda_{t_1} \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_{t_1} + \lambda_{t_2} \mathbf{P}_{t_2} \otimes \mathbf{J}_{\nu\nu}.\end{aligned}$$

Sabemos que  $\text{vec}(\Theta_{\text{DLS}})$  satisface

$$\mathbb{E}_{F_X}(\mathcal{X}_{t,t} + \mathcal{P}) \text{vec}(\Theta_{\text{DLS}}) = \mathbb{E}_{F_X}(\mathcal{X}_{t,t+1}).$$

Sea  $F_{Y_{t,t+1}}$  la distribución conjunta del proceso bivariado,  $(Y_t, Y_{t+1})$ , entonces la distribución bivariada de la contaminación es

$$F_{Y_{t,t+1}} = (1 - \varepsilon) F_{X_{t,t+1}} + \varepsilon F_{V_{t,t+1}}.$$

Denotemos por  $\text{vec}(\Theta_{\text{DLS}}^\varepsilon)$  el correspondiente estimador usando el proceso  $Y_t^\varepsilon$ . Por lo tanto, la ecuación  $\mathbb{E}_{F_Y}(\mathcal{X}_{t,t} + \mathcal{P}) \text{vec}(\Theta_{\text{DLS}}^\varepsilon) = \mathbb{E}_{F_{Y_{t,t+1}}}(\mathcal{X}_{t,t+1})$  es equivalente a

$$\left\{ (1 - \varepsilon) \mathbb{E}_{F_{X_{t,t+1}}}(\mathcal{X}_{t,t} + \mathcal{P}) + \varepsilon \mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1} \mathcal{V}_{t,t}) \right\} \text{vec}(\Theta_{\text{DLS}}^\varepsilon) = (1 - \varepsilon) \mathbb{E}_{F_{X_{t,t+1}}}(\mathcal{X}_{t,t+1}) + \varepsilon \mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1} \mathcal{V}_{t,t+1}).$$

Diferenciando con respecto a  $\varepsilon$  y evaluando en  $\varepsilon = 0$ :

$$\mathbb{E}_{F_{X_{t,t+1}}}(\mathcal{X}_{t,t} + \mathcal{P}) \frac{\partial}{\partial \varepsilon} \text{vec}(\Theta_{\text{DLS}}^\varepsilon) \Big|_{\varepsilon=0} + \mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1} \mathcal{V}_{t,t}) \text{vec}(\Theta_{\text{DLS}}) = \mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1} \mathcal{V}_{t,t+1}).$$

Observemos que

$$\frac{\partial}{\partial \varepsilon} \text{vec}(\Theta_{\text{DLS}}^\varepsilon) \Big|_{\varepsilon=0} = \text{vec} \left( \frac{\partial}{\partial \varepsilon} \Theta_{\text{DLS}}^\varepsilon \Big|_{\varepsilon=0} \right) = \text{vec}(\text{IF}(F_{X,Z,V}; \Theta_{\text{DLS}})) = \text{IF}(F_{X,Z,V}; \text{vec}(\Theta_{\text{DLS}})).$$

Así,

$$\begin{aligned}\text{IF}(F_{X,Z,V}; \text{vec}(\Theta_{\text{DLS}})) &= \\ &= \left\{ \mathbb{E}_{F_{X_{t,t+1}}}(\mathcal{X}_{t,t} + \mathcal{P}) \right\}^{-1} \left\{ \mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1} \mathcal{V}_{t,t+1}) - \mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1} \mathcal{V}_{t,t}) \text{vec}(\Theta_{\text{DLS}}) \right\}.\end{aligned}$$

La contaminación es solamente en una componente, i.e., la distribución conjunta  $F_{V_{t,t+1}}$  es tal que  $F_{V_{t,t+1}} = \delta_{(V_i(s), X_{t+1}(s))}$ , donde  $\delta_{(v_1, v_2)}(x, y)$  es la distribución que

asigna masa uno a  $(v_1, v_2)$ . Para probar que el correspondiente IF es acotado, notemos que por definición  $w_{t,t+1} = dO(V_t(s), F_X)dO(X_{t+1}(s), F_X)$  y  $dO = 1/(1 + FO)$ , por lo tanto  $w_{t,t+1} = O(1/\zeta^2)$ . Dado que  $\mathcal{V}_{t,t} = \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{v}_t \mathbf{v}_t^T \mathbf{J}_{\phi\nu}$ , y  $F_{V_{t,t+1}} = \delta_{(V_t(s), X_{t+1}(s))}$ , tenemos que

$$\mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1}\mathcal{V}_{t,t}) = w_{t,t+1}\mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{v}_t \mathbf{v}_t^T \mathbf{J}_{\phi\nu} = O(1/\zeta^2)O(\zeta^2) = O(1),$$

y

$$\mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1}\mathcal{V}_{t,t+1}) = w_{t,t+1}\mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{v}_t \mathbf{x}_{t+1}^T \mathbf{J}_{\phi\nu} = O(1/\zeta^2)O(\zeta) = O(1/\zeta).$$

Consecuentemente,

$$\mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1}\mathcal{V}_{t,t+1}) - \mathbb{E}_{F_{V_{t,t+1}}}(w_{t,t+1}\mathcal{V}_{t,t})\text{vec}(\Theta_{\text{DLS}}) = O(1),$$

y concluimos que  $\text{IF}(F_{X,Z,V}; \text{vec}(\Theta_{\text{DLS}}))$  está acotado.  $\square$

#### 4.4.2. Función de influencia empírica

En la práctica, se estima la versión empírica de IF (EIF) para cuantificar la influencia de una observación  $i$  contaminada en la estimación de  $\hat{\beta}(t, s)$ .

Sea  $\mathbf{X} = \{X_i(s); s \in D, i = 1, \dots, N\}$  una serie de tiempo FAR(1) de tamaño  $N$ , y sea  $\hat{\beta}(\mathbf{X})$  el estimador del kernel basado en las observaciones  $\mathbf{X}$ . La función de influencia empírica de  $\hat{\beta}(t, s)$  en el tiempo  $i$  es definido como

$$\text{EIF}_i(\hat{\beta}, \mathbf{X}, \zeta) = N \left\{ \hat{\beta}(\mathbf{X}[i, \zeta]) - \hat{\beta}(\mathbf{X}) \right\}, \quad (4.15)$$

donde  $\mathbf{X}[i, \zeta]$  es igual a  $\mathbf{X}$ , excepto en la  $i$ -ésima curva donde esto es reemplazado por la curva contaminada  $V_i(\zeta, s)$ , con tamaño de contaminación  $\zeta$  y  $s \in D$ ; ver sección 4.5.1 para los modelos de contaminación.

Observar que si no hay dependencia temporal, entonces el EIF es independiente del índice de observación. En el caso de series de tiempo se tiene tal dependencia, y por lo tanto se obtienen diferentes EIF de acuerdo a cuál observación es contaminada.

En este caso cuantificamos el EIF considerando el promedio de  $\text{EIF}_i$  sobre  $i = 1, \dots, N$  y lo denotamos por MEIF,

$$\text{MEIF}_N(\hat{\beta}, \mathbf{X}, \zeta) = \sum_{i=1}^N \left\{ \hat{\beta}(\mathbf{X}[i, \zeta]) - \hat{\beta}(\mathbf{X}) \right\}. \quad (4.16)$$

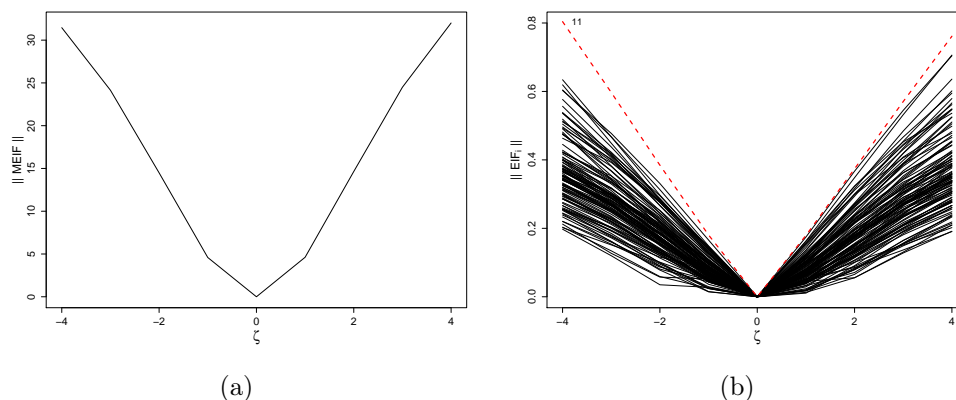


Figura 4.5: (a): Norma de MEIF, (b): hair-plot funcional para el estimador  $\hat{\beta}_{\text{PeLS}}$  con modelo 1 para la contaminación. La observación 11 está representada por líneas discontinuas.

Para visualizar el EIF consideramos el hair-plot. El hair-plot es una herramienta de visualización que fue propuesta por [Genton and Ruiz-Gazen \(2010\)](#) para observaciones en  $\mathbb{R}$ , con el propósito de visualizar las influencias de las observaciones en datos dependientes. Para el caso cuando se tienen variables aleatorias funcionales, graficamos la norma  $\|\text{EIF}_i(\hat{\beta}, \mathbf{X}, \zeta)\|_{\mathcal{S}}$  contra  $\zeta$ . Y así se tiene una visualización similar al de hair-plot en  $\mathbb{R}$ . Similarmente con MEIF, graficamos  $\|\text{MEIF}(\hat{\beta}, \mathbf{X}, \zeta)\|_{\mathcal{S}}$  contra  $\zeta$ .

La Figura 4.5 (a) y Figura 4.5 (b) muestra el hair-plot funcional de  $\hat{\beta}_{\text{PeLS}}(t, s)$  con  $\beta(t, s) = c \exp\{-(t^2 + s^2)/2\}$ , donde  $c$  es tal que  $\|\Phi\|_{\mathcal{S}} = 0.5$ , y el modelo de contaminación es  $V_i(\zeta, s) = X_i(s) + \zeta$ . Observamos que MEIF, vemos que si  $|\zeta|$  incrementa, entonces  $\|\text{MEIF}\|_{\mathcal{S}}$  también incrementa. Con respecto al hair-plot funcional, vemos que  $\|\text{EIF}_i\|_{\mathcal{S}}$  es diferente en cada curva contaminada, en este caso, la contaminación en la curva 11 resulta una observación con la mayor influencia. Más adelante se presenta un estudio de simulación comparando la IF de  $\hat{\beta}_{\text{DLS}}(t, s)$  y la IF de  $\hat{\beta}_{\text{PeLS}}(t, s)$ ;  $\hat{\beta}_{\text{PeLS}}(t, s)$  está definido en la sección 4.5.

## 4.5. Estudio de simulación Monte Carlo

En esta sección, presentamos estudios de simulación para evaluar el desempeño del estimador DLS (definido en la sección 4.3). Consideramos diferentes tipos de



kernel  $\beta(t, s)$  y también diferentes niveles y tipos de contaminación.

Comparamos los resultados de la propuesta con los resultados de los siguientes estimadores.

1. Estimador PeLS:  $\hat{\beta}_{\text{PeLS}}(t, s) = \boldsymbol{\nu}^T(s) \hat{\Theta}_{\text{PeLS}} \boldsymbol{\eta}(t)$ , donde  $\hat{\Theta}_{\text{PeLS}}$  es solución a la ecuación (4.7) y  $w_{i-1,i} = 1$  para toda  $i$ .
2. Estimador PCA: consiste en aproximar  $\hat{\Gamma}^{-1}$  truncando el número de las autofunciones,

$$\hat{\beta}_{\text{PCA}}(t, s) = \frac{1}{N-1} \sum_{l=1}^{N-1} \sum_{j=1}^p \sum_{i=1}^p \hat{\lambda}_j^{-1} \langle X_l, \hat{v}_j \rangle \langle X_{l+1}, \hat{v}_i \rangle \hat{v}_j(s) \hat{v}_i(t),$$

donde  $\hat{v}_i(t)$  es autofunción de  $\hat{\Gamma}_0$ .

3. Estimador RPCA: Similar al estimador PCA, reemplazando  $\hat{v}_i(t)$  por  $\hat{v}_i^R(t)$ , donde  $\hat{v}_i^R(t)$  es una la versión robusta de  $\hat{v}_i(t)$  propuesto en Hyndman and Ullah (2007). El algoritmo consiste de dos pasos, primero se calculan las autofunciones robustas  $\{\hat{u}_i\}$  como en Hubert et al. (2002) usando la medida de dispersión  $Q_n(z_1, \dots, z_N) = 2.2219c_n \{ |z_i - z_j|; i < j \}_{(k)}$ . Luego se expresa cada dato funcional usando esta base de autofunciones,  $X_i = \sum_{k=1}^K a_{i,k} \hat{u}_k$ . En el segundo paso se calcula el error cuadrático integrado de cada curva  $i$ ,  $w_i^* = \int_t \left( X_i - \sum_{k=1}^K a_{i,k} \hat{u}_k \right)^2$ . El valor de  $w_i^*$  debe ser grande si  $X_i$  es una curva atípica. Estos valores son usados para definir ponderaciones,  $w_i$ , de cada dato funcional. Finalmente  $\hat{v}_i^R$  es obtenido maximizando  $w_i \langle v_i, X_i \rangle$ , con condiciones tal que  $v_i$  es autofunción.

Para el cómputo y manipulación de las curvas, usamos el paquete *fda* y *ftsa* de R Core Team (2017).

### 4.5.1. Esquema de la simulación

Simulamos series de tiempo funcional  $\{X_i(s); s \in [0, 1], i = 1, \dots, N\}$  usando el modelo FAR(1) definido en (4.4) con dos tipos de kernel,  $\beta(t, s)$ , y dos tipos de ruido funcional,  $W^i(t)$ . Los kernels a considerar son (a) kernel Gaussiano,  $\beta_1(t, s) = c \exp\{-(t^2 + s^2)/2\}$  donde  $c$  es tal que  $\|\Phi\|_{\mathcal{S}} = 0.5$ , y (b) kernel lineal,  $\beta_2(t, s) = ct$ , donde  $c$  es tal que  $\|\Phi\|_{\mathcal{S}} = 0.5$ , ver Figura 4.6. Los ruidos funcionales son (a) movimiento Browniano,  $W^1(t) = W(t)$ ,  $t \in [0, 1]$ , y (b) puente Browniano,  $W^2(t) = W(t) - tW(1)$ , donde  $W(t)$  es un movimiento Browniano en  $[0, 1]$ .

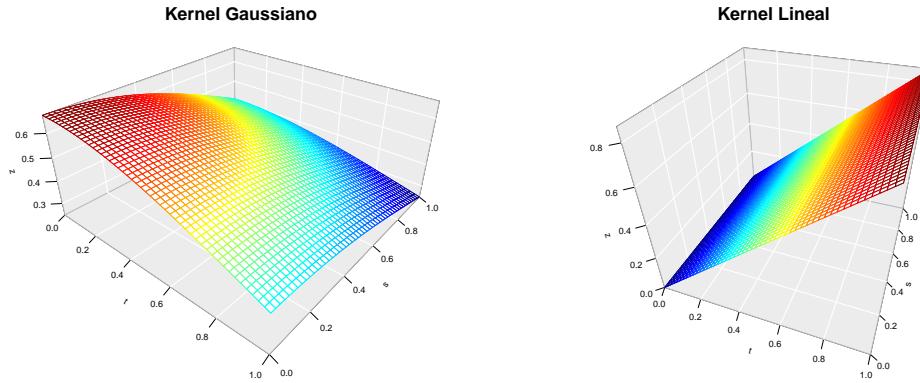


Figura 4.6: Tipos de kernel considerados en la simulación.

Consideramos tamaños de muestra  $N = 200$  y  $N = 300$ , evaluados en  $m = 100$  puntos equiespaciados en el intervalo  $[0, 1]$  con  $R = 1000$  réplicas. Para el ajuste de curvas usamos una base de spline cúbicos para cada curva simulada, tomando en cuenta 15 componentes en la base, i.e,  $X_i(t) = \sum_{r=1}^{15} a_{r,i} \psi_r(t)$ . Similarmente, consideramos 15 bases para  $\nu$  y  $\eta$  en la representación tensorial de  $\beta(t, s)$ . La función de penalización (4.8) que se considera en la estimación es el siguiente

$$\mathbf{P}_t = \int \left\{ \frac{\partial^2}{\partial s^2} \nu(s) \right\} \left\{ \frac{\partial^2}{\partial s^2} \nu^T(s) \right\} ds \text{ y } \mathbf{P}_s = \int \left\{ \frac{\partial^2}{\partial s^2} \eta(s) \right\} \left\{ \frac{\partial^2}{\partial s^2} \eta^T(s) \right\} ds. \quad (4.17)$$

Para determinar el parámetro de ajuste  $\lambda_t$  y  $\lambda_s$  se usa validación cruzada.

Consideramos los mismos modelos de contaminación que en [Sun and Genton \(2011\)](#) para generar datos atípicos en el proceso  $Y_n^\varepsilon = (1 - Z_n^\varepsilon)X_n + Z_n^\varepsilon V_n$ , con  $Z_n^\varepsilon$  un proceso Bernoulli tal que  $\mathbb{P}(Z_n^\varepsilon = 1) = \varepsilon$ . A continuación se enlistan estos modelos:

1. Modelo 1 (sin contaminación):  $V_n(\zeta, t) = X_n(t)$ ;
2. Modelo 2 (shifted outlier):  $V_n(\zeta, t) = X_n(t) + \sigma\zeta$ , donde  $\zeta$  es la magnitud de la contaminación y  $\sigma$  tiene distribución Bernoulli en  $\{-1, 1\}$  con probabilidad  $1/2$ ;
3. Modelo 3 (parcialmente contaminado):  $V_n(\zeta, t) = X_n(t) + \sigma\zeta 1_{[U,1]}(t)$ , con los mismos parámetros que en el Modelo 2, donde  $U$  tiene distribución uniforme en  $[0, 1]$ ; y

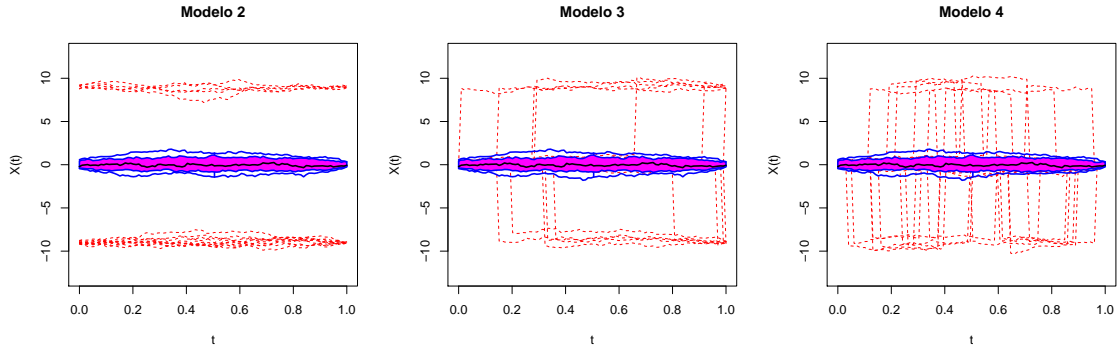


Figura 4.7: Muestra de curvas con contaminación. Las curvas discontinuas y en color rojo son curvas detectadas como atípicas usando boxplot funcional basado en la profundidad dO definido en (4.11).

4. Modelo 4 (contaminación a picos):  $V_n(\zeta, t) = X_n(t) + \sigma\zeta 1_{[U, U+l]}(t)$ , con los mismos parámetros que en el Modelo 2, donde  $U$  tiene distribución uniforme en  $[0, 1 - l]$ ;

Fijamos  $\zeta = 8$ ,  $l = 1/4$  y variamos el nivel de la contaminación,  $\varepsilon = 0.05$ ,  $\varepsilon = 0.10$ , y  $\varepsilon = 0.15$ . La Figura 4.7 muestra los tipos de contaminación usando los modelos anteriores. Las curvas discontinuas y en color rojo representan las curvas atípicas. Para evaluar el desempeño de los estimadores en esta configuración de datos, calculamos el error cuadrático integrado (ISE),

$$\text{ISE} = \sqrt{\int_D \int_D \left\{ \beta(t, s) - \hat{\beta}(t, s) \right\}^2 dt ds}. \quad (4.18)$$

### 4.5.2. Resultados de Simulación

A continuación presentamos los resultados obtenidos de las simulaciones.

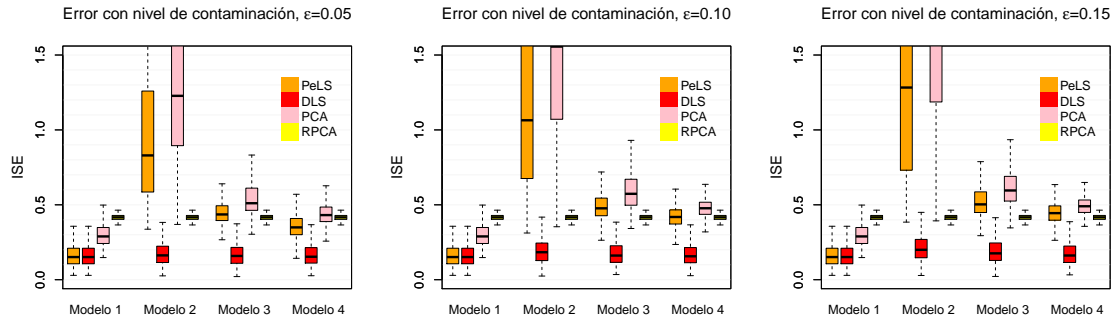
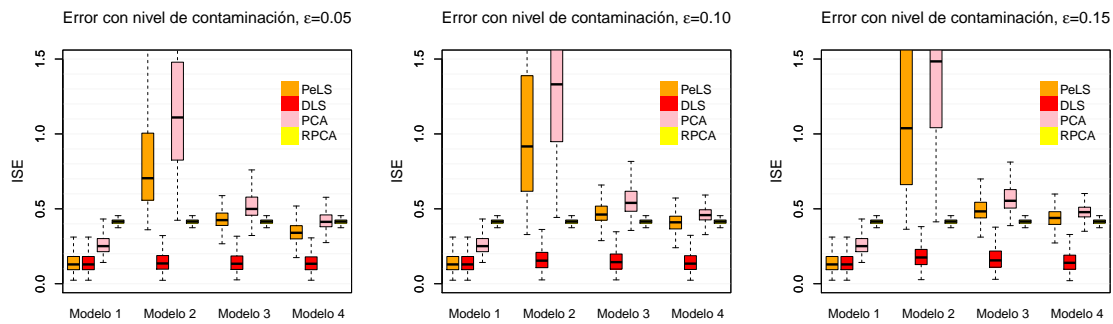
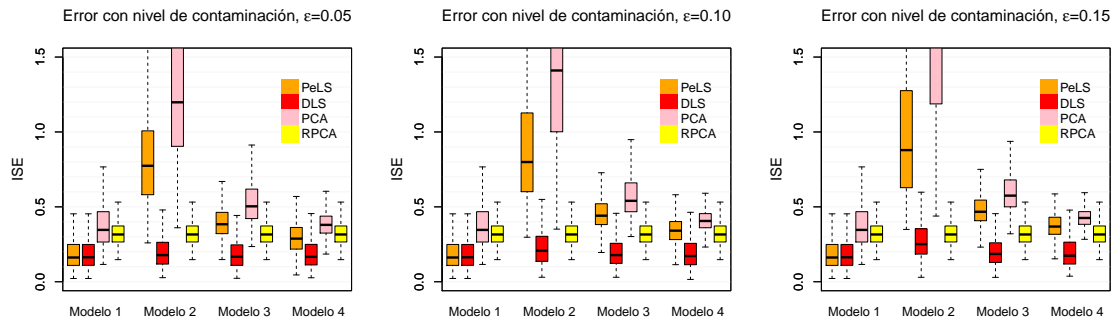
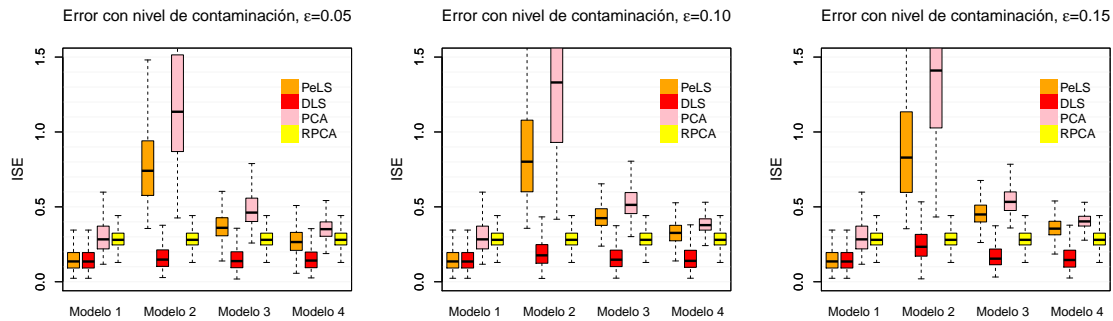
(a)  $N = 200$  WN=Puerto Browniano(b)  $N = 300$  WN=Puerto Browniano(c)  $N = 200$  WN=Mov Browniano(d)  $N = 300$ , WN=Mov Browniano

Figura 4.8: Boxplot de los valores de ISE para cada simulación con tamaños de muestra  $N = 200$  y  $N = 300$ .  $\beta(t, s)$  es el kernel Gaussiano, y el nivel de contaminación varía de 0.05 a 0.15 (de izquierda a derecha). (a) y (b) son usando puente Browniano como ruido funcional y (c) y (d) un movimiento Browniano.

La Figura 4.8 muestra el boxplot de los valores de ISE obtenidos en cada réplica con diferentes configuraciones de simulación y  $\beta(t, s)$  kernel Gaussiano. Para cada modelo de contaminación, graficamos el boxplot correspondiente a los estimadores PeLS, DLS, PCA y RPCA. Observemos que los resultados son similares en ambos ruidos funcionales, así que describimos los resultados sin distinción de estos ruidos.

Tabla 4.1: Valores de MISE de la simulación Monte Carlo correspondiente al kernel Gaussiano ( $\beta(t, s)$ ) con dos ruidos diferentes, puentes Browniano (Bb) y movimiento Browniano (Bm). Modelo 1, 2, 3, y 4 indican los diferentes tipos de contaminación (definidos en la sección 4.5.1); el nivel de la contaminación varía de 0.05 a 0.15 y los tamaños de muestra son  $N = 200$  and  $N = 300$ . Los números en negrita resaltan el valor con mejor desempeño.

Método	Modelo 1		Modelo 2		Modelo 3		Modelo 4	
$N = 200$	Bb	Bm	Bb	Bm	Bb	Bm	Bb	Bm
$\varepsilon = 0.05$								
PeLS	<b>0.166</b>	<b>0.191</b>	0.989	0.830	0.457	0.408	0.356	0.293
DLS	<b>0.166</b>	<b>0.191</b>	<b>0.177</b>	<b>0.205</b>	<b>0.171</b>	<b>0.191</b>	<b>0.170</b>	<b>0.194</b>
PCA	0.305	0.384	1.314	1.316	0.569	0.552	0.443	0.389
RPCA	0.416	0.327	0.416	0.327	0.416	0.327	0.416	0.327
$\varepsilon = 0.10$								
PeLS	<b>0.166</b>	<b>0.191</b>	1.285	0.902	0.501	0.469	0.420	0.345
DLS	<b>0.166</b>	<b>0.191</b>	<b>0.195</b>	<b>0.233</b>	<b>0.178</b>	<b>0.199</b>	<b>0.173</b>	<b>0.200</b>
PCA	0.305	0.384	1.679	1.536	0.608	0.587	0.481	0.410
RPCA	0.416	0.327	0.416	0.327	0.416	0.327	0.416	0.327
$\varepsilon = 0.15$								
PeLS	<b>0.166</b>	<b>0.191</b>	1.511	1.003	0.530	0.490	0.446	0.375
DLS	<b>0.166</b>	<b>0.191</b>	0.216	<b>0.282</b>	<b>0.194</b>	<b>0.206</b>	<b>0.176</b>	<b>0.202</b>
PCA	0.305	0.384	1.915	1.804	0.627	0.603	0.493	0.429
RPCA	0.416	0.327	0.416	0.327	0.416	0.327	0.416	0.327

Método	Modelo 1		Modelo 2		Modelo 3		Modelo 4	
$N = 300$	Bb	Bm	Bb	Bm	Bb	Bm	Bb	Bm
$\varepsilon = 0.05$								
PeLS	<b>0.142</b>	<b>0.153</b>	0.838	0.781	0.440	0.377	0.344	0.273
DLS	<b>0.142</b>	<b>0.153</b>	<b>0.149</b>	<b>0.167</b>	<b>0.145</b>	<b>0.155</b>	<b>0.144</b>	<b>0.156</b>
PCA	0.263	0.314	1.186	1.220	0.538	0.498	0.422	0.357
RPCA	0.415	0.292	0.415	0.292	0.415	0.292	0.415	0.292
$\varepsilon = 0.10$								
PeLS	<b>0.142</b>	<b>0.153</b>	1.092	0.869	0.481	0.439	0.410	0.329
DLS	<b>0.142</b>	<b>0.153</b>	<b>0.164</b>	<b>0.194</b>	<b>0.156</b>	<b>0.163</b>	<b>0.147</b>	<b>0.160</b>
PCA	0.263	0.314	1.436	1.408	0.568	0.537	0.462	0.383
RPCA	0.415	0.292	0.415	0.292	0.415	0.292	0.415	0.292
$\varepsilon = 0.15$								
PeLS	<b>0.142</b>	<b>0.153</b>	1.252	0.917	0.503	0.464	0.439	0.359
DLS	<b>0.142</b>	<b>0.153</b>	<b>0.186</b>	<b>0.251</b>	<b>0.170</b>	<b>0.174</b>	<b>0.152</b>	<b>0.163</b>
PCA	0.263	0.314	1.637	1.543	0.578	0.549	0.480	0.405
RPCA	0.415	0.292	0.415	0.292	0.415	0.292	0.415	0.292

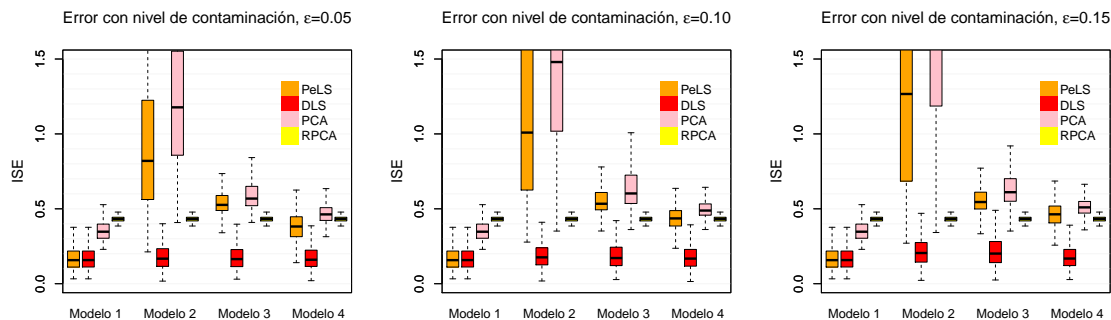
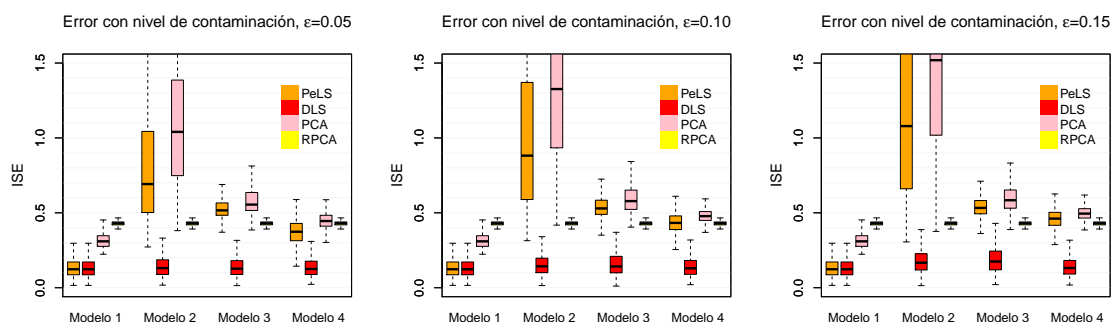
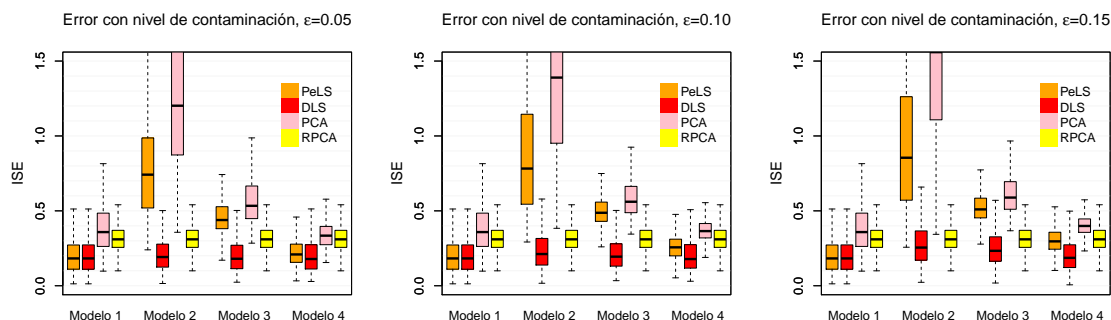
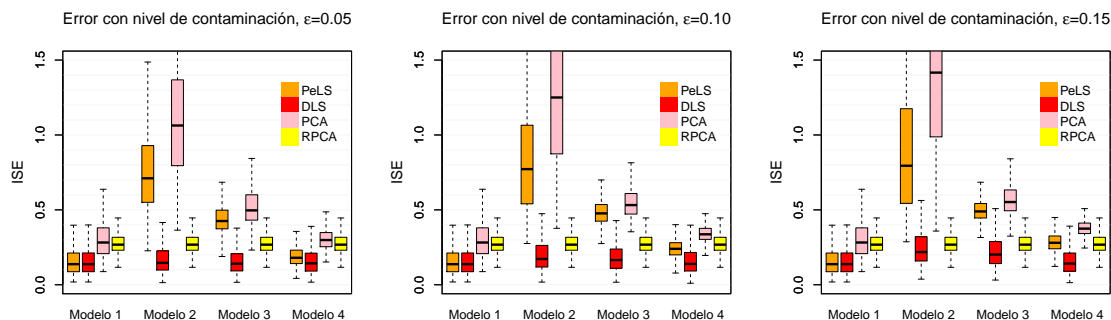
(a)  $N = 200$  WN=Puente Browniano(b)  $N = 300$  WN=Puente Browniano(c)  $N = 200$  WN=Mov Browniano(d)  $N = 300$ , WN=Mov Browniano

Figura 4.9: Boxplot de los valores de ISE para cada simulación con tamaños de muestra  $N = 200$  y  $N = 300$ .  $\beta(t, s)$  es el kernel lineal, y el nivel de contaminación varía de 0.05 a 0.15 (de izquierda a derecha). (a) y (b) son usando puente Browniano como ruido funcional y (c) y (d) un movimiento Browniano.

En la Figura 4.8 podemos observar que la presencia de curvas atípicas en los datos hace que incremente los valores de ISE, estos valores varían de acuerdo al tipo de contaminación. Los estimadores no robustos PeLS y PCA, presentan los valores más altos de ISE comparados con los otros estimadores, especialmente cuando la contaminación proviene del Modelo 2. El estimador RPCA presenta mejores

Tabla 4.2: Valores de MISE de la simulación Monte Carlo correspondiente al kernel lineal  $(\beta(t, s))$  con dos ruidos diferentes, puentes Browniano (Bb) y movimiento Browniano (Bm). Modelo 1, 2, 3, y 4 indican los diferentes tipos de contaminación (definidos en la sección 4.5.1); el nivel de la contaminación varía de 0.05 a 0.15 y los tamaños de muestra son  $N = 200$  y  $N = 300$ . Los números en negrita resaltan el valor con mejor desempeño.

Método	Modelo 1		Modelo 2		Modelo 3		Modelo 4	
$N = 200$								
	Bb	Bm	Bb	Bm	Bb	Bm	Bb	Bm
q=0.05								
PLS	<b>0.172</b>	<b>0.206</b>	0.949	0.791	0.555	0.471	0.382	0.224
DLS	<b>0.172</b>	<b>0.206</b>	<b>0.182</b>	<b>0.217</b>	<b>0.178</b>	<b>0.205</b>	<b>0.174</b>	<b>0.208</b>
PCA	0.360	0.390	1.253	1.275	0.631	0.580	0.473	0.347
RPCA	0.432	0.323	0.432	0.323	0.432	0.323	0.432	0.323
q=0.10								
PLS	<b>0.172</b>	<b>0.206</b>	1.258	0.872	0.567	0.507	0.438	0.259
DLS	<b>0.172</b>	<b>0.206</b>	<b>0.192</b>	<b>0.243</b>	<b>0.190</b>	<b>0.219</b>	<b>0.182</b>	<b>0.206</b>
PCA	0.360	0.390	1.633	1.499	0.656	0.596	0.497	0.369
RPCA	0.432	0.323	0.432	0.323	0.432	0.323	0.432	0.323
q=0.15								
PLS	<b>0.172</b>	<b>0.206</b>	1.468	0.962	0.564	0.527	0.464	0.302
DLS	<b>0.172</b>	<b>0.206</b>	<b>0.218</b>	<b>0.289</b>	<b>0.217</b>	<b>0.254</b>	<b>0.183</b>	<b>0.209</b>
PCA	0.360	0.390	1.937	1.711	0.645	0.620	0.513	0.403
RPCA	0.432	0.323	0.432	0.323	0.432	0.323	0.432	0.323
$N = 300$								
	Bb	Bm	Bb	Bm	Bb	Bm	Bb	Bm
q=0.05								
PLS	<b>0.137</b>	<b>0.158</b>	0.826	0.759	0.534	0.446	0.372	0.189
DLS	<b>0.137</b>	<b>0.158</b>	<b>0.145</b>	<b>0.170</b>	<b>0.143</b>	<b>0.160</b>	<b>0.139</b>	<b>0.160</b>
PCA	0.320	0.307	1.120	1.128	0.596	0.537	0.451	0.306
RPCA	0.430	0.279	0.430	0.279	0.430	0.279	0.430	0.279
q=0.10								
PLS	<b>0.137</b>	<b>0.158</b>	1.060	0.836	0.541	0.488	0.433	0.242
DLS	<b>0.137</b>	<b>0.158</b>	<b>0.157</b>	<b>0.198</b>	<b>0.159</b>	<b>0.184</b>	<b>0.143</b>	<b>0.161</b>
PCA	0.320	0.307	1.424	1.327	0.603	0.557	0.482	0.342
RPCA	0.430	0.279	0.430	0.279	0.430	0.279	0.430	0.279
q=0.15								
PLS	<b>0.137</b>	<b>0.158</b>	1.279	0.897	0.544	0.501	0.460	0.285
DLS	<b>0.137</b>	<b>0.158</b>	<b>0.177</b>	<b>0.247</b>	<b>0.190</b>	<b>0.223</b>	<b>0.146</b>	<b>0.162</b>
PCA	0.320	0.307	1.630	1.500	0.608	0.576	0.498	0.378
RPCA	0.430	0.279	0.430	0.279	0.430	0.279	0.430	0.279

resultados comparados con PeLS y PCA pero aún así no son satisfactorios, además el uso de autofunciones robustas no es del todo correcto por el cambio en los autovalores que se discutió al inicio del capítulo. En contraste, no se observa un aumento significativo en los valores de ISE al usar el estimador DLS. Podemos observar que estos resultados de DLS son consistentes a través de los diferentes modelos de contaminación y en las diferentes configuraciones de simulación, incluyendo los diferentes tamaños de muestra  $N$ .

La Tabla 4.1 presenta un resumen de los resultados de la media de los valores de ISE (MISE). Como hemos notado, el estimador DLS es el que tiene el mejor desempeño comparado con los demás estimadores, teniendo valores pequeños de MISE y esto sucede en todos los casos. Más aún, notemos que cuando no hay contaminación, el estimador DLS presenta los mismo valores que el estimador PeLS, indicando que el método de detección de curvas atípicas es eficiente.

La Figura 4.9, muestra los resultados correspondientes al kernel lineal, en este caso tenemos las mismas conclusiones que en el caso kernel Gaussiano. La Tabla 4.2 muestra un resumen de estos resultados.

Un hecho importante en la estimación del operador  $\Phi$ , es que debe tener la norma correcta. Por ejemplo, bajo el supuesto de estacionariedad del proceso FAR(1) (Bosq, 2000), la norma del operador correspondiente debe ser menor que uno, i.e.,  $\|\Phi\|_{\mathcal{S}} < 1$ . En nuestras simulaciones, el operador es tal que  $\|\Phi\|_{\mathcal{S}} = 0.5$ . La Figura 4.10 y 4.11 muestran los boxplot de las normas de los estimadores del operador en cada réplica, usando los diferentes estimadores PeLS, DLS, PCA, y RPCA, y las diferentes configuraciones de simulación.

Como es de esperarse, los boxplot correspondientes al Modelo 1 (no contaminación) son similares, con valores cercano a la norma original, 0.5, excepto las normas del estimador usando RPCA. Sin embargo cuando los datos son contaminados, los valores de las normas son afectados. Similarmente a los resultados de ISE, la norma de los estimadores no robustos, PeLS and PCA, son los más afectados, especialmente cuando el modelo de contaminación es el Modelo 2, se obtienen estimadores con norma mayor a uno. Por otro lado, los boxplot de las normas de los operadores estimados con DLS contienen siempre el valor verdadero de la norma.



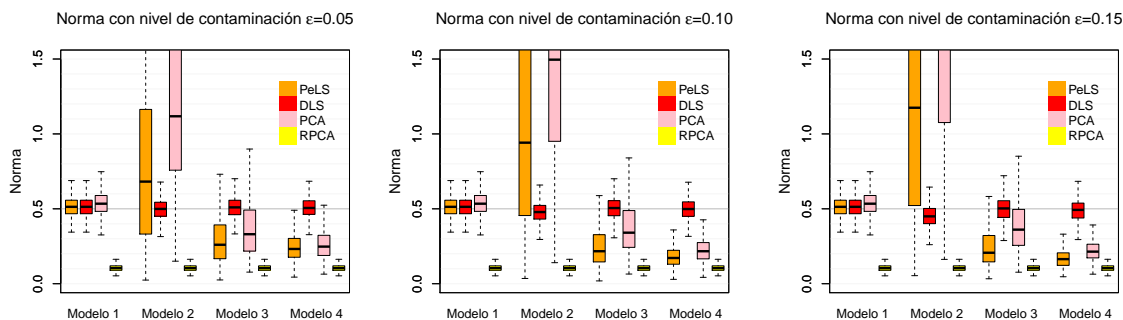
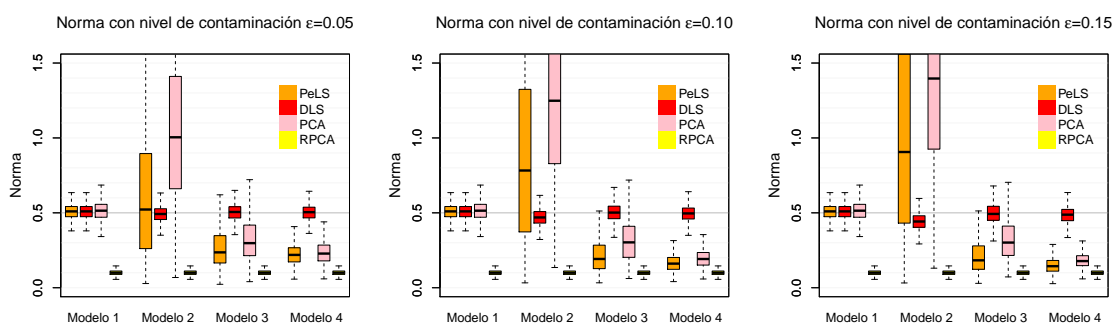
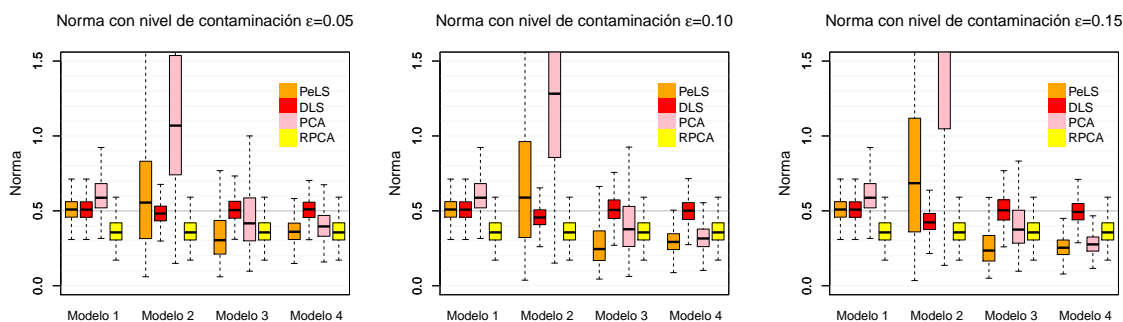
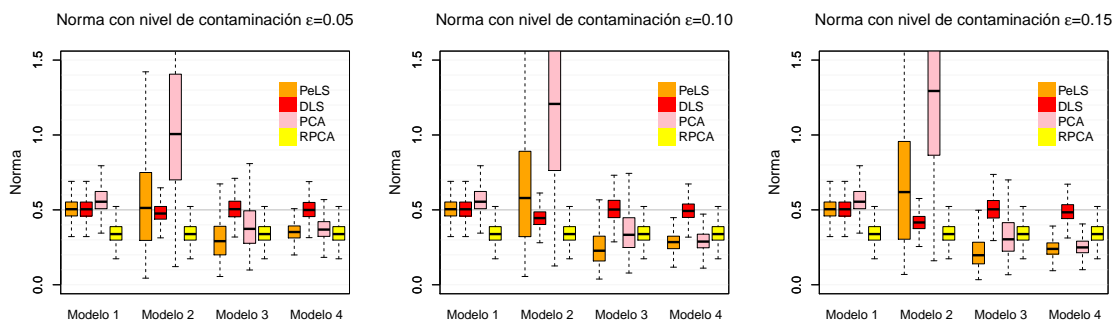
(a)  $N = 200$ ,  $WN=\text{Puente Browniano}$ (b)  $N = 300$ ,  $WN=\text{Puente Browniano}$ (c)  $N = 200$ ,  $WN=\text{Movimiento Browniano}$ (d)  $N = 300$ ,  $WN=\text{Movimiento Browniano}$ 

Figura 4.10: Boxplot de las normas de los operadores estimados en cada réplica con tamaños de muestra  $N = 200$  y  $N = 300$ .  $\beta(t, s)$  es el kernel Gaussiano y es tal que  $\|\Phi\|_{\mathcal{S}} = 0.5$  y los niveles de contaminación varían de 0.05 a 0.15 (de izquierda a derecha). (a) y (b) son usando puente Browniano como ruido funcional y (c) y (d) un movimiento Browniano.

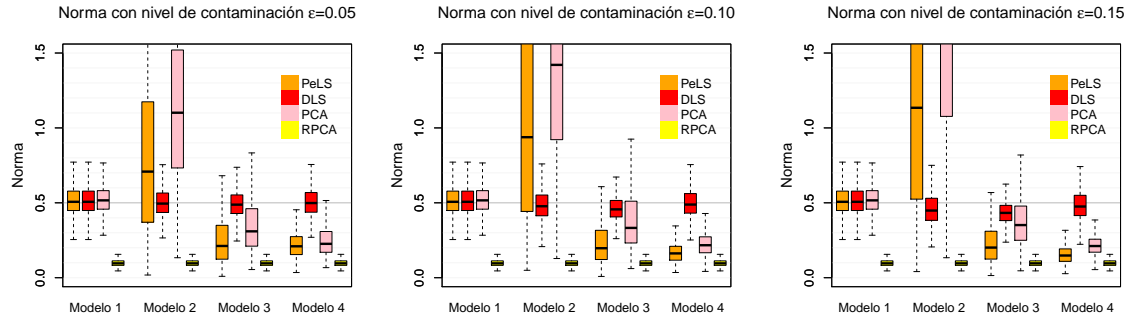
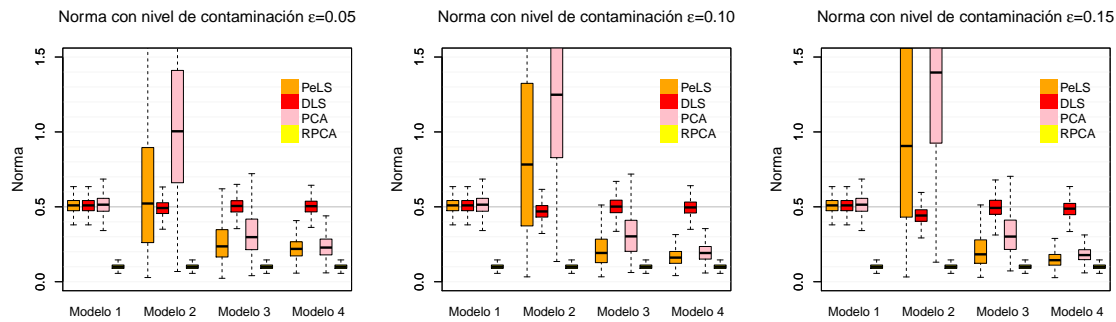
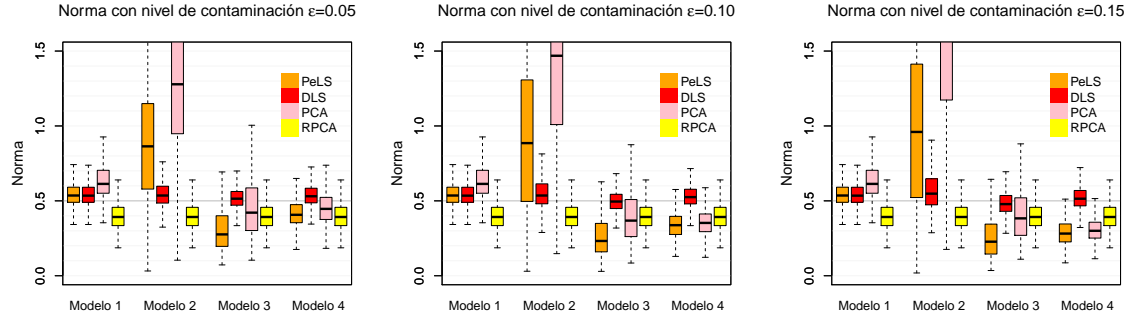
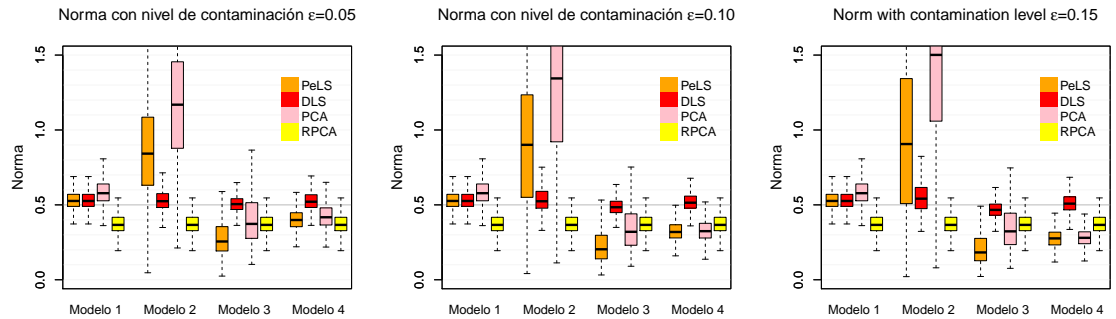
(a)  $N = 200$ , WN=Puente Browniano(b)  $N = 300$ , WN=Puente Browniano(c)  $N = 200$ , WN=Movimiento Browniano(d)  $N = 300$ , WN=Movimiento Browniano

Figura 4.11: Boxplot de las normas de los operadores estimados en cada réplica con tamaños de muestra  $N = 200$  y  $N = 300$ .  $\beta(t, s)$  es el kernel lineal y es tal que  $\|\Phi\|_{\mathcal{S}} = 0.5$  y los niveles de contaminación varían de 0.05 a 0.15 (de izquierda a derecha). (a) y (b) son usando puente Browniano como ruido funcional y (c) y (d) un movimiento Browniano.

### 4.5.3. Función de influencia empírica

En esta sección mostramos los resultados de simulación correspondientes a la función de influencia. Para saber de la robusticidad del estimador DLS, calculamos la Función de Influencia Empírica. Para explorar la influencia de la contaminación en todos los posibles tiempos de observación, mostramos la EIF promedio (MEIF) sobre el tiempo de observación. También mostramos los resultados correspondientes al estimador PeLS con el objeto de compararlos.

Para ilustrar el desempeño en términos de MEIF, simulamos series de tiempo funcional FAR(1) con kernel Gaussiano,  $\beta(t, s) = c \exp\{-(t^2 + s^2)/2\}$ , donde  $c$  es tal que  $\|\Phi\|_{\mathcal{S}} = 0.5$  con tamaño de muestra  $N = 200$ . Para cada muestra,  $\{X_i(s)\}_{i=1}^N$ , calculamos MEIF de la ecuación (4.16) con diferentes modelos de contaminación, esto es,  $V_i =$  Modelo 2, Modelo 3, y Modelo 4. Variando  $\zeta$  de  $-14$  a  $14$  y realizamos un estudio de simulación Monte Carlo con 100 réplicas.

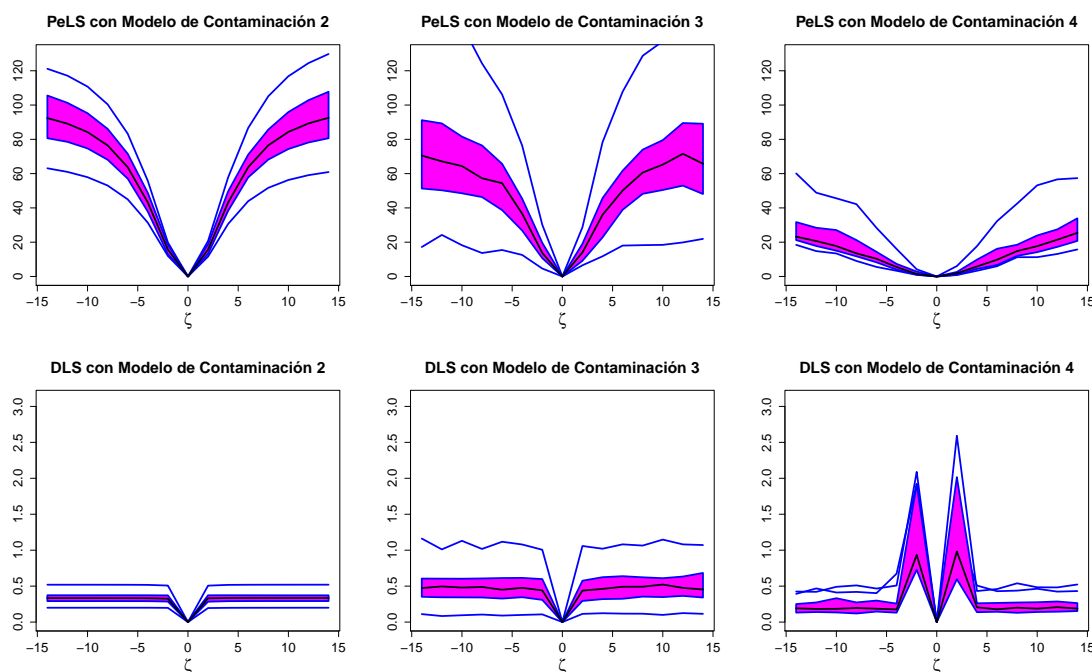


Figura 4.12: Boxplot funcional de la norma  $\|\text{MEIF}\|_{\mathcal{S}}$  para cada simulación de  $\hat{\beta}_{\text{PeLS}}(t, s)$  (Izquierda) y  $\hat{\beta}_{\text{DLS}}(t, s)$  (derecha). En este caso  $N = 200$ ,  $\beta(t, s)$  es el kernel Gaussiano, y  $W(t)$  es el puente Browniano.

La Figura 4.12 muestra la norma de Hilbert-Schmidt de cada MEIF ( $\|\text{MEIF}\|_{\mathcal{S}}$ ) como función de  $\zeta$  para cada muestra. Observemos que el  $\|\text{MEIF}\|_{\mathcal{S}}$  correspondiente al estimador DLS permanece constante cuando el valor de  $|\zeta|$  incrementa. Por otro lado, el  $\|\text{MEIF}\|_{\mathcal{S}}$ , correspondiente al estimador PeLS, incrementa cuando  $|\zeta|$  incrementa, reafirmando que el IF de DLS es acotado, propiedad que es deseable.

# Conclusiones

En esta tesis se abordaron dos problemas en el contexto de series de tiempo. Primero se estudió el problema de memoria larga en series de tiempo no lineales. Se adoptó la información mutua como una medida de dependencia para la definición de memoria larga, considerando esta medida sobre vectores a pares  $(X_t, X_{t-h})$ . Así, usando la entropía del proceso, se propuso una prueba de hipótesis basada en la prueba envolvente, usando un enfoque funcional. Con los estudios de simulación se mostró el desempeño del método donde las conclusiones fueron muy satisfactorias. Este trabajo abre posibilidades para el estudio de cointegración no lineal, generalizando las ideas del contexto lineal.

También se estudió el problema del estimador robusto para el modelo autoregresivo funcional de orden uno. Se propuso usar el estimador de mínimos cuadrados ponderando las observaciones usando la noción de profundidad de curvas. Se estudió su función de influencia como una medida de la robusticidad, se demostró teóricamente que esto está acotado. Cabe mencionar que el estimador propuesto solo cubre un tipo específico de datos atípicos. También se propuso extender la idea de hair-plot en el caso de series de tiempo como una herramienta de visualización de las influencias de cada observación sobre el estimador. Se mostraron resultados de simulación donde las conclusiones fueron muy satisfactorias.

En la actualidad, las series de tiempo funcionales son cada vez más comunes en la práctica, tanto así que muchos trabajos de investigación se han conducido en esta área. Como hemos detallado durante la tesis, el enfoque funcional ofrece alternativas para lidiar con muchos problemas presentados en el caso multivariado.

Otro trabajo realizado durante el doctorado es la propuesta del modelo factorial para datos funcionales, no se detalló en esta tesis pero se puede leer en [Martínez-Hernández et al. \(2017b\)](#). Se propone un modelo factorial para datos funcionales, se habla de los estimadores correspondientes, los factores y los coeficientes de los factores, cargas factoriales. La idea nace de explicar los factores comunes en un contexto de series de tiempo cointegrado, similar a [Gonzalo and Granger \(1995\)](#), donde el espacio de las observaciones es particionada en dos subespacios; subespacio

donde la serie es estacionaria y otro subespacio donde no lo es.

Así como este último trabajo, existen muchas extensiones que se pueden desarrollar a partir de los trabajos presentados en esta tesis, como cointegración funcional, estimadores robustos para el caso multivariado funcional y para el caso orden  $p$ , así también como modelos funcionales de espacio-estado. Concluimos que los problemas desarrollados en esta tesis dan origen a muchas áreas de investigación.

# Bibliografía

- Aparicio, F. M. and Escribano, A. (1998). Information-theoretic analysis of serial dependence and cointegration. *Stud. Nonlinear Dyn. Econom.*, 3(3):119–140 (electronic).
- Arribas-Gil, A. and Romo, J. (2015). Discussion of “Multivariate functional outlier detection”. *Stat. Methods Appl.*, 24(2):263–267.
- Azzouni, A. and Pujolle, G. (2017). *arXiv:1705.05690*.
- Bali, J. L., Boente, G., Tyler, D. E., and Wang, J.-L. (2011). Robust functional principal components: a projection-pursuit approach. *Ann. Statist.*, 39(6):2852–2882.
- Berenguer-Rico, V. and Gonzalo, J. (2014). Summability of stochastic processes—a generalization of integration for non-linear processes. *J. Econometrics*, 178(part 2-3):331–341.
- Bhattacharya, R. N., Gupta, V. K., and Waymire, E. (1983). The Hurst effect under trends. *J. Appl. Probab.*, 20(3):649–662.
- Bickel, P. J. and Bühlmann, P. (1996). What is a linear process. *Statistics*, 93:12128–12131.
- Blumentritt, T. and Schmid, F. (2012). Mutual information as a measure of multivariate association: analytical properties and statistical estimation. *J. Stat. Comput. Simul.*, 82(9):1257–1274.
- Boente, G. and Vahnovan, A. (2017). Robust estimators in semi-functional partial linear regression models. *J. Multivariate Anal.*, 154:59–84.
- Bosq, D. (2000). *Linear processes in function spaces*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York. Theory and applications.
- Bouezmarni, T. and Rolin, J.-M. (2003). Consistency of the beta kernel density function estimator. *Canad. J. Statist.*, 31(1):89–98.

- Bradley, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in probability and statistics (Oberwolfach, 1985)*, volume 11 of *Progr. Probab. Statist.*, pages 165–192. Birkhäuser Boston, Boston, MA.
- Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2:107–144. Update of, and a supplement to, the 1986 original.
- Bradley, Jr., R. C. (1983). Information regularity and the central limit question. *Rocky Mountain J. Math.*, 13(1):77–97.
- Brillinger, D. R. (1975). *Time series*. Holt, Rinehart and Winston, Inc., New York-Montreal, Que.-London. Data analysis and theory, International Series in Decision Processes.
- Brockwell, P. J. and Davis, R. A. (2006). *Time series: theory and methods*. Springer Series in Statistics. Springer, New York. Reprint of the second (1991) edition.
- Brooks, C. and Hinich, M. J. (2001). Bicorrelations and Cross-Bicorrelations As Non-linearity Tests and Tools for Exchange Rate Forecasting. *Journal of Forecasting*, 20(3):181–196.
- Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statist. Probab. Lett.*, 45(1):11–22.
- Chan, K. S., Petrucci, J. D., Tong, H., and Woolford, S. W. (1985). A multiple-threshold AR(1) model. *J. Appl. Probab.*, 22(2):267–279.
- Chen, S. X. (1999). Beta kernel estimators for density functions. *Comput. Statist. Data Anal.*, 31(2):131–145.
- Chen, X., Wu, W. B., and Yi, Y. (2009). Efficient estimation of copula-based semiparametric Markov models. *Ann. Statist.*, 37(6B):4214–4253.
- Crambes, C. and Mas, A. (2013). Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, 19(5B):2627–2651.
- Dai, W. and Genton, M. G. (2017). Directional outlyingness for multivariate functional data. *arXiv:1612.04615*.
- Darsow, W. F., Nguyen, B., and Olsen, E. T. (1992). Copulas and Markov processes. *Illinois J. Math.*, 36(4):600–642.
- Davidson, J. (2009). When is a time-series  $I(0)$ ? In *The methodology and practice of econometrics*, pages 322–342. Oxford Univ. Press, Oxford.



- Dufour, J.-M. (2006). Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, 133(2):443 – 477.
- Dufrénot, G. and Mignon, V. (2002). *Recent developments in nonlinear cointegration with applications to macroeconomics and finance*. Kluwer Academic Publishers, Boston, MA.
- Dürre, A., Fried, R., and Vogel, D. (2016). The spatial sign covariance matrix and its application for robust correlation estimation. *arXiv:1606.02274*.
- Escanciano, J. C. and Hualde, J. (2009). Persistence in nonlinear time series: A nonparametric approach. *CAEPR Working Paper*, (2009-003).
- Escribano, A. (2004). Nonlinear error correction: The case of money demand in the united kingdom (1878–2000). *Macroeconomic Dynamics*, 8(1):76–116.
- Franses, P. H. and McAleer, M. (1998). Testing for unit roots and non-linear transformations. *J. Time Ser. Anal.*, 19(2):147–164.
- Genton, M. G. and Ruiz-Gazen, A. (2010). Visualizing influential observations in dependent data. *J. Comput. Graph. Statist.*, 19(4):808–825. Supplementary materials available online.
- Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600.
- Gervini, D. (2012). Functional robust regression for longitudinal data. *arXiv:1211.7332*.
- Giannerini, S., Maasoumi, E., and Bee Dagum, E. (2015). Entropy testing for nonlinear serial dependence in time series. *Biometrika*, 102(3):661–675.
- Gneiting, T. and Schlather, M. (2004). Stochastic models that separate fractal dimension and the hurst effect. *SIAM Review*, 46(2):269–282.
- Gonzalo, J. and Granger, C. (1995). Estimation of common long-memory components in cointegrated systems. *J. Bus. Econom. Statist.*, 13(1):27–35.
- Granger, C. and Lin, J. L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *J. Time Ser. Anal.*, 15(4):371–384.
- Granger, C. W. J. (1995). Modelling nonlinear relationships between extended-memory variables. *Econometrica*, 63(2):265–279.
- Granger, C. W. J. and Hallman, J. (1991). Long memory series with attractors. *Oxford Bulletin of Economics and Statistics*, 53(1):11–26.

- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.
- Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Phys. D*, 31(2):277–283.
- Hinich, M. J. (1982). Testing for Gaussianity and linearity of a stationary time series. *J. Time Ser. Anal.*, 3(3):169–176.
- Hinich, M. J. (1996). Testing for dependence in the input to a linear time series model. *J. Nonparametr. Statist.*, 6(2-3):205–221.
- Huang, H. and Sun, Y. (2016). Total variation depth for functional data. *arXiv:1611.04913*.
- Hubert, M., Rousseeuw, P., and Segaeert, P. (2015). Multivariate functional outlier detection. *Stat. Methods Appl.*, 24(2):269–277.
- Hubert, M., Rousseeuw, P. J., and Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1):101 – 111. Fourth International Conference on Environ metrics and Chemometrics held in Las Vegas, NV, USA, 18-20 September 2000.
- Hurst, H. E. (1951). The long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Eng.*, 116:770–808.
- Hurst, H. E. (1956). Methods of using long-term storage in reservoirs. *Proceedings of the Institution of Civil Engineers*, 5:519–543.
- Hyndman, R. J. and Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *J. Comput. Graph. Statist.*, 19(1):29–45.
- Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Comput. Statist. Data Anal.*, 51(10):4942–4956.
- Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Teor. Veroyatnost. i Primenen.*, 7:361–392.
- Ibragimov, I. A. (1975). A note on the central limit theorems for dependent random variables. *Theory of Probability & Its Applications*, 20(1):135–141.
- Ibragimov, R. and Lentzas, G. (2008). Copulas and long memory.

- Ivanescu, A. E., Staicu, A.-M., Scheipl, F., and Greven, S. (2015). Penalized function-on-function regression. *Comput. Statist.*, 30(2):539–568.
- Joe, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.*, 84(405):157–164.
- Joe, H. (2015). *Dependence modeling with copulas*, volume 134 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.
- Kokoszka, P. and Reimherr, M. (2013). Determining the order of the functional autoregressive model. *J. Time Series Anal.*, 34(1):116–129.
- Lagerås, A. N. (2010). Copulas for Markovian dependence. *Bernoulli*, 16(2):331–342.
- Lee, S., Shin, H., and Billor, N. (2013).  $M$ -type smoothing spline estimators for principal functions. *Comput. Statist. Data Anal.*, 66:89–100.
- Leland, W. E., Willinger, W., Taqqu, M. S., and Wilson, D. V. (1995). On the self-similar nature of ethernet traffic. *SIGCOMM Comput. Commun. Rev.*, 25(1):202–213.
- Liu, J. and Brockwell, P. J. (1988). On the general bilinear time series model. *J. Appl. Probab.*, 25(3):553–564.
- Lo, A. W. (1991). Long-term memory in stock market prices. *J. Econometrica*, 59(part 5):1279–1313.
- Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999). Robust principal component analysis for functional data. *TEST*, 8(1):1–73.
- Long, J. P. and Huang, J. Z. (2016). A study of functional depths. *arXiv:1506.01332v3*.
- López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *J. Amer. Statist. Assoc.*, 104(486):718–734.
- Maddala, G. S. and Kim, I.-M. (1998). *Unit roots, cointegration and structural change*. Themes in Modern Econometrics. Cambridge University Press, Cambridge.
- Malliaropulos, D. (2000). A note on nonstationarity, structural breaks, and the fisher effect. *Journal of Banking & Finance*, 24(5):695 – 707.
- Mandelbrot, B. (1965). Une classe processus stochastiques homothétiques à soi; application à la loi climatologique H. E. Hurst. *C. R. Acad. Sci. Paris*, 260:3274–3277.
- Mandelbrot, B. B. and Wallis, J. R. (1968). Noah, joseph, and operational hydrology. *Water Resources Research*, 4(5):909–918.

- Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.
- Maronna, R. A. and Yohai, V. J. (2013). Robust functional linear regression based on splines. *Computational Statistics & Data Analysis*, 65:46 – 55.
- Martin, R. D. and Yohai, V. c. J. (1986). Influence functionals for time series. *Ann. Statist.*, 14(3):781–855.
- Martínez-Hernández, I., Genton, M. G., and González-Farías, G. (2017a). Robust depth-based estimation of the functional autoregressive model. *On the revision*.
- Martínez-Hernández, I., Gonzalo, J., and González-Farías, G. (2017b). Dynamic factor model for functional data. *Manuscript*.
- Martínez-Hernández, I. and González-Farías, G. (2017). Nonparametric test for long memory. *Manuscript*.
- Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory*, 11(2):258–289.
- McElroy, T. and Politis, D. N. (2007). Computer-intensive rate estimation, diverging statistics and scanning. *Ann. Statist.*, 35(4):1827–1848.
- Moran, P. A. P. (1964). On the range of cumulative sums. *Ann. Inst. Statist. Math.*, 16:109–112.
- Müller, U. K. (2008). The impossibility of consistent discrimination between  $I(0)$  and  $I(1)$  processes. *Econometric Theory*, 24(3):616–630.
- Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(2):381–404.
- Nelsen, R. B. (2006). *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition.
- Nieto-Reyes, A. and Battey, H. (2015). A topologically valid definition of depth for functional data. *arXiv:1410.5686v2*.
- Nieto-Reyes, A. and Battey, H. (2016). A topologically valid definition of depth for functional data. *Statist. Sci.*, 31(1):61–79.
- Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253.

- Phillips, P. C. B. and Xiao, Z. (1998). A primer on unit root testing. *Journal of Economic Surveys*, 12(5):423–470.
- Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4):379–396.
- Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B*, 53(3):539–572.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition.
- Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A.*, 42:43–47.
- Samorodnitsky, G. (2016). *Stochastic processes and long range dependence*. Springer Series in Operations Research and Financial Engineering. Springer, Cham.
- Scheipl, F. and Greven, S. (2016). Identifiability in penalized function-on-function regression models. *Electron. J. Stat.*, 10(1):495–526.
- Sercu, P., Uppal, R., and Van Hulle, C. (1995). The exchange rate in the presence of transaction costs: Implications for tests of purchasing power parity. *The Journal of Finance*, 50(4):1309–1319.
- Serfling, R. and Wijesuriya, U. (2017). Depth-based nonparametric description of functional data, with emphasis on use of spatial depth. *Comput. Statist. Data Anal.*, 105:24–45.
- Shao, H. and Soong, B. H. (2016). Traffic flow prediction with long short-term memory networks (lstm). In *2016 IEEE Region 10 Conference (TENCON)*, pages 2986–2989.
- Stokes, H. H. and Hinich, M. (2011). Detecting and modeling nonlinearity in the gas furnace data. *Computational Statistics*, 26(1):77–93.
- Subba Rao, T. and Gabr, M. M. (1984). *An introduction to bispectral analysis and bilinear time series models*, volume 24 of *Lecture Notes in Statistics*. Springer-Verlag, New York.
- Sun, Y. and Genton, M. G. (2011). Functional boxplots. *J. Comput. Graph. Statist.*, 20(2):316–334.

- Sun, Y. and Genton, M. G. (2012). Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics*, 23(1):54–64.
- Taqqu, M. S., Teverovsky, V., and Willinger, W. (1995). Estimators for long-range dependence: An empirical study. *Fractals*, 3:785–798.
- Teräsvirta, T., Tjøstheim, D., and Granger, C. W. J. (2010). *Modelling nonlinear economic time series*. Advanced Texts in Econometrics. Oxford University Press, Oxford.
- Terdik, G. (1999). *Bilinear stochastic models and related problems of nonlinear time series analysis*, volume 142 of *Lecture Notes in Statistics*. Springer-Verlag, New York. A frequency domain approach.
- Tong, H. (2015). Threshold models in time series analysis—some reflections. *J. Econometrics*, 189(2):485–491.
- Tuan, P. D. (1986). The mixing property of bilinear and generalised random coefficient autoregressive models. *Stochastic Process. Appl.*, 23(2):291–300.
- Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
- Walters-Williams, J. and Li, Y. (2009). *Estimation of Mutual Information: A Survey*, pages 389–396. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.
- Wu, S. and Müller, H.-G. (2011). Response-adaptive regression for longitudinal data. *Biometrics*, 67(3):852–860.
- Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482.