



Universidade de Vigo

Trabajo Fin de Máster

Técnicas de remuestreo en modelos de predicción semiparamétricos de series de tiempo

Pablo Carballo Fraguas

Máster en Técnicas Estadísticas

Curso 2023-2024

Propuesta de Trabajo Fin de Máster

<p>Título en galego: Técnicas de remostraxe en modelos de predicción semiparamétricos de series de tempo</p>
<p>Título en español: Técnicas de remuestreo en modelos de predicción semiparamétricos de series de tiempo</p>
<p>English title: Resampling techniques in semiparametric time series forecasting models</p>
<p>Modalidad: Modalidad A</p>
<p>Autor: Pablo Carballo Fraguas, Universidad de Santiago de Compostela</p>
<p>Director: Wenceslao González Manteiga, Universidad de Santiago de Compostela; Manuel Febrero Bande, Universidad de Santiago de Compostela</p>
<p>Breve resumen del trabajo:</p> <p>En los últimos 30 años se han desarrollado modelos de predicción de SO_2 y NO_x para la Central Térmica de As Pontes bajo distintos planteamientos metodológicos de regresión general. Estos modelos son, en general, de tipo semiparamétrico incluyendo en la parte no paramétrica estimaciones de la tendencia realizadas con modelos aditivos, redes neuronales o datos funcionales entre otros. Con esta propuesta de trabajo de fin de máster se pretende abordar los siguientes objetivos:</p> <ol style="list-style-type: none"> a) Una revisión general actualizada de los modelos implementados para las variables antes mencionadas en la utilidad desarrollada para la empresa. b) Una revisión de los mecanismos de remuestreo desarrollados en las diversas técnicas de predicción, con el objeto de generar regiones de confianza predictiva. c) Una ilustración comparativa de los distintos procedimientos con datos reales medioambientales o simulados.
<p>Recomendaciones: Se recomienda haber cursado cursos del máster en los tópicos “Modelos de Regresión”, “Datos Funcionales”, “Estadística Matemática”, “Series de Tiempo” y “Análisis Multivariante”.</p>
<p>Otras observaciones: Los tutores lideran el proyecto de predicción en la Central Térmica de As Pontes cuyos aspectos teóricos y prácticos conforman el hilo motivador del trabajo.</p>

Don Wenceslao González Manteiga, Catedrático de universidad de la Universidad de Santiago de Compostela, y don Manuel Febrero Bande, Catedrático de universidad de la Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

**Técnicas de remuestreo en modelos
de predicción semiparamétricos de
series de tiempo**

fue realizado bajo su dirección por don Pablo Carballo Fraguas para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 3 de junio de 2024.

El director:
Don Wenceslao González Manteiga

El director:
Don Manuel Febrero Bande

El autor:
Don Pablo Carballo Fraguas

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, [Disposición 2978 del BOE núm. 48 de 2022](#)), **el autor declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas, ...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración, ... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Resumen	IX
Índice de Notación	XI
Prefacio	XV
0.1. La Unidad de Producción Térmica de As Pontes	XV
0.2. El problema medioambiental	XVI
0.3. Naturaleza de los datos	XVII
0.4. Introducción a las series de tiempo	XVIII
1. Modelos paramétricos de series de tiempo	1
1.1. Introducción a la metodología Box-Jenkins	1
1.2. Estacionariedad, causalidad e invertibilidad de un proceso ARMA	2
1.3. Extendiendo los modelos ARMA	7
1.4. Predicción en modelos autorregresivos	9
1.5. Aplicación a los datos medioambientales	10
2. El método Bootstrap	13
2.1. Breve introducción al método bootstrap	13
2.2. Bootstrap en la estimación con datos dependientes	18
2.3. Bootstrap en la predicción con datos dependientes	30
2.4. Bootstrap en modelos temporales semiparamétricos	42
2.5. Formulación de modelos en el contexto medioambiental	47
3. Técnicas de remuestreo en otros modelos empleados	49
3.1. Modelo parcialmente lineal	49

3.2. Modelo de redes neuronales	51
3.3. Modelo con datos funcionales	53
3.4. Otras aproximaciones	56
4. Test bootstrap para estructuras simples	59
4.1. Comparación con el estadístico de Härdle y Mammen	61
4.2. Calibrado del test	62
4.3. Validez del bootstrap y ejemplo de aplicación	64
5. Ilustración con datos reales	65
5.1. Modelo paramétrico	65
5.2. Modelo no paramétrico	66
5.3. Modelo semiparamétrico	67
5.4. Modelo de redes neuronales	68
5.5. Conclusiones finales y <i>future work</i>	68
Apéndices	73
A. Series de tiempo	73
B. Algunos conceptos de Teoría de la Probabilidad	75
C. Código de R empleado	81
D. Distribuciones notables	91
Bibliografía	95

Resumen

Resumen en español

Disponer de un sistema de control y predicción en tiempo real de episodios de contaminación en una industria donde se producen emisiones de diferentes agentes contaminantes es una herramienta de gran utilidad en el contexto medioambiental y de salud pública. Su utilización permite a la empresa no solo conocer en cada instante la situación en la que se encuentran las emisiones sino también anticiparse a posibles picos de emisión, ajustándose a los límites legales establecidos y evitando pérdidas tanto económicas como medioambientales y sanitarias.

El objetivo que se persigue en este trabajo es presentar, con un enfoque fundamentalmente metodológico, algunos de los principales modelos de predicción propuestos en un problema medioambiental determinado. En concreto, se hará especial hincapié en un modelo semiparamétrico, para lo cual se comenzará presentando con detalle diferentes herramientas en el contexto de series de tiempo que subyacen a la construcción del mismo y que ayudarán a la comprensión de su naturaleza. Luego, se introducirá la metodología bootstrap y se presentará una amplia variedad de técnicas de remuestreo en este contexto de datos dependientes. A continuación, se tratarán algunos modelos alternativos al semiparamétrico que fueron surgiendo a raíz de diferentes enfoques sobre el problema, así como de cambios estructurales en la industria. Después, se pondrá el foco en la importancia de la especificación de los modelos y se presentará un contraste de especificación para algunas familias simples de ellos. Por último, con una pequeña ilustración de algunos de los modelos presentados con datos reales.

English abstract

Having a real-time control and prediction system for pollution episodes in an industry where emissions of different pollutants occur is a highly useful tool in the environmental and public health context. Its use allows the company not only to know the status of emissions at any given moment but also to anticipate possible emission peaks, complying with established legal limits and avoiding economic, environmental, and health losses.

The objective of this work is to present, with a fundamentally theoretical approach, some of the main prediction models proposed for a specific environmental problem. Specifically, emphasis will be placed on a semiparametric model, for which different tools in the context of time series that underlie its construction will be presented in detail to aid in understanding its nature. Then, the bootstrap methodology will be introduced, and a wide variety of resampling techniques in this context of dependent data will be presented. Subsequently, some alternative models to the semiparametric model that emerged from different approaches to the problem, as well as structural changes in the industry, will be discussed. Next, the focus will be on the importance of model specification, and a specification test for some simple model families will be presented. Finally, a small illustration of some of the presented models with real data will be provided.

Índice de Notación

La lista que se presenta a continuación recoge gran parte de la notación que será empleada más adelante en el cuerpo de este documento.

$\alpha(\cdot, \cdot)$	Función de autocorrelaciones parciales
\mathbf{I}_n	Matriz identidad de dimensiones $n \times n$
$\gamma(\cdot, \cdot)$	Función de autocovarianzas
$(\Omega, \mathcal{F}, \mathbb{P})$	Espacio de probabilidad
$\{X_t, t \in \mathbb{T}\}$	Proceso estocástico general
$[\cdot]$	Función techo
\mathbb{C}	Conjunto de números complejos
$\text{Corr}[\cdot, \cdot]$	Correlación
$\text{Cov}[\cdot, \cdot]$	Covarianza
$\mathbb{E}[\cdot]$	Esperanza
$\mathbb{I}\{\cdot\}$	Variable indicadora
\mathbb{N}	Conjunto de números naturales
\mathbb{P}	Probabilidad
\mathbb{R}	Conjunto de números reales
\mathbb{R}^+	Conjunto de números reales positivos
\mathbb{T}	Conjunto de índices
$\text{Var}[\cdot]$	Varianza
\mathbb{Z}	Conjunto de números enteros
\mathbf{B}	Operador retardo
\mathbf{F}	Operador de avance
\mathbf{X}	Vector aleatorio
\mathcal{A}, \mathcal{F}	σ -álgebra de sucesos

\mathcal{B}	σ -álgebra de Borel
\mathcal{K}	Operador de suavizado
\mathcal{X}	Variable aleatoria funcional
μ	Media
μ_t	Función de medias
∇	Operador de diferenciación regular
Ω	Conjunto de sucesos
ω	Suceso
$\overline{\mathbb{R}}$	Recta real ampliada
\bar{X}	Media muestral
\xrightarrow{d}	Convergencia en distribución
$\Phi(\cdot)$	Función de distribución de una Normal estándar
ϕ_1, \dots, ϕ_p	Coefficientes autorregresivos
ρ	Correlación
$\rho(\cdot, \cdot)$	Función de autocorrelaciones simples
σ	Desviación típica
$\sigma(\cdot)$	Función de varianzas
$\theta_1, \dots, \theta_q$	Coefficientes de medias móviles
$\varphi_X(t)$	Función característica de X
\hat{F}	Función de distribución estimada
\hat{f}	Función de densidad estimada
$AR(p)$	Modelo autorregresivo de orden p
F	Función de distribución
f	Función de densidad
$f(\lambda)$	Función de densidad espectral
F_n	Función de distribución empírica
K	Función de tipo núcleo
$m(\cdot)$	Función de regresión
M_t	Matriz histórica
$MA(q)$	Modelo de medias móviles de orden q
$N(\mu, \sigma)$	Distribución Normal de media μ y desviación típica σ
$N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Distribución Normal multivariante de vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$

$O(\cdot)$	O grande de Landau
$O_P(\cdot)$	O grande en probabilidad
$P_{\mathcal{M}}[\cdot]$	Operador proyección sobre el subespacio \mathcal{M}
$PMSE[\cdot]$	Error cuadrático medio de predicción
R	Estadístico general
S_n^2	Varianza muestral
$sp\{\cdot\}$	Sistema lineal generado por un conjunto
X	Variable aleatoria escalar
$ARIMA(p, d, q)$	Modelo ARIMA de órdenes p , d y q
$ARMA(p, q)$	Modelo ARMA de órdenes p y q

Prefacio

En este capítulo de introducción se comenzará realizando una breve presentación de la central térmica de As Pontes y del problema medioambiental asociado a ella. Asimismo, se pondrá de manifiesto la necesidad de disponer de herramientas matemáticas que permitan abordar el problema, para lo cual se llevará a cabo un breve repaso de algunos de los conceptos básicos de las series de tiempo.

0.1. La Unidad de Producción Térmica de As Pontes

La presente sección tiene por objetivo presentar a la Unidad de Producción Térmica (U.P.T.) de As Pontes de García Rodríguez. La información sobre ella que aquí se recoge ha sido obtenida a partir de [Wikipedia \(2024\)](#) y [Cámara Oficial Minera de Galicia \(2024a\)](#).

La U.P.T. de As Pontes está situada al norte de la provincia gallega de A Coruña y es uno de los centros de producción de energía eléctrica que posee en la Península Ibérica la empresa *Endesa Generation, S.A.* Está formada por una central térmica formada por 4 ciclos convencionales alimentados mediante carbón y una central de ciclo combinado alimentado mediante gas natural. La ubicación de esta y otros elementos destacables a su alrededor pueden verse en la [Figura 0.1](#).

En cuanto a su historia, la central térmica comenzó a construirse en el año 1972, no siendo hasta el año 1976 cuando fue puesta en funcionamiento. En un principio estaba diseñada para ser alimentada mediante el lignito extraído en las explotaciones mineras locales, el cual se caracterizaba por un alto contenido en azufre. Sin embargo, entre los años 1993 y 1996 se llevaron a cabo diferentes modificaciones como respuesta a las nuevas exigencias medioambientales para evitar altas emisiones de SO_2 , adaptando las instalaciones para el consumo de una mezcla formada por un 50% del carbón local y un 50% de carbón de importación. Este último posee muy poco contenido de azufre, por lo que al quemarlo genera poca cantidad de SO_2 , al contrario de lo que ocurría con el carbón local. Esto provocó que la mina local redujese su producción a prácticamente la mitad de lo que venía haciendo hasta entonces hasta que, años más tarde, el agotamiento de las reservas del yacimiento se produjo prácticamente de manera simultánea con la transposición de la Directiva Comunitaria de Grandes Instalaciones de Combustión, añadiendo una mayor exigencia al control de las emisiones, y desembocando, por tanto, en una transformación de la central para consumir tan solo carbón de importación a partir de enero de 2008¹. Esto tuvo como consecuencia el cierre de la explotación minera de 5 por 2 km, la cual había estado trabajando con intensidad entre 1977 y 2007, produciendo en dicho periodo un total de 260 millones de toneladas de lignito y convirtiéndose, de este modo, en la mayor explotación de este tipo de carbón de Galicia, seguida por los 93 millones extraídos de la mina de Meirama (véase [Cámara Oficial Minera de Galicia 2024b](#)). En ese mismo año en que finalizó el uso de lignito local (2008) se puso en marcha el grupo de ciclo combinado, el cual estaba alimentado mediante gas natural.

¹ Motivados por el fin de la extracción de lignito local, en los últimos tiempos la central se fue adaptando también al consumo de hulla, un tipo de carbón mineral cuya composición de carbono se encuentra entre un 80 y un 90%.

El 27 de diciembre de 2019 Endesa formalizó ante el Ministerio de Transición Ecológica la solicitud de cierre de los 4 grupos de carbón debido al incremento sustancial del precio de los derechos de CO₂ y a una fuerte caída del precio del gas (véase [Monforte 2019](#)). En septiembre de 2022 se autorizó el cierre de 2 de ellos, manteniendo la actividad de los otros 2 a causa de la incertidumbre con respecto al abastecimiento de gas ruso derivada del conflicto armado entre Ucrania y Rusia. Finalmente, en agosto del año 2023 se autorizó el cierre definitivo de todos los grupos de carbón, siendo publicado en el BOE el 19 de agosto de 2023.

Por último, en cuanto a las instalaciones, merecen especial atención su parque de carbones, con una superficie de 10 ha (aproximadamente 12 campos de fútbol) y capacidad para 250 000 toneladas y, sobre todo, la chimenea, conocida como *Endesa Termic*, con una altura de 356 m y un diámetro de 36 m en la base y de 18 m en la cima².



Figura 0.1: Lago artificial de As Pontes (1), rotopalpas (2), escombrera exterior (3) y U.P.T (4). Extraído de [Cámara Oficial Minera de Galicia \(2024a\)](#).



Figura 0.2: Chimenea de la central térmica de As Pontes. Extraída de [Wikipedia-PepedoCouto \(2009\)](#).

0.2. El problema medioambiental

La actividad de la central eléctrica tiene como consecuencia la liberación a la atmósfera de una serie de agentes contaminantes. Debe destacarse que la central se encuentra muy próxima a enclaves naturales de alto valor ecológico, por lo que es necesario tener controladas las emisiones para mitigar la contaminación, tanto para la protección del medio ambiente como para la propia salud humana, acorde a los límites establecidos por la legalidad vigente.

Como consecuencia de ello, y de acuerdo con los requisitos establecidos por la Consellería de Medio Ambiente, la central posee una red de vigilancia de la calidad atmosférica formada por varias estaciones de medición (hasta 17, como se ve en la Figura 0.3) que tienen 7 analizadores automáticos localizados en varios puntos en un radio de 30 km alrededor de la central para medir dióxido de azufre (SO₂), óxidos de nitrógeno (NO_x), partículas en suspensión en el ambiente, temperatura y oxígeno. Además, hay una estación meteorológica que, junto con las otras estaciones, proporciona información de manera continua a un ordenador situado en la central para ayudar a medir y predecir la contaminación. Los datos que allí se reciben son empleados para evaluar la situación en los alrededores de la planta en tiempo real y así poder intervenir, en caso de que fuese necesario, realizando operaciones específicas orientadas a la reducción de las emisiones para evitar que se produzca un episodio de contaminación³.

Además, desde el año 1992 la central dispone de un sistema de ayuda al control de la contaminación atmosférica, el cual fue desarrollado por el Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela en colaboración con el Departamento de Medio Ambiente de la central. Dicho sistema trabaja con la serie de tiempo formada por la concentración media de las

² En el momento de su construcción era la más alta de Europa.

³ Conviene destacar que en la actual situación de actividad de la central estos episodios son muy improbables

últimas 2 horas de SO_2 ⁴ en cada estación de control. En concreto, realiza predicciones a un horizonte de 30 minutos (luego fue ampliado) debido a que este era el tiempo que tardaban, aproximadamente, los operarios de la central en implementar las contramedidas oportunas y compensar mediante otros contribuidores de la red eléctrica nacional el efecto de estas sobre la producción energética.

Para alcanzar llevar a cabo esas predicciones de los niveles de SO_2 , a lo largo de los años se han ido desarrollando diferentes modelos estadísticos que han sido añadidos al sistema que se encarga de mostrarle al equipo de operarios de la central las predicciones. A comienzos de los 90 se comenzó empleando un modelo semiparamétrico para predecir a un horizonte de 30 minutos pero debido a cambios estructurales en la central el personal necesitaba más tiempo para realizar los cambios oportunos para evitar el episodio de contaminación y fue necesario ampliarlo hasta 1 hora. Además, el paradigma cambió significativamente en los años 2007 y 2008 debido a dos hechos. Por un lado, estaba el consumo exclusivo de carbón de importación, lo cual provocó que las emisiones de SO_2 se redujesen alrededor de un 95 %. Por otro, el hecho de que comenzase su actividad la estación de ciclo combinado, lo cual provocó que fuese necesario predecir también la concentración de NO_x .

Por último, comentar que otro objetivo era prevenir episodios en la calidad del aire originados por altas cantidades de SO_2 en la tierra y para ello era necesario conocer también datos relativos al tiempo ya que, bajo condiciones meteorológicas adversas (importa el dónde y el cuándo), los niveles pueden verse afectados drásticamente. Sin embargo, es necesario tener en cuenta también la manera en que son recogidos algunos de estos datos meteorológicos, ya que, por ejemplo, la chimenea está a una altura de 356 m pero la temperatura y dirección y velocidad del viento se medían a 10 y 80 m.

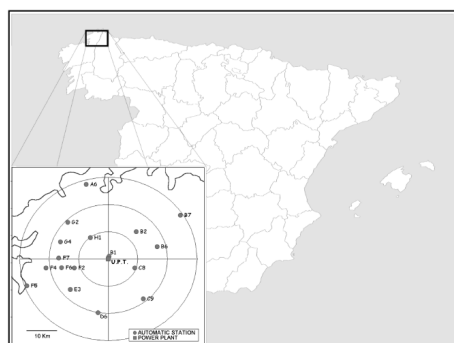


Figura 0.3: Mapa de España y localización de la central térmica de As Pontes junto con el sistema de control de calidad atmosférica. Extraído de [Fernández-de Castro et al. \(2003\)](#).

0.3. Naturaleza de los datos

Es conveniente realizar algunos comentarios acerca de la naturaleza de las series temporales de las concentraciones de SO_2 y NO_x . En primer lugar, cabe mencionar que los datos sobre estas concentraciones se reciben, normalmente en tiempo real, cada minuto y cada cinco minutos. Además, para ambos agentes contaminantes se obtiene la media bihoraria para llevar a cabo predicciones de valores futuros. La serie de medias bihorarias de la concentración de SO_2 presenta un comportamiento bastante característico, influenciado en gran medida por las condiciones meteorológicas y la topografía del lugar. Más concretamente, los valores de esta serie se encuentran, por lo general, muy cercanos a 0 durante largos periodos de tiempo pero pueden producirse incrementos bruscos repentinos (denominados *episodios* o *incidentes*), estando estos ampliamente separados en el tiempo. Por último, cabe mencionar que, en cuanto al comportamiento de la serie de las medias bihorarias de NO_x , este es muy similar al comentado acerca de la del SO_2 , con la salvedad de que ahora la escala es menor.

⁴ Se considera esta medida sobre la sustancia contaminante porque es aquella sobre la cual la legalidad vigente de entonces establecía los límites.

0.4. Introducción a las series de tiempo

En esta sección se llevará a cabo la presentación de algunos de los conceptos básicos del ámbito de las series de tiempo, los cuales serán fundamentales en los modelos estadísticos que se presentarán a lo largo del trabajo y que buscarán generalizar conceptos como los de media, varianza o correlación conocidos en la estadística usual, donde los datos son escalares y no guardan una relación de dependencia temporal.

Definición 0.1 (Proceso estocástico y trayectoria del proceso, [Bobrowski 2005](#), p. 123). Una familia de variables aleatorias $\{X_t, t \in \mathbb{T}\}$ definidas en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, donde \mathbb{T} es un conjunto de índices abstracto llamado *tiempo*, se denomina *proceso estocástico*. Si $\mathbb{T} = \mathbb{N}$ (o \mathbb{Z}) se dice que el proceso es en *tiempo discreto* y si $\mathbb{T} = \mathbb{R}$ (o \mathbb{R}^+), se dice que es en *tiempo continuo*. Además, para cada $\omega \in \Omega$, la función $t \mapsto X_t(\omega)$ se denomina *trayectoria* o *realización del proceso estocástico*.

Observación 0.1. Dado un t fijo se tiene que X_t es una variable aleatoria sobre $(\Omega, \mathcal{F}, \mathbb{P})$. Desde otro punto de vista, fijado un $\omega \in \Omega$, al variar el tiempo lo que se obtiene es un proceso que va evolucionando y que describe una trayectoria cuando se representa gráficamente.

Observación 0.2 ([Politis y McElroy 2020](#), p. 31). Un proceso estocástico puede describirse de manera alternativa como la colección de todas las posibles trayectorias que se pueden obtener acorde a la ley de probabilidad subyacente,

Definición 0.2 (Serie de tiempo, [Politis y McElroy 2020](#), p. 1). Una *serie de tiempo* es una colección de observaciones de una variable X tomadas en tiempo discreto, es decir, es un proceso estocástico en tiempo discreto⁵.

Definición 0.3 (Operador retardo⁶, [Shumway y Stoffer 2017](#), p. 56). El *operador retardo* se define como $\mathbf{B}X_t = X_{t-1}$. De manera análoga, este se extiende a potencias de mayor orden como, por ejemplo, $\mathbf{B}^2X_t = \mathbf{B}(\mathbf{B}X_t) = \mathbf{B}(X_{t-1}) = X_{t-2}$. Así, en general, se tiene que $\mathbf{B}^kX_t = X_{t-k}$.

Definición 0.4 (Filtro general y filtro lineal). Un *filtro general* ψ es una aplicación que tomando como entrada la serie de tiempo $\{X_t\}$ genera como salida otra serie de tiempo $\{Y_t\}$, es decir, establece la correspondencia $\{X_t\} \mapsto \{Y_t\}$. Se dice que el filtro ψ es un *filtro lineal* si se cumple que si $\{X_t\} \mapsto \{Y_t\}$ y $\{X'_t\} \mapsto \{Y'_t\}$ entonces para cualesquiera escalares a y b se cumple que $\{aX_t + bX'_t\} \mapsto \{aY_t + bY'_t\}$.



Figura 0.4: El concepto de filtro.
Extraído de [Politis y McElroy 2020](#), p. 57.

Observación 0.3 (Representación de un filtro lineal). El concepto de filtro lineal puede ser visto como una generalización de las transformaciones lineales de un espacio vectorial n -dimensional al caso de un espacio infinito dimensional. Para ello, considérense los vectores $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ y $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ junto con la matriz $A \in \mathbb{R}^{n \times n}$ asociada a la transformación lineal, esto es, $\mathbf{Y} = \mathbf{A}\mathbf{X}$. Si se considera la fila t -ésima de la ecuación anterior y se denota por $a_{i,j}$ al elemento de la matriz A que ocupa la fila i y la columna j , se tiene que $Y_t = \sum_{j=1}^n a_{t,j}X_j$. Considérese ahora el cambio de notación en la variable contadora dado por $j = t - k$. Se verifica que $Y_t = \sum_{k=t-n}^{t-1} a_{t,t-k}X_{t-k}$. Así pues, en el caso de los filtros lineales, que llevan la serie de tiempo $\{X_t\}_{t \in \mathbb{Z}}$ en $\{Y_t\}_{t \in \mathbb{Z}}$ puede pensarse en la representación dada por

$$Y_t = \sum_{k=-\infty}^{\infty} a_{t,t-k}X_{t-k},$$

⁵ Típicamente se tendrá que $\mathbb{T} = \mathbb{Z}$.

⁶ También existe el conocido como *operador de avance*, que se define como $\mathbf{F}X_t = X_{t+1}$. Es inmediato ver que es el inverso del operador retardo, es decir, $\mathbf{F} = \mathbf{B}^{-1}$.

para una matriz infinito dimensional A cuyas entradas son $a_{t,t-k}$, las cuales se conocen como *coeficientes del filtro lineal*. Nótese que, en general, los coeficientes de esta representación dependen del instante temporal t . En el caso en que $a_{t,t-k}$ depende solamente de k y no de t se dice que el filtro lineal es invariante respecto al tiempo, tal y como se recoge en la siguiente definición.

Definición 0.5 (Filtro lineal invariante respecto al tiempo, [Politis y McElroy 2020](#), p. 57). Se denomina *filtro lineal invariante respecto al tiempo* a una aplicación que, tomando como entrada una serie de tiempo $\{X_t\}_{k \in \mathbb{Z}}$ y generando como salida otra serie de tiempo $\{Y_t\}_{k \in \mathbb{Z}}$, tiene asociada una sucesión de coeficientes $\{a_k\}_{k \in \mathbb{Z}}$ tales que se cumple que

$$Y_t = \sum_{k=-\infty}^{\infty} a_k X_{t-k}, \quad \forall t \in \mathbb{Z}.$$

Observación 0.4. Si se denota por $A(\mathbf{B}) = \sum_{k=-\infty}^{\infty} a_k \mathbf{B}^k$ se verifica que para un filtro lineal invariante respecto al tiempo se tiene que

$$Y_t = \sum_{k=-\infty}^{\infty} a_k X_{t-k} = \sum_{k=-\infty}^{\infty} a_k \mathbf{B}^k X_t = A(\mathbf{B}) X_t, \quad \forall t \in \mathbb{Z},$$

lo cual justifica todavía más la analogía con respecto a las transformaciones lineales en espacios vectoriales finito dimensionales que fue establecida en la Observación 0.3.

Definición 0.6 (Función de medias, [Brockwell y Davis 2016](#), p. 13). Sea $\{X_t\}$ una serie de tiempo. La *función de medias* de $\{X_t\}$ viene dada por $\mu_t = \mathbb{E}[X_t]$.

Definición 0.7 (Función de autocovarianzas, [Brockwell y Davis 2016](#), p. 13). Sea $\{X_t\}$ una serie de tiempo con $\mathbb{E}[X_t^2] < \infty$. La *función de autocovarianzas* de $\{X_t\}$ viene dada por

$$\gamma(r, s) = \text{Cov}[X_r, X_s] = \mathbb{E}[(X_r - \mu_r)(X_s - \mu_s)], \quad \forall r, s \in \mathbb{Z}.$$

La *función de autocovarianzas a retardo h* se define como $\gamma(h) = \text{Cov}[X_t, X_{t+h}]$.

Definición 0.8 (Función de autocorrelaciones simples, [Brockwell y Davis 2016](#), p. 13). Sea $\{X_t\}$ una serie de tiempo. Se define la *función de autocorrelaciones simples* (abreviadamente, *fas*) de $\{X_t\}$ como

$$\rho(r, s) = \frac{\gamma(r, s)}{\sigma(r)\sigma(s)}, \quad \forall r, s \in \mathbb{Z},$$

donde $\sigma(t) = \sqrt{\text{Var}[X_t]}$. La *fas a retardo h* se define como $\rho_X(h) = \rho(t, t+h)$.

Observación 0.5 ([Shumway y Stoffer 2017](#), p. 96). Sean X, Y y Z variables aleatorias. La correlación parcial entre X e Y dada Z se obtiene haciendo primero una regresión de X sobre Z , para obtener la predicción \hat{X} , y luego haciendo una regresión de Y sobre Z , para obtener \hat{Y} , para luego calcular

$$\rho_{X,Y|Z} = \text{Corr}[X - \hat{X}, Y - \hat{Y}] \in [-1, 1].$$

Así, $\rho_{X,Y|Z}$ mide la correlación entre X e Y una vez ha sido eliminado de ellas el efecto de Z .

Definición 0.9 (Función de autocorrelaciones parciales, [Aneiros 2023](#), p. 30). Sea $\{X_t\}$ una serie de tiempo estacionaria. Se define la *función de autocorrelaciones parciales* (abreviadamente, *fap*) de $\{X_t\}$ como aquella función que otorga a los instantes de tiempo r y s el coeficiente de correlación lineal entre las observaciones X_r y X_s cuando se elimina de ellas la dependencia lineal debida a los valores intermedios, es decir,

$$\alpha(r, s) = \frac{\text{Cov}[X_r - \hat{X}_r^{(r,s)}, X_s - \hat{X}_s^{(r,s)}]}{\sqrt{\text{Var}[X_r - \hat{X}_r^{(r,s)}] \text{Var}[X_s - \hat{X}_s^{(r,s)}]}},$$

donde $\hat{X}_j^{(r,s)}$ denota al mejor predictor lineal de X_j construido mediante las variables en los instantes comprendidos estrictamente entre r y s . La *fap a retardo h* se define como $\alpha(h) = \alpha(t, t+h)$.

Definición 0.10 (Ruido blanco, [Brockwell y Davis 2016](#), p. 14). Se dice que un proceso estocástico $\{X_t\}$ es *ruido blanco* si para todo t se verifica que

$$\mu_t = 0, \quad \text{Var}[X_t] = \sigma^2, \quad \gamma(t, t+k) = 0, \quad \forall k \in \mathbb{Z} \setminus \{0\}.$$

Ejemplo 0.1 (Filtro lineal de ruido blanco invariante respecto al tiempo). Sea $\{a_t\}$ un proceso de ruido blanco. Se tiene que el proceso definido como

$$X_t = \sum_{k=0}^q \theta_k a_{t-k} = a_t + \theta_1 a_{t-1} + \cdots + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q},$$

donde $\theta_0 = 1$, es un filtro lineal del ruido blanco $\{a_t\}$ invariante respecto al tiempo, ya que basta considerar $\theta_k = 0$ para $k < 0$ y $k > q$. La serie $\{X_t\}$ que ha sido definida se denomina proceso de medias móviles de orden q , abreviadamente $\text{MA}(q)$, y será tratada con detalle en el Capítulo 1.

0.4.1. Estacionariedad de una serie de tiempo

Definición 0.11 (Distribuciones marginales, [Politis y McElroy 2020](#), p. 33). Sea $\{X_t\}$ un proceso estocástico. Para cualquier $m \geq 1$, la distribución conjunta de cualquier m -tupla de variables aleatorias de dicho proceso $\{X_{t_1}, \dots, X_{t_m}\}$ se conoce como *distribución marginal m -dimensional*.

Definición 0.12 (Estacionariedad estricta o fuerte, [Politis y McElroy 2020](#), p. 35). Una serie de tiempo $\{X_t\}$ se dice que es *estrictamente o fuertemente estacionaria* si la distribución de cada m -tupla de variables es la misma cuando los índices de todas ellas se desplazan en el tiempo, es decir, la distribución de cualquier vector aleatorio $(X_{t+1}, X_{t+2}, \dots, X_{t+m})$ es la misma que la de (X_1, X_2, \dots, X_m) para todo tiempo t y para cualquier $m \geq 1$.

Proposición 0.1 ([Politis y McElroy 2020](#), p. 36). Sea $\{X_t\}$ una serie de tiempo estrictamente estacionaria con momento de orden 2 finito, i.e., $\mathbb{E}[X_t^2] < \infty$ para todo $t \in \mathbb{Z}$. Entonces se tiene que $\mathbb{E}[X_t]$ no depende de t y $\gamma(t, t+k) = \text{Cov}[X_t, X_{t+k}]$ es una función de k que no depende de t .

Definición 0.13 (Estacionariedad débil, [Brockwell y Davis 2016](#), p. 13). Sea $\{X_t\}$ una serie de tiempo. Se dice que la serie es *débilmente estacionaria* si para todo t se cumple que

- 1) μ_t es finita e independiente de t , y
- 2) $\gamma(t, t+h)$ es finita e independiente de t para cada h .

Observación 0.6. Existen numerosas obras en la literatura relativa a las series de tiempo que al carácter débilmente estacionario lo denominan estacionariedad covariante o estacionariedad de segundo orden. Esta segunda denominación tiene su origen en considerar $m = 2$ en la estacionariedad estricta.

Ejemplo 0.2 (El ruido blanco es débilmente estacionario). A partir de la Definición 0.10 resulta inmediato ver que los procesos de ruido blanco son débilmente estacionarios.

Observación 0.7. La definición de estacionariedad débil implica una estabilidad en la función de medias y en la estructura de covarianzas a lo largo del tiempo, ya que establece que la media es constante y que la función de autocovarianzas entre dos instantes toma un valor que solamente depende de la separación que hay entre ellos. Además, como $\text{Var}[X_t] = \gamma(t, t)$, esta condición está imponiendo también que la varianza de las observaciones debe ser constante.

Tal y como están definidas la estacionariedad débil y la estricta, la Proposición 0.1 establece que toda serie de tiempo estrictamente estacionaria, con momento de orden 2 finito, es también débilmente estacionaria. No obstante, el recíproco no es cierto, como se verá en el siguiente ejemplo.

Ejemplo 0.3. Sea la serie de tiempo $\{X_t\}_{t \in \mathbb{Z}}$ dada por

$$X_t = \begin{cases} Z_t, & \text{si } t \text{ es par,} \\ \frac{1}{\sqrt{2}} (Z_{t-1}^2 - 1), & \text{si } t \text{ es impar,} \end{cases}$$

donde las variables aleatorias Z_t son independientes y están idénticamente distribuidas como una $N(0, 1)$. En cuanto al carácter débilmente estacionario, en lo relativo a la función de medias es sencillo ver que $\mathbb{E}[X_t] = 0$ para todo $t \in \mathbb{Z}$. Por otro lado, se tiene que la función de varianzas viene dada por

$$\gamma(0) = \text{Var}[X_t] = \begin{cases} \text{Var}[Z_t] = 1, & \text{si } t \text{ es par,} \\ \frac{1}{2} \text{Var}[Z_{t-1}^2] = 1, & \text{si } t \text{ es impar,} \end{cases}$$

donde en el caso impar se ha hecho uso de que una variable aleatoria gaussiana estándar al cuadrado se distribuye como una $\chi^2(1)$, cuya varianza es igual a 2⁷. Por otro lado, en el caso general de $\gamma(h)$, para $h \neq 0$, la independencia de las variables aleatorias Z_t implica que $\gamma(h) = \text{Cov}[X_t, X_{t+h}] = 0$. En consecuencia, por todo lo visto se tiene que el proceso $\{X_t\}_{t \in \mathbb{Z}}$ es débilmente estacionario (de hecho, es ruido blanco). En cuanto a la estacionariedad estricta, considerando $m = 1$ y para un $x_t \in \mathbb{R}$ fijo:

$$\mathbb{P}\{X_t \leq x_t\} = \begin{cases} \Phi(x_t), & \text{si } t \text{ es par,} \\ 2\Phi(\sqrt{\sqrt{2}}x_t + 1) - 1, & \text{si } t \text{ es impar,} \end{cases}$$

donde $\Phi(\cdot)$ representa a la función de distribución de una Normal estándar. Así pues, las distribuciones marginales 1-dimensionales cuando t es par son diferentes que cuando t es impar y, en consecuencia, el proceso $\{X_t\}_{t \in \mathbb{Z}}$ no es estrictamente estacionario.

Definición 0.14 (Serie de tiempo gaussiana, Politis y McElroy 2020, p. 34). Una serie de tiempo es gaussiana si para todo $m \geq 1$ la distribución marginal m -dimensional es una gaussiana multivariante⁸.

Observación 0.8 (Bajo gaussianidad, la estacionariedad débil equivale a la fuerte). Sea $\{X_t\}_{t \in \mathbb{Z}}$ una serie de tiempo gaussiana débilmente estacionaria y para un $m \geq 1$ considérese la m -tupla dada por $\mathbf{X}_0 = (X_1, X_2, \dots, X_m)'$, cuya distribución conjunta es una $N_m(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Debido al carácter débilmente estacionario se tiene que $\mu_t = \mathbb{E}[X_t] = \mu < \infty$, por lo que $\boldsymbol{\mu}_0 = (\mu, \mu, \dots, \mu)'$. Por otro lado, sea $\mathbf{X}_t = (X_{t+1}, X_{t+2}, \dots, X_{t+m})'$, que debido al carácter gaussiano sigue una distribución $N_m(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. La estacionariedad débil de la serie implica que

$$\begin{aligned} \boldsymbol{\mu}_t &= (\mu, \mu, \dots, \mu)' = \boldsymbol{\mu}_0, \\ (\boldsymbol{\Sigma}_t)_{ij} &= \text{Cov}[X_{t+i}, X_{t+j}] = \text{Cov}[X_i, X_j] = (\boldsymbol{\Sigma}_0)_{ij}, \quad \forall i, j \in \{1, 2, \dots, m\}. \end{aligned}$$

En consecuencia, como la distribución normal multivariante queda totalmente especificada a partir del vector de medias y la matriz de covarianzas se tiene que la distribución de \mathbf{X}_0 es la misma que la de \mathbf{X}_t , llegando así a que la serie $\{X_t\}_{t \in \mathbb{Z}}$ es estrictamente estacionaria.

0.4.2. Descomposición de Wold

Para finalizar este prefacio se va a dar una caracterización de la estacionariedad débil, para lo cual se comenzará primero introduciendo los llamados *procesos determinísticos* o *predecibles* a través de un sencillo ejemplo. A continuación de este, se comentará algunas definiciones relativas a la geometría del espacio de Hilbert $\{\mathbb{L}_2, \langle \cdot, \cdot \rangle\}$, cuya definición y algunas de sus características pueden ser consultadas en la Sección B.1, y que permitirán definir la predictibilidad de un proceso formalmente.

⁷ Para conocer algunas de las principales características de estas distribuciones puede consultarse el Apéndice D.

⁸ En la Sección D.2 pueden encontrarse algunas propiedades relativas a esta distribución.

Ejemplo 0.4 (Proceso determinístico). Considérese el proceso estocástico dado por

$$X_t = Y \cos(\omega t) + Z \sin(\omega t),$$

donde $\omega \in (0, \pi)$ es una constante e Y y Z son variables aleatorias de media 0 que están incorreladas y de varianza común igual a σ^2 . En primer lugar, unos sencillos cálculos permiten ver que el proceso $\{X_t\}$ es débilmente estacionario. Por otro lado, empleando algunas fórmulas trigonométricas básicas puede verse que $X_t = 2 \cos(\omega) X_{t-1} - X_{t-2}$. Esta propiedad de poder predecir de manera lineal y con error 0 el valor del proceso en un instante a partir de los valores pasados del proceso hasta dicho instante⁹ es lo que se conoce como predictibilidad del proceso.

En el ejemplo anterior ha quedado reflejada la idea de lo que es un proceso predecible. No obstante, es preciso definir este concepto de manera formal, lo cual requiere de algunos conceptos previos.

Definición 0.15 (Sistema lineal generado por un conjunto finito en $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$). Sean X_1, X_2, \dots, X_n variables aleatorias en el espacio de Hilbert $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$. El *sistema lineal generado* por esas variables es el conjunto de todas sus posibles combinaciones lineales y se denota como $sp\{X_1, X_2, \dots, X_n\}$.

Proposición 0.2 (Politis y McElroy (2020), p. 104). Sean X_1, X_2, \dots, X_n variables aleatorias en el espacio de Hilbert $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$. Se tiene que $sp\{X_1, \dots, X_n\}$ es un subespacio lineal cerrado de \mathbb{L}_2 .

La noción de sistema lineal generado por un conjunto finito de variables aleatorias en $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$ puede ser extendida al caso en que la colección de varias aleatorias sea infinita. No obstante, en ese caso no está garantizado que tal subespacio lineal sea cerrado, por lo que la definición en dicho caso se va a realizar de manera diferente.

Definición 0.16 (Sistema lineal cerrado generado por un conjunto infinito en $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$). Sea $\{X_t\}_{t \in \mathbb{T}}$ un conjunto de variables aleatorias en $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$. El *sistema lineal cerrado generado* por esas variables aleatorias se define como la clausura de $sp\{X_t, t \in \mathbb{T}\}$ y se denota como $\overline{sp}\{X_t, t \in \mathbb{T}\}$. Dicho de otro modo, consiste en todas las combinaciones lineales de las variables aleatorias, junto con los límites de las sucesiones de dicho subespacio lineal.

Notación. En lo que resta de Prefacio, el mejor predictor lineal de X_t , en el sentido de tener un menor $PMSE$ ¹⁰, basado en la muestra X_1, X_2, \dots, X_n se denotará como $P_{sp\{X_1, X_2, \dots, X_n\}}[X_t]$. El motivo por el cual se emplea esta notación en lugar de la habitual, \widehat{X}_t , es que existe un resultado que establece que el mejor predictor lineal de X_t basado en la muestra X_1, X_2, \dots, X_n es su proyección ortogonal sobre el espacio $sp\{X_1, X_2, \dots, X_n\}$. Para ver este resultado y profundizar acerca de la proyección de variables aleatorias puede consultarse la Sección B.4.3.

Definición 0.17 (Innovación, Politis y McElroy 2020, p. 226). Sea $\mathcal{M}_t = \overline{sp}\{X_s, s \leq t\}$. Se define el *error de predicción dados todos los valores pasados* en el instante t como

$$\varepsilon_t^{(\infty)} = X_t - P_{\mathcal{M}_{t-1}}[X_t].$$

Análogamente, se define la *innovación en el instante t o error de predicción a horizonte 1* dadas las pasadas n observaciones como $\varepsilon_t^{(n)} = X_t - P_{sp\{X_s, t-n \leq s < t\}}[X_t]$.

Definición 0.18 (Serie de tiempo predecible, Politis y McElroy 2020, p. 229). Sea $\{X_t\}$ una serie de tiempo débilmente estacionaria. Se dice que $\{X_t\}$ es *predecible* o *determinística* si se cumple que

$$\text{Var} \left[\varepsilon_t^{(\infty)} \right] = 0.$$

⁹ Aunque en este caso basta conocer el de los 2 instantes anteriores.

¹⁰ Siglas de *prediction mean squared error* o error cuadrático medio de predicción, que para un estimador \widehat{X}_t de X_t se define como

$$PMSE \left[\widehat{X}_t \right] = \mathbb{E} \left[\left(\widehat{X}_t - X_t \right)^2 \right].$$

Finalmente, se está ya en condiciones de enunciar la conocida como descomposición de Wold, la cual descompone las series de tiempo no determinísticas en un parte predecible y otra que, con la notación comentada en el Ejemplo 0.1, es un proceso MA(∞).

Teorema 0.3 (Descomposición de Wold¹¹, Politis y McElroy 2020, p. 230). *Sea $\{X_t\}_{t \in \mathbb{Z}}$ una serie de tiempo débilmente estacionaria de media 0 que no es predecible. Se tiene que la serie $\{X_t\}_{t \in \mathbb{Z}}$ puede descomponerse en la suma de dos series de tiempo $\{Y_t\}_{t \in \mathbb{Z}}$ y $\{U_t\}_{t \in \mathbb{Z}}$, es decir,*

$$X_t = Y_t + U_t, \quad \forall t \in \mathbb{Z},$$

tales que

- 1) $\{Y_t\}_{t \in \mathbb{Z}}$ y $\{U_t\}_{t \in \mathbb{Z}}$ están mutuamente incorreladas,
- 2) $\{Y_t\}_{t \in \mathbb{Z}}$ admite una representación MA(∞), esto es, $Y_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$, con $\psi_0 = 1$, para alguna serie de tiempo $\{a_t\}_{t \in \mathbb{Z}}$ que sea un proceso de ruido blanco de varianza σ^2 ,
- 3) a_t representa el error de predicción dados todos los valores pasados de Y_t (y por tanto la de X_t) en el instante t , es decir,

$$a_t = Y_t - P_{\overline{sp}\{Y_s, s < t\}} [Y_t] = X_t - P_{\overline{sp}\{X_s, s < t\}} [X_t];$$

por lo que σ^2 es la varianza del error de predicción, y

- 4) $\{U_t\}_{t \in \mathbb{Z}}$ es predecible.

Así pues, la descomposición de Wold establece que cualquier serie de tiempo débilmente estacionaria es la suma de un proceso de medias móviles de orden potencialmente infinito y una tendencia determinística. Más aún, en el caso de que dicha serie no tenga componentes determinísticas podrá buscarse el mejor predictor lineal en el marco de los procesos MA(∞), los cuales pueden ser aproximados por los modelos ARMA, que serán estudiados con detalle en el Capítulo 1.

NOTA. Se hará referencia al carácter débilmente estacionario simplemente como estacionario.

0.4.3. Predicción en series de tiempo

En esta sección se comenzará definiendo el concepto de intervalo de predicción para, a continuación, presentar un algoritmo recursivo, aplicable a cualquier serie de tiempo con momentos de orden 2 finitos, para realizar predicciones. La principal referencia que se ha consultado ha sido la Sección 2.5.4 de Brockwell y Davis (2016).

Definición 0.19 (Intervalo de predicción). Se entenderá por *intervalo de predicción* y se denotará por $\mathbb{I}(\mathbf{X})$ a un intervalo aleatorio que depende de las observaciones $\mathbf{X} \in \mathbb{R}^n$ y que contiene a un valor futuro $X_f \in \mathbb{R}$ con probabilidad fijada y conocida como *cobertura nominal*. Si se denota a esta por $1 - \alpha$ e $\mathbb{I}(\mathbf{X}) = (L(\mathbf{X}), U(\mathbf{X}))$ entonces se tendrá que

$$\mathbb{P}\{X_f \in \mathbb{I}(\mathbf{X})\} = \mathbb{P}\{L(\mathbf{X}) < X_f < U(\mathbf{X})\} = 1 - \alpha.$$

Sea $\{X_t\}$ una serie de tiempo con función $\mu_t = 0$ y $\mathbb{E}[X_t^2] < \infty$ para todo $t \in \mathbb{Z}$. Además, se define $\mathcal{M}_n = sp\{X_1, X_2, \dots, X_n\}$ y se denota el mejor predictor lineal de X_n a horizonte 1 basado en X_1, X_2, \dots, X_n y a su PMSE como

$$\nu_n = \mathbb{E}\left[\left(X_{n+1} - \widehat{X}_{n+1}\right)^2\right], \quad \text{con } \widehat{X}_n = \begin{cases} 0, & \text{si } n = 1, \\ P_{\mathcal{M}_{n-1}}[X_n], & \text{si } n = 2, 3, \dots \end{cases}$$

¹¹ Para estudiar con mayor detalle la descomposición general propuesta en Wold (1954) puede consultarse la Sección 7.6 de Politis y McElroy (2020).

Por otro lado, las innovaciones se denotarán como U_n , esto es, $U_n = X_n - \widehat{X}_n$. Con notación vectorial, definiendo $\mathbf{U}_n = (U_1, \dots, U_n)'$ y $\mathbf{X}_n = (X_1, \dots, X_n)'$ se tiene que la ecuación anterior puede reescribirse como $\mathbf{U}_n = A_n \mathbf{X}_n$, donde A_n tiene la forma

$$A_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{1,1} & 1 & 0 & \cdots & 0 \\ a_{2,2} & a_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1,n-1} & a_{n-1,n-2} & a_{n-1,n-3} & \cdots & 1 \end{pmatrix}, \quad A_n^{-1} = C_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1 \end{pmatrix},$$

obteniéndose los coeficientes de A_n a partir de las representaciones lineales¹² de los \widehat{X}_n y donde, en vista de la forma de las matrices A_n , son invertibles, con inversa de la forma C_n . Así, el vector de predictores lineales a horizonte 1 dado por $\widehat{\mathbf{X}}_n = (X_1, P_{\mathcal{M}_1}[X_2], \dots, P_{\mathcal{M}_{n-1}}[X_n])$ puede ser expresado como

$$\widehat{\mathbf{X}}_n = \mathbf{X}_n - \mathbf{U}_n = C_n \mathbf{U}_n - \mathbf{U}_n = \Theta_n \mathbf{U}_n = \Theta_n (\mathbf{X}_n - \widehat{\mathbf{X}}_n), \quad (1)$$

donde $\Theta_n = C_n - I_n$. Reescribiendo la ecuación (1) sin notación vectorial se tiene que

$$\widehat{X}_{n+1} = \begin{cases} 0, & \text{si } n = 0, \\ \sum_{j=1}^n \theta_{n,j} (X_{n+1-j} - \widehat{X}_{n+1-j}), & \text{si } n = 1, 2, \dots \end{cases}$$

En definitiva, toda vez que los coeficientes $\theta_{n,j}$ han sido determinados, los predictores a horizonte 1 pueden ser obtenidos de manera recursiva. Este procedimiento se recoge en el Algoritmo 1, en el cual se generan esos coeficientes junto con sus PMSE, ν_n , a partir de las covarianzas $\gamma(i, j)$. Así, mediante este algoritmo se calcularía sucesivamente $\nu_0; \theta_{1,1}, \nu_1; \theta_{2,2}, \theta_{2,1}, \nu_2; \theta_{3,3}, \theta_{3,2}, \theta_{3,1}, \nu_3; \dots$

Observación 0.9. En el Algoritmo 1 se entiende que cuando el límite superior del sumatorio, $j - 1$, es negativo, entonces dicho sumatorio es vacío.

Algoritmo 1 Algoritmo de innovaciones, [Brockwell y Davis 2016](#)

- 1: **Calcular:** $\nu_0 = \gamma(1, 1)$.
 - 2: **Repetir** para $k = 1, \dots, n$:
 - 3: **Repetir** para $j = 0, \dots, k - 1$:
 - 4: **Calcular:** $\theta_{k,k-j} = \nu_j^{-1} \left[\gamma(k+1, j+1) - \sum_{i=0}^{j-1} \theta_{j,j-i} \theta_{k,k-i} \nu_i \right]$.
 - 5: **Calcular:** $\nu_k = \gamma(k+1, k+1) - \sum_{i=0}^{k-1} \theta_{k,k-i}^2 \nu_i$.
-

Observación 0.10 (Predicción lineal a horizonte h). En la Sección 2.5.5 de [Brockwell y Davis \(2016\)](#) puede verse cómo el mejor predictor lineal de X_n a horizonte h basado en X_1, X_2, \dots, X_n es

$$P_{\mathcal{M}_n}[X_{n+h}] = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j} (X_{n+h-j} - \widehat{X}_{n+h-j}),$$

donde los coeficientes $\theta_{n,j}$ se obtienen de nuevo a partir del Algoritmo 1. Además, se tiene que el PMSE de este estimador viene dado por

$$PMSE[P_{\mathcal{M}_n}[X_{n+h}]] = \gamma(n+h, n+h) - \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}^2 \nu_{n+h-j-1}.$$

¹² En general, se tendrá que $\widehat{X}_n = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$. Por ejemplo, $\widehat{X}_2 = -a_{1,1} X_1$.

Capítulo 1

Modelos paramétricos de series de tiempo

1.1. Introducción a la metodología Box-Jenkins

Habiendo introducido en el Prefacio algunos conceptos básicos del ámbito de las series de tiempo, se presentan a continuación los modelos ARMA (*AR*, *autorregresivo*; *MA*, *moving average*), los cuales constituirán la base de los denominados modelos *ARIMA* (*Autoregressive Integrated Moving Average*), que serán presentados más adelante. Para ello, se comenzará introduciendo los modelos autorregresivos y los modelos de medias móviles.

Definición 1.1 (Proceso Autoregresivo, [Shumway y Stoffer 2017](#), p. 76). Un proceso estacionario $\{X_t\}$ que admite una representación de la forma

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t, \quad (1.1)$$

donde c, ϕ_1, \dots, ϕ_p ($\phi_p \neq 0$) son constantes y $\{a_t\}$ es un proceso de ruido blanco tal que a_t es independiente de X_{t-1}, X_{t-2}, \dots , se conoce con el nombre de *proceso autorregresivo de orden p* y se denota como $AR(p)$.

Observación 1.1. Si la media del proceso estacionario $\{X_t\}$ es μ se tiene que $c = \mu(1 - \phi_1 - \cdots - \phi_p)$.

Definición 1.2 (Proceso de Medias Móviles, [Shumway y Stoffer 2017](#), p. 81). Un proceso $\{X_t\}$ que admite una representación de la forma

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}, \quad (1.2)$$

donde $c, \theta_1, \dots, \theta_q$ ($\theta_q \neq 0$) son constantes y $\{a_t\}$ es un proceso de ruido blanco se conoce con el nombre de *proceso de medias móviles de orden q* ($MA(q)$).

Definición 1.3 (Proceso ARMA, [Shumway y Stoffer 2017](#), p. 83). Un proceso $\{X_t\}$ se dice que es un modelo *ARMA*(p, q) si es estacionario y admite una representación de la forma

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} \\ + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q},$$

donde $c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$ ($\phi_p \neq 0 \neq \theta_q$) son constantes y $\{a_t\}$ es un proceso de ruido blanco. Los parámetros p y q se conocen como órdenes autorregresiva y de medias móviles, respectivamente.

NOTA. La ecuación de un proceso $\{X_t\}$ que sigue un modelo ARMA(p, q) se suele escribir de forma abreviada como

$$\phi(\mathbf{B}) X_t = \theta(\mathbf{B}) a_t, \quad \forall t \in \mathbb{Z}, \quad (1.3)$$

donde

$$\begin{aligned} \phi(\mathbf{B}) &= (1 - \phi_1 \mathbf{B} - \phi_2 \mathbf{B}^2 - \dots - \phi_p \mathbf{B}^p), \\ \theta(\mathbf{B}) &= (1 + \theta_1 \mathbf{B} + \theta_2 \mathbf{B}^2 + \dots + \theta_q \mathbf{B}^q). \end{aligned}$$

Observación 1.2. En un modelo ARMA(p, q), cuando $q = 0$ se obtiene un proceso autorregresivo de orden p y, análogamente, cuando $p = 0$ se obtiene un proceso de medias móviles de orden q .

1.2. Estacionariedad, causalidad e invertibilidad de un proceso ARMA

1.2.1. Estacionariedad

Sea $\{X_t\}$ un proceso que sigue un modelo MA(q), esto es, que admite una representación de la forma

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q},$$

donde $c, \theta_1, \dots, \theta_q$ ($\theta_q \neq 0$) son constantes y $\{a_t\}$ es un proceso de ruido blanco de media 0 y varianza σ^2 . Uno sencillos cálculos permiten ver que su función de medias es nula para todo $t \in \mathbb{Z}$. En cuanto a su función de covarianzas, sea $h \in \mathbb{Z}$, con $h > 0$ ¹. Se tiene que

$$\begin{aligned} \gamma(h) &= \text{Cov}[X_t, X_{t+h}] = \mathbb{E}[X_t X_{t+h}] - \mathbb{E}[X_t] \mathbb{E}[X_{t+h}] \\ &= \mathbb{E} \left[\left(\sum_{j=0}^q \theta_j a_{t-j} \right) \left(\sum_{k=0}^q \theta_k a_{t+h-k} \right) \right] \\ &= \sum_{j,k=0}^q \theta_j \theta_k \sigma^2 \mathbb{I}\{k = h + j\} \\ &= \sigma^2 \sum_{j=0}^{q-h} \theta_j \theta_{h+j} < \infty, \end{aligned}$$

donde se ha considerado $\theta_0 = 1$ y se ha asumido que $q - h \geq 0$, ya que en caso contrario se tendría un sumando vacío y la función de autocovarianzas sería nula². De este modo, se ha visto que la función de autocovarianzas es finita e independiente de t para cada h , lo cual, junto con lo indicado sobre la función de medias, tiene como consecuencia que los procesos MA(q) son estacionarios.

Observación 1.3. Indirectamente, en el análisis anterior se ha visto que en un proceso MA(q) la función de autocorrelaciones simples se anula para todo retardo mayor que q , lo cual puede ser una herramienta muy útil a la hora de identificar el orden del proceso de forma gráfica.

Sea ahora $\{X_t\}$ un proceso que verifica la ecuación autorregresiva dada por (1.1). A diferencia de lo que ocurría en el caso de los procesos MA(q), en este tipo de modelos no se tendrá que, en general, el proceso $\{X_t\}$ sea estacionario.

¹ Dado que la función de autocorrelaciones simples es simétrica, es decir, $\gamma(-h) = \gamma(h)$, no se ha perdido generalidad con esta suposición.

² Esto se puede ver directamente a partir de la ecuación (1.2), ya que en la ecuación de X_t solo están involucrados los q valores pasados a partir de t y, en consecuencia, si $h \geq q$ se tendría que en las expresiones de X_t y de X_{t+h} no habría ningún término común.

Ejemplo 1.1. Sea $\{X_t\}$ un proceso que verifica la ecuación dada por $X_t = X_{t-1} + a_t$, donde $\{a_t\}$ es un proceso de ruido blanco de media 0 y varianza σ^2 e independiente del pasado de X_t . Supóngase que X_t es un proceso estacionario. En particular, se tendrá que

$$\sigma_X^2 = \gamma(0) = \text{Var}[X_t] = \text{Var}[X_0], \quad \forall t \in \mathbb{Z}. \quad (1.4)$$

No obstante, se cumple que

$$\sigma_X^2 = \text{Var}[X_t] = \text{Var}[X_{t-1} + a_t] = \text{Var}[X_{t-2}] + 2\sigma^2 = \dots = \text{Var}[X_0] + t\sigma^2,$$

lo cual es una contradicción con (1.4). En consecuencia, $\{X_t\}$ no es un proceso estacionario.

Observación 1.4. Aunque no se pueden establecer un resultado análogo al comentado en la Observación 1.3, en la página 98 de Shumway y Stoffer (2017) se puede ver que en un proceso $\text{AR}(p)$ se cumple que la función de autocorrelaciones parciales se anula para todo retardo mayor que p pero no en el retardo p , lo cual también será muy útil a la hora de identificar el orden del proceso autorregresivo de manera gráfica.

En definitiva, en términos de estacionariedad, se ha visto que los procesos de medias móviles siempre tienen esta propiedad, mientras que en el caso de los modelos autorregresivos esto no es cierto en general. Además, así como en el caso del modelo $\text{AR}(1)$ resulta sencillo ver que el proceso es estacionario si $|\phi| < 1$, existe un resultado que generaliza esto a los modelos $\text{ARMA}(p, q)$.

Teorema 1.1 (Existencia y unicidad de solución estacionaria, Brockwell y Davis 2016, pp. 74-75). *Sea la ecuación dada por*

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = a_t + \theta_1 a_{t-1} + \dots + \theta_q a_{t-q}.$$

Se tiene que existe una solución estacionaria, $\{X_t\}$, y es la única estacionaria, si, y solo si, se cumple que

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \neq 0, \quad \forall z \in \mathbb{C} \text{ tal que } |z| = 1.$$

1.2.2. Causalidad

Definición 1.4 (Causalidad, Politis y McElroy 2020, p.140). Un proceso $\{X_t\}$ que sigue un modelo $\text{ARMA}(p, q)$ se dice que es *causal* con respecto a sus entradas $\{a_t\}$ si existe una sucesión $\{\psi_j\}$ para $j \geq 0$ tal que

$$X_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}, \quad (1.5)$$

en cuyo caso la expresión anterior se denomina representación causal $MA(\infty)$ del proceso $\{X_t\}$.

Observación 1.5. Es inmediato ver que los procesos causales se trata de procesos aleatorios cuyos valores presentes no dependen de valores futuros. Además, nótese que aunque no se ha destacado de manera explícita, la igualdad de la definición involucra la convergencia de la serie que esta a la derecha.

Observación 1.6 (Convergencia cuadrática vs. Convergencia absoluta, Politis y McElroy 2020, pp. 140-141). En primer lugar conviene destacar que

$$\text{Var} \left[\sum_{j=0}^{\infty} \psi_j a_{t-j} \right] = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2.$$

Así pues, si la serie de los coeficientes ψ_j no converge de forma cuadrática, esto es, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, entonces la varianza es infinita y, en consecuencia, $\sum_{j=0}^{\infty} \psi_j a_{t-j} \notin \mathbb{L}_2$. De este modo, no podría darse

la igualdad (1.5), ya que X_t debe ser estacionario. Una condición necesaria más fuerte que se le podría exigir a la serie de los coeficientes es la convergencia absoluta, la cual implica la convergencia cuadrática. No obstante, el recíproco no es cierto, tal y como puede comprobarse con la serie $\sum_{k=1}^{\infty} \frac{1}{k^2}$.

Ejemplo 1.2 (Causalidad de los procesos MA(q)). Sea $\{X_t\}$ un proceso de MA(q), esto es,

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q},$$

donde $c, \theta_1, \dots, \theta_q$ ($\theta_q \neq 0$) son constantes y $\{a_t\}$ es un proceso de ruido blanco. Resulta inmediato ver que el proceso $\{X_t\}$ es causal, ya que basta tomar la sucesión $\{\psi_j\}_{j \geq 0}$ dada por

$$\psi_0 = 1, \quad \psi_j = \theta_j, \quad j = 1, 2, \dots, q, \quad \psi_j = 0, \quad j > q.$$

Ejemplo 1.3 (Causalidad de los procesos AR(1), Politis y McElroy 2020, pp. 137.139). Sea $\{X_t\}$ un proceso autorregresivo de orden 1 y de media 0, esto es, es un proceso estacionario que verifica la ecuación dada por

$$X_t = \phi_1 X_{t-1} + a_t, \tag{1.6}$$

donde ϕ_1 es constante y $\{a_t\}$ es un proceso de ruido blanco de varianza σ^2 . Se distinguen varios casos:

1) $|\phi_1| < 1$

Empleando recursivamente la ecuación (1.6) se tiene que $X_t = \phi_1^k X_{t-k} + \sum_{j=0}^{k-1} \phi_1^j a_{t-j}$. Ahora bien, como el proceso $\{X_t\}$ es estacionario, se tiene que su función de autocovarianzas es constante y finita. Además, como se ha supuesto que es un proceso de media cero esto implica que $\|X_t\|^2 = \mathbb{E}[X_t^2]$ es constante y finita³. En consecuencia, dado que en este caso $|\phi_1| < 1$ se cumple que

$$\lim_{k \rightarrow \infty} (\|\phi_1^k X_{t-k}\|) = \lim_{k \rightarrow \infty} (|\phi_1^k| \|X_{t-k}\|) = 0$$

Además, se cumple que

$$\sum_{j=0}^{\infty} \text{Var} [\phi_1^j a_{t-j}] = \sum_{j=0}^{\infty} \phi_1^{2j} \sigma^2 = \frac{\sigma^2}{1 - \phi_1^2} < \infty.$$

En consecuencia, por el Teorema B.3 se tiene que $\sum_{j=0}^{\infty} \phi_1^j a_{t-j}$ converge en \mathbb{L}_2 . Por lo tanto se tiene que $X_t = \sum_{j=0}^{\infty} \phi_1^j a_{t-j}$, es decir, el proceso $\{X_t\}$ es causal.

2) $|\phi_1| = 1$

Considérese el caso en que $\phi_1 = 1$ (el caso $\phi = -1$ sería tratado de manera análoga). La ecuación (1.6) en este caso se transforma en $X_t = X_{t-1} + a_t$. Supóngase que existe un proceso $\{X_t\}$ estacionario que satisface tal ecuación. En ese caso, tomando como condición inicial X_0 se cumple que para $h \in \mathbb{N}$

$$\text{Var} [X_h - X_0] = \text{Var} [X_{h-1} + a_h - X_0] = \cdots = \text{Var} \left[\sum_{k=1}^h a_k \right] = h\sigma^2.$$

Por otro lado, también se tiene que⁴

$$\text{Var} [X_h - X_0] = \text{Var} [X_h] + \text{Var} [X_0] - 2 \text{Cov} [X_h, X_0] = 2\gamma(0) - 2\gamma(h).$$

³ La norma que se está considerando es en el espacio de variables aleatorias con momento de orden 2 finito y puede consultarse su definición en el Apéndice B.1.

⁴ Véase la notación establecida en la Sección 0.4.

Así pues, se ha llegado a que

$$2\gamma(0) - 2\gamma(h) = h\sigma^2 \implies 2(1 - \rho(h)) = \frac{h\sigma^2}{\gamma(0)} \iff \rho(h) = 1 - \frac{h\sigma^2}{2\gamma(0)}$$

Ahora bien, como $\gamma(0) < \infty$ se tiene que la función $\rho(h)$ será menor que -1 siempre que h sea mayor que $\frac{4\gamma(0)}{\sigma^2}$, lo cual entra en contradicción con que $|\rho(h)| \leq 1$ para todo h .

3) $|\phi_1| > 1$

Dado que se busca una solución para todo $t \in \mathbb{Z}$ esto permite transformar la ecuación «hacia adelante» (1.6) en la ecuación equivalente dada por

$$X_t = \phi_1^{-1}X_{t+1} - \phi_1^{-1}a_{t+1}.$$

Ahora bien, como $|\phi_1^{-1}| < 1$ se está en un caso como el primero. Así pues, empleando argumentos similares se tiene que una solución estacionaria será la dada por

$$X_t = -\sum_{j=1}^{\infty} \phi_1^{-j} a_{t+j},$$

la cual depende únicamente de valores futuros del proceso de ruido blanco. Por ello, este proceso se acostumbra a llamar *anti-causal*.

Teorema 1.2 (Causalidad de un proceso ARMA, Politis y McElroy 2020, p. 141). *Sea $\{X_t\}$ un proceso ARMA(p, q) que satisface la ecuación (1.3) tal que los polinomios ϕ y θ no tienen raíces comunes. Entonces se tiene que $\{X_t\}$ es causal con respecto al proceso de ruido blanco $\{a_t\}$ si, y solo si, $\phi(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| \leq 1$. En tal caso, la solución estacionaria de la ecuación del proceso es la dada por*

$$X_t = \sum_{j=0}^{\infty} \psi_j a_{t-j},$$

donde los coeficientes ψ_j son tales que $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$, para todo $z \in \mathbb{C}$ tal que $|z| \leq 1$.

Observación 1.7. La ecuación que satisfacen los ψ_j es equivalente a que $\phi(z)\psi(z) = \theta(z)$. En consecuencia, igualando los coeficientes de z^j a ambos lados se tiene la fórmula recursiva dada por $\psi_j - \sum_{k=1}^p \phi_k \psi_{j-k} = \theta_j$, con $j = 0, 1, \dots$, donde $\theta_0 = 1, \theta_j = 0$ para $j > q$ y $\psi_j = 0$ para $j < 0$.

Corolario 1.3 (Politis y McElroy 2020, p. 143). *Sea $\{X_t\}$ un proceso ARMA(p, q) que satisface la ecuación (1.3) tal que los polinomios ϕ y θ no tienen raíces comunes. Si $\phi(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| = 1$, la única solución estacionaria de la ecuación del proceso es la dada por*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j a_{t-j},$$

donde los ψ_j son tales que $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$, para todo $z \in \mathbb{C}$ tal que $1/r < |z| < r$ para algún $r > 1$.

1.2.3. Invertibilidad

El Teorema 1.2 establece unas condiciones bajo las cuales un proceso que sigue un modelo ARMA puede ser expresado en términos del proceso de ruido blanco como un proceso MA(∞). De manera similar, podría pensarse si existe alguna manera de «invertir» el razonamiento y poder expresar cada elemento del proceso de ruido blanco como una serie infinita en función de los elementos del proceso original. Este nuevo enfoque da lugar a la propiedad conocida como *invertibilidad*.

Definición 1.5 (Invertibilidad, Politis y McElroy 2020, p. 144). Un proceso $\{X_t\}$ que sigue un modelo ARMA(p, q) que satisface la ecuación (1.3) se dice que es *invertible* con respecto a sus entradas $\{a_t\}$ si existe una sucesión $\{\pi_j\}_{j \geq 0}$ tal que

$$a_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

en cuyo caso la expresión anterior se denomina representación AR(∞) del proceso $\{X_t\}$.

Observación 1.8. Se emplea el término *invertible* debido a que a partir de la ecuación del proceso ARMA es inmediato ver que la serie de tiempo $\{X_t\}$ es la salida de un filtro lineal que tiene a la serie $\{a_t\}$ como entrada y, si es posible invertir ese filtro, entonces se puede recuperar la serie $\{a_t\}$ a partir de $\{X_t\}$, tal y como se ilustra en la Figura 1.1. La utilidad de disponer de un proceso invertible se puede ver en la práctica cuando es posible observar una muestra finita de la serie de tiempo $\{X_t\}$ pero el proceso de ruido blanco no es observable. En esa situación, si dicha serie es invertible es posible capturar el comportamiento de a_t únicamente a partir de los valores presentes y pasados de X_t .

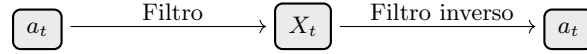


Figura 1.1: El concepto de invertibilidad.
Inspirado en Politis y McElroy 2020, p. 145.

Observación 1.9. En el Teorema 1.2 lo que se hace, en esencia, es despejar X_t en la ecuación (1.3) multiplicando por la inversa de $\phi(\mathbf{B})$ en el caso en que esté bien definida y obteniendo como solución

$$X_t = \frac{\theta(\mathbf{B})}{\phi(\mathbf{B})} a_t.$$

No obstante, la ecuación (1.3) es formalmente simétrica, en el sentido de que a ambos lados de la igualdad aparece una serie de tiempo multiplicada cada una de ellas por un polinomio diferente. En consecuencia, uno podría pensar en aplicar un razonamiento análogo al realizado en el caso de la causalidad y multiplicar dicha ecuación por la inversa de $\theta(\mathbf{B})$, supuesto que esta esté bien definida, es decir, $a_t = \theta(\mathbf{B})^{-1} \phi(\mathbf{B}) X_t$. Este resultado se recoge en el siguiente teorema, a continuación del cual aparece la Figura 1.2, en la cual se muestra de manera explícita el filtro y la inversa a los que se hacía referencia en la Figura 1.1.

Teorema 1.4 (Invertibilidad de un proceso ARMA, Politis y McElroy 2020, p. 144). Sea $\{X_t\}$ un proceso ARMA(p, q) que satisface la ecuación (1.3) tal que los polinomios ϕ y θ no tienen raíces comunes. Entonces se tiene que $\{X_t\}$ es invertible con respecto al proceso de ruido blanco $\{a_t\}$ si, y solo si, $\theta(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| \leq 1$. En tal caso, se tiene que

$$a_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

donde los coeficientes π_j son tales que $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}$, para todo $z \in \mathbb{C}$ tal que $|z| \leq 1$.

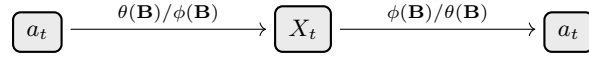


Figura 1.2: Filtro lineal de un proceso ARMA y su inversa.
Inspirado en Politis y McElroy 2020, p. 146.

Corolario 1.5 (Politis y McElroy 2020, p. 145). Sea $\{X_t\}$ un proceso ARMA(p, q) que satisface la ecuación (1.3) tal que los polinomios ϕ y θ no tienen raíces comunes. Si $\theta(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| = 1$, entonces puede expresarse

$$a_t = \sum_{j=-\infty}^{\infty} \pi_j X_{t-j},$$

donde los ψ_j son tales que $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$, para todo $z \in \mathbb{C}$ tal que $1/r < |z| < r$ para algún $r > 1$.

Por último, cabe mencionar que a partir de los resultados que han sido presentados es posible establecer conjuntamente condiciones que garanticen tanto el carácter causal como el invertible.

Corolario 1.6 (Causalidad e invertibilidad de un proceso ARMA, Politis y McElroy 2020, p. 146). Sea $\{X_t\}$ un proceso ARMA(p, q) que satisface la ecuación (1.3) tal que los polinomios ϕ y θ no tienen raíces comunes. Se tiene que $\{X_t\}$ es un proceso causal e invertible con respecto a $\{a_t\}$ si, y solo si, se cumple que

$$\theta(z)\phi(z) \neq 0, \quad \forall z \in \mathbb{C} \text{ tal que } |z| \leq 1.$$

En tal caso se tiene que

$$X_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}, \quad a_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

donde los coeficientes ψ_j y π_j satisfacen las ecuaciones

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad \pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad \forall z \in \mathbb{C} \text{ tal que } |z| \leq 1.$$

1.3. Extendiendo los modelos ARMA

Los modelos ARMA que se han presentado tienen como peculiaridad que se trata de modelos estacionarios. Sin embargo, en la realidad no suele ser habitual encontrarse con procesos de esa naturaleza (presencia de tendencias, patrones repetitivos...) y por ese motivo es necesario disponer de modelos apropiados para modelar dichas situaciones.

En muchos de esos ejemplos en los cuales no se tiene estacionariedad esta puede obtenerse sin más que *diferenciar la serie regularmente*⁵, que es un proceso que consiste en restarle a la serie el valor en el instante anterior, es decir, si se denota por ∇ a tal operación se tiene que para todo t

$$\nabla X_t = X_t - X_{t-1}.$$

⁵ En algunas ocasiones bastará con realizar esta operación una vez y en otras será necesario realizarla varias veces.

Observación 1.10. Empleando el operador retardo definido anteriormente puede reescribirse la operación de diferenciar regularmente d veces un proceso como $\nabla^d X_t = (1 - \mathbf{B})^d X_t$.

Observación 1.11. Esta operación diferencia tiene como utilidades que permite eliminar, por un lado, raíces unitarias y, por otro, tendencias determinísticas polinómicas (en general, diferenciar regularmente d veces ayuda a eliminar tendencias de grado d).

Así pues, habiendo introducido todos los conceptos básicos necesarios y aclarando, donde era necesario, algunos aspectos de los mismos, se está ya en condiciones de definir los conocidos como modelos ARIMA.

Definición 1.6 (Procesos ARIMA, [Shumway y Stoffer 2017](#), p. 132). Un proceso $\{X_t\}$ se dice que es un proceso ARIMA(p, d, q) si tras aplicarle d diferencias regulares se convierte en un proceso ARMA(p, q), es decir, si se tiene que $\nabla^d X_t = (1 - \mathbf{B})^d X_t$ es un proceso ARMA(p, q).

Ahora que ya se han introducido los modelos ARIMA, el siguiente paso es disponer de herramientas adecuadas para poder ajustarlos a diferentes conjuntos de datos. Para finalizar esta sección, se recogen a continuación las etapas que se distinguen en este proceso de ajuste y en las cuales, por una cuestión de extensión del trabajo no se profundizará demasiado

1) Representación secuencial de los datos.

2) Posible transformación del conjunto de datos.

Si hay problemas con la homocedasticidad, puede pensarse en aplicar una transformación logarítmica o de tipo Box-Cox para estabilizar la varianza.

3) Identificación de los órdenes del modelo.

En primer lugar, conviene recordar lo que se ha comentado acerca de la *fas* de los modelos $MA(q)$ y de la *fap* de los modelos $AR(p)$. Dado un conjunto de datos, se pueden representar las correspondientes funciones de autocorrelaciones muestrales y analizar su comportamiento, en el sentido de ver en qué retardo se tiene que la correspondiente función es significativamente distinta de cero⁶. Si tal retardo es el r , nos indicará que se trata de un modelo $AR(k)$ o $MA(k)$ según se haya considerado la función de autocorrelaciones muestral simple o parcial.

Además, cabe mencionar que para poder identificar modelos ARIMA más complejos de órdenes p y q que no sean necesariamente uno de ellos nulo existen diferentes métodos basados en diferentes criterios (criterio de información de Akaike, criterio de información de Bayes...). Para profundizar más acerca de estos procedimientos puede consultarse la Sección 5.5 de [Brockwell y Davis \(2016\)](#).

4) Estimación de los parámetros del modelo propuesto en 3).

Suele hacerse por máxima verosimilitud. Para conocer los detalles véase la Sección 5.2 de [Brockwell y Davis \(2016\)](#).

5) Diagnóstico del modelo propuesto en 3).

En la Sección 5.3 de [Brockwell y Davis \(2016\)](#) pueden consultarse los detalles del proceso de diagnóstico.

6) Selección del modelo tras finalizar las diagnósticos de los modelos propuestos.

⁶ Nótese que aquí de manera indirecta se ha hecho referencia a un contraste de hipótesis acerca de las variables aleatorias que representan la *fas* y la *fap* muestrales. En la Sección 3.5.2 de [Peña \(2005\)](#) puede verse que, si se denota por T al mayor instante temporal en el que ha sido observada la serie, bajo condiciones bastante generales se tiene que sus distribuciones, bajo el modelo correspondiente, son una $N\left(0, 1/\sqrt{T}\right)$. Esto permite construir en los gráficos secuenciales de ambas funciones de autocorrelaciones las bandas de confianza con un nivel aproximado del 5% dadas por $\left(-1,96/\sqrt{T}, 1,96/\sqrt{T}\right)$.

1.4. Predicción en modelos autorregresivos

En primer lugar, conviene destacar que en la Sección 3.3 de [Brockwell y Davis \(2016\)](#) puede verse cómo en modelos $\text{ARMA}(p, q)$ causales, la aplicación del Algoritmo 1 es más sencilla, ya que, por ejemplo, en los procesos $\text{MA}(q)$ se tiene que $\gamma(h) = 0$ para $h > q$, por lo que algunos de los coeficientes $\theta_{n,j}$ se anulan. No obstante, en lugar de optar por la aplicación del algoritmo de innovaciones, en esta sección se llevará a cabo un enfoque diferente basado directamente en la esperanza condicionada.

Para obtener las ecuaciones de las predicciones en los modelos autorregresivos supóngase, en primer lugar, que se está ante un modelo $\text{AR}(1)$, donde el parámetro ϕ_1 se supone conocido. Fijando primeramente un horizonte temporal de predicción $h = 1$, el objetivo que se persigue es obtener un predictor lineal óptimo⁷ para X_{t+1} dados X_t, X_{t-1}, \dots, X_1 . Para ello, el criterio de optimalidad a considerar es minimizar el error cuadrático medio de predicción.

En consecuencia, en base a los resultados de la Sección B.4.3, el predictor lineal óptimo se podrá obtener mediante el cálculo de la esperanza condicionada dada por $\mathbb{E}[X_{t+1} | X_t, X_{t-1}, \dots, X_1]$. Así pues, en virtud de la ecuación autorregresiva que sigue el proceso se cumple que

$$\mathbb{E}[X_{t+1} | X_t, X_{t-1}, \dots, X_1] = \phi_1 X_t,$$

Así pues, el predictor lineal óptimo será el dado por $\widehat{X}_{t+1}(X_1, \dots, X_t) = \phi_1 X_t$, cuyo error cuadrático medio de predicción es

$$PMSE \left[\widehat{X}_{t+1}(X_1, \dots, X_t) \right] = \mathbb{E} \left[\left(X_{t+1} - \widehat{X}_{t+1}(X_1, \dots, X_t) \right)^2 \right] = \sigma^2.$$

Considérese ahora un horizonte temporal h . En primer lugar, recuérdese que empleando recursivamente la ecuación del modelo se tiene que $X_{t+h} = \phi_1^h X_t + \sum_{j=0}^{h-1} \phi_1^j a_{t+h-j}$. En consecuencia, calculando la esperanza condicionada se tiene que el predictor lineal que se considerará en este caso es $\widehat{X}_{t+h}(X_1, \dots, X_t) = \phi_1^h X_t$.

Las ideas que se han ilustrado en el caso del modelo autorregresivo de orden 1 pueden generalizarse de manera inmediata al caso de los modelos $\text{AR}(p)$, con $p \geq 2$. Así, para $h = 1$ en este caso se tiene que

$$\widehat{X}_{t+1}(X_1, \dots, X_t) = \mathbb{E}[X_{t+1} | X_t, X_{t-1}, \dots, X_1] = \sum_{k=1}^p \phi_k X_{t+1-k}.$$

Sin embargo, la mayor complejidad de la estructura autorregresiva lleva a un cambio en la expresión del predictor ya en el caso en que $h = 2$. Para verificar este hecho obsérvese primero la expresión de X_{t+2} siguiente:

$$\begin{aligned} X_{t+2} &= \phi_1 X_{t+1} + \sum_{k=2}^p \phi_k X_{t+2-k} + a_{t+2} \\ &= \phi_1 \left(\sum_{k=1}^p \phi_k X_{t+1-k} + a_{t+1} \right) + \sum_{k=2}^p \phi_k X_{t+2-k} + a_{t+2} \\ &= \phi_1 \left(\sum_{k=1}^p \phi_k X_{t+1-k} \right) + \sum_{j=1}^{p-1} \phi_{j+1} X_{t+1-j} + \phi_1 a_{t+1} + a_{t+2} \\ &= \sum_{k=1}^p (\phi_1 \phi_k + \phi_{k+1}) X_{t+1-k} + \phi_1 a_{t+1} + a_{t+2}, \end{aligned}$$

⁷ En este caso se ha restringido la búsqueda de predictores óptimos a la clase de predictores lineales pero podría haberse realizado otra restricción sobre los mismos.

donde, en adelante, se tendrá que $\phi_{p+k} = 0$ para $k = 1, 2, \dots$ ⁸ Calculando la esperanza condicionada de la última de las igualdades se llega a que en este caso el predictor lineal óptimo es el dado por

$$\begin{aligned}\widehat{X}_{t+2}(X_1, \dots, X_t) &= \sum_{k=1}^p (\phi_1 \phi_k + \phi_{k+1}) X_{t+1-k} \\ &= \phi_1 \sum_{k=1}^p \phi_k X_{t+1-k} + \sum_{k=1}^p \phi_{k+1} X_{t+1-k} \\ &= \phi_1 \widehat{X}_{t+1}(X_1, \dots, X_t) + \sum_{k=2}^p \phi_k X_{t+2-k}\end{aligned}$$

En definitiva, a partir de esto último puede probarse que el mejor predictor lineal de X_{t+h} en el caso general de un modelo AR(p), con $p \geq 2$, es el dado por

$$\widehat{X}_{t+h}(X_1, \dots, X_t) = \begin{cases} \sum_{k=1}^{h-1} \phi_k \widehat{X}_{t+h-k}(X_1, \dots, X_t) + \sum_{k=h}^p \phi_k X_{t+h-k}, & \text{si } h = 2, 3, \dots, p, \\ \sum_{k=1}^p \phi_k \widehat{X}_{t+h-k}(X_1, \dots, X_t), & \text{si } h = p+1, p+2, \dots \end{cases}$$

1.5. Aplicación a los datos medioambientales

En los primeros años de desarrollo de mecanismos de predicción para la central térmica de As Pontes la concentración media de SO₂ observada durante un lapso de tiempo de 2 horas y cuya frecuencia de transmisión de datos era cada 5 minutos no podía superar ciertos umbrales establecidos por la legislación vigente en aquel entonces. Así pues, para mantener el control sobre las concentraciones medias bihorarias de esta sustancia la central eléctrica debía tomar decisiones con antelación suficiente y con los recursos de los que disponían se requerían alrededor de 30 minutos desde la toma de la decisión hasta que esta se hacía efectiva.

Con este planteamiento, los primeros modelos predictivos que se consideraron para este agente contaminante trabajaban con series de medias bihorarias, esto es, denotando como SO₂(t) a la concentración de SO₂ en tiempo t (pentaminutal) medida en $\mu\text{g}/\text{m}^3$, con

$$X_t = \frac{1}{24} \sum_{i=0}^{23} SO_2(t-i).$$

Así, en cada instante t se recibe una nueva observación de la serie X_t y se debe predecir X_{t+6} conociendo X_t, X_{t-1}, \dots . Dicho de otro modo, el problema que se pretende abordar consiste en realizar para cada serie de SO₂ una predicción cada 5 minutos de la concentración media bihoraria que se tendrá 6 unidades de tiempo (30 minutos) más tarde.

Para llevar a cabo esa predicción empleando un modelo paramétrico de tipo Box-Jenkins los investigadores decidieron que era suficiente con emplear 72 medias bihorarias sucesivas (6 horas) y, además, también se tomó la decisión de que los parámetros del modelo ARIMA (p, d y q) se actualizaran cada 5 minutos. Además, definieron en cada instante de tiempo lo que se conoce como «serie activa» y que no es más que la serie de tiempo muestral formada por las 72 observaciones de las últimas 6 horas.

⁸ Este convenio se ha tomado con el fin de compactar la escritura de las fórmulas de los predictores lineales.

No obstante, tal y como fue comentado en el Prefacio, la serie X_t posee ciertas características que provocan que las predicciones obtenidas mediante los modelos Box-Jenkins no se comporten adecuadamente. De entre ellas cabría destacar que esta serie toma valores cercanos a 0 durante largos periodos de tiempo (días o semanas) y luego estos periodos de «comportamiento estacionario» se ven interrumpidos de repente por los llamados incidentes, que se corresponden con un valor de la serie superior a los $300 \mu\text{g}/\text{m}^3$ ⁹, y cuya duración acostumbra a ser de entre 3 y 4 horas. Además, estos episodios son muy diferentes entre sí en cuanto a duración, máximos y, en general, la forma que presentan a la hora de representarlos en un gráfico temporal.

Teniendo en cuenta el objetivo que se ha fijado, lo deseable sería disponer de una muestra lo suficientemente grande como para contener información de una cantidad representativa de incidentes. Analizando lo ocurrido en el pasado con los tiempos entre incidentes hasta entonces, en [Prada-Sánchez y Febrero-Bande \(1997\)](#) vieron que para esto sería necesario considerar una muestra de al menos 1 año, lo cual se corresponde con un tamaño de muestra de algo más de 105 000. Esto conduciría no solo a trabajar con muestras computacionalmente muy costosas sino que, además, la mayor parte de sus valores estarían cercanos a 0, por lo que se estaría realizando un gran esfuerzo por información que no aporta nada a la predicción de los incidentes.

Para solventar esto, fue desarrollada una metodología diferente a la hora de almacenar los datos de la concentración de SO_2 y crearon la llamada *matriz histórica*, la cual consiste en un registro de 500 vectores¹⁰ de la forma (X_{t-1}, X_t, X_{t+6}) y que será tratada con mayor detalle en el siguiente apartado.

1.5.1. La matriz histórica

En este apartado se va a especificar tanto la construcción inicial de la matriz histórica como el proceso mediante el cual se va actualizando. La principal referencia que ha sido consultada para la elaboración de esta sección ha sido [Prada-Sánchez y Febrero-Bande \(1997\)](#).

Matriz semilla y mecanismo de actualización

Por todo lo comentado en el apartado anterior se tomó la decisión de no considerar muestras al estilo usual, sino emplear un mecanismo conocido como *matriz histórica*, el cual consiste en condensar la información de un largo periodo de tiempo en una muestra de tamaño razonablemente menor empleando información incompleta aunque representativa. Así, a grandes rasgos lo que se pretendía hacer era, ya que se tenía la intención de emplear 2 variables predictoras, emplear como apoyo para la predicción el conjunto de 1000 ternas dado por

$$\{((x_1^1, x_2^1), x_3^1), \dots, ((x_1^{1000}, x_2^{1000}), x_3^{1000})\},$$

donde cada x_3^i es un «resumen» o «valor representativo» del rango de medias bihorarias no cercanas a 0 observadas durante un largo periodo y es incluido en la terna junto con el «par predictivo» (x_1^i, x_2^i) .

Más concretamente, el primero de los pasos consiste en determinar el rango de valores de medias bihorarias no cercanas a 0 durante los 2 años anteriores. A continuación, este rango es dividido en 10 estratos, de manera que cada uno de ellos contenga aproximadamente el mismo número de datos. Luego, se seleccionan de manera aleatoria 100 valores x_3^i de cada uno de los estratos y con ellos se forman las ternas $((x_1^i, x_2^i), x_3^i)$, donde x_1^i y x_2^i son, respectivamente, las medias bihorarias observadas 35 y 30 minutos antes del instante asociado a x_3^i .

⁹ Se pueden alcanzar máximos de $1200 \mu\text{g}/\text{m}^3$ o incluso 2400 en los casos más extremos. No obstante, es muy poco frecuente que se den en 2 estaciones de medición o más a la vez, lo cual se traduce en que no sea necesario trabajar de manera conjunta con todas ellas.

¹⁰ Aunque al comienzo esta era la dimensión, en versiones posteriores llega a tener 2000 entradas.

Así pues, con el procedimiento que ha sido descrito se construye la semilla M_0 de la matriz histórica M_t . Una vez definida dicha semilla, la actualización de la matriz histórica cuando sucede una media bihoraria x_t distinta de 0 se lleva a cabo como sigue:

- 1) Se determina a cuál de los 10 estratos pertenece x_t .
- 2) Se elimina la terna más antigua del estrato determinado en el paso anterior en la matriz histórica M_{t-1} .
- 3) Se añade la terna $((x_{t-7}, x_{t-6}), x_t)$ al estrato de M_{t-1} determinado en el primer paso para obtener la nueva matriz histórica M_t .

Capítulo 2

El método Bootstrap

2.1. Breve introducción al método bootstrap

En esta sección, que es un breve resumen del Capítulo 1 de [Cao y Fernández \(2023\)](#), se llevará a cabo una introducción a la metodología bootstrap, poniendo de manifiesto la analogía existente entre esta y el paradigma inferencial clásico.

Una primera aproximación hacia lo que es el método bootstrap es que se trata de un procedimiento estadístico en el cual se lleva a cabo un proceso de remuestreo, esto es, se obtienen nuevas muestras empleando algún mecanismo aleatorio que emplee la verdadera muestra original, con el fin de aproximar la distribución de algún estadístico. Dos de los principales motivos por los cuales este método ha alcanzado una gran expansión en cuanto a sus aplicaciones en diferentes contextos estadísticos (datos funcionales, series de tiempo, contrastes de hipótesis...) son, por un lado, el hecho de que no requiere ninguna hipótesis sobre el mecanismo generador de los datos¹ y, por otro, que su implementación en ordenador suele resultar sencilla en comparación con otros métodos. No obstante, relacionado con esta segunda ventaja, se encuentra también su principal inconveniente y es que, tal y como se verá más adelante, este método necesita en muchas ocasiones del uso de técnicas de Monte Carlo, lo cual debido a la naturaleza de este método redundaba en la necesidad de mayor computación intensiva. Es por ello que durante algunos años la aplicación de este método, si bien se encontraba en auge, estaba limitada por la capacidad de cálculo de los ordenadores de la época. Hecho que, por el contrario, no suele ser un problema con los equipos que existen hoy en día.

Descrito someramente el método bootstrap, a continuación se llevará a cabo una comparativa entre el paradigma inferencial clásico y su análogo bootstrap, con el fin de clarificar la utilidad del mismo.

Paradigma inferencial clásico

Supóngase que $\mathbf{X} = (X_1, \dots, X_n)$ es una muestra aleatoria simple de una población que sigue una distribución F y que se tiene interés en realizar inferencia sobre cierto parámetro poblacional $\theta = \theta(F)$. Para ello, se considera un estadístico que es función de la muestra y de la propia distribución poblacional como, por ejemplo

$$R(\mathbf{X}, F) = \theta(F_n) - \theta(F) = \hat{\theta} - \theta,$$

donde F_n representa a la función de distribución empírica. Para poder realizar inferencia sobre $\theta(F)$ sería interesante conocer la distribución en el muestreo del estadístico $R(\mathbf{X}, F)$, que en algunas ocasio-

¹ Aunque sí que son necesarias algunas, que acostumbran a ser más relajadas, para obtener propiedades asintóticas del procedimiento bootstrap.

nes concretas es posible calcular directamente y en otras puede ser aproximada de manera asintótica.

Escenario bootstrap

En el contexto que ha sido descrito en el apartado anterior, la metodología bootstrap comienza reemplazando la desconocida distribución poblacional, F , por una estimación de la misma, que se denotará por \hat{F} (e.g. $\hat{F} = F_n$, en el caso del *bootstrap naïve* o *uniforme* o $\hat{F} = F_{\hat{\theta}}$, en el caso del *bootstrap paramétrico*). A partir de esta distribución aproximada es posible generar, de manera condicionada a la muestra observada, lo que se denominarán *remuestras bootstrap*:

$$\mathbf{X}^* = (X_1^*, \dots, X_n^*),$$

donde toda $X_i^*, i = 1, \dots, n$, tiene distribución \hat{F} . Así pues, se puede hablar de la *distribución en el remuestreo* o *distribución bootstrap* de $R^* = R(\mathbf{X}^*, \hat{F})$.

La idea original de la metodología bootstrap (véase [Efron 1979](#)) consiste en que la distribución de θ^* en torno a $\hat{\theta}$ aproxima la distribución de $\hat{\theta}$ en torno a θ . De este modo, se busca aproximar la distribución en el muestreo de R mediante la distribución en el remuestreo de R^* . No obstante, la distribución bootstrap de R^* es poco frecuente que se pueda calcular de manera directa, por lo que casi siempre suele aproximarse mediante técnicas de Monte Carlo.

2.1.1. Consideraciones sobre los métodos bootstrap

Tal y como se afirma en [Cao \(1999\)](#), a la hora de llevar a cabo la implementación práctica de un procedimiento bootstrap es necesario considerar dos cuestiones:

- 1) ¿Cuán buena es la aproximación bootstrap? ¿Es asintóticamente correcta²?
- 2) ¿Cómo puede ser calculada (o, al menos aproximada) la distribución bootstrap del estadístico?

La primera de las preguntas es lo que se conoce como *validez del bootstrap* y establece la consistencia del método. Típicamente, se considera alguna distancia entre la distribución bootstrap de R^* y la distribución en el muestreo de R y se comprueba si dicha distancia tiende a 0 a medida que el tamaño de muestra tiende a infinito. Además, otro punto importante en este contexto es analizar la velocidad de convergencia del método, ya que en muchas ocasiones existen metodologías alternativas con las que resulta muy interesante comparar sus rendimientos.

Con el fin de proporcionar al lector algunas herramientas para cubrir los aspectos mencionados en el párrafo anterior, se presentan a continuación dos resultados que emplean los denominados *desarrollos de Edgeworth* de un estadístico. No obstante, antes de introducir el primero de ellos, conviene recordar el conocido como *Teorema Central del Límite*.

Teorema 2.1 (Teorema de Lévy-Lindeberg, [Petrov y Mordecki \(2008\)](#), p. 157). *Sean X_1, \dots, X_n, \dots variables aleatorias independientes e idénticamente distribuidas, con $E[X_1] = \mu$ y $\text{Var}[X_1] = \sigma^2 < \infty$. Entonces se tiene que*

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} N(0, 1).$$

Así pues, denotando

$$R = R(\mathbf{X}, F) = \sqrt{n} \frac{\bar{X} - \mu}{\sigma},$$

² Signifique lo que signifique según el contexto de aplicación.

lo que establece el Teorema 2.1 es que la distribución asintótica de R es una gaussiana estándar. Ahora bien, ¿cómo de buena es esta aproximación? o, dicho de otro modo, ¿cómo de rápida es la convergencia a la Normal? La respuesta a estas preguntas fue dada por el matemático sueco Harald Cramér en el año 1928 mediante un resultado que permite descomponer la distribución del estadístico en una suma de términos que facilitan el análisis de su comportamiento asintótico³.

Teorema 2.2 (Cramér 1928a). Sean $X_1, X_2, \dots, X_n, \dots$ variables aleatorias independientes e idénticamente distribuidas con distribución común F , con media μ y desviación típica σ . Supóngase que existe cierto $j \in \mathbb{N}$ tal que $\mathbb{E}[|X|^{j+2}] < \infty$ y que, denotando como $\varphi_X(t) = \mathbb{E}[e^{itX}]$ la función característica poblacional, se cumple que $\lim_{|t| \rightarrow \infty} |\varphi_X(t)| < 1$. Bajo tales condiciones se tiene que

$$\begin{aligned} \mathbb{P}\{R \leq u\} &= \mathbb{P}\left\{\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \leq u\right\} \\ &= \Phi(u) + n^{-1/2} p_1(u) \phi(u) + \dots + n^{-(j-1)/2} p_{j-1}(u) \phi(u) + O\left(n^{-j/2}\right), \end{aligned}$$

donde $\Phi(u)$ es la función de distribución de una distribución Normal estándar, $\phi(u)$ es su densidad asociada y los $p_i(u)$ representan a polinomios de grado $3i - 1$ con paridad alternada, esto es, p_1 es simétrico, p_2 es antisimétrico, p_3 es simétrico... Además, sus coeficientes dependen de los momentos de X de orden menor o igual que $i + 2$. En particular,

$$\begin{aligned} p_1(u) &= -\frac{1}{6} \frac{k_3}{\sigma^3} (u^2 - 1), \\ p_2(u) &= -u \left[\frac{1}{24} \frac{k_4}{\sigma^4} (u^2 - 3) + \frac{1}{72} \left(\frac{k_3}{\sigma^3}\right)^2 (u^4 - 10u^2 + 15) \right], \end{aligned}$$

donde k_j representa el j -ésimo cumulante, esto es, el término que acompaña a $\frac{(it)^j}{j!}$ en el desarrollo en serie del logaritmo de la función característica:

$$\log(\varphi_X(t)) = \sum_{j=1}^{\infty} k_j \frac{(it)^j}{j!}.$$

Ejemplo 2.1 (Grado de aproximación de la distribución de R a la Normal). A partir del Teorema 2.2 resulta inmediato ver que

$$\mathbb{P}\{R \leq u\} = \Phi(u) + O\left(n^{-1/2}\right).$$

Así pues, el grado de aproximación de la distribución de R a la Normal estándar es de $O(n^{-1/2})$.

Ejemplo 2.2 (Grado de aproximación de la distribución de R a la de R^*). En primer lugar, un desarrollo de Edgeworth para la distribución, ahora en el remuestreo, de R^* conduce a que

$$\mathbb{P}^*\{R^* \leq u\} = \Phi(u) + n^{-1/2} \hat{p}_1(u) \phi(u) + \dots + n^{-(j-1)/2} \hat{p}_{j-1}(u) \phi(u) + O_P\left(n^{-j/2}\right) \quad (2.1)$$

donde los polinomios $\hat{p}_i(u)$ son de la misma forma que los $p_i(u)$ solo que sustituyendo los cumulantes teóricos por los empíricos y la desviación típica teórica por la empírica. De este modo se tiene que

$$\hat{p}_i(u) - p_i(u) = O_P\left(n^{-1/2}\right). \quad (2.2)$$

De este modo, a partir de (2.1), (2.2) y el Teorema 2.2 se tiene que

$$\begin{aligned} \mathbb{P}\{R \leq u\} - \mathbb{P}^*\{R^* \leq u\} &= n^{-1/2} [p_1(u) - \hat{p}_1(u)] \phi(u) + O_P\left(n^{-1}\right) \\ &= O_P\left(n^{-1}\right). \end{aligned}$$

³ En Cramér (1928b) pueden verse algunas de las primeras aplicaciones estadísticas de este resultado propuestas por el propio Cramér.

En consecuencia, ha quedado demostrado que el orden de aproximación de la distribución en el muestreo de R a la distribución en el remuestreo de R^* es mejor que el orden de aproximación a la Normal estándar límite.

Observación 2.1. Nótese que los órdenes de aproximación que han sido calculados en el Ejemplo 2.1 y en el Ejemplo 2.2 han sido obtenidos en el caso general en que se desconoce detalle alguno de la naturaleza de F . Si, por ejemplo, se tuviese conocimiento de que F es simétrica habría que analizar de nuevo la situación e incluso podrían considerarse mecanismos bootstrap diferentes al naïve⁴. En general, la idea que hay detrás del bootstrap sugiere que hay que buscar imitar de la mejor manera posible la distribución poblacional con todas las características de la misma que sean conocidas.

El Teorema 2.2 tiene como limitación el hecho de que está enfocado únicamente al caso en el cual el estadístico es la media muestral. Sin embargo, 50 años después de que Cramér demostrase ese primer resultado los matemáticos indios Rabintra Nath Bhattacharya y Jayanta Kumar Ghosh llevaron a cabo una generalización de los desarrollos de Edgeworth a otros estadísticos, estandarizados o estudentizados, obtenidos para otros estimadores arbitrarios, $\hat{\theta}$, no necesariamente la media muestral.

Teorema 2.3 (Bhattacharya y Ghosh 1978, pp. 436-437). Sean $X_1, X_2, \dots, X_n, \dots$ variables aleatorias independientes e idénticamente distribuidas con distribución común F . Sean, además, $\theta = \theta(F)$ un parámetro de dicha distribución y $\hat{\theta}$, un estimador de dicho parámetro. Supóngase, además, que

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma_\theta^2).$$

Entonces, bajo ciertas condiciones de regularidad, se tiene que

$$\begin{aligned} \mathbb{P}\left\{\sqrt{n}\frac{\hat{\theta} - \theta}{\sigma_\theta} \leq u\right\} &= \Phi(u) + n^{-1/2}p_1(u)\phi(u) + \dots + n^{-(j-1)/2}p_{j-1}(u)\phi(u) + O(n^{-j/2}), \\ \mathbb{P}\left\{\sqrt{n}\frac{\hat{\theta} - \theta}{\hat{\sigma}_\theta} \leq u\right\} &= \Phi(u) + n^{-1/2}q_1(u)\phi(u) + \dots + n^{-(j-1)/2}q_{j-1}(u)\phi(u) + O(n^{-j/2}), \end{aligned}$$

donde los $p_i(u)$ y $q_i(u)$ representan a polinomios de grado $3i - 1$ con paridad alternada, esto es, p_1 y q_1 son simétricos, p_2 y q_2 son antisimétricos, p_3 y q_3 son simétricos...

Ejemplo 2.3 (Error de cobertura de un intervalo de confianza). Una de las aplicaciones que tiene el Teorema 2.3 es en el ámbito de los intervalos de confianza. Por ejemplo, si se desea construir un intervalo de confianza de nivel $1 - \alpha$ para la media, μ , en el caso de que la desviación típica, σ , es desconocida este puede obtenerse empleando la aproximación por la Normal asintótica y una técnica *plug-in*⁵, llegando a que el intervalo de confianza resulta ser

$$\hat{I} = \left(\bar{X} - \frac{S_n}{\sqrt{n}} z_{1-\alpha/2}, \bar{X} + \frac{S_n}{\sqrt{n}} z_{1-\alpha/2} \right), \quad (2.3)$$

donde $z_{1-\alpha/2}$ es el cuantil $1 - \alpha/2$ de la Normal estándar y S_n es la desviación típica muestral⁶. Así pues, el estadístico en el que está basado el procedimiento inferencial consiste en

$$R_1 = \sqrt{n} \frac{\bar{X} - \mu}{S_n}.$$

⁴ En este contexto se puede emplear una metodología conocida como *bootstrap simetrizado*, que no será tratada en este trabajo pero puede consultarse en las páginas 121 y 122 de Davison y Hinkley (1997). En cuanto a los grados de aproximación, se tiene que en este caso la Normal es una $O(n^{-1})$, el bootstrap uniforme es una $O_P(n^{-1})$ y el bootstrap simetrizado es una $O_P(n^{-3/2})$.

⁵ En esencia, consiste en obtener el intervalo utilizando el Teorema 2.1 en el caso de que σ es conocida y luego sustituir σ por un estimador de la misma.

⁶ $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$.

Aplicando un desarrollo de Edgeworth como el obtenido en el Teorema 2.3 puede obtenerse una acotación del error de cobertura del intervalo de confianza (2.3):

$$\begin{aligned}
\mathbb{P}\left\{\mu \in \hat{I}\right\} - (1 - \alpha) &= \mathbb{P}\left\{\bar{X} - \frac{S_n}{\sqrt{n}} z_{1-\alpha/2} < \mu < \bar{X} + \frac{S_n}{\sqrt{n}} z_{1-\alpha/2}\right\} - (1 - \alpha) \\
&= \mathbb{P}\left\{\mu < \bar{X} + \frac{S_n}{\sqrt{n}} z_{1-\alpha/2}\right\} - \mathbb{P}\left\{\mu < \bar{X} - \frac{S_n}{\sqrt{n}} z_{1-\alpha/2}\right\} - (1 - \alpha) \\
&= \mathbb{P}\{-z_{1-\alpha/2} < R_1\} - \mathbb{P}\{z_{1-\alpha/2} < R_1\} - (1 - \alpha) \\
&= \mathbb{P}\{R_1 < z_{1-\alpha/2}\} - \mathbb{P}\{R_1 < -z_{1-\alpha/2}\} - (\Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2})) \\
&= n^{-1/2} q_1(z_{1-\alpha/2}) \phi(z_{1-\alpha/2}) + O(n^{-1}) - \left(n^{-1/2} q_1(-z_{1-\alpha/2}) \phi(-z_{1-\alpha/2}) + O(n^{-1})\right) \\
&= O(n^{-1}),
\end{aligned}$$

donde la última igualdad se debe al carácter par del polinomio $q_1(u)$ y de la densidad de la Normal.

NOTA. Para analizar en mayor profundidad el concepto de los desarrollos de Edgeworth se recomienda al lector la consulta, en primer lugar, del Capítulo 3 de Kolassa (2006). En él pueden encontrarse, por ejemplo, algunas motivaciones de naturaleza heurística para estos desarrollos e incluso se introducen otro tipo de desarrollos en serie relacionados con los de Edgeworth, como es el caso de los de Cornish-Fisher. Por otro lado, se recomienda también al lector consultar el libro Hall (2013), en el cual, además de estudiar con detalle los desarrollos de Edgeworth, se hace especial hincapié en su utilidad a la hora de demostrar diferentes propiedades de la metodología bootstrap.

En cuanto a la segunda de las preguntas que fueron formuladas al comienzo de este apartado, relativa a cómo puede ser calculada la distribución en el remuestreo del estadístico bootstrap, considérese, por ejemplo, el caso del bootstrap uniforme, el cual fue propuesto en Efron (1979) y se emplea en el caso de disponer de datos independientes. En esencia, esta metodología consiste en generar réplicas bootstrap en las que evaluar el estadístico a partir de la función de distribución empírica, esto es, arrojar un X_i^* tal que $\mathbb{P}\{X_i = X_j\} = n^{-1}$, para $j = 1, \dots, n$, donde n es el tamaño de la muestra original. La clave está en que las remuestras están siendo obtenidas mediante un sorteo equiprobable en una población finita y este procedimiento da lugar a cantidad total de remuestras de n^n , que es una cantidad finita que, a priori, permitiría conocer exactamente la distribución en el remuestreo del estadístico bootstrap. El problema que se desprende de esta situación puede verse muy claramente en la Tabla 1. En ella se observa que el número de posibles remuestras es extremadamente grande incluso en el caso de tamaños de muestra originales razonablemente pequeños.

Así pues, si bien la distribución en el remuestreo sería, en principio, calculable directamente al tratarse de una distribución discreta, la realidad es que el cómputo de todos los posibles valores del estadístico bootstrap junto con sus probabilidades asociadas resulta imposible en la práctica. En consecuencia, lo que se acostumbra a hacer es emplear una aproximación de Monte Carlo, cuya idea principal subyacente en este caso consiste en tomar una gran cantidad, B , de remuestras bootstrap $\mathbf{X}^{*1}, \mathbf{X}^{*2}, \dots, \mathbf{X}^{*B}$ y emplear las evaluaciones del estadístico bootstrap en todas ellas para aproximar su distribución en el remuestreo. La ventaja ahora reside en que se dispone del mecanismo generador de remuestras y se podrán obtener tantas como se desee aumentando el tamaño de B , lo cual se traducirá también en un mayor coste computacional.

n	5	10	15	20	25	30
n^n	$3,125 \times 10^3$	1×10^{10}	$4,379 \times 10^{17}$	$1,049 \times 10^{26}$	$8,882 \times 10^{34}$	$2,059 \times 10^{44}$

Tabla 1: Cantidad de posibles remuestras para diferentes tamaños muestrales.

2.1.2. Implementación del bootstrap uniforme

Para finalizar esta primera parte de introducción a la metodología bootstrap se va a ilustrar la implementación de un método bootstrap. En concreto, se va a introducir un procedimiento algorítmico para el bootstrap uniforme y, además, dado que, por lo general, la distribución en el remuestreo del estadístico bootstrap no será calculable, se incluye también la aproximación por Monte Carlo.

Algoritmo 2 Bootstrap uniforme o naïve, [Efron 1979](#)

1: **Obtener la función de distribución empírica**, F_n , asociada a la muestra X_1, \dots, X_n :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq x\}.$$

2: **Repetir** para $b = 1, \dots, B$:

3: **Arrojar una remuestra bootstrap a partir de F_n** , $\mathbf{X}^{*b} = (X_1^{*b}, \dots, X_n^{*b})$, donde

$$\mathbb{P}\{X_i^{*b} = X_j\} = \frac{1}{n}, \quad j = 1, \dots, n.$$

4: **Calcular la versión bootstrap del estadístico**:

$$R^{*b} = R(\mathbf{X}^{*b}, F_n).$$

5: **Aproximar la distribución** en el muestreo de R mediante las réplicas R^{*1}, \dots, R^{*B} .

Observación 2.2. Nótese que, así como en el Paso 1 del Algoritmo 2 se considera la función de distribución empírica, en el caso en que se dispusiese de cierta información sobre F , esta podría incluirse adecuadamente en el mecanismo que se emplea para generar las remuestras bootstrap, dando lugar así a otros métodos conocidos como *bootstrap paramétrico* ($F \in \{F_\theta\}_{\theta \in \Theta}$), *bootstrap simetrizado* (F es simétrica), *bootstrap suavizado* (F es continua⁷)... No obstante, como los datos que subyacen a este trabajo se encuentran en el contexto de series de tiempo, no se comentarán estos mecanismos de remuestreo sino que se comentarán aquellos que son empleados en el contexto de datos dependientes.

2.2. Bootstrap en la estimación con datos dependientes

En la sección anterior se presentó la idea general de la metodología bootstrap y se ilustró su implementación en el caso de datos independientes e idénticamente distribuidos con el bootstrap uniforme. En la presente sección, basada en el artículo [Cao \(1999\)](#), el paradigma será diferente, ya que los datos se supondrán dependientes. No obstante, en este contexto puede estarse ante dos casos bien diferenciados en función de si se supone algún modelo de dependencia especificado sobre los datos o no.

En el primer caso, se asumirá un modelo sobre la estructura de dependencia de los datos, típicamente alguno de los modelos del Capítulo 1. Por otro lado, el segundo escenario consiste en no asumir ninguna especificación acerca de la estructura de dependencia, más allá de condiciones como la estacionariedad y el carácter *mixing*, un concepto relacionado con la dependencia que será definido luego. En concreto, se tratará el carácter fuertemente mixing⁸, que indica cómo la dependencia entre las variables aleatorias se va atenuando a medida que aumenta la distancia temporal entre ellas.

⁷ En consecuencia, por ser una función de distribución continua, existirá una función de densidad asociada, $f(x)$, cuya relación viene dada por $F'(x) = f(x)$.

⁸ Si el lector quiere profundizar más acerca de la estadística con datos dependientes puede consultar [Bertail et al. \(2006\)](#), en donde se detallan algunos desarrollos recientes en dicho campo junto con algunas aplicaciones.

NOTA. Aunque en este trabajo solamente se trata la propiedad fuertemente mixing, existen muchas otras maneras de medir la dependencia, en general, entre dos σ -álgebras. Por ejemplo, en [Bradley \(2005\)](#) se consideran en total 5 medidas diferentes y se incluyen, además, algunas de sus propiedades básicas y algunos resultados teóricos relacionados con ellas.

Definición 2.1 (Fuertemente mixing o α -mixing⁹, [Rio 2017](#), p. XV). Sea $\{X_t\}_{t \in \mathbb{Z}}$ una serie de tiempo y sea \mathcal{F}_s^t la σ -álgebra generada por X_s, \dots, X_t . Se dice que la serie $\{X_t\}_{t \in \mathbb{Z}}$ es *fuertemente mixing* o *α -mixing* si cumple que

$$\sup_{A \subset \mathcal{F}_1^n, B \subset \mathcal{F}_{n+k}^\infty} |\mathbb{P}\{A \cap B\} - \mathbb{P}\{A\} \cdot \mathbb{P}\{B\}| \leq \alpha_k, \quad \lim_{k \rightarrow \infty} \alpha_k = 0.$$

Así pues, establecidos los marcos teóricos en cada uno de los dos escenarios, en los siguientes apartados de esta sección se irán introduciendo algunos de los diferentes procedimientos bootstrap en el contexto de la estimación con datos dependientes, comenzando primero por aquellos que son utilizados cuando se dispone de una estructura explícita de dependencia.

2.2.1. Método bootstrap propuesto por Stute en 1995

En este apartado se introducirá la metodología bootstrap propuesta en [Stute \(1995\)](#) en el contexto de los modelos de regresión lineal con errores que siguen un modelo AR(1). Más concretamente, las técnicas de remuestreo son empleadas para aproximar la distribución de los estimadores de los coeficientes de la regresión obtenidos mediante un criterio de mínimos cuadrados generalizados en 2 etapas. Así, comiencese considerando el modelo lineal dado por

$$Y = X\beta + \varepsilon, \quad (2.4)$$

donde

$$Y = (Y_1, Y_2, \dots, Y_n)', \quad \beta = (\beta_1, \beta_2, \dots, \beta_p)', \quad \varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)',$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Si los errores son homocedásticos y están incorrelacionados el método de mínimos cuadrados ordinario (OLS, *ordinary least squares*) conduce al estimador del vector de coeficientes de la regresión dado por $\hat{\beta}_n^0 = (X'X)^{-1} X'Y$.

Por otro lado, si la matriz de covarianzas de ε es igual a $\sigma^2 W$ y la matriz W no es la identidad pero es conocida, aplicando el método de mínimos cuadrados generalizados (GLS, *generalized least squares*) se llega al estimador dado por

$$\hat{\beta}_n^1 = (X'W^{-1}X)^{-1} X'W^{-1}Y$$

.

No obstante, lo habitual es desconocer W , por lo que es necesario aplicar un procedimiento de mínimos cuadrados generalizado en 2 etapas, dando lugar al estimador

$$\hat{\beta}_n^2 = \left(X' \widehat{W}^{-1} X \right)^{-1} X' \widehat{W}^{-1} Y.$$

⁹ El carácter α -mixing fue introducido en [Rosenblatt \(1956\)](#).

Considérese ahora un modelo lineal como (2.4) donde los errores siguen un modelo AR(1) de media 0, es decir,

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \delta_t, \quad \forall t \in \mathbb{Z}, \quad (2.5)$$

donde $|\phi_1| < 1$ y es desconocido y los δ_t son independientes y tienen distribución común F , con vector de medias $\mathbf{0}$ y matriz de covarianzas $\sigma^2 W$. Se tiene que, bajo condiciones bastante generales,

$$\sqrt{n} \left(\hat{\beta}_n^2 - \beta \right) \xrightarrow{d} N(\mathbf{0}, \Sigma). \quad (2.6)$$

El procedimiento que Stute propone para aproximar la distribución del estadístico que figura en (2.6) es el que se recoge en el Algoritmo 3, donde además se incluye un procedimiento de tipo Monte Carlo que ayudará a aproximar la distribución en el remuestreo del estadístico bootstrap.

Observación 2.3. Nótese que en el modelo (2.5) es sencillo comprobar que

$$W^{-1} = \begin{pmatrix} 1 & -\phi_1 & 0 & \cdots & 0 \\ -\phi_1 & 1 + \phi_1^2 & -\phi_1 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -\phi_1 & 1 + \phi_1^2 & -\phi_1 \\ 0 & 0 & \cdots & -\phi_1 & 1 \end{pmatrix}. \quad (2.7)$$

Además, W^{-1} es una matriz simétrica y definida positiva, por lo que por el teorema de factorización de Cholesky¹⁰ existe una matriz P triangular inferior tal que $P'P = W^{-1}$. En este caso, se tiene que

$$P = \begin{pmatrix} \sqrt{1 - \phi_1^2} & 0 & 0 & \cdots & 0 \\ -\phi_1 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\phi_1 & 1 \end{pmatrix}. \quad (2.8)$$

Por último, antes de introducir el algoritmo propuesto por Stute, se recoge el resultado teórico en el cual descansa la validez o consistencia del método bootstrap empleado.

Teorema 2.4 (Stute 1995, p. 399). *Considérese el modelo lineal dado por $Y = X\beta + \varepsilon$, donde $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ y $\varepsilon \in \mathbb{R}^n$ y tiene una matriz de covarianzas igual a $\sigma^2 W$. Supóngase, además, que la matriz de diseño, X , verifica las siguientes condiciones:*

1) Para algunas matrices definidas positivas $V_0, V_1 \in \mathbb{R}^{p \times p}$ se cumple que

$$n^{-1} X'X \longrightarrow V_0, \quad n^{-1} X'W^{-1}X \longrightarrow V_1.$$

2) Para algún $\alpha \in (0, 1/2]$ se cumple que $\|X\| = \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} |x_{ij}| = O(n^{1/2-\alpha})$.

Entonces se tiene que, con probabilidad 1,

$$\sup_{\mathbf{x} \in \mathbb{R}^p} \left| \mathbb{P} \left\{ \sqrt{n} \left(\hat{\beta}_n^2 - \beta \right) \leq \mathbf{x} \right\} - \mathbb{P}^* \left\{ \sqrt{n} \left(\hat{\beta}_n^{2*} - \hat{\beta}_n^2 \right) \leq \mathbf{x} \right\} \right| \longrightarrow 0.$$

¹⁰ Para profundizar en esta factorización véase la página 147 de Ciarlet et al. (1989).

Algoritmo 3 Estimación bootstrap de la distribución de $\hat{\beta}_n^2$, [Stute 1995](#).

- 1: **Calcular el estimador** de β por OLS: $\hat{\beta}_n^0$.
- 2: **Construir los residuos**: $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n) = Y - X\hat{\beta}_n^0$.
- 3: **Calcular el estimador** de ϕ_1 mediante OLS a partir de $\hat{\varepsilon}$:

$$\hat{\phi}_1 = \frac{\sum_{j=1}^{n-1} \hat{\varepsilon}_j \hat{\varepsilon}_{j+1}}{\sum_{j=1}^{n-1} \hat{\varepsilon}_j^2}.$$

- 4: **Construir la matriz** \widehat{W}^{-1} , obtenida al sustituir ϕ_1 por $\hat{\phi}_1$ en (2.7).
- 5: **Calcular el estimador** de β aplicando el método GLS en 2 etapas: $\hat{\beta}_n^2$.
- 6: **Construir los nuevos residuos**: $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \dots, \tilde{\varepsilon}_n)' = Y - X\hat{\beta}_n^2$.
- 7: **Construir la matriz** \tilde{P} , obtenida al sustituir ϕ_1 por $\hat{\phi}_1$ en (2.8).
- 8: **Calcular** $\tilde{\delta}$, con

$$\tilde{\delta} = \tilde{P}\tilde{\varepsilon} = \left(\sqrt{1 - \hat{\phi}_1^2} \tilde{\varepsilon}_1, \tilde{\delta}_2, \dots, \tilde{\delta}_n \right)'.$$

- 9: **Definir la distribución empírica de los $\tilde{\delta}$ centrados**:

$$\hat{F}_n^{\tilde{\delta}}(x) = \frac{1}{n-1} \sum_{i=2}^n \mathbb{I} \{ \hat{\delta}_i \leq x \}, \quad \hat{\delta}_i = \tilde{\delta}_i - \bar{\delta}, \quad \bar{\delta} = \frac{1}{n-1} \sum_{j=2}^n \tilde{\delta}_j.$$

- 10: **Repetir** para $b = 1, \dots, B$:
- 11: **Obtener una sucesión**¹¹ de elementos independientes e idénticamente distribuidos a partir de $\hat{F}_n^{\tilde{\delta}}$: $\hat{\delta}_n^{*b}, \hat{\delta}_{n-1}^{*b}, \hat{\delta}_{n-2}^{*b}, \dots$
- 12: **Construir los análogos bootstrap de los errores** ε_i : $\varepsilon_i^{*b} = \sum_{k=0}^{\infty} \hat{\phi}_1^k \hat{\delta}_{i-k}^{*b}$, $1 \leq i \leq n$.
- 13: **Construir las réplicas bootstrap de la respuesta** mediante la ecuación del modelo:

$$Y^{*b} = X\hat{\beta}_n^2 + \varepsilon^{*b}, \quad \varepsilon^{*b} = (\varepsilon_1^{*b}, \varepsilon_2^{*b}, \dots, \varepsilon_n^{*b})'$$

- 14: **Calcular** $\hat{\phi}_1^{*b}$, la versión bootstrap de $\hat{\phi}_1$. Para ello, repetir los pasos del 1 al 4 considerando Y^{*b} en lugar de Y .
 - 15: **Construir la matriz** \widehat{W}^{*b} , obtenida al sustituir ϕ_1 por $\hat{\phi}_1^{*b}$ en (2.7).
 - 16: **Definir el análogo bootstrap de $\hat{\beta}_n^2$** : $\hat{\beta}_n^{2*b} = \left(X' \widehat{W}^{*b} X \right)^{-1} X' \widehat{W}^{*b-1} Y^{*b}$.
 - 17: **Aproximar la distribución** de $\hat{\beta}_n^2 - \beta$ mediante la de su análogo bootstrap: $\hat{\beta}_n^{2*} - \hat{\beta}_n^2$.
-

2.2.2. Método bootstrap propuesto por Franke, Kreiss y Mammen en 2002

En este apartado se van a presentar diferentes mecanismos de remuestreo propuestos en [Franke et al. \(2002\)](#) en el contexto de la estimación de la distribución de los estimadores de tipo núcleo en modelos autorregresivos no paramétricos donde, motivados por las posibles aplicaciones en el ámbito

¹¹ En la práctica la sucesión, y por extensión la serie asociada, quedará reemplazada por una muestra y una suma finita de un tamaño adecuado.

econométrico, los errores puede ser heterocedásticos. Considérese una muestra X_0, \dots, X_T del modelo

$$X_t = m(X_{t-1}, \dots, X_{t-p}) + \sigma(X_{t-1}, \dots, X_{t-q})\varepsilon_t, \quad t = 0, 1, \dots,$$

donde los $\{\varepsilon_t\}$ son independientes e idénticamente distribuidos con media 0 y varianza 1 y donde las funciones de suavizado $m(\cdot)$ y $\sigma(\cdot)$ son desconocidas. Por simplicidad en la notación, se considerará únicamente el caso $p = q = 1$, aunque la generalización a modelos de mayor orden es inmediata.

En cuanto a la estimación de las funciones $m(\cdot)$ y $\sigma(\cdot)$, va a realizarse empleando estimadores tipo núcleo de Nadaraya-Watson^{12 13}. Más concretamente, se considerarán los siguientes estimadores:

$$\begin{aligned} \widehat{m}_{h_1}(x) &= \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{K_{h_1}(x - X_t) X_{t+1}}{\widehat{f}_{h_1}(x)}, \\ \widehat{\sigma}_{1,h_2}^2(x) &= \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{K_{h_2}(x - X_t) X_{t+1}^2 - \widehat{m}_{h_2}^2(x)}{\widehat{f}_{h_2}(x)}, \\ \widehat{\sigma}_{2,h_2}^2(x) &= \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{K_{h_2}(x - X_t) \widehat{r}_{t+1}^2}{\widehat{f}_{h_2}(x)}, \end{aligned} \quad (2.9)$$

donde se tiene que $K_h(\cdot) = h^{-1}K(\cdot/h)$ para alguna función núcleo K , $\widehat{r}_{t+1} = X_{t+1} - \widehat{m}_{h_2}(X_t)$ y $\widehat{f}_h(x)$ es una estimación tipo núcleo de la densidad univariante de la serie $\{X_t\}$, es decir,

$$\widehat{f}_h(x) = \frac{1}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t). \quad (2.10)$$

Si bien es cierto que, bajo ciertas condiciones sobre $\{X_t\}$, los tres estimadores considerados son asintóticamente gaussianos¹⁴, en Franke et al. (2002) se demuestra la consistencia de varios procedimientos bootstrap para aproximar sus distribuciones y se ven ejemplos donde para tamaños de muestra pequeños se mejora la aproximación que ofrece la normal asintótica.

Bootstrap autorregresivo

Este primer método de remuestreo, construido en base a las propuestas realizadas en Franke y Wendel (1992) y Kreutzberger (1993), tiene un planteamiento muy similar al procedimiento bootstrap basado en los residuos en modelos ARMA(p, q) que fue analizado en la Sección 2.3.4. De este modo, el primer paso consiste en generar un proceso bootstrap

$$X_t^* = \widetilde{m}(X_{t-1}^*) + \widetilde{\sigma}(X_{t-1}^*)\varepsilon_t^*, \quad t = 1, 2, \dots \quad (2.11)$$

donde $\widetilde{m}(\cdot)$ y $\widetilde{\sigma}(\cdot)$ son estimadores piloto de $m(\cdot)$ y $\sigma(\cdot)$, respectivamente, y donde $\varepsilon_1^*, \dots, \varepsilon_T^*$ es una remuestra de variables aleatorias independientes e idénticamente distribuidas con distribución condicional $\widetilde{P}_\varepsilon$. El siguiente paso, obtenida la remuestra X_0^*, \dots, X_T^* a partir de (2.11), es obtener los análogos bootstrap de (2.9) y (2.10), es decir, $\widehat{m}_{h_1}^*(x)$, $\widehat{\sigma}_{1,h_2}^{2*}(x)$, $\widehat{\sigma}_{2,h_2}^{2*}(x)$ y $\widehat{f}_h^*(x)$. Finalmente, se aproximaría, respectivamente, la distribución de los estadísticos

$$\sqrt{Th}[\widehat{m}_h(x_0) - m(x_0)], \quad \sqrt{Th}[\widehat{\sigma}_{j,h}^2(x_0) - \sigma^2(x_0)], \quad j = 1, 2,$$

¹² Una definición formal de este tipo de estimadores puede encontrarse en la página 71 de Wasserman (2006).

¹³ Consúltase la Sección 2.6 de Franke et al. (2002) para ver cómo adaptar el bootstrap de este apartado al caso en que los estimadores son de tipo polinómico local.

¹⁴ Esto puede verse en Robinson (1983) y en Masry (1996).

para algún $x_0 \in \mathbb{R}$, mediante la distribución condicional de los estadístico bootstrap dados por

$$\sqrt{Th} [\widehat{m}_h^*(x_0) - \widetilde{m}(x_0)], \quad \sqrt{Th} [\widehat{\sigma}_{j,h}^{2*}(x_0) - \widetilde{\sigma}^2(x_0)], \quad j = 1, 2.$$

En cuanto a la elección de los estimadores piloto, se consideran estimadores de tipo núcleo de Nadaraya-Watson $\widehat{m}_{g_1}(\cdot)$ y $\widehat{\sigma}_{j,g_2}^2(x)$, donde $j = 1$ o $j = 2$ y donde g_1 y g_2 son parámetros ventana, generalmente distintos de h_1 y h_2 ¹⁵. En lo relativo a $\widetilde{P}_\varepsilon$, se va a tomar la función de distribución empírica de los residuos centrados. En el Algoritmo 4 ha sido resumido el esquema de remuestreo tomando $\widehat{\sigma}_{2,g_2}^2(\cdot)$ como estimador de $\sigma^2(\cdot)$. Por simplicidad en la notación se denota como $\widehat{\sigma}_{g_2}^2(\cdot)$.

Algoritmo 4 Bootstrap autorregresivo, Franke et al. 2002

- 1: **Construir los estimadores** de $m(\cdot)$ y $\sigma^2(\cdot)$ a partir de la muestra: $\widehat{m}_{h_1}(\cdot), \widehat{\sigma}_{h_2}^2(\cdot)$.
- 2: **Construir los estimadores piloto** de $m(\cdot)$ y $\sigma^2(\cdot)$ a partir de la muestra: $\widehat{m}_{g_1}(\cdot), \widehat{\sigma}_{g_2}^2(\cdot)$.
- 3: **Construir los residuos:** $\widehat{\varepsilon}_j = \frac{X_j - \widehat{m}_{g_1}(X_{j-1})}{\widehat{\sigma}_{g_2}^2(X_{j-1})}$, $j = 1, \dots, T$.
- 4: **Definir la distribución empírica de los residuos centrados:**

$$\widehat{F}_T^{\widetilde{\varepsilon}}(x) = \frac{1}{T} \sum_{j=1}^T \mathbb{I}\{\widetilde{\varepsilon}_j \leq x\}, \quad \text{donde } \widetilde{\varepsilon}_j = \widehat{\varepsilon}_j - \bar{\varepsilon}, \quad \bar{\varepsilon} = \frac{1}{T} \sum_{t=1}^T \widehat{\varepsilon}_t.$$

- 5: **Repetir** para $b = 1, \dots, B$:
- 6: **Obtener a partir de $\widehat{F}_T^{\widetilde{\varepsilon}}$ una remuestra de los residuos:** $\varepsilon_1^{*b}, \dots, \varepsilon_T^{*b}$.
- 7: **Construir la remuestra bootstrap:** fijando $X_0^{*b} = X_0$ se define

$$X_t^{*b} = \widehat{m}_{g_1}(X_{t-1}^{*b}) + \widehat{\sigma}_{j,g_2}^2(X_{t-1}^{*b}) \varepsilon_t^*, \quad t = 1, \dots, T.$$

- 8: **Construir las versiones bootstrap de los estimadores** a partir de la remuestra $X_0^{*b}, \dots, X_T^{*b}$: $\widehat{m}_{h_1}^{*b}(x), \widehat{\sigma}_{h_2}^{2*b}(x)$.
- 9: **Aproximar la distribución de $\sqrt{Th_1}[\widehat{m}_{h_1}(x_0) - m(x_0)]$ y $\sqrt{Th_2}[\widehat{\sigma}_{h_2}^2(x_0) - \sigma^2(x_0)]$** mediante

$$\left\{ \sqrt{Th_1} [\widehat{m}_{h_1}^{*b}(x_0) - \widehat{m}_{g_1}(x_0)] \right\}_{b=1}^B, \quad \left\{ \sqrt{Th_2} [\widehat{\sigma}_{h_2}^{2*b}(x_0) - \widehat{\sigma}_{g_2}^2(x_0)] \right\}_{b=1}^B.$$

Bootstrap regresivo

En este procedimiento bootstrap se genera el modelo de regresión con diseño fijo condicionado a X_0, \dots, X_{T-1} dado por

$$X_t^* = \widetilde{m}(X_{t-1}) + \widetilde{\sigma}(X_{t-1}) \varepsilon_t^*, \quad t = 1, 2, \dots$$

donde de nuevo $\widetilde{m}(\cdot)$ y $\widetilde{\sigma}(\cdot)$ son estimaciones de las funciones $m(\cdot)$ y $\sigma(\cdot)$, respectivamente, y donde $\varepsilon_1^*, \dots, \varepsilon_T^*$ es una remuestra de variables aleatorias independientes e idénticamente distribuidas con distribución condicional $\widetilde{P}_\varepsilon$.

A diferencia de lo que ocurría en la ecuación (2.11), ahora en el lado derecho de la ecuación se emplea el proceso original en lugar de la réplica bootstrap. Dicho de otro modo, ahora se remuestra un modelo de regresión no paramétrica, lo cual motiva el nombre dado a esta metodología.

¹⁵ En la Sección 2.5 de Franke et al. (2002) pueden verse algunas condiciones para la elección de estas ventanas piloto.

Observación 2.4. Existe alguna variación del método bootstrap regresivo como, por ejemplo, el *bootstrap local* presentado en Paparoditis y Politis (2000) y Bühlmann et al. (2002). En él, se remuestrea sobre un modelo donde la distribución de $\sigma(X_{t-1})\varepsilon_t$ condicionada a los valores pasados se supone que depende de manera suave del valor X_{t-1} , lo cual dota de mayor complejidad a la estructura de remuestreo con respecto al bootstrap regresivo, donde solo se ha supuesto dependiente de X_{t-1} la varianza condicional de $\sigma(X_{t-1})\varepsilon_t$.

Para la presentación de este método y el wild bootstrap, que será introducido después, se van a considerar ambos en una clase más general de modelos. Así, supóngase que se tiene un proceso estocástico estacionario $\{(X_t, Y_t)\}$ y que se desea estimar la media y la varianza condicionales, esto es, $m(x) = \mathbb{E}[Y | X = x]$ y $\sigma^2(x) = \mathbb{E}[(Y - m(x))^2 | X = x]$. Los estimadores $\widehat{m}_h(\cdot)$, $\widehat{\sigma}_{1,h}^2(\cdot)$ y $\widehat{\sigma}_{2,h}^2(\cdot)$ se definen ahora como

$$\begin{aligned}\widehat{m}_h(x) &= \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{K_h(x - X_t) Y_t}{\widehat{f}_h(x)}, \\ \widehat{\sigma}_{1,h}^2(x) &= \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{K_h(x - X_t) Y_t^2 - \widehat{m}_h^2(x)}{\widehat{f}_h(x)}, \\ \widehat{\sigma}_{2,h}^2(x) &= \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{K_h(x - X_t) \widehat{r}_{t+1}^2}{\widehat{f}_h(x)},\end{aligned}\tag{2.12}$$

donde, al igual que antes, $\widehat{r}_{t+1} = Y_t - \widehat{m}_h(X_t)$ ¹⁶ y, además, $\varepsilon_{t+1} = \sigma(X_t)^{-1} (Y_t - m(X_t))$.

Volviendo a la metodología de remuestreo, en el bootstrap regresivo el procedimiento consiste en generar una remuestra bootstrap $\varepsilon_1^*, \dots, \varepsilon_T^*$ y construir, empleando la muestra original como diseño fijo del modelo, la remuestra dada por

$$Y_t^* = \widetilde{m}(X_{t-1}) + \widetilde{\sigma}(X_{t-1})\varepsilon_t^*,$$

donde $\widetilde{m}(\cdot)$ y $\widetilde{\sigma}(\cdot)$ son estimadores piloto. Luego, a partir de esta remuestra se llevarían a cabo los mismos pasos que los realizados en el caso del bootstrap autorregresivo. En definitiva, el esquema de remuestreo sería muy similar al presentado en el Algoritmo 4 pero cambiando la definición de los residuos y la ecuación del proceso que genera las remuestras acorde a lo que se ha comentado.

Observación 2.5. La configuración considerada en el bootstrap autorregresivo también puede enmarcarse dentro de esta clase de modelos más general sin más que considerar $Y_t = X_{t+1}$.

Wild bootstrap

En este tercer procedimiento se genera de nuevo un modelo de regresión con diseño fijo condicionado a X_0, \dots, X_T y que en este caso viene dado por

$$X_t^* = \widetilde{m}(X_{t-1}) + \eta_t^*, \quad t = 1, 2, \dots$$

donde $\widetilde{m}(\cdot)$ es una estimación de la función $m(\cdot)$ y donde $\eta_1^*, \dots, \eta_T^*$ es una remuestra de variables aleatorias donde la media de η_t^* es 0 y su varianza es r_t^2 , con $r_t = X_t - \widehat{m}_h(X_{t-1})$.

El primer paso a realizar en esta metodología consiste en obtener una muestra $\overline{\eta}_1, \dots, \overline{\eta}_T$ de variables aleatorias independientes e idénticamente distribuidas con $\mathbb{E}[\overline{\eta}_t] = 0$ y $\mathbb{E}[\overline{\eta}_t^2] = \mathbb{E}[\overline{\eta}_t^3] = 1$.

¹⁶ Se emplea el subíndice $t + 1$ para mantener una consistencia en la notación con respecto al bootstrap autorregresivo.

Luego, el siguiente paso es definir $\eta_t^* = r_t \bar{\eta}_t$, $t = 1, \dots, T$. Finalmente, la remuestra generada con el método wild bootstrap se define como

$$X_t^* = \tilde{m}(X_{t-1}) + \eta_t^*, \quad t = 1, \dots, T,$$

donde $\tilde{m}(\cdot)$ es un estimador piloto de $m(\cdot)$. En cuanto al resto del mecanismo de remuestreo, se procedería de manera análoga a los 2 casos anteriores.

Por último, cabe destacar que dado que este procedimiento bootstrap va a ser empleado más adelante en otros apartados del presente trabajo, el esquema de remuestreo ha sido recogido en el Algoritmo 5 en el contexto del modelo general propuesto en el apartado dedicado al bootstrap regresivo, esto es, aquel donde dado un proceso estocástico estacionario $\{(X_t, Y_t)\}$ se considera el modelo

$$Y_t = m(X_t) + \varepsilon_t, \quad \mathbb{E}[\varepsilon_t | X_t] = 0, \quad t = 1, 2, \dots$$

Observación 2.6. En el modelo general considerado la única hipótesis sobre los errores es que la media condicional sea 0, permitiendo por tanto que estos puedan no ser independientes. En este tipo de modelos, o en aquellos en los que la función $\sigma(\cdot)$ sea una función irregular que actúe como un parámetro de ruido y en los que el principal objetivo sea la estimación de $m(\cdot)$, el wild bootstrap es un método apropiado, ya que no busca que los errores bootstrap generados sean independientes e idénticamente distribuidos.

Observación 2.7. Este mecanismo de remuestreo fue originalmente propuesto en Wu (1986), aunque el «bautizo» con el nombre de wild bootstrap fue realizado posteriormente en los artículos Mammen (1992) y Hardle y Mammen (1993). El motivo de este nombre es que para estimar la distribución del error $\varepsilon_j = Y_j - m(X_j)$ condicionada a X_j se emplea tan solo el residuo $r_j = Y_j - \hat{m}_h(X_j)$.

Observación 2.8. Tal y como ha sido comentado anteriormente, el paso 2 del Algoritmo 5 puede llevarse a cabo encontrando una variable aleatoria V^* que cumpla que $\mathbb{E}^*[V^*] = 0$, $\mathbb{E}^*[V^{*2}] = 1$ y que $\mathbb{E}^*[V^{*3}] = 1$ y luego definiendo $\hat{\varepsilon}_i^* = r_i V_i^*$. Por ejemplo, tomando como V^* una variable aleatoria discreta con masa de probabilidad en dos puntos, al imponer las condiciones se tiene que tales puntos son $a = 0,5(1 - \sqrt{5})$ y $b = 0,5(1 + \sqrt{5})$ y, además, se tiene que $p = \mathbb{P}^*\{V^* = a\} = \frac{5+\sqrt{5}}{10}$.

Algoritmo 5 Wild bootstrap, Franke et al. 2002

- 1: **Construir el estimador** de $m(\cdot)$ a partir de la muestra $\{(X_t, Y_t)\}_{t=0}^T$: $\hat{m}_h(\cdot)$.
- 2: **Construir los residuos** $r_t = Y_t - \hat{m}_h(X_t)$, $t = 0, 1, \dots, T$.
- 3: **Repetir** para $b = 1, \dots, B$:
- 4: **Generar errores bootstrap** $\hat{\varepsilon}_t^{*b}$, $t = 0, 1, \dots, T$, condicionalmente a la muestra observada a partir de una distribución que cumpla que $\mathbb{E}^*[\hat{\varepsilon}_t^*] = 0$, $\mathbb{E}^*[\hat{\varepsilon}_t^{*2}] = r_t^2$ y $\mathbb{E}^*[\hat{\varepsilon}_t^{*3}] = r_t^3$.
- 5: **Generar análogos bootstrap de la variable respuesta** empleando un estimador piloto

$$Y_t^{*b} = \hat{m}_g(X_t) + \hat{\varepsilon}_t^{*b}, \quad t = 0, 1, \dots, T.$$

- 6: **Construir el análogo bootstrap de $\hat{m}_h(\cdot)$** a partir de $\{(X_t, Y_t^*)\}_{t=0}^T$:

$$\hat{m}_h^{*b}(x) = \sum_{i=0}^T \frac{K_h(x - X_i)}{\sum_{j=1}^n K_h(x - X_j)} Y_i^{*b}.$$

- 7: **Aproximar la distribución en el muestreo** de $\sqrt{n\bar{h}}[\hat{m}_h(x) - m(x)]$ por la distribución en el remuestreo de $\sqrt{n\bar{h}}[\hat{m}_h^*(x) - \hat{m}_g(x)]$.
-

2.2.3. El bootstrap por bloques

En este apartado se va a presentar una de las primeras propuestas de procedimientos de remuestreo realizadas en el contexto de disponer de datos dependientes sobre los cuales no se asume ninguna estructura explícita de dependencia. Esta metodología recibe el nombre de *moving blocks bootstrap* (MBB) o *bootstrap por bloques* y fue propuesta de manera independiente en [Kunsch \(1989\)](#) y en [Liu y Singh \(1992\)](#).

Algoritmo 6 El bootstrap por bloques. Esquema de remuestreo extraído de [Cao 1999](#)

- 1: **Fijar el tamaño del bloque**, b , con $b \in \mathbb{N}$.
- 2: **Calcular el número de bloques** $k = \lceil n/b \rceil$ ¹⁷.
- 3: **Definir los bloques**: $B_{i,b} = (X_i, X_{i+1}, \dots, X_{i+b-1})$, con $i = 1, 2, \dots, q$ ($q = n - b + 1$).
- 4: **Arrojar k bloques** ξ_1, \dots, ξ_k con distribución equiprobable sobre $\{B_1, B_2, \dots, B_q\}$, donde

$$\xi_i = (\xi_{i,1}, \dots, \xi_{i,b})$$

- 5: **Definir el vector de remuestras bootstrap** \mathbf{X}^* como el formado por las n primeras componentes de

$$(\xi_{1,1}, \dots, \xi_{1,b}, \xi_{2,1}, \dots, \xi_{2,b}, \dots, \xi_{k,1}, \dots, \xi_{k,b}).$$

El esquema de remuestreo de este método ha sido recogido en el Algoritmo 6. A partir de él resulta sencillo ver que si se considera un tamaño de bloque $b = 1$ se tiene que $k = n$, dando lugar bootstrap uniforme ordinario. Por el contrario, si $b = n$ se tiene que $k = 1$ y entonces el procedimiento bootstrap sería un proceso de remuestreo degenerado, en el sentido de que todas las posibles remuestras coinciden con la muestra original.

En base al comentario realizado en el párrafo anterior, parece claro que una cuestión de relevancia en este método es la elección del tamaño del bloque, b . En ese sentido, existen artículos como [Hall et al. \(1995\)](#), donde se obtiene una expresión asintótica para el error cuadrático medio en el contexto concreto de la estimación mediante técnicas bootstrap del sesgo y de la varianza que permite afirmar que el tamaño óptimo del bloque en ese caso es del orden $n^{1/3}$.

Para finalizar esta sección se va a analizar un ejemplo de aplicación de este método que ilustrará un hecho a tener en cuenta a la hora de aplicarlo: la serie bootstrap generada no es estacionaria.

Ejemplo 2.4 ([Cao 1999](#), p. 105). Considérese una muestra X_1, X_2, \dots, X_n de una serie de tiempo, con $n = 100$. Con el fin de generar remuestras se va a aplicar el bootstrap por bloques con un tamaño de bloque $b = 10$. A partir de este esquema de remuestreo pueden analizarse las distribuciones bootstrap conjuntas de (X_9^*, X_{10}^*) y (X_{10}^*, X_{11}^*) . Se tiene que

$$\begin{aligned} \mathbb{P}^* \{(X_9^*, X_{10}^*) = (X_i, X_j)\} &= \frac{1}{91} && \text{para } i = 9, 10, \dots, 99, \\ & && j = i + 1; \\ \mathbb{P}^* \{(X_{10}^*, X_{11}^*) = (X_i, X_j)\} &= \frac{1}{91^2} && \text{para } i = 10, 11, \dots, 100, \\ & && j = 1, 2, \dots, 91. \end{aligned}$$

De este modo, como ambas distribuciones bootstrap son diferentes el proceso generado mediante el bootstrap por bloques no es estrictamente estacionario, aunque la serie original lo hubiera sido.

Observación 2.9. Cometiendo un abuso de lenguaje podría decirse que el método de remuestreo por bloques no es estacionario.

¹⁷ La función $\lceil x \rceil$ devuelve el menor entero mayor o igual que x .

2.2.4. El bootstrap estacionario

En esta sección, motivados por la falta de estacionariedad del bootstrap por bloques, se va a presentar un método de remuestreo de naturaleza similar pero donde ahora se va a remuestrear sobre bloques de tamaño aleatorio, donde esa longitud de bloque se va a distribuir de acuerdo a una distribución geométrica. Esta técnica bootstrap se conoce con el nombre de bootstrap estacionario y fue propuesta en [Politis y Romano \(1994b\)](#).

Supóngase que se tiene una serie de tiempo $\{X_t\}$ estrictamente estacionaria y débilmente dependiente¹⁸ y que θ es un parámetro desconocido de la distribución conjunta (infinito dimensional) de la serie. Considerando una muestra X_1, \dots, X_n de tal serie, el objetivo consiste en realizar inferencia sobre θ a partir de algún estimador $T_n = T_n(X_1, \dots, X_n)$. Más concretamente, se desea construir una región de confianza para el parámetro θ , para lo cual será necesario conocer la distribución en el muestreo de T_n .

En estos casos suele ser habitual que para la construcción de la región de confianza se emplee un pivote $R_n = R_n(X_1, \dots, X_n; \theta)$, que podría ser, por ejemplo, de la forma $R_n = T_n - \theta$ o una versión studentizada del mismo. Así, a partir de este pivote la idea es que, de ser conocida la verdadera distribución en el muestreo de R_n , las afirmaciones sobre la función de distribución de R_n podrían invertirse para dar lugar a afirmaciones sobre regiones de confianza sobre θ . No obstante, dado que habitualmente la distribución en el muestreo de R_n no será conocida, lo que se hará es aproximarla empleando el bootstrap estacionario del siguiente modo. Fijado un número de remuestras B , se empleará este método para generar dichas remuestras y obtener el análogo bootstrap del pivote en cada una de ellas R_n^{*b} , con $b = 1, \dots, B$. Finalmente, se aproximará la distribución en el muestreo de R_n mediante la distribución empírica de $\{R_n^{*b}\}_{b=1}^B$.

En cuanto a la consistencia de este método de remuestreo, en [Politis y Romano \(1994b\)](#) se considera el caso en que $\theta = \mu = \mathbb{E}[X_t]$ y $R_n = \sqrt{n}(\bar{X}_n - \mu)$, donde $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Sea γ la función de autocovarianzas de la serie de tiempo y sean α_k los coeficientes mixing. Supóngase que se cumple que $\gamma(0) + \sum_{r=1}^{\infty} |\gamma(r)| < \infty$, que para algún $d > 0$ se tiene que $\mathbb{E}[|X_t|^{d+2}] < \infty$ y que los coeficientes mixing verifican la siguiente condición: $\sum_{k=1}^{\infty} \alpha_k^{\frac{d}{d+2}} < \infty$. En tal caso se demuestra que cuando $p \rightarrow 0$ y $np \rightarrow \infty$ se tiene que

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}^* \left\{ \sqrt{n} (\bar{X}_n^* - \bar{X}_n) \leq x \right\} - \mathbb{P} \left\{ \sqrt{n} (\bar{X}_n - \mu) \leq x \right\} \right| \xrightarrow{p} 0.$$

Observación 2.10. En [Politis y Romano \(1994b\)](#) se va un paso más allá e incluso se da la idea de la generalización del resultado anterior al caso de estadísticos funcionales¹⁹ $T(F_n)$, donde T es un funcional Fréchet diferenciable.

El esquema de remuestreo de este procedimiento puede ser consultado en el Algoritmo 7, donde se han definido los bloques circulares de b observaciones comenzando en X_i , con $i = 1, \dots, n$, como $B_{i,b} = (X_i, X_{i+1}, \dots, X_{i+b-1})$, $i = 1, 2, \dots$, donde si $j > n$ se toma $X_j = X_{((j-1) \bmod n)+1}$.

Observación 2.11. Nótese que los bloques circulares son significativamente diferentes a los definidos en el bootstrap por bloques, siendo la principal diferencia que, dado que en la muestra no hay datos en instantes posteriores al n , en el bootstrap por bloques no se define ningún bloque de longitud b ($b > 1$) que comience en X_n , mientras que en el bootstrap estacionario se «envuelven los datos circularmente» de forma que el dato X_1 vaya a continuación de X_n . Este es uno de los motivos por los cuales con este procedimiento sí que se consigue preservar la estacionariedad de la remuestra.

¹⁸ Se dice que una serie de tiempo es débilmente estacionaria si se tiene que $\lim_{h \rightarrow \infty} (\text{Corr}[X_t, X_{t+h}]) = 0$.

¹⁹ Un estadístico se dice funcional si es función de la distribución empírica, es decir, si su valor no depende del orden de los datos sobre los que está evaluado.

Algoritmo 7 Bootstrap estacionario, [Politis y Romano 1994b](#)

- 1: **Fijar un número** $p \in [0, 1]$.
- 2: **Obtener las longitudes aleatorias de los bloques** como realizaciones independientes e idénticamente distribuidas L_1, L_2, \dots , con distribución geométrica de parámetro p , es decir,

$$\mathbb{P}\{L_1 = m\} = (1 - p)^m p, \quad m = 1, 2, \dots$$

- 3: **Obtener los valores iniciales aleatorios de los bloques** como realizaciones independientes I_1, I_2, \dots , con distribución equiprobable sobre $\{1, \dots, n\}$.
- 4: **Definir la remuestra bootstrap** X_1^*, \dots, X_n^* como los n primeros valores de

$$(B_{I_1, L_1}, B_{I_2, L_2}, \dots).$$

Observación 2.12. A partir del Algoritmo 7 resulta inmediato ver que el mínimo número de bloques, k , necesario para que la longitud del conjunto de bloques que define a la remuestra sea de al menos n es el primer k que cumple que $\sum_{i=1}^k L_i \geq n$.

Observación 2.13. Del Algoritmo 7 se desprende que el tamaño medio del bloque es $1/p$. En consecuencia, el papel del parámetro p puede verse como el inverso del que jugaba b en el MBB. Así, $p = 1$ es comparable con el caso en que $b = 1$ en el MBB y $p = 0$, con $b \rightarrow \infty$.

Una posible generalización de este método consiste en considerar una distribución de los L_i que no sea geométrica y una distribución de los I_j que no sea necesariamente equiprobable. Ahora bien, en tales casos sería necesario tener especial cuidado en preservar el carácter estacionario en las remuestras, que no está garantizado que se conserve, en general, para cualesquiera distribuciones sobre los L_i y los I_j . De hecho, el MBB puede pensarse en un caso particular, sin más que considerar $q = n - b + 1$ y

$$\mathbb{P}\{L_1 = m\} = \begin{cases} 1, & \text{si } m = b, \\ 0, & \text{si } m \neq b, \end{cases} \quad \mathbb{P}\{I_1 = j\} = \begin{cases} 1/q, & \text{si } j = 1, 2, \dots, q, \\ 0, & \text{si } j = q + 1, q + 2, \dots, n. \end{cases}$$

Por último, para finalizar esta sección se presenta en el Algoritmo 8 un esquema de remuestreo equivalente al Algoritmo 7 pero más simple.

Algoritmo 8 Bootstrap estacionario, [Politis y Romano 1994b](#). Esquema alternativo

- 1: **Fijar un número** $p \in [0, 1]$.
- 2: **Arrojar el primer valor de la remuestra** como una realización de F_n , la función de distribución empírica de la muestra original X_1, \dots, X_n :

$$X_1^* = X_i, \quad i \in \{1, \dots, n\}.$$

- 3: **Repetir** para $j = 2, \dots, n$:
 - 4: Dado $X_{j-1}^* = X_i$ **definir** X_j^* como una realización de F_n con probabilidad p y como X_{i+1} con probabilidad $1 - p$.
-

Observación 2.14. A partir del Algoritmo 8 se tiene que si se toma $p = 1$ el procedimiento bootstrap que se obtiene es el bootstrap uniforme. Por otro lado, si se toma $p = 0$ se obtiene una remuestra que no es más que una permutación circular de la muestra original, lo cual da lugar a que si el pivote considerado, R_n , es funcional su distribución bootstrap sea degenerada.

2.2.5. El método del submuestreo

En este apartado se va a presentar el procedimiento propuesto en [Politis y Romano \(1994a\)](#) conocido con el nombre de método de submuestreo. Esta metodología surge en el ámbito de la construcción de regiones de confianza para parámetros, para lo cual es necesario conocer (o en este caso, aproximar) la distribución en el muestreo de algún estadístico. En concreto, el objetivo que se persigue en el citado artículo consiste en construir regiones de confianza asintóticamente válidas bajo condiciones minimales. Sea X_1, \dots, X_n una muestra de variables aleatorias consistente en²⁰

- a) realizaciones independientes e idénticamente distribuidas con distribución común F , o
- b) realizaciones de un proceso estocástico fuertemente mixing.

Considérese un parámetro $\theta = \theta(F)$ y un estimador del mismo dado por $T_n = T_n(X_1, \dots, X_n)$. Con el fin de realizar inferencia sobre el parámetro $\theta(F)$ se busca poder estimar la verdadera distribución en el muestreo de T_n . Para ello, defínase $J_n(F)$ como la distribución en el muestreo del estadístico $\tau_n(T_n - \theta(F))$, cuya función de distribución asociada se denota como $J_n(\cdot, F)$. Además, fíjese un número $b \in \mathbb{N}$ tal que $b < n$ y, para los casos a) y b) antes considerados defínase

- a) $S_{n,i} = T_b(Y_i)$, con $i = 1, 2, \dots, N$, donde Y_1, \dots, Y_N son todas las $N = \binom{n}{b}$ posibles submuestras de tamaño b que pueden obtenerse sin reemplazamiento a partir de la muestra original.
- b) $S_{n,i} = T_b(B_{i,b})$, con $i = 1, 2, \dots, N$, donde $N = n - b + 1$ es el número total de posibles bloques de tamaño b y donde $B_{i,b} = (X_i, X_{i+1}, \dots, X_{i+b-1})$.

Así, la propuesta del método, cuyo esquema en el caso b) se ha recogido en el Algoritmo 9, consiste en emplear como aproximación de $J_n(\cdot, F)$ la función de distribución empírica de los valores $\tau_b(S_{n,i} - T_n)$:

$$L_n(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{\tau_b(S_{n,i} - T_n) \leq x\}.$$

La consistencia del método descansa sobre el siguiente resultado que, en esencia, establece que el método de submuestreo es asintóticamente válido bajo condiciones minimales sobre el tamaño del bloque y supuesto que el estadístico de interés tenga una distribución límite.

Teorema 2.5 ([Politis y Romano 1994a](#), pp. 2033-2034). *Con la notación establecida supóngase que $\tau_b/\tau_n \rightarrow 0$, $b \rightarrow \infty$ y $b/n \rightarrow 0$ cuando n tiende a ∞ . Además, supóngase que $J_n(F) \xrightarrow{d} J(F)$. Sea x un punto de continuidad de $J(\cdot, F)$. Se tiene que*

- 1) $L_n(x) \xrightarrow{p} J(x, F)$.
- 2) Si $J(\cdot, F)$ es continua, entonces $\sup_{x \in \mathbb{R}} |L_n(x) - J_n(x, F)| \xrightarrow{p} 0$.
- 3) Sea $c_n(1 - \alpha) = \inf_{x \in \mathbb{R}} \{L_n(x) \geq 1 - \alpha\}$ y análogamente sea $c(1 - \alpha, F) = \inf_{x \in \mathbb{R}} \{J(x, F) \geq 1 - \alpha\}$. Si $J(\cdot, F)$ es continua en $c(1 - \alpha, F)$, entonces

$$\mathbb{P}\{\tau_n(T_n - \theta(F)) \leq c_n(1 - \alpha)\} \rightarrow 1 - \alpha,$$

y la cobertura asintótica del intervalo $[T_n - \tau_n^{-1}c_n(1 - \alpha), \infty)$ es $1 - \alpha$.

- 4) Si para cada $d > 0$ se tiene que $\sum_{n=1}^{\infty} e^{-d(n/b)} < \infty$ y $\tau_b(T_n - \theta(F)) \rightarrow 0$ de forma casi segura, entonces la convergencia de los apartados 1) y 2) se da con probabilidad 1.

²⁰ Aunque los métodos bootstrap de esta sección son para datos dependientes, se va a incluir también la versión del método en el caso de datos independientes e idénticamente distribuidos.

Observación 2.15. Para comprender intuitivamente por qué el método funciona bien bajo hipótesis tan débiles, nótese que cada subconjunto de tamaño b es de hecho una muestra de tamaño b del verdadero modelo. En consecuencia, bajo la hipótesis de que existe una distribución límite, parece razonable pensar que las distribuciones en el muestreo basadas en muestras de tamaño b y n deberían estar próximas.

Por último, cabe mencionar que en algún caso podría ocurrir que al aplicar el método de submuestreo en el caso de datos independientes e idénticamente distribuidos el número total de posibles submuestras fuese demasiado grande. En ese caso, podría realizarse la aproximación estocástica siguiente. Fijado un $s < N$, sean I_1, \dots, I_s escogidos aleatoriamente con o sin reemplazamiento del conjunto $\{1, 2, \dots, N\}$. Se tiene el siguiente resultado.

Corolario 2.6 (Politis y Romano 1994a, p. 2037). *Bajo las hipótesis del Teorema 2.5 y suponiendo que $s \rightarrow \infty$ cuando n tiende a ∞ , las tesis del Teorema 2.5 se mantienen si se sustituye $L_n(x)$ por*

$$\hat{L}_n(x) = \frac{1}{s} \sum_{i=1}^s \mathbb{I}\{\tau_b(S_{n,I_i} - T_n) \leq x\}.$$

Algoritmo 9 Método de submuestreo, Politis y Romano 1994a

- 1: **Evaluar el estadístico sobre la muestra original:** $T_n = T_n(X_1, \dots, X_n)$.
- 2: **Fijar un tamaño de bloque** $b \in \mathbb{N}$.
- 3: **Definir los bloques** $B_{i,b} = (X_i, X_{i+1}, \dots, X_{i+b-1})$, $i = 1, 2, \dots, N$ ($N = n - b + 1$).
- 4: **Evaluar el estadístico en cada bloque** $S_{n,i} = T_b(B_{i,b})$, $i = 1, 2, \dots, N$.
- 5: **Aproximar la distribución en el muestreo** de $\tau_n(T_n - \theta(F))$ mediante la distribución empírica dada por

$$L_n(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{\tau_b(S_{n,i} - T_n) \leq x\}.$$

2.3. Bootstrap en la predicción con datos dependientes

Considerando la misma situación que la presentada en la Sección 2.2, en la presente sección se introducirán métodos de remuestreo para la predicción con datos dependientes, tanto en el caso de tener especificada una estructura de dependencia como en casos más generales. La principal referencia consultada ha sido el artículo Cao (1999)

2.3.1. Método bootstrap propuesto por Stine en 1987

En este apartado se introducirá el mecanismo de remuestreo presentado en Stine (1987) en el contexto de la estimación del PMSE del mejor predictor lineal estimado en modelos autorregresivos donde la distribución del proceso de ruido blanco no se supone gaussiana.

Sea $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ una muestra finita de un proceso AR(p), con p conocido, y de media 0, es decir, que el proceso es estacionario y se cumple que

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + a_t, \quad (2.13)$$

donde los errores $\{a_t\}$ son independientes con distribución común F , media 0 y varianza σ^2 y, debido a la estacionariedad, se tiene que el polinomio $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| = 1$. Además, se supone que a_t es independiente de $X_t, X_{t-1}, X_{t-2}, \dots$

Si se denota $\mathbf{X}_t = (X_t, X_{t-1}, \dots, X_{t-p+1})'$, puede escribirse de manera matricial la ecuación (2.13) de la manera siguiente:

$$\mathbf{X}_t = \mathbf{A} \mathbf{X}_{t-1} + \mathbf{J} a_t,$$

donde se tiene que

$$\mathbf{A} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \in \mathbb{R}^{p \times p}, \quad \mathbf{J} = (1, 0, \dots, 0)' \in \mathbb{R}^p.$$

Sea ahora un valor futuro X_{n+h} . A partir de los cálculos realizados en la Sección 1.4 es sencillo ver que el predictor lineal óptimo es el dado por $\hat{X}_{t+h} = \mathbf{A}_1^h \mathbf{X}_t$, donde \mathbf{A}_1^h es la primera fila de \mathbf{A}^h . No obstante este estimador tiene el inconveniente de que es necesario conocer la especificación de los parámetros del modelo autorregresivo.

Una posible estrategia para paliar esta desventaja podría ser aplicar un método *plug-in*, esto es, obtener un estimador (por ejemplo, mediante el mínimos cuadrados) del vector de coeficientes autorregresivos y sustituir el mismo en la matriz \mathbf{A} , de modo que el predictor estimado sería

$$\tilde{X}_{t+h} = \tilde{\mathbf{A}}_1^h \mathbf{X}_t,$$

donde $\tilde{\mathbf{A}}$ denota a la matriz obtenida al sustituir la primera fila de \mathbf{A} por el vector de coeficientes estimados. No obstante, este procedimiento *plug-in* tiene como consecuencia un aumento en el error cuadrático medio de predicción. Más concretamente, se tiene que²¹

$$PMSE \left[\tilde{X}_{t+h} \right] = PMSE \left[\hat{X}_{t+h} \right] + O(n^{-1}).$$

En cuanto a la creación de un intervalo de predicción, en la página 1073 de [Stine \(1987\)](#) se trata primero el caso en el los errores a_t se distribuyen como una distribución gaussiana de media 0 y varianza σ^2 . Bajo esta suposición, se propone un método basado en un desarrollo en serie de Taylor de orden 2 que evita que el estimador del PMSE sea sesgado y que, además, da lugar a un intervalo de predicción cuya cobertura media se aproxima a la cobertura nominal. Sin embargo, los intervalos de predicción construidos bajo la suposición de gaussianidad no se comportan adecuadamente (i.e., no se obtiene la cobertura nominal fijada) cuando la distribución de los errores, a_t , no es gaussiana. Por ello, se propone una metodología más general cuya única suposición acerca de la distribución de los errores es que esta es simétrica y con momentos finitos.

La idea principal en la que se basa el mecanismo de remuestreo en ese caso más general consiste en generar réplicas $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_n^*)$ de los datos observados, \mathbf{X} , y luego extender estas réplicas hacia el «futuro», de manera que se puedan comparar las predicciones a partir de \mathbf{X}^* con dichas extensiones de la remuestra hacia el futuro. De este modo, se comparará \tilde{X}_{n+1}^* con X_{n+1}^* , \tilde{X}_{n+2}^* con X_{n+2}^* , y así sucesivamente. Así pues, esos errores de predicción observables serán los que se emplearán para estimar el PMSE y construir intervalos de predicción. El procedimiento que ha sido descrito se recoge en el Algoritmo 10, donde únicamente se recoge cómo aproximar el PMSE. La extensión de esta metodología para construir intervalos de predicción será comentada a continuación.

En primer lugar, nótese que si uno conociese la distribución H de los errores de predicción dados por $X_{n+h} - \tilde{X}_{n+h}$, un intervalo de predicción de cobertura nominal $1 - \alpha$ para X_{n+h} podría ser

$$\left(\tilde{X}_{n+h} + x_{(1-\alpha)/2}, \tilde{X}_{n+h} + x_{(1+\alpha)/2} \right), \quad (2.14)$$

²¹ Este y otros detalles acerca del aumento del PMSE pueden consultarse en la página 1073 de [Stine \(1987\)](#).

donde $x_\beta = \inf \{x \in \mathbb{R} \mid H(x) \geq \beta\}$. Así pues, el intervalo de predicción bootstrap obvio consistiría en sustituir en (2.14) la distribución H por su estimación empírica mediante el bootstrap, es decir, por

$$H_B^*(x) = \frac{1}{B} \sum_{b=1}^B \mathbb{I} \left\{ \tilde{X}_{n+h}^{*b} - X_{n+h}^{*b} \leq x \right\}.$$

No obstante, esta estimación de H ignora la estructura que tienen los errores de predicción y, por ello, en la página 1074 de [Stine \(1987\)](#) se analiza esta estructura, obteniendo un nuevo estimador de H que la respeta²². Denotando por x_β^* al percentil β de dicha estimación se tiene que el intervalo de predicción bootstrap será el dado por

$$\left(\tilde{X}_{n+h} + x_{(1-\alpha)/2}^*, \tilde{X}_{n+h} + x_{(1+\alpha)/2}^* \right).$$

Algoritmo 10 Estimación bootstrap del PMSE en modelos $AR(p)$, [Stine 1987](#)

- 1: **Estimar los coeficientes autorregresivos** por mínimos cuadrados: $\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p$.
- 2: **Construir los residuos**: $\hat{a}_i = X_i - \hat{\phi}_1 X_{i-1} - \hat{\phi}_2 X_{i-2} - \dots - \hat{\phi}_p X_{i-p}$, $i = p+1, p+2, \dots, n$.
- 3: **Definir la distribución empírica de los residuos centrados**:

$$\hat{F}_n^{\hat{a}}(x) = \frac{1}{n-p} \sum_{i=p+1}^n \mathbb{I} \{ \hat{a}'_i \leq x \}, \quad \hat{a}'_i = \hat{a}_i - \bar{a}, \quad \bar{a} = \frac{1}{n-p} \sum_{i=p+1}^n \hat{a}_i.$$

- 4: **Repetir** para $b = 1, \dots, B$:
- 5: **Obtener una remuestra de los residuos** \hat{a}_j^{*b} , con $j = 1, \dots, n$, a partir de $\hat{F}_n^{\hat{a}}$.
- 6: **Fijar los p valores iniciales** de la remuestra iguales a 0²³:

$$X_1^{*b} = X_2^{*b} = \dots = X_p^{*b} = 0.$$

- 7: **Calcular los $n+h-p$ valores restantes** mediante la ecuación del modelo:

$$X_j^{*b} = \hat{\phi}_1 X_{j-1}^{*b} + \hat{\phi}_2 X_{j-2}^{*b} + \dots + \hat{\phi}_p X_{j-p}^{*b} + \hat{a}_j^{*b}, \quad j = p+1, p+2, \dots, n+h.$$

- 8: **Calcular la versión bootstrap de los estimadores de los coeficientes autorregresivos** a partir de la remuestra hasta el instante n : $\hat{\phi}_1^{*b}, \hat{\phi}_2^{*b}, \dots, \hat{\phi}_p^{*b}$.
- 9: **Obtener \tilde{X}_{n+h}^{*b}** , la predicción de X_{n+h}^{*b} a partir de los coeficientes bootstrap y las últimas observaciones de la remuestra.
- 10: **Aproximar el PMSE** mediante su análogo bootstrap aproximado por Monte Carlo:

$$\widehat{PMSE}^* \left[\tilde{X}_{n+h}^{*b} \right] = \frac{1}{B} \sum_{b=1}^B \left(\tilde{X}_{n+h}^{*b} - X_{n+h}^{*b} \right)^2.$$

Observación 2.16. En el Algoritmo 10, el $PMSE$ bootstrap aproximado por Monte Carlo busca estimar el error cuadrático medio de predicción de la estimación del mejor predictor lineal de X_{n+h} a partir de la muestra X_1, X_2, \dots, X_n . Con la notación introducida al comienzo de este apartado, se tiene que $\widehat{PMSE}^* \left[\tilde{X}_{n+h}^{*b} \right]$ es un estimador de $PMSE \left[\tilde{X}_{n+h} \right]$.

²² En el universo bootstrap una de las premisas principales es imitar la estructura y propiedades que tiene el estadístico «clásico», de tal forma que el método de remuestreo empleado proporcione resultados consistentes con la realidad.

²³ Una alternativa este paso es realizar un sorteo uniforme de los $n-p+1$ posibles bloques de observaciones consecutivas de la muestra original de la serie.

2.3.2. Métodos bootstrap propuestos por Thombs y Schucany en 1990

En el presente apartado se introducirán los dos métodos presentados en [Thombs y Schucany \(1990\)](#) para generar remuestras de un modelo $AR(p)$. Una vez presentados dichos procedimientos, se mostrará también una manera de proceder para obtener intervalos de predicción bootstrap para este tipo de modelos y se dará un resultado teórico que garantizará la validez del método empleado.

Formulación hacia adelante

Supóngase que la serie de tiempo $\{X_t\}$ sigue un modelo $AR(p)$. En tal caso se tiene que

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t, \quad t \in \mathbb{Z}, \quad (2.15)$$

donde $\{a_t\}$ es un proceso de ruido blanco independiente del pasado de X_t .

Los dos mecanismos de generación de remuestras que se van a presentar a lo largo de esta sección están basados en los residuos de dos formulaciones del modelo autoregresivo de orden p , siendo la presentada en (2.15) la llamada «formulación hacia adelante». El motivo por el cual recibe este nombre es que la ecuación que rige el modelo necesita disponer de p valores fijados o conocidos de la serie para, a continuación, ir generando el valor del instante $p+1, p+2, \dots$. Dicho de otro modo, para poder aplicar el modelo es necesario conocer la dinámica temporal de la serie al menos durante p instantes para luego ir generando los valores en instantes sucesivos de manera recursiva de acuerdo con la ecuación dada por (2.15).

Observación 2.17. Nótese que será muy importante que el mecanismo de remuestreo que se emplee en el proceso preserve el carácter iid de los errores a_t .

Sea X_{t-n+1}, \dots, X_t una muestra de tamaño n del modelo $AR(p)$ y sean $\hat{\phi}_0, \hat{\phi}_1, \dots, \hat{\phi}_p$ los parámetros del modelo estimados por mínimos cuadrados. Se define el residuo i -ésimo hacia adelante como²⁴

$$\hat{a}_i = X_i - \hat{X}_i = X_i - \hat{\phi}_0 - \hat{\phi}_1 X_{i-1} - \cdots - \hat{\phi}_p X_{i-p}, \quad i = t-n+p+1, t-n+p+2, \dots, t. \quad (2.16)$$

Los residuos que acaban de definirse desempeñarán el papel de los errores del modelo en las remuestras, por lo que será importante que proporcionen una buena estimación de la distribución F_a , que en ningún momento se ha asumido que deba ser gaussiana.

Debido a lo ya comentado con anterioridad, para generar réplicas bootstrap serán necesarios p valores iniciales, los cuales pueden ser seleccionados tomando los primeros p valores de la serie (véase [Efron y Tibshirani 1986](#)) o escogiendo de forma aleatoria un bloque de p valores observados adyacentes (véase [Stine 1982](#)). Así pues, dados esos p valores iniciales, los $n-p$ elementos restantes de la remuestra se obtendrán a través de la ecuación del modelo, de manera que

$$X_j^* = \hat{\phi}_0 + \hat{\phi}_1 X_{j-1}^* + \cdots + \hat{\phi}_p X_{j-p}^* + \hat{a}_j^*, \quad j = t-n+p+1, t-n+p+2, \dots, t,$$

donde \hat{a}_j^* es una elección aleatoria a partir de \hat{F}_a y los parámetros $\hat{\phi}_0, \hat{\phi}_1, \dots, \hat{\phi}_p$ son los parámetros del modelo $AR(p)$ estimados por mínimos cuadrados.

Observación 2.18 ([Thombs y Schucany 1990](#), p. 487). Si en la estimación de los parámetros se emplea otro criterio diferente al de mínimos cuadrados entonces los residuos resultantes podrían no estar centrados y sería necesario definir $\tilde{a}_t = \hat{a}_t - \bar{a}$, donde $\bar{a} = \frac{1}{n-p} \sum_{i=t-n+p+1}^t \hat{a}_i$.

²⁴ Nótese que los índices para los que es posible calcular los residuos dependen tanto del tamaño de la muestra como del número de parámetros del modelo. En concreto, el menor instante observado en la muestra es $t-n+1$ y para que $i-p$ (el menor instante del que depende el residuo) tome ese valor se tiene que i debe ser igual a $t-n+p+1$.

Formulación hacia atrás

El método de generación de remuestras del apartado anterior, si bien funciona razonablemente bien para estimar los errores típicos de los parámetros $\hat{\phi}_i$, este no es coherente cuando se utiliza para estimar la distribución condicional de X_{t+k} ²⁵. Para serlo debería conseguir que las muestras bootstrap generadas imitasen la estructura de correlaciones de la serie que se quiere predecir y que fijase condicionalmente los últimos p valores, i.e., para cada réplica

$$X_t^* = X_t, X_{t-1}^* = X_{t-1}, \dots, X_{t-p+1}^* = X_{t-p+1}.$$

Por ese motivo, se introduce otro procedimiento de remuestreo basado en otra formulación de los modelos $AR(p)$. En las páginas 222 y 223 de [Box et al. \(2008\)](#) puede verse que estos procesos también admiten una expresión como combinación lineal de valores futuros más un término de error, manteniendo incluso la estructura de correlaciones del proceso. La expresión del modelo en estos términos, conocida como «formulación hacia atrás», es la siguiente:

$$X_t = \phi_0 + \phi_1 X_{t+1} + \dots + \phi_p X_{t+p} + e_t.$$

A diferencia de lo que ocurría con el modelo (2.15), esta reformulación alternativa requiere que las variables X_t sean generadas «hacia atrás» en el tiempo. Por ello, una elección natural como valores iniciales de las remuestras son las últimas p observaciones de la serie²⁶. En el Algoritmo 11 puede verse el nuevo esquema de remuestreo, cuya naturaleza es análoga al anterior.

Observación 2.19 ([Thombs y Schucany 1990](#), p. 487). Las verdaderas distribuciones de los errores en las expresiones hacia atrás y hacia adelante del modelo $AR(p)$, bajo normalidad, son iguales. Por el contrario, si las series no son gaussianas no solo no son iguales sino que los residuos hacia atrás \hat{e}_i solo son incorrelados. Por ello, no parece que sea adecuado emplear técnicas de remuestreo basadas en la distribución empírica de los \hat{e}_i . Sin embargo, el uso de la distribución empírica funciona razonablemente bien en la práctica, por lo que se pasará por alto esta falta de coherencia.

Algoritmo 11 Remuestras bootstrap en modelo $AR(p)$ hacia atrás, [Thombs y Schucany 1990](#)

- 1: **Estimar los parámetros** por mínimos cuadrados: $\hat{\phi}_0, \hat{\phi}_1, \dots, \hat{\phi}_p$.
- 2: **Construir los residuos hacia atrás** \hat{e}_i y calcular su versión centrada. Sea \hat{F}_e la distribución empírica de tales residuos.
- 3: **Repetir** para $b = 1, \dots, B$:
- 4: **Fijar los p valores finales** de la remuestra iguales a los p últimos valores de la serie:

$$X_t^{*b} = X_t, X_{t-1}^{*b} = X_{t-1}, \dots, X_{t-p+1}^{*b} = X_{t-p+1}.$$

- 5: **Obtener los residuos hacia atrás** \hat{e}_i^{*b} , con $i = t-p, \dots, t-n+1$ a partir de su distribución empírica \hat{F}_e .
- 6: **Obtener los $n-p$ valores restantes** de la remuestra bootstrap mediante la ecuación hacia atrás del modelo:

$$X_i^{*b} = \hat{\phi}_0 + \hat{\phi}_1 X_{i+1}^{*b} + \dots + \hat{\phi}_p X_{i+p}^{*b} + \hat{e}_i^{*b}, \quad i = t-p, \dots, t-n+1.$$

- 7: **Definir la remuestra como** $X_{t-n+1}^{*b}, X_{t-n+2}^{*b}, \dots, X_{t-p}^{*b}, X_{t-p+1}, \dots, X_t$.
-

²⁵ Nótese que si X_t sigue un modelo $AR(p)$ la distribución de X_{t+k} condicionada a todos los valores del pasado es la misma que la distribución condicionada únicamente a los p últimos.

²⁶ Así pues, las remuestras bootstrap tendrán como peculiaridad que todas tienen los mismos p últimos valores de la serie y la misma estructura de correlaciones que la serie que se pretende predecir.

Intervalos de predicción bootstrap

Sea X_{t-n+1}, \dots, X_t una muestra de tamaño n de un modelo $AR(p)$ y para $b = 1, 2, \dots, B$ sean las remuestras hacia atrás dadas por

$$X_{t-n+1}^{*b}, \dots, X_{t-p}^{*b}, X_{t-p+1}, \dots, X_t.$$

El mecanismo que se va a emplear consiste en calcular el valor futuro X_{t+k}^{*b} en cada remuestra aplicando la formulación hacia adelante del modelo. Así, para $b = 1, 2, \dots, B$ se tendrá que

$$X_{t+k}^{*b} = \widehat{\phi}_0^{*b} + \widehat{\phi}_1^{*b} X_{t+k-1}^{*b} + \dots + \widehat{\phi}_p^{*b} X_{t+k-p}^{*b} + \widehat{a}_{t+k}^{*b}, \quad (2.17)$$

donde los parámetros $\widehat{\phi}_0^*, \dots, \widehat{\phi}_p^*$ son estimados por mínimos cuadrados en la remuestra bootstrap y \widehat{a}_{t+k}^{*b} es una realización aleatoria de \widehat{F}_a . Así pues, habiendo obtenido el conjunto de B valores futuros bootstrap $X_{t+k}^{*1}, \dots, X_{t+k}^{*B}$, los límites del intervalo de predicción se definen como

$$\left(z_{t+k}^{*(\alpha/2)}, z_{t+k}^{*(1-\alpha/2)} \right),$$

donde $z_{t+k}^{*(\alpha/2)}$ y $z_{t+k}^{*(1-\alpha/2)}$ son los cuantiles $\alpha/2$ y $1 - \alpha/2$, respectivamente, de la estimación por Monte Carlo de la distribución bootstrap de X_{t+k}^* .

Observación 2.20. Para un horizonte $k > 1$ las predicciones deben ser calculadas de manera recursiva, obteniendo primero X_{t+1}^{*b} , luego X_{t+2}^{*b} y así sucesivamente hasta X_{t+k}^{*b} . Por lo tanto, cuanto mayor sea k , mayor será la cantidad de elementos que incrementan la varianza de la predicción²⁷.

Observación 2.21. Nótese que tal y como se han sido construidas las remuestras, dependiendo de cuál sea el horizonte de predicción k los valores X_{t+k-i}^{*b} que aparecen en (2.17) pueden ser valores futuros que han sido generados por bootstrap o bien algunos de los p últimos valores de la remuestra (y, por lo tanto, de la muestra original). En este último caso, debido a cómo ha sido construida la remuestra, se tiene que ese X_{t+k-i}^{*b} será el mismo para todos los b y, en consecuencia, la variabilidad de la predicción vendrá dada únicamente por algunos X_{t+k-j}^{*b} ²⁸, los parámetros $\widehat{\phi}_i^{*b}$ y los \widehat{a}_{t+k}^{*b} .

Algoritmo 12 Intervalo de predicción bootstrap en modelo $AR(p)$, Thombs y Schucany 1990

- 1: **Calcular los residuos hacia adelante, \widehat{a}_j y hacia atrás, \widehat{e}_j** y obtener sus versiones centradas. Sean \widehat{F}_a y \widehat{F}_e las distribuciones empíricas de tales residuos.
- 2: **Repetir** para $b = 1, \dots, B$:
- 3: **Generar una réplica bootstrap** a partir de la **formulación hacia atrás**:

$$X_{t-n+1}^{*b}, X_{t-n+2}^{*b}, \dots, X_{t-p}^{*b}, X_{t-p+1}, \dots, X_t.$$

- 4: **Estimar los parámetros $\widehat{\phi}_0^{*b}, \dots, \widehat{\phi}_p^{*b}$** a partir de la remuestra generada en el Paso 3.
- 5: **Generar un valor futuro bootstrap** a partir de la **formulación hacia adelante**:

$$X_{t+k}^{*b} = \widehat{\phi}_0^{*b} + \widehat{\phi}_1^{*b} X_{t+k-1}^{*b} + \dots + \widehat{\phi}_p^{*b} X_{t+k-p}^{*b} + \widehat{a}_{t+k}^{*b}$$

donde \widehat{a}_{t+k}^{*b} es una realización aleatoria de \widehat{F}_a .

- 6: **Obtener los límites del intervalo de predicción** a partir de los correspondientes cuantiles empíricos de $X_{t+k}^{*1}, \dots, X_{t+k}^{*B}$.
-

²⁷ Esto parece razonable, ya que lo que cabe esperar de una predicción que se realiza con información hasta cierto instante es que su varianza sea mayor a medida que aumenta el horizonte al que se quiere predecir o, dicho de otro modo, que la predicción sea menos precisa cuanto mayor sea ese horizonte.

²⁸ Aquellos que no cumplan que $X_{t+k-i}^{*b} = X_{t+k-i}$.

Por último, se va a presentar el resultado teórico en el que descansa la validez del procedimiento bootstrap para el cálculo de intervalos de predicción que ha sido presentado.

Teorema 2.7 (Thombs y Schucany 1990, p. 489). *Sea $\{X_t\}$ un proceso estacionario autorregresivo con $\mathbb{E}[a_t] = 0$ y $\mathbb{E}[|a_t^\alpha|] < \infty$ para algún $\alpha > 2$. Sea, además (X_{t-n+1}, \dots, X_t) una muestra de $\{X_t\}$. Cuando n tiende a ∞ se tiene que a lo largo de casi todas las muestras se cumple que*

- (1) $(\hat{\phi}_0^*, \hat{\phi}_1^*, \dots, \hat{\phi}_p^*) \longrightarrow (\hat{\phi}_0, \hat{\phi}_1, \dots, \hat{\phi}_p)$, en probabilidad condicionada, y
- (2) $X_{t+k}^* \xrightarrow{d} X_{t+k}$.

2.3.3. El bootstrap condicional propuesto por Cao y otros en 1997

En esta sección se van a presentar los mecanismos de remuestreo propuestos en Cao et al. (1997) y conocidos ambos como métodos bootstrap condicionales. Estos procedimientos surgen en el mismo contexto que en el que fue presentado en Thombs y Schucany (1990), es decir, en la construcción de intervalos de predicción bootstrap en modelos autorregresivos. No obstante, los métodos bootstrap condicionales que van a ser presentados tienen como principal ventaja con respecto al propuesto en Thombs y Schucany (1990) que son computacionalmente más rápidos.

Sea X_1, \dots, X_t una muestra de un modelo AR(p) estacionario dado por

$$X_s = \phi_0 + \phi_1 X_{s-1} + \dots + \phi_p X_{s-p} + a_s, \quad s \in \mathbb{Z}, \quad (2.18)$$

donde los errores a_s son variables aleatorias independientes de media 0 y varianza σ_a^2 con distribución común F_a y las raíces del polinomio $\phi(z) = 1 - \sum_{i=1}^p \phi_i z^i$ caen fuera del círculo complejo unidad. El objetivo que se va a perseguir es construir un intervalo de predicción para el valor futuro de la serie dado por X_{t+k} para un cierto $k > 0$.

Una primera aproximación podría ser emplear el Algoritmo 10 (los 9 primeros pasos), propuesto en Stine (1987) y comentado en la Sección 2.3.1. No obstante, esta estrategia de remuestreo no busca replicar la distribución condicional de X_{t+k} dada la muestra observada²⁹ y por ello no es adecuada para el objetivo que se persigue en este apartado.

Por otro lado, otra estrategia podría ser emplear el algoritmo propuesto en Thombs y Schucany (1990), que sí está diseñado para replicar la estructura de la distribución condicional de X_{t+k} . No obstante, este procedimiento es bastante costoso computacionalmente. El motivo de esto es que en cada réplica bootstrap, X_{t+k}^* , es necesario construir una remuestra hacia atrás que tiene que ser modelada acorde con una estructura AR(p) y luego es necesario volver a estimar los parámetros de nuevo en dicha remuestra. Si, aún encima, el número de remuestras que se pretende calcular es elevado, el tiempo de cómputo del intervalo de predicción puede ser muy elevado.

Para paliar este comportamiento, en Cao et al. (1997) se proponen dos métodos en los cuales no se construyen remuestras hacia atrás, lo cual además provoca que no se añada a la predicción la variabilidad procedente de la estimación de los parámetros. El esquema de remuestreo del primero de ellos ha sido recogido en el Algoritmo 13, donde puede verse que el mecanismo de remuestreo solamente busca generar valores futuros de la serie dada la muestra observada y para ello no se construye una réplica bootstrap de la serie original. En cuanto al segundo de los métodos propuestos en el artículo, este consiste en sustituir la aproximación de F_a mediante la distribución empírica de los residuos recentrados por un estimador tipo núcleo, $\hat{F}_{a,h}$, en el caso de suponer que la distribución de los errores

²⁹ De hecho, al tratarse de una muestra de un modelo AR(p) esta distribución condicional es la misma que la condicionada a los p últimos valores de la muestra.

es continua. Más aún, en caso de tener algún tipo de información acerca de dicha distribución podría incorporarse al mecanismo de remuestreo que ha sido propuesto.

La idea intuitiva que subyace al procedimiento bootstrap presentado es la siguiente. A partir de la ecuación (2.18) uno podría ver X_{t+k} como una suma de términos que únicamente dependen de la muestra observada y los verdaderos parámetros y otros términos que son combinaciones de los errores futuros a_{t+1}, \dots, a_{t+k} , cuyos coeficientes dependen también los parámetros desconocidos. En consecuencia, la distribución condicional de X_{t+k} es esencialmente una convolución de k distribuciones obtenidas a partir de reescalar la distribución de los errores F_a . Luego conociendo F_a y los parámetros, se conocería también la distribución condicional de X_{t+k} . Por lo tanto, un estimador razonable de esta distribución condicional consistiría en reemplazar los verdaderos parámetros por sus estimaciones mediante el método de mínimos cuadrados y la distribución de los errores por una estimación, \widehat{F}_a .

Algoritmo 13 Intervalo de predicción bootstrap en modelo AR(p), Cao et al. 1997

- 1: **Estimar los coeficientes autorregresivos** por mínimos cuadrados: $\widehat{\phi}_0, \widehat{\phi}_1, \dots, \widehat{\phi}_p$.
- 2: **Construir los residuos hacia adelante:**

$$\widehat{a}_i = X_i - \widehat{\phi}_0 - \widehat{\phi}_1 X_{i-1} - \widehat{\phi}_2 X_{i-2} - \dots - \widehat{\phi}_p X_{i-p}, \quad i = p+1, p+2, \dots, t.$$

- 3: **Definir la distribución empírica de los residuos centrados:**

$$\widehat{F}_n^{\widehat{a}}(x) = \frac{1}{n-p} \sum_{i=p+1}^t \mathbb{I}\{\widehat{a}'_i \leq x\}, \quad \widehat{a}'_i = \widehat{a}_i - \bar{a}, \quad \bar{a} = \frac{1}{t-p} \sum_{i=p+1}^t \widehat{a}_i.$$

- 4: **Repetir** para $b = 1, \dots, B$:
- 5: **Obtener una remuestra de los residuos** \widehat{a}_s^{*b} , con $s = t+1, \dots, t+k$, a partir de $\widehat{F}_n^{\widehat{a}}$.
- 6: **Fijar p valores iniciales** iguales a la muestra original:

$$X_s^{*b} = X_s, \quad s = t-p+1, \dots, t.$$

- 7: **Calcular los k futuros valores** mediante la ecuación del modelo:

$$X_s^{*b} = \widehat{\phi}_0 + \widehat{\phi}_1 X_{s-1}^{*b} + \dots + \widehat{\phi}_p X_{s-p}^{*b} + \widehat{a}_s^{*b}, \quad s = t+1, \dots, t+k.$$

- 8: **Aproximar la distribución** de X_{t+k} mediante la aproximación por Monte Carlo de la distribución condicional de X_{t+k}^* dada la muestra original.
-

Observación 2.22. En Cao et al. (1997) se demuestra que tanto el procedimiento bootstrap propuesto por ellos como el de Thombs y Schucany (1990) asintóticamente se comportan igual.

Observación 2.23. Si el horizonte de predicción k es pequeño ($k = 1, 2$) los cálculos de la distribución bootstrap exacta pueden ser menos costosos computacionalmente que la aproximación por Monte Carlo.

Por último, se presenta el resultado que garantiza la consistencia del método bootstrap que ha sido presentado, ya que al demostrar la convergencia en distribución condicional de la réplica bootstrap X_{t+k}^* a X_{t+k} , establece que la cobertura asintótica del intervalo coincide con el valor nominal fijado.

Teorema 2.8 (Cao et al. 1997, p. 970). *Sea $\{X_s\}$ una serie de tiempo siguiendo un proceso AR(p) como el indicado en (2.18) donde $\phi(z) \neq 0$ para todo $z \in \mathbb{Z}$ tal que $|z| \leq 1$. Además, supóngase que $\mathbb{E}[a_s] = 0$ y que $\mathbb{E}[|a_s|^\alpha] < \infty$ para algún $\alpha > 2$. Entonces, bajo el mecanismo de remuestreo presentado en esta sección (la versión empírica o la suavizada) se tiene que si h tiende a 0 se cumple que para casi todas las muestras*

$$X_{t+k}^* \xrightarrow{d} X_{t+k}.$$

2.3.4. Método bootstrap propuesto por Franke y Kreiss en 1992

A lo largo de esta sección se presentará la metodología bootstrap propuesta en [Kreiss y Franke \(1992\)](#) en el ámbito de la estimación de la distribución en el muestreo de M -estimadores de los parámetros de un modelo $\text{ARMA}(p, q)$. La idea principal del método de remuestreo es similar a la de [Stine \(1987\)](#), en el sentido de generar réplicas de los residuos y luego aplicar la ecuación del modelo para obtener remuestras bootstrap sobre las que evaluar el estimador, solo que para ello ahora será necesario estimar también los coeficientes de la parte relativa al proceso de medias móviles para poder obtener dichos residuos.

Sea una serie de tiempo $\{X_t\}$ que sigue un modelo $\text{ARMA}(p, q)$, es decir, que es estacionaria y verifica la ecuación dada por

$$\phi(\mathbf{B}) X_t = \theta(\mathbf{B}) a_t, \quad \forall t \in \mathbb{Z},$$

donde $\boldsymbol{\theta} = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q) \in \Theta \subset \mathbb{R}^{p+q}$, con $\phi_p \neq 0$ y $\theta_q \neq 0$, es el vector de parámetros desconocido del modelo ARMA y el proceso de ruido blanco $\{a_t\}$ lo constituyen variables aleatorias independientes y de distribución común F de media 0 y varianza finita σ^2 .

Supóngase, además, que los polinomios $\phi(z)$ y $\theta(z)$ no tienen raíces en común y que el espacio de parámetros Θ es tal que se cumplen las condiciones de causalidad e invertibilidad, es decir, por lo visto en el [Corolario 1.6](#), que se cumple que

$$\theta(z)\phi(z) \neq 0, \quad \forall z \in \mathbb{C} \text{ tal que } |z| \leq 1.$$

La idea principal del procedimiento bootstrap presentado en [Kreiss y Franke \(1992\)](#) es la siguiente. En primer lugar, considérese una muestra finita de la serie de tiempo $\{X_t\}$ y denótese esta como $\mathbf{X} = (X_{1-p}, \dots, X_n)'$. El primer paso consiste en obtener, a partir de estos datos observados, una estimación del vector de parámetros

$$\hat{\boldsymbol{\theta}}_n = (\hat{\phi}_{1,n}, \dots, \hat{\phi}_{p,n}, \hat{\theta}_{1,n}, \dots, \hat{\theta}_{q,n}) \in \Theta,$$

de manera que a partir de esta se pueda generar una estimación de los residuos.

Denotando por \hat{F}_n a la distribución empírica de los residuos centrados, el siguiente paso consiste en obtener una remuestra (con técnicas de Monte Carlo, se generarían B) de variables independientes e idénticamente distribuidas como \hat{F}_n para luego transformarlas en «datos bootstrap» empleando la ecuación del modelo, solo que tomando como vector de parámetros $\hat{\boldsymbol{\theta}}_n$. Esto permitiría poder volver a evaluar el estimador $\hat{\boldsymbol{\theta}}_n$ en la remuestra obtenida y así, repitiendo este proceso un número razonable de veces, podría aproximarse su distribución.

Conviene destacar en relación a la estimación del vector de parámetros $\boldsymbol{\theta}$, que en el artículo se centran en un tipo especial de estimadores llamados M -estimadores, los cuales se definen a continuación.

Definición 2.2 (M -estimador, [Van der Vaart y Wellner 1996](#), p. 284). Se denomina M -estimador a aquel que maximiza (o minimiza) un cierto criterio en forma de función.

Observación 2.24. En el caso de observaciones independientes e idénticamente distribuidas X_1, \dots, X_n un criterio habitual es el de la forma

$$\ell(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m_{\boldsymbol{\theta}}(X_i),$$

para ciertas funciones conocidas $m_{\boldsymbol{\theta}}$ en el espacio muestral. Por ejemplo, el método de máxima verosimilitud se corresponde con la elección de $m_{\boldsymbol{\theta}} = \log(f_{\boldsymbol{\theta}})$, donde $f_{\boldsymbol{\theta}}$ es la densidad de las observaciones.

Más concretamente, en este contexto de los modelos ARMA(p, q) se definen los M -estimadores³⁰ del vector de parámetros $\boldsymbol{\theta}$ como aplicaciones medibles $\widehat{\boldsymbol{\theta}}_n^M : \mathbb{R}^{p+q} \rightarrow \boldsymbol{\Theta}$ que son soluciones de

$$\Psi_n(\widehat{\boldsymbol{\theta}}_n^M) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \psi(a_j(\widehat{\boldsymbol{\theta}}_n^M)) \mathbf{Z}(j-1; \widehat{\boldsymbol{\theta}}_n^M) = 0, \quad (2.19)$$

donde ψ es una función de pesos que depende del método para construir el M -estimador. Además, en la ecuación (2.19) se han empleado las siguientes abreviaturas para compactar la escritura:

$$\begin{aligned} a_j(\boldsymbol{\theta}) &= \sum_{k=0}^{j-1} \beta_k(\boldsymbol{\theta}) \left(X_{j-k} - \sum_{i=1}^p \phi_i X_{j-k-i} \right), \quad j = 1, \dots, n, \\ \mathbf{Z}(j-1; \boldsymbol{\theta}) &= \sum_{k=0}^{j-1} \beta_k(\boldsymbol{\theta}) (X(j-1-k)', A(j-1-k; \boldsymbol{\theta})')', \end{aligned} \quad (2.20)$$

donde

$$\begin{aligned} \sum_{k=0}^{\infty} \beta_k(\boldsymbol{\theta}) z^k &= \left(1 + \sum_{j=1}^q \theta_j z^j \right)^{-1}, \quad \forall z \in \mathbb{C} \text{ tal que } |z| \leq 1, \\ X(j-1) &= (X_{j-1}, \dots, X_{j-p})', \\ A(j-1; \boldsymbol{\theta}) &= (a_{j-1}(\boldsymbol{\theta}), \dots, a_{j-q}(\boldsymbol{\theta}))'. \end{aligned}$$

A continuación se recogen dos resultados teóricos que establecen propiedades asintóticas acerca de los estimadores \sqrt{n} -consistentes³¹. La utilidad de estos resultados radicarà luego en poder aproximar mediante técnicas bootstrap la distribución de $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^M - \boldsymbol{\theta})$, con $\widehat{\boldsymbol{\theta}}_n^M$ un M -estimador \sqrt{n} -consistente, sin necesidad de aplicar las técnicas de remuestreo a un método explícito para construir el M -estimador (mínimos cuadrados, máxima verosimilitud, etc.).

Teorema 2.9 (Kreiss y Franke 1992, pp. 301-302). *Sea una función $\psi : \mathbb{R} \rightarrow \mathbb{R}$ dos veces continuamente diferenciable con derivadas acotadas ψ' y ψ'' . Si $\{\widehat{\boldsymbol{\theta}}_n\}_{n \in \mathbb{N}} \subset \boldsymbol{\Theta}$ denota a una sucesión de estimadores \sqrt{n} -consistentes del vector de parámetros $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ y se cumple que $\mathbb{E}[a_t^4] < \infty$ entonces se tiene que*

$$\Psi_n(\widehat{\boldsymbol{\theta}}_n) - \bar{\Psi}_n + \Gamma_n \sqrt{n} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = O_P(n^{-1/2}),$$

donde

$$\begin{aligned} \bar{\Psi}_n &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \psi(a_j) \mathbf{Z}_n(j-1), \\ \Gamma_n &= \frac{1}{n} \sum_{j=1}^n \psi'(a_j) \mathbf{Z}_n(j-1) \mathbf{Z}_n(j-1)', \\ \mathbf{Z}_n(j-1) &= \sum_{k=0}^{j-1} \beta_k(\widehat{\boldsymbol{\theta}}_n) (X(j-1-k)', A(j-1-k)')', \\ A(j-1) &= (a_{j-1}, \dots, a_{j-q})'. \end{aligned}$$

³⁰ Para profundizar acerca de los M -estimadores en modelos ARMA(p, q) puede consultarse Kreiss (1985).

³¹ Un estimador $\widehat{\boldsymbol{\theta}}_n$ de $\boldsymbol{\theta}$ se dice que es \sqrt{n} consistente si, y solo si, se tiene que

$$\sqrt{n} |\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| = O_P(1).$$

Teorema 2.10 (Kreiss y Franke 1992, p. 303). *Bajo las mismas condiciones que en el Teorema 2.9 y siendo $\{\widehat{\boldsymbol{\theta}}_n^M\}_{n \in \mathbb{N}} \subset \boldsymbol{\Theta}$ una sucesión de M -estimadores \sqrt{n} -consistentes del vector de parámetros $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ se tiene que*

$$\sqrt{n} \left(\widehat{\boldsymbol{\theta}}_n^M - \boldsymbol{\theta} \right) = \Gamma_n^{-1} \bar{\Psi}_n + O_P \left(n^{-1/2} \right).$$

Así pues, en virtud del Teorema 2.10, bastará con aplicar la metodología bootstrap al estadístico $\Gamma_n^{-1} \bar{\Psi}_n$, donde habrá que conocer la función de pesos ψ asociada al método que construye el M -estimador³². No obstante, antes de introducir el método de remuestreo se va a presentar un resultado fundamental, ya que establece la validez de la estimación de la distribución de los residuos.

Teorema 2.11 (Kreiss y Franke 1992, pp. 305-306). *Sea $\{\widehat{\boldsymbol{\theta}}_n\}_{n \in \mathbb{N}}$ una sucesión de estimadores \sqrt{n} -consistentes de $\boldsymbol{\theta}$, con $\widehat{\boldsymbol{\theta}}_n = (\widehat{\phi}_{1,n}, \dots, \widehat{\phi}_{p,n}, \widehat{\theta}_{1,n}, \dots, \widehat{\theta}_{q,n})$ y considérense los residuos dados por*

$$\widehat{a}_{j,n} = \sum_{k=0}^{j-1} \widehat{\beta}_{k,n} \left(X_{j-k} - \sum_{i=1}^p \widehat{\phi}_{i,n} X_{j-k-i} \right), \quad j = 1, \dots, n, \quad (2.21)$$

donde

$$\sum_{k=0}^{\infty} \widehat{\beta}_{k,n} z^k = \left(1 + \sum_{j=1}^q \widehat{\theta}_{j,n} z^j \right)^{-1}, \quad \forall z \in \mathbb{C} \text{ tal que } |z| \leq 1.$$

Sea, además, \widehat{F}_n la función de distribución empírica de los residuos definidos en (2.21) centrados. Denotando por $d_2(\cdot, \cdot)$ a la distancia de Mallows³³ se tiene que cuando n tiende a infinito $d_2(F, \widehat{F}_n)$ converge en probabilidad a 0.

Ya que es posible aproximar razonablemente bien el comportamiento de los residuos mediante \widehat{F}_n y solamente se dispone de un mecanismo de remuestreo para ellos, el siguiente paso consistirá en reescribir el estadístico $\Gamma_n^{-1} \bar{\Psi}_n$ de manera que dependa únicamente de los residuos. Para ello, nótese que la serie de tiempo $\{X_t\}$ es causal por hipótesis, se tiene que

$$X_t = \sum_{k=0}^{\infty} \rho_k(\boldsymbol{\theta}) a_{t-k}, \quad (2.22)$$

donde los coeficientes $\rho_k(\boldsymbol{\theta})$ pueden ser obtenidos a partir de la fórmula recursiva presentada en la Observación 1.7. Así pues, dado un M -estimador $\widehat{\boldsymbol{\theta}}_n^M$ y denotando $r = \max\{p, q\}$, considérese una remuestra a_{1-r}^*, \dots, a_n^* , obtenida a partir de la distribución empírica \widehat{F}_n^M de los residuos $a_j(\widehat{\boldsymbol{\theta}}_n^M)$ centrados, donde dichos residuos se obtienen sustituyendo $\boldsymbol{\theta}$ y ϕ_i en (2.20) por $\widehat{\boldsymbol{\theta}}_n^M$ y $\phi_{i,n}^M$, respectivamente. A partir de la ecuación (2.22) la remuestra de la serie de tiempo puede definirse como

$$X_t^* = \sum_{k=1}^p \widehat{\phi}_{k,n}^M X_{t-k}^* + \sum_{k=1}^q \widehat{\theta}_{k,n}^M a_{t-k}^* + a_t^* = \sum_{l=0}^{t+p-1} \rho_l(\widehat{\boldsymbol{\theta}}_n^M) a_{t-l}^*, \quad t = 1-p, \dots, n, \quad (2.23)$$

³² En el caso del método de mínimos cuadrados ordinarios se tiene que $\psi(x) = x$ para todo $x \in \mathbb{R}$.

³³ Dadas dos distribuciones P y Q se define la distancia de Mallows entre ellas como

$$d_2(P, Q) = \inf \left\{ \sqrt{\mathbb{E}[(X - Y)^2]} \right\},$$

donde el ínfimo se toma sobre todos los pares de variables aleatorias (X, Y) con X e Y distribuidos acorde a P y a Q , respectivamente.

donde, por conveniencia, se tiene que $X_j^* = 0$ y $a_j^* = 0$ para $j \leq -r$.

Se definen, ahora las versiones bootstrap de $\bar{\Psi}_n$ y de Γ_n :

$$\begin{aligned}\Gamma_n^* &= \frac{1}{n} \sum_{j=1}^n \psi'(a_j^*) \mathbf{Z}_n^*(j-1) \mathbf{Z}_n^*(j-1)', \\ \mathbf{Z}_n^*(j-1) &= \sum_{k=0}^{j-1} \beta_k \left(\widehat{\boldsymbol{\theta}}_n^M \right) (X_{j-1-k}^*, \dots, X_{j-p-k}^*, a_{j-1-k}^*, \dots, a_{j-q-k}^*)', \\ \bar{\Psi}_n^* &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \psi(a_j^*) \mathbf{Z}_n^*(j-1).\end{aligned}$$

Teorema 2.12 (Kreiss y Franke 1992, pp. 307-308). *Con la notación anterior, sea ψ una función dos veces diferenciable con derivadas acotadas para todo $x \in \mathbb{R}$ y sea $\{\widehat{\boldsymbol{\theta}}_n^M\}_{n \in \mathbb{N}} \subset \Theta$ una sucesión de estimadores \sqrt{n} -consistentes de $\boldsymbol{\theta}$. Entonces se tiene que*

$$d_2(\Gamma_n, \Gamma_n^*) \longrightarrow 0, \quad d_2(\bar{\Psi}_n, \bar{\Psi}_n^*) \longrightarrow 0,$$

donde la convergencia es en probabilidad y donde, por abuso de notación, se ha empleado $d_2(U, V)$ para denotar la distancia de Mallows entre F_U y F_V , las distribuciones de U y V , respectivamente.

Observación 2.25. En la página 307 de Kreiss y Franke (1992) se demuestra que los estadísticos $\Gamma_n^{-1} \bar{\Psi}_n$ y $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^M - \boldsymbol{\theta})$ convergen en distribución a la misma normal asintótica. En consecuencia, la validez del procedimiento bootstrap empleado queda demostrada a partir de ese resultado y del Teorema (2.12).

Algoritmo 14 Estimación bootstrap de la distribución de M -estimadores en modelos ARMA(p, q), Kreiss y Franke 1992

- 1: **Construir $\widehat{\boldsymbol{\theta}}_n^M$ un M -estimador \sqrt{n} -consistente de $\boldsymbol{\theta}$ a partir de la muestra X_{1-p}, \dots, X_n .**
- 2: **Construir los residuos:**

$$\widehat{a}_{j,n}(\boldsymbol{\theta}_n^M) = \sum_{k=0}^{j-1} \beta_k \left(\widehat{\boldsymbol{\theta}}_n^M \right) \left(X_{j-k} - \sum_{i=1}^p \widehat{\phi}_{i,n}^M X_{j-k-i} \right), \quad j = 1, \dots, n,$$

- 3: **Definir la distribución empírica de los residuos centrados:**

$$\widehat{F}_n^{\widetilde{a}}(x) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}\{\widetilde{a}_j \leq x\}, \quad \text{donde } \widetilde{a}_j = \widehat{a}_{j,n}(\boldsymbol{\theta}_n^M) - \overline{\widehat{a}_{j,n}(\boldsymbol{\theta}_n^M)}, \quad \overline{\widehat{a}_{j,n}(\boldsymbol{\theta}_n^M)} = \frac{1}{n} \sum_{k=1}^n \widehat{a}_{k,n}(\boldsymbol{\theta}_n^M)$$

- 4: **Repetir para $b = 1, \dots, B$:**
- 5: **Obtener a partir de $\widehat{F}_n^{\widetilde{a}}$ una remuestra de los residuos: $a_{1-r}^{*b}, \dots, a_n^{*b}$.**
- 6: **Construir la remuestra bootstrap:** fijando $X_j^{*b} = 0$ y $a_j^{*b} = 0$ para $j \leq -r$ se define

$$X_t^{*b} = \sum_{k=1}^p \widehat{\phi}_{k,n}^M X_{t-k}^{*b} + \sum_{k=1}^q \widehat{\theta}_{k,n}^M a_{t-k}^{*b} + a_t^{*b}, \quad t = 1-p, \dots, n.$$

- 7: **Construir $(\Gamma_n^{*b})^{-1} \bar{\Psi}_n^{*b}$ a partir de $X_{1-p}^{*b}, \dots, X_n^{*b}$ y de $a_{1-q}^{*b}, \dots, a_n^{*b}$.**
 - 8: **Aproximar la distribución de $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^M - \boldsymbol{\theta})$ mediante $(\Gamma_n^{*1})^{-1} \bar{\Psi}_n^{*1}, \dots, (\Gamma_n^{*B})^{-1} \bar{\Psi}_n^{*B}$.**
-

2.3.5. Métodos bootstrap en procesos estacionarios generales

En el caso de no disponer de una estructura de dependencia explícita, en el apartado de los métodos bootstrap para la estimación se vio que en ese contexto sigue siendo posible disponer de métodos de remuestreo. No obstante, estos métodos no son válidos en el contexto de la predicción, ya que no son consistentes a la hora de estimar la distribución condicional de X_{n+k}^* dada la muestra X_1, \dots, X_n . De este modo, para estimar dicha distribución es necesario realizar alguna suposición acerca del tipo de dependencia como, por ejemplo, que el proceso sea un proceso de Markov de orden p , es decir, que la distribución de X_{n+k} condicionada a la muestra X_1, \dots, X_n sea la misma que la condicionada a X_{n-p+1}, \dots, X_n .

En caso de que la serie de tiempo $\{X_t\}$ sea un proceso de Markov ($p = 1$), en [Cao \(1999\)](#) se propone el estimador tipo núcleo para $F_k(y | \mathbf{x}) = \mathbb{P}\{X_{n+k} \leq y | X_{n-p+1} = x_1, \dots, X_n = x_p\}$ dado por

$$\widehat{F}_{k,h}(y | x) = \sum_{i=1}^{n-k} \frac{K_h(x - X_i)}{\sum_{j=1}^{n-k} K_h(x - X_j)} \mathbb{I}\{X_{i+k} \leq y\},$$

donde $K_h(u) = h^{-1}K(u/h)$, con $h > 0$, y $K(\cdot)$ es una función de tipo núcleo. Ahora bien, emplear este estimador para calcular el intervalo de predicción con cobertura nominal $1 - \alpha$ dado por $(\widehat{F}_{k,h}^{-1}(\alpha/2 | x), \widehat{F}_{k,h}^{-1}(1 - \alpha/2 | x))$ es equivalente a llevar a cabo un procedimiento bootstrap tal que

$$\mathbb{P}^*\{X_{n+k}^* = X_{i+k}\} = \widehat{p}_i = \frac{K_h(x - X_i)}{\sum_{j=1}^{n-k} K_h(x - X_j)}, \quad i = 1, 2, \dots, n - k.$$

El esquema de dicho procedimiento puede verse en el Algoritmo 15, en el cual han sido incluidas también técnicas de Monte Carlo para aproximar la distribución bootstrap de X_{n+k}^* .

Algoritmo 15 Intervalos de predicción bootstrap en procesos de Markov, [Cao 1999](#)

- 1: **Construir los bloques muestrales** de tamaño $k + 1$: $B_{i,k+1}$, con $i = 1, \dots, n - k$.
 - 2: **Calcular las probabilidades** \widehat{p}_i , con $i = 1, \dots, n - k$.
 - 3: **Repetir** para $b = 1, \dots, B$:
 - 4: **Obtener un bloque** $B_{j,k+1}^{*b}$ a partir de un sorteo con probabilidades \widehat{p}_i sobre el conjunto $\{B_{1,k+1}, \dots, B_{n-k,k+1}\}$.
 - 5: **Definir la réplica bootstrap** X_{n+k}^{*b} como la última observación de $B_{j,k+1}^{*b}$.
 - 6: **Construir el intervalo de predicción bootstrap** mediante los cuantiles empíricos de $\{X_{n+k}^{*b}\}_{b=1}^B$.
-

Observación 2.26. La consistencia del mecanismo bootstrap propuesto se demuestra bajo ciertas suposiciones como el carácter α -mixing de la serie de tiempo $\{X_t\}$ o condiciones de regularidad acerca de la función tipo núcleo. Estas condiciones, junto con la referencia al resultado teórico que garantiza la consistencia pueden consultarse en las páginas 113 y 114 de [Cao \(1999\)](#).

2.4. Bootstrap en modelos temporales semiparamétricos

En este apartado se llevará a cabo una introducción al procedimiento de remuestreo bajo un modelo semiparamétrico para series de tiempo presentado en [García-Jurado et al. \(1995\)](#), el cual puede verse como una generalización de los modelos Box-Jenkins, ya que consiste en una parte no paramétrica

que estima la tendencia, una predicción bajo un modelo Box-Jenkins de la serie de residuos y algunas técnicas bootstrap para construir intervalos de predicción. En cuanto a la estructura de esta sección, primero se comentará la situación general en la cual se enmarca el modelo propuesto. Después, se verán los enfoques en este tipo de problemas, por un lado, de la metodología Box-Jenkins y, por otro, de un enfoque no paramétrico para, a continuación, extender este último enfoque a un contexto semiparamétrico. Por último, se desarrollará la metodología bootstrap para construir intervalos de predicción en modelos semiparamétricos cuando el término de error sigue un modelo $ARI(p, d)$.

2.4.1. Contexto clásico: enfoque paramétrico y no paramétrico

Sea $\{(\mathbf{Z}_l, Y_l)\}_{l \in \mathbb{Z}}$ una serie de tiempo estrictamente estacionaria, donde \mathbf{Z}_l es una serie de tiempo r -dimensional e Y_l es una serie de tiempo unidimensional. Se pretende estimar mediante una serie de tiempo muestral de longitud n la función dada por

$$m(\mathbf{z}_l^0) = m(F(\cdot | \mathbf{Z}_l = \mathbf{z}_l^0)),$$

donde $F(\cdot | \mathbf{Z}_l = \mathbf{z}_l^0)$ es la distribución condicional de Y_l dado $\mathbf{Z}_l = \mathbf{z}_l^0$. Si se toma m como la media funcional y, además, se considera $Y_l = X_{l+k}$, con $k \geq 1$, y $\mathbf{Z}_l = (X_l, \dots, X_{l-r+1})$, donde $\{X_l\}$ es una serie de tiempo estacionaria, lo que se está buscando estimar es la función de autorregresión de orden k , es decir,

$$m(x_1^0, \dots, x_r^0) = \mathbb{E}[X_{l+k} | (X_l, \dots, X_{l-r+1}) = (x_l^0, \dots, x_{l-r+1}^0)], \quad (2.24)$$

a partir de una muestra de tamaño n dada por X_{t-n+1}, \dots, X_t . Para abordar esta cuestión a lo largo de los años se han utilizado principalmente dos enfoques:

1) Metodología Box-Jenkins.

Sea, por ejemplo, un modelo $AR(p)$, i.e.,

$$X_l = \phi_0 + \phi_1 X_{l-1} + \dots + \phi_p X_{l-p} + a_l, \quad (2.25)$$

donde $\{a_l\}$ es un proceso de ruido blanco independiente del pasado de X_l . Con la notación que ha sido introducida al comienzo de la sección en este caso se tiene que $r = p$ y que la función de autorregresión de orden k viene dada por

$$m(x_1^0, \dots, x_p^0) = \phi_0^{(k)} + \phi_1^{(k)} x_1^0 + \dots + \phi_p^{(k)} x_p^0,$$

donde los coeficientes $\phi_i^{(k)}$ se pueden obtener de manera recursiva a partir de (2.25). Nótese que en el caso de un modelo $ARMA(p, q)$ no existe la función de autorregresión lineal de orden k .

2) Enfoque no paramétrico.

En este caso la función $m(\mathbf{z}_l^0) = \mathbb{E}[Y_l | \mathbf{Z}_l = \mathbf{z}_l^0]$ se estima directamente sin imponer ninguna hipótesis de tipo paramétrico sobre ella. En general, dada una muestra de tamaño n un estimador no paramétrico de m es el dado por

$$\hat{m}_n(\mathbf{z}_l^0) = \sum_{i=1}^n W_{ni}(\mathbf{z}_l^0, (\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)) Y_i,$$

donde los W_{ni} son pesos que pueden venir dados por funciones de tipo núcleo (véase [Yakowitz 1985](#)), k_n vecinos más próximos, etc.

2.4.2. Formulación del modelo semiparamétrico

Lo que se pretende en esta sección es generalizar este segundo enfoque hacia un modelo semiparamétrico, ya que en la experiencia personal de los autores del artículo con datos reales relacionados con la concentración de SO_2 los residuos de la fase de predicción no paramétrica habitualmente no resultaban ser ruido blanco. En concreto, con la notación anterior, se propone el modelo dado por

$$Y_t = m(\mathbf{Z}_t) + e_t, \quad (2.26)$$

donde $\{e_t\}$ sigue un modelo $\text{ARMA}(p, q)$ y es independiente de $\{\mathbf{Z}_t\}$.

El objetivo que se perseguirá a lo largo de la sección será hacer predicciones sobre la serie $\{Y_t\}$ en el instante t tras haber observado dicha serie hasta el instante $t - k$ ³⁴ y la serie $\{\mathbf{Z}_t\}$ hasta el instante t . En concreto, la predicción en el instante t a través de la muestra de tamaño n dada por $(\mathbf{Z}_{t-k-n+1}, Y_{t-k-n+1}), \dots, (\mathbf{Z}_{t-k}, Y_{t-k})$ bajo el modelo (2.26) vendrá dada por

$$\hat{Y}_t = \hat{m}_n(\mathbf{Z}_t) + \hat{e}_t,$$

donde \hat{m}_n es la estimación no paramétrica mencionada en el apartado anterior (por ejemplo, con pesos de tipo núcleo) y \hat{e}_t es la predicción a horizonte k realizada a partir del modelo ARMA estimado de la serie de residuos $\hat{e}_t = Y_t - \hat{m}_n(\mathbf{Z}_t)$.

2.4.3. Comparación de los intervalos de predicción

Intervalos usando la metodología Box-Jenkins

En primer lugar, cabe destacar que como los e_t no son observables parece natural trabajar con la predicción \hat{e}_t , construida a partir de los residuos del modelo ARMA estimado, es decir,

$$\hat{e}_{t-(n+k)+1} = Y_{t-(n+k)+1} - \hat{m}_n(\mathbf{Z}_{t-(n+k)+1}), \dots, \hat{e}_{t-k} = Y_{t-k} - \hat{m}_n(\mathbf{Z}_{t-k}).$$

Cabe destacar que los parámetros de este modelo ARMA se estiman de manera consistente bajo ciertas condiciones (véanse las páginas 308 y 309 de [García-Jurado et al. 1995](#)). Además, teniendo en cuenta la metodología clásica de los modelos Box-Jenkins, puede construirse un intervalo de predicción asintótico de nivel α para Y_t , obteniéndose

$$\left(\hat{m}_n(\mathbf{Z}_t) + \hat{e}_t - z_{\alpha/2} \left(\hat{\sigma}^2 \sum_{j=0}^{k-1} \hat{\pi}_j^2 \right)^{1/2}, \hat{m}_n(\mathbf{Z}_t) + \hat{e}_t + z_{\alpha/2} \left(\hat{\sigma}^2 \sum_{j=0}^{k-1} \hat{\pi}_j^2 \right)^{1/2} \right), \quad (2.27)$$

donde $z_{\alpha/2}$ es el cuantil $1 - \alpha/2$ de la normal estándar, $\hat{\sigma}^2$ es la estimación usual de la varianza asociada al proceso de ruido blanco del modelo ARMA para e_t y los $\hat{\pi}_j$ son los coeficientes estimados de los polinomios π_j obtenidos de la relación

$$\pi(B) = \frac{\theta(B)}{\phi(B)(1-B)^d},$$

donde los coeficientes de ϕ y de θ están estimados de manera consistente a partir de la parte ARMA $\{\hat{e}_t\}$. Además, se incluye el factor $(1-B)^d$ para albergar el caso más general en el cual el modelo que se supone sobre $\{e_t\}$ es un $\text{ARIMA}(p, d, q)$.

³⁴ Por respeto a los autores se ha mantenido la notación empleada por ellos para referirse al instante de predicción. Conviene recordar que anteriormente lo que se ha hecho es observar la serie hasta el instante t y predecir a horizonte k , es decir, en el instante $t + k$. Ahora se predecirá al mismo horizonte pero al observar solamente hasta el instante $t - k$, el instante de predicción será t .

Intervalos usando la metodología bootstrap

Conviene destacar que el intervalo de predicción que se ha presentado en el apartado anterior está basado en asumir que los $\{a_l\}$ que aparecen en el modelo ARIMA son un proceso de ruido blanco. Sin embargo, en diversas ocasiones esta hipótesis no es cierta y en tales casos el comportamiento de dicho intervalo es muy pobre. Por ese motivo, se propone una alternativa que emplea técnicas bootstrap.

En primer lugar, cabe mencionar que, en los términos del modelo general semiparamétrico propuesto bajo forma (2.26), si se considera $m = 0$ (ausencia de componente no paramétrica) y se asume que la serie $\{e_l\}$ sigue un modelo autorregresivo de orden p , la construcción de intervalos de predicción mediante técnicas bootstrap es la ya vista en el apartado anterior mediante el método de [Thombs y Schucany \(1990\)](#)³⁵.

Así pues, lo que se propone en el artículo [García-Jurado et al. \(1995\)](#) es una adaptación o generalización del mecanismo propuesto en [Thombs y Schucany \(1990\)](#) que puede ser empleada en modelos $ARI(p, d)$. Supóngase que la serie $\{e_t\}$ en el modelo (2.26) sigue un modelo $ARI(p, d)$, esto es, un modelo que tras aplicarle d diferencias regulares se convierte en un proceso $AR(p)$. En consecuencia, se tiene que

$$\phi(B)(1-B)^d e_l = a_l,$$

siendo $\{a_l\}$ un proceso de ruido blanco. En otras palabras, la serie $\tilde{e}_l = \nabla^d e_l$ sigue un modelo $AR(p)$.

Aplicando el método bootstrap propuesto en [Thombs y Schucany \(1990\)](#) (véase el Algoritmo 11) pueden obtenerse las remuestras bootstrap siguientes³⁶:

$$\tilde{e}_{t-k-n+d+1}^*, \dots, \tilde{e}_{t-k-p}^*, \tilde{e}_{t-k-p+1}, \dots, \tilde{e}_{t-k}.$$

De este modo, con cada una de estas remuestras pueden construirse predicciones en los instantes $t-k+1, t-k+2, \dots, t$. En el Algoritmo 12 puede verse tal procedimiento y, en definitiva, las series bootstrap que se obtienen serían

$$\tilde{e}_{t-k-n+d+1}^*, \dots, \tilde{e}_{t-k-p}^*, \tilde{e}_{t-k-p+1}, \dots, \tilde{e}_{t-k}, \tilde{e}_{t-k+1}^*, \dots, \tilde{e}_t^*.$$

Sin embargo, no debe olvidarse que el objetivo que se persigue es obtener remuestras y predicciones bootstrap de la serie $\{e_l\}$ y lo que se acaba de obtener es referente a $\{\tilde{e}_l\}$. No obstante, puede utilizarse la relación que hay entre estas dos series para obtener las series buscadas.

En concreto, mediante el método de inducción en d puede probarse que el sistema lineal de $k+d$ ecuaciones con $k+d$ incógnitas dado por³⁷

$$\begin{aligned} e_{j-k}^* &= e_{j-k}, & j &= t-d+1, \dots, t, \\ \nabla^d e_{t-k+i}^* &= \tilde{e}_{t-k+i}^*, & i &= 1, \dots, k, \end{aligned}$$

tiene una única solución en las variables e_j^* con $j = t-k-d+1, \dots, t$. Como consecuencia de esto se tiene que e_t^* puede ser expresado en términos de $e_{t-k-d+1}, \dots, e_{t-k}, \tilde{e}_{t-k+1}^*, \dots, \tilde{e}_t^*$.

³⁵ Más concretamente, lo que se ha visto es un mecanismo bootstrap para aproximar la distribución condicionada de e_t dados $e_{t-k}, e_{t-k-1}, \dots, e_{t-(n+k)+1}$.

³⁶ Nótese que debido a las d diferencias regulares el primer índice de la remuestra debe ser $t-k-n+d+1$. Así, por un lado, los p valores finales de la remuestra se fijan tomándolos iguales a los de la serie \tilde{e}_t , mientras que los $n-p-d$ valores restantes son generados a partir de la distribución empírica de los residuos corregidos correspondientes. Con las remuestras bootstrap así obtenidas luego se construyen las predicciones en $t-k+1, t-k+2, \dots, t$.

³⁷ La serie $\{e_l\}$ se observa hasta el instante $t-k$ y en las primeras d ecuaciones se toman los valores correspondientes de dicha serie como los d valores fijos que se necesitan debido a las diferenciaciones regulares. En cuanto a las otras k , son las propias expresiones de las diferenciaciones desde el instante $t-k+1$ hasta el último instante en que se ha obtenido la remuestra de $\{\tilde{e}_l\}$, es decir, t .

Empleando el procedimiento de remuestreo anterior varias veces puede obtenerse la siguiente aproximación del intervalo de predicción a k retardos para e_t

$$\left(q_t^{*(\alpha/2)}, q_t^{*(1-\alpha/2)} \right), \quad (2.28)$$

donde $q_t^{*(\alpha/2)}$ y $q_t^{*(1-\alpha/2)}$ son los cuantiles $\alpha/2$ y $1-\alpha/2$ de la distribución bootstrap de e_t^* . Así, bajo el modelo semiparamétrico dado por (2.26), un intervalo de predicción para Y_t empleando el intervalo de predicción bootstrap puede ser dado por

$$\left(\widehat{m}_n(\mathbf{Z}_t) + \widehat{q}_t^{*(\alpha/2)}, \widehat{m}_n(\mathbf{Z}_t) + \widehat{q}_t^{*(1-\alpha/2)} \right), \quad (2.29)$$

donde los cuantiles bootstrap \widehat{q}^* se obtienen de la componente ARMA, es decir, $\{\widehat{e}_l\}$.

Algoritmo 16 Intervalo de predicción bootstrap en modelo $ARI(p, d)$, [García-Jurado et al. 1995](#)

- 1: **Calcular la serie de diferencias** de la parte paramétrica del modelo: $\widetilde{e}_l = \nabla^d e_l$.
- 2: **Repetir** para $b = 1, \dots, B$:
- 3: **Aplicar el Algoritmo 12** a la serie \widetilde{e}_l para obtener la serie bootstrap:

$$\widetilde{e}_{t-k-n+d+1}^{*b}, \dots, \widetilde{e}_{t-k-p}^{*b}, \widetilde{e}_{t-k-p+1}, \dots, \widetilde{e}_{t-k}, \widetilde{e}_{t-k+1}^{*b}, \dots, \widetilde{e}_t^{*b}.$$

- 4: **Construir la predicción bootstrap** e_t^{*b} a partir de la serie del Paso 3:

$$\begin{aligned} e_{j-k}^{*b} &= e_{j-k}, & j &= t-d+1, \dots, t, \\ \nabla^d e_{t-k+i}^{*b} &= \widetilde{e}_{t-k+i}^{*b}, & i &= 1, \dots, k. \end{aligned}$$

- 5: **Obtener los límites del intervalo de predicción** a partir de los correspondientes cuantiles empíricos de $e_t^{*1}, \dots, e_t^{*B}$.
-

2.4.4. Validez del modelo

Así como la validez del intervalo (2.28) se demostrará por medio de un teorema, que puede verse como una extensión del Teorema 2.7, en el caso del intervalo (2.29), en el artículo [García-Jurado et al. \(1995\)](#) se ve que su comportamiento en el estudio de simulación que allí se presenta es competitivo con respecto a la alternativa dada por el intervalo (2.27) pero no hay resultados teóricos³⁸ que garanticen la consistencia del método. De este modo, no se han encontrado evidencias en contra de la consistencia del método cuando $m \neq 0$ pero es importante mencionar que esto por sí solo no constituye ningún tipo de demostración, sino que puede verse como una especie de incentivo para tratar de demostrar la consistencia de manera teórica.

Teorema 2.13 ([García-Jurado et al. 1995](#), pp. 309-310). *Considérese el modelo sobre la serie $\{e_l\}$ dado por $\phi(B)(1-B)^d e_l = a_l$, con $\mathbb{E}[a_l] = 0$ y $\mathbb{E}[|a_l|^\alpha] < \infty$ para algún $\alpha > 2$. Entonces se tiene que de forma casi segura $e_t^* \xrightarrow{d} e_t$, es decir, de forma casi segura se tiene que*

$$\lim_{n \rightarrow \infty} (\mathbb{P}^* \{e_t^* \leq z \mid e_{t-k}, \dots, e_{t-k-n+1}\} - \mathbb{P} \{e_t \leq z \mid e_{t-k}, \dots, e_{t-k-n+1}\}) = 0.$$

³⁸ Al menos que el autor del presente trabajo haya podido encontrar.

2.5. Formulación de modelos en el contexto medioambiental

En esta sección se va a presentar la aplicación de diferentes metodologías en el contexto del problema medioambiental que fue introducido al comienzo del trabajo. Para ello se van a tener cuenta las observaciones ya realizadas en la Sección 1.5, en la cual se introdujo el concepto de matriz histórica y se comentaron los malos resultados de la aplicación de modelos paramétricos de tipo Box-Jenkins.

2.5.1. Modelo no paramétrico

En el contexto de los modelos de predicción no paramétricos la única hipótesis que se asume sobre la serie de tiempo $\{X_t\}$ es su carácter estacionario. Además, tal y como se mencionó en la Sección 2.4, en estos casos la estimación de la función de regresión es una función no lineal de las variables predictoras que en su forma más general engloba los caso del estimador de tipo Nadaraya-Watson, el estimador basado en splines o el basado en vecinos más próximos.

En la Sección 3 de Prada-Sánchez y Febrero-Bande (1997) se ilustra el caso concreto del estimador no paramétrico empleando la aproximación tipo núcleo de tipo Nadaraya-Watson, la cual en el caso general de la estimación de la función de regresión de una variable aleatoria Y sobre una variable r -dimensional \mathbf{X} dada una muestra de n observaciones independientes $\{(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^n, y^n)\}$ es la dada por

$$\hat{m}(\mathbf{x}) = \sum_{i=1}^n \frac{K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}^i)}{\sum_{j=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}^j)} y^i, \quad \mathbf{x} \in \mathbb{R}^r, \quad (2.30)$$

donde $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2} \mathbf{x})$, siendo el núcleo $K(\cdot)$ una función de densidad r -dimensional y \mathbf{H} una matriz $r \times r$ simétrica y definida positiva conocida como matriz de suavizado. Más concretamente, en el citado artículo emplean como función núcleo una densidad gaussiana y para la selección de la matriz de suavizado emplean un criterio basado en validación cruzada³⁹.

Adaptando la situación al contexto de la estimación de la función de autorregresión de orden k definida en (2.24) se tiene que, en primer lugar, la muestra de tamaño n , cuya obtención se hará a través de la llamada matriz histórica, es la dada por

$$\{((x_1^1, \dots, x_r^1), x_{r+k}^1), \dots, ((x_1^n, \dots, x_r^n), x_{r+k}^n)\}.$$

Por otro lado, el estimador (2.30) ahora tendrá la expresión siguiente:

$$\hat{m}(\mathbf{x}) = \sum_{i=1}^n \frac{K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}^i)}{\sum_{j=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}^j)} x_{r+k}^i.$$

Cabe mencionar, además, que en la Sección 3.2 de Prada-Sánchez y Febrero-Bande (1997) se lleva a cabo un estudio para la selección del mínimo número de variables predictoras (i.e., determinar el valor de r) que deberían emplearse para una predicción adecuada de la serie de SO_2 , llegando a la conclusión de que con los resultados que obtienen parece razonable emplear solamente 2 variables.

Con todas las consideraciones que han sido comentadas los investigadores llevaron a cabo una serie de predicciones a partir de valores pasados de la serie y compararon los resultados obtenidos con

³⁹ En el artículo Härdle y Vieu (1992) puede verse que esta elección de la matriz de suavizado en el contexto de series de tiempo minimiza el error cuadrático medio de predicción y, aún más, en el caso de observaciones independientes es asintóticamente óptima.

los verdaderos valores de la misma. No obstante, al igual que había ocurrido en el caso del modelo paramétrico que fue comentado en la Sección 1.5, los resultados que se obtuvieron no fueron satisfactorios, en el sentido de que las predicciones no se ajustaban razonablemente bien a los valores reales de la serie.

2.5.2. Modelo semiparamétrico

En [García-Jurado et al. \(1995\)](#) se consideró un modelo semiparamétrico del estilo del indicado en (2.26), donde con la notación allí empleada y sustituyendo los índices l por t , ahora se tendrá que

$$Y_l \longrightarrow X_{t+6}, \quad \mathbf{Z}_l \longrightarrow (X_t, X_{t-1}), \quad m(\mathbf{Z}_l) = \mathbb{E}[Y_l | \mathbf{Z}_l] \longrightarrow \mathbb{E}[X_{t+6} | X_t, X_{t-1}], \quad e_l \longrightarrow e_t,$$

donde $\{e_t\}$ sigue un modelo ARIMA(p, d, q) y es independiente de $\{(X_t, X_{t-1})\}$. Así pues, se tendrá que la ecuación de este modelo semiparamétrico será la dada por

$$X_{t+6} = \mathbb{E}[X_{t+6} | X_t, X_{t-1}] + e_{t+6}. \quad (2.31)$$

Para obtener una predicción a un horizonte de media hora, el primer paso que llevaron a cabo en [García-Jurado et al. \(1995\)](#) consistió en comenzar estimando la parte no paramétrica del modelo (2.31), esto es, $\mathbb{E}[X_{t+6} | X_t, X_{t-1}]$, mediante un estimador de tipo Nadaraya-Watson empleando un núcleo gaussiano y un criterio de selección de ventana basado en validación cruzada. Para tal estimación, en cada instante de tiempo se consideró como serie activa la serie de tiempo muestral formada por las mediciones de las últimas 6 horas (i.e., 72 observaciones).

Estimada $\mathbb{E}[X_{t+6} | X_t, X_{t-1}]$ no paraméricamente con la información provista por una matriz histórica y denotando a tal aproximación como $\hat{m}(X_t, X_{t-1})$, el siguiente paso consistió en construir la serie de tiempo de los residuos relativos a las últimas 6 horas, esto es, si para cada i se define $\hat{e}_i = X_i - \hat{\mathbb{E}}[X_i | X_{i-6}, X_{i-7}]$, la serie dada por

$$\hat{e}_{t-64}, \hat{e}_{t-63}, \dots, \hat{e}_t.$$

A continuación, empleando la metodología Box-Jenkins descrita en la Sección 1.1, se ajusta un modelo ARIMA adecuado a la serie de residuos anterior y con dicho modelo se calcula la predicción \hat{e}_{t+6} . Por último, con la estimación no paramétrica y la predicción bajo el modelo ARIMA calculadas, se obtiene la predicción puntual bajo el modelo semiparamétrico (2.31), la cual viene dada por

$$\hat{X}_{t+6} = \hat{m}(X_t, X_{t-1}) + \hat{e}_{t+6}.$$

Para finalizar esta sección, cabe mencionar que, si bien en el presente trabajo no se incluye, en la Sección 4 de [García-Jurado et al. \(1995\)](#) se lleva a cabo un estudio comparativo pormenorizado entre las predicciones puntuales empleando directamente la metodología clásica de Box-Jenkins, un modelo no paramétrico y el modelo semiparamétrico que se ha comentado anteriormente para el caso de datos reales de concentración de SO₂ en la central térmica de As Pontes. Además, con esos mismos datos también se detallan los resultados obtenidos al construir intervalos de predicción mediante un modelo Box-Jenkins y mediante bootstrap, de la manera en que se ha especificado su construcción a lo largo de las anteriores secciones.

Observación 2.27. Para un estudio de simulación comparativo más detallado de las técnicas paramétricas, no paramétricas y semiparamétricas que han sido presentadas en el contexto de la predicción puede consultarse [Cao \(1994\)](#).

Capítulo 3

Técnicas de remuestreo en otros modelos empleados

En este capítulo se van a presentar otros modelos que fueron considerados a lo largo de los años para resolver el problema medioambiental que ha sido considerado durante este trabajo. No obstante, a diferencia de lo ocurrido con los modelos Box-Jenkins y la metodología bootstrap, no se entrará en tanto detalle en los aspectos teóricos, si bien se proporcionarán al lector diversas referencias para profundizar en ello.

La situación general en la que se van a desarrollar todos los modelos es la misma que la presentada en la Sección 2.4, es decir, dada una muestra de una serie de tiempo $\{(\mathbf{Z}_l, Y_l)\}_{l=1}^n$, donde \mathbf{Z}_l es una serie de tiempo r -dimensional e Y_l es unidimensional, se busca predecir el valor de Y_{n+h} , con $h > 0$, en función de \mathbf{Z}_{n+h} a partir de dicha muestra. Por ejemplo, podría predecirse mediante $\hat{Y}_{n+h} = \hat{m}(\mathbf{Z}_{n+h})$, donde \hat{m} es una estimación, basada en la muestra, de la función de regresión dada por $m(\mathbf{Z}_l) = \mathbb{E}[Y_l | \mathbf{Z}_l]$, con $l \in \mathbb{Z}$.

Una situación muy común, y en la cual se van a basar los modelos que van a ser presentados, es aquella en la que $\mathbf{Z}_l = (X_l, \dots, X_{l-r+1})$ e $Y_l = X_{l+h}$, siendo $\{X_l\}$ una serie de tiempo unidimensional, en cuyo caso se buscaría estimar la media condicional $\mathbb{E}[X_{n+h} | X_n, \dots, X_{n-r+1}]$ a partir de una muestra X_{1-r+1}, \dots, X_n .

3.1. Modelo parcialmente lineal

Hasta algunos meses antes de la publicación de Prada-Sánchez et al. (2000) se había empleado el modelo semiparamétrico que ha sido tratado en la Sección 2.4.2, obteniendo un rendimiento satisfactorio. No obstante, cambios operacionales en la central provocaron que fuese necesario extender el horizonte de predicción a 1 hora para que los operarios tuviesen tiempo suficiente para tomar las acciones oportunas y reducir los niveles de SO_2 en caso de producirse un episodio y, en consecuencia, las predicciones comenzaron a dejar de comportarse de manera adecuada. El motivo era que se había aumentado considerablemente el horizonte de predicción ($h = 12$) pero en el modelo semiparamétrico solamente se consideraba información del pasado de la serie. Por este motivo, los investigadores decidieron incluir en el modelo información adicional que viniese dada por variables meteorológicas o de emisión, añadiendo así una combinación lineal de variables exógenas explicativas, cuyos coeficientes sería necesario estimar también en el proceso de predicción.

De este modo, en lugar de la serie $\{(\mathbf{Z}_l, Y_l)\}_{l \in \mathbb{Z}}$, se pasó a considerar la serie $\{(\mathbf{V}_l, \mathbf{Z}_l, Y_l)\}_{l \in \mathbb{Z}}$, donde \mathbf{V}_l es una serie de tiempo q -dimensional de variables exógenas explicativas, y fue diseñado el modelo parcialmente lineal dado por

$$Y_l = \mathbf{V}_l' \beta + m(\mathbf{Z}_l) + \varepsilon_l, \quad l \in \mathbb{Z} \quad (3.1)$$

donde $\beta \in \mathbb{R}^q$ es desconocido y los ε_l son términos de error de media 0. De esta manera, el modelo consiste en un término «autoexplicativo» y una combinación lineal de variables exógenas.

Observación 3.1. Las variables exógenas se incluyen como un término lineal en el modelo (3.1) para poder interpretar los términos del coeficiente β y también por su simplicidad computacional.

En cuanto a la selección de las variables exógenas, en la sección 3 del citado artículo se definen hasta 13 variables aunque finalmente, por cuestiones de fiabilidad, estabilidad y redundancia, las que se incluyeron fueron las siguientes: $\Delta_{10}^{80} T_l$, la diferencia de temperatura entre 80 y 10 m sobre el nivel del mar, en $^{\circ}C$; T_l^1 , la temperatura a 10 m sobre el nivel del mar, en $^{\circ}C$; V_l^8 , la velocidad del viento a 80 m sobre el nivel del mar, en m/s ; R_l , la radiación solar, en langley; H_l , la humedad, en %; y E_l , la emisión de SO_2 de la central térmica, en $\mu g/m^3$. Además, con el fin de modelizar tanto el efecto instantáneo como el retardado de estas variables, y dado que la predicción se realiza a horizonte $h = 12$, se tuvieron en cuenta sus valores en el instante l y en el $l - 12$. Así, los 2 siguientes vectores de variables exógenas fueron considerados en los diferentes modelos:

$$\begin{aligned} \mathbf{V}_l^1 &= (\Delta_{10}^{80} T_{l-12}, \Delta_{10}^{80} T_l, T_{l-12}^1, T_l^1, V_{l-12}^8, V_l^8, R_{l-12}, R_l, H_{l-12}, H_l, E_{l-12}, E_l), \\ \mathbf{V}_l^2 &= (\Delta_{10}^{80} T_l - \Delta_{10}^{80} T_{l-12}, T_l^1 - T_{l-12}^1, V_l^8 - V_{l-12}^8, R_l - R_{l-12}, H_l - H_{l-12}, E_l - E_{l-12}), \end{aligned}$$

donde en \mathbf{V}_l^2 se está considerando únicamente su efecto en términos de sus incrementos horarios.

En lo relativo a las muestras, se emplearon 1000 filas de una matriz histórica creada como fue indicado en la Sección 1.5.1 y considerando datos de la forma $(\mathbf{V}_l^i, \mathbf{Z}_l, Y_l)$, donde \mathbf{V}_l^i es uno de los vectores descritos anteriormente, $\mathbf{Z}_l = (X_l, X_{l-3})$ e $Y_l = X_{l+12}$, siendo $\{X_l\}$ la serie de medias bihorarias de la concentración de SO_2 .

Observación 3.2. El vector \mathbf{Z}_l posee dos componentes para que el predictor pueda distinguir entre la fase ascendente de un episodio de contaminación (i.e., $X_{l-3} < X_l$) y la fase descendente ($X_{l-3} > X_l$). No se consideran valores adyacentes de la serie, es decir, separados 5 minutos, debido a que $h = 12$ y la experiencia de los investigadores les llevó a que el intervalo óptimo entre los valores era de 15 minutos.

En el artículo mencionado llevan a cabo un estudio comparativo de 4 predictores diferentes¹ en un caso simulado y también con datos reales de un episodio de contaminación ocurrido el 23 de agosto de 1995, obteniendo, por un lado, que se obtienen mejores predicciones empleando \mathbf{V}^2 que \mathbf{V}_1 y por otro, que, aunque algunos obtuvieron buenos resultados durante el pico del episodio, predijeron el inicio de este más tarde de lo que había sucedido en realidad. Esto, suponiendo que el modelo (3.1) estaba bien especificado, sugería que las variables exógenas disponibles por aquel entonces estaban poco correladas con \mathbf{Z}_l y la mayor parte de la información para la predicción venía dada por el pasado de la serie. Además, las variables que cabía esperar que mejorasen la predicción solamente podían ser medidas esporádicamente con los recursos de los que se disponía. En consecuencia, tomaron la decisión de emplear uno de los 2 predictores que se van a introducir para cerrar este apartado.

Los predictores propuestos surgían del modelo de regresión lineal múltiple² que se obtiene al tomar $m(\mathbf{Z}_l) = \mathbf{Z}_l' \gamma$, con $\gamma \in \mathbb{R}^r$. Considerando una muestra de tamaño n este puede ser escrito como

$$\mathbf{Y} = \mathbf{V}\beta + \mathbf{Z}\gamma + \boldsymbol{\varepsilon}, \quad \gamma \in \mathbb{R}^r, \beta \in \mathbb{R}^q, \quad (3.2)$$

¹ Son casos particulares de (3.1) empleando diferentes estimaciones de β y $m(\cdot)$.

² Este modelo también fue considerado pero no ofreció buenos resultados en términos predictivos.

donde $\mathbf{Y}, \boldsymbol{\varepsilon} \in \mathbb{R}^k$ y \mathbf{V} y \mathbf{Z} son matrices $n \times q$ y $n \times r$ cuyas filas vienen dadas por \mathbf{V}'_i y \mathbf{Z}'_i , respectivamente. Los estimadores por mínimos cuadrados en 2 pasos de β y γ en el modelo (3.2) son

$$\begin{aligned}\widehat{\beta}_0 &= (\mathbf{V}'(\mathbf{I} - \mathbf{P}_z)\mathbf{V})^{-1} \mathbf{V}'(\mathbf{I} - \mathbf{P}_z)\mathbf{Y}, \\ \mathbf{Z}'\widehat{\gamma}_0 &= \mathbf{P}_z(\mathbf{Y} - \mathbf{V}\widehat{\beta}_0),\end{aligned}$$

donde \mathbf{I} es la matriz identidad y $\mathbf{P}_z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ es la matriz de proyección en el espacio generado por las columnas de \mathbf{Z} . Así, un predictor natural de Y_{n+h} en función de $(\mathbf{V}_{n+h}, \mathbf{Z}_{n+h})$ a partir de la muestra $\{(\mathbf{V}_i, \mathbf{Z}_i, Y_i)\}_{i=1}^n$ es el dado por $p_n^0(\mathbf{V}_{n+h}, \mathbf{Z}_{n+h}) = \mathbf{V}'_{n+h}\widehat{\beta}_0 + \mathbf{Z}'_{n+h}\widehat{\gamma}_0$.

Una generalización inmediata consiste en reemplazar \mathbf{P}_z por la matriz de estimación no paramétrica \mathbf{Z}_H , con $(\mathbf{Z}_H)_{ij} = w_j^{\mathbf{H},n}(\mathbf{Z}_i, (\mathbf{Z}_1, \dots, \mathbf{Z}_n))$, que son un conjunto de pesos generados por una función núcleo y donde $\mathbf{H} \in \mathbb{R}^{r \times r}$ es una matriz definida positiva conocida como matriz de suavizado. El predictor así obtenido sería³

$$\begin{aligned}p_n^1(\mathbf{V}_{n+h}, \mathbf{Z}_{n+h}) &= \mathbf{V}'_{n+h}\widehat{\beta}_1 + \sum_{j=1}^n w_j^{\mathbf{H},n}(\mathbf{Z}_{n+h}, (\mathbf{Z}_1, \dots, \mathbf{Z}_n)) (Y_j - \mathbf{V}'_j\widehat{\beta}_1), \\ \widehat{\beta}_1 &= (\mathbf{V}'(\mathbf{I} - \mathbf{Z}_H)\mathbf{V})^{-1} \mathbf{V}'(\mathbf{I} - \mathbf{Z}_H)\mathbf{Y}.\end{aligned}$$

El segundo predictor propuesto, surgió en [Speckman \(1988\)](#) de la idea de restar $\mathbb{E}[Y_l | \mathbf{Z}_l]$ a ambos lados en (3.1), obteniéndose así el nuevo modelo de regresión lineal dado por

$$Y_l - \mathbb{E}[Y_l | \mathbf{Z}_l] = (\mathbf{V}_l - \mathbb{E}[\mathbf{V}_l | \mathbf{Z}_l])' \beta + \varepsilon_l.$$

Definiendo $\widetilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{Z}_H)\mathbf{Y}$ y $\widetilde{\mathbf{V}} = (\mathbf{I} - \mathbf{Z}_H)\mathbf{V}$ se llega al predictor dado por

$$\begin{aligned}p_n^2(\mathbf{V}_{n+h}, \mathbf{Z}_{n+h}) &= \mathbf{V}'_{n+h}\widehat{\beta}_2 + \sum_{j=1}^n w_j^{\mathbf{H},n}(\mathbf{Z}_{n+h}, (\mathbf{Z}_1, \dots, \mathbf{Z}_n)) (Y_j - \mathbf{V}'_j\widehat{\beta}_2), \\ \widehat{\beta}_2 &= (\widetilde{\mathbf{V}}'\widetilde{\mathbf{V}})^{-1} \widetilde{\mathbf{V}}'\widetilde{\mathbf{Y}}.\end{aligned}$$

Observación 3.3. Cada predictor puede ser corregido añadiendo el predictor de tipo Box-Jenkins del residuo asociado, al estilo de lo realizado en el modelo semiparamétrico de la Sección 2.4.2.

3.2. Modelo de redes neuronales

Debido a la entrada en vigor de la directriz del Consejo Europeo 1999/30/CE publicada el 22 de abril de 1999 relativa a la limitación de cantidad de SO_2 , NO_x , partículas y plomo en el ambiente, fue necesario introducir algunas modificaciones para ajustar el sistema de predicción desarrollado a las necesidades de la nueva situación. Así, por ejemplo, se pasaba a requerir el control sobre la media horaria de la concentración de SO_2 (en lugar de bihoraria) y los nuevos límites establecidos eran mucho más restrictivos que los marcados por la central con el beneplácito de la Xunta de Galicia. De este modo, el nuevo objetivo pasaba a ser obtener predicciones precisas a horizonte $h = 12$ durante los episodios de contaminación de la serie de tiempo dada por

$$X_t = \frac{1}{12} \sum_{i=0}^{11} \text{SO}_2(t-i),$$

donde $\text{SO}_2(t)$ representa la concentración en el instante t (pentaminutal), medida en $\mu\text{g}/\text{m}^3$. Sin embargo, aunque en una primera aproximación se adaptó el sistema de predicción basado en el modelo

³ Pueden consultarse [Green et al. \(1985\)](#) o [Speckman \(1988\)](#) para profundizar acerca de las propiedades de $\widehat{\beta}_1$.

semiparamétrico presentado en la Sección 2.4.2, debido a la inestabilidad de la serie los resultados obtenidos fueron razonablemente peores que los obtenidos con la serie bihoraria aumentando sensiblemente la variabilidad de las predicciones. Así, motivados por mejorar la respuesta dada por el Sistema de Predicción Estadística de Inmisión (SIPEI), y conscientes de que ya se habían empleado redes neuronales en el ámbito de la contaminación del aire (véase Pérez et al. 2000) e incluso para predecir niveles de SO_2 (véase Boznar et al. 1993), en Fernández-de Castro et al. (2003) fue presentado un modelo basado en ellas para predecir X_{t+6} .

3.2.1. Metodología general

Una red neuronal consiste en un conjunto de nodos repartidos en capas interconectadas que buscan «aprender de los datos» imitando el proceso de aprendizaje de un cerebro humano a partir de diversos casos de ejemplo. En cuanto a su funcionamiento, cada nodo (neurona) tiene una función de activación asociada que le permite procesar la información que llega a él a través de otros nodos, teniendo cada conexión con él un peso asociado. De este modo, este proceso se va trasladando de la primera capa (capa de entrada) a la última (capa de salida), obteniéndose la salida de la red neuronal.

En lo relativo al tipo de red considerado, fue una *Backpropagation Network*, que recibe este nombre debido al procedimiento que emplea para ajustar los pesos. Está formada por una capa de entrada de M nodos, una capa oculta de L , una capa de salida de N y, potencialmente, algún término de tendencia en la capa oculta y en la de salida. Notacionalmente, se va a identificar h con los elementos de la capa oculta y o , con los de la de salida; w_{ji}^h , con el peso asociado al arco que va del nodo j de la capa h al nodo i de la capa anterior; f_j^h , con la función de activación del nodo j de la capa h ; θ_j^h , con la tendencia del nodo j de la capa h ; y o_j^h , con la salida del nodo j de la capa h . Análogamente, se definen w_{kj}^o , f_k^o , θ_k^o y o_k . De este modo, la salida de cada nodo de la capa de salida será

$$o_k = f_k^o \left[\theta_k^o + \sum_{j=1}^L w_{kj}^o f_j^h \left(\theta_j^h + \sum_{i=1}^M w_{ji}^h x_i \right) \right], \quad k = 1, \dots, N.$$

Observación 3.4. Así como los pesos son el resultado del entrenamiento de la red neuronal, su diseño (funciones de activación, número de capas...) debe ser adaptado al contexto al que se aplica.

El proceso de entrenamiento, consiste en introducir un vector de entradas $\mathbf{X} = (x_1, x_2, \dots, x_M)'$ cuyo vector de respuestas asociado $\mathbf{Y} = (y_1, y_2, \dots, y_N)'$ es conocido e ir comparando con este último la salida de la red para, en función del error observado, ir modificando los pesos hasta que este sea menor que una tolerancia prefijada.

Observación 3.5. El proceso de entrenamiento de una red neuronal puede durar mucho tiempo debido al número de veces que deben ajustarse los pesos hasta no superar el umbral de error. Además, es muy importante disponer de datos de entrenamiento adecuados a los intereses buscados, ya que la red neuronal va a imitar los patrones incluidos en ellos (sean estos buenos para esos intereses o no).

3.2.2. Construcción en el problema medioambiental

El diseño de la red que ha sido considerado para la predicción puntual de X_{t+6} consiste en una capa de entrada, una capa oculta con función de activación logística y una capa de salida con función de activación igual a la identidad y con un número de nodos igual a la dimensión de la respuesta que se quiere obtener, i.e., $N = 1$. Además, debido a lo comentado en la Sección 3.1 sobre la baja calidad de las variables meteorológicas, como entrada de la red se consideraron solamente valores pasados de la serie. Más concretamente, se tomó $\mathbf{X} = (X_{t-3}, X_t)'$, es decir, $M = 2$, quedando por lo tanto un diseño

como el indicado en la Figura 3.1. Por otro lado, los conjuntos de entrenamiento eran vectores de la forma $(X_{t-3}, X_t, X_{t+6})'$, escogidos de una matriz histórica construida según lo descrito en la Sección 1.5.1 a partir de datos reales de la central del año 1999. Así, el predictor construido mediante esta red neuronal sería el dado por

$$\hat{X}_{t+6} = o_1 = \sum_{j=1}^L w_{1j}^o f_j^h(\theta_j^h + w_{j1}^h X_{t-3} + w_{j2}^h X_t), \quad f_j^h(z) = \frac{1}{1 + e^{-z}}, \quad j = 1, \dots, L,$$

donde tanto los pesos como las tendencias son determinados durante el proceso de entrenamiento, el cual está detallado en el apéndice de Fernández-de Castro et al. (2003). En relación a esto, fueron entrenadas diferentes redes neuronales considerando diferentes nodos en la capa oculta (valores de L entre 40 y 60) y empleando 1169 vectores de entrenamiento para ajustar un total de $4L$ parámetros⁴. Una vez entrenada, para comprobar su buen funcionamiento se obtenía con ella una predicción para un episodio que no había sido incluido en el conjunto de entrenamiento.

Observación 3.6. Dado que había 17 estaciones en las que realizar predicciones cada 5 minutos y no era factible entrenar 17 redes neuronales en ese tiempo, se optó por incluir en el sistema de predicción una red ya entrenada, convirtiéndose en fundamental controlar el buen funcionamiento de esta.

Por último, en el estudio comparativo del artículo puede verse que la red neuronal captura mejor la tendencia en cada instante y, en consecuencia, las predicciones son más estables (menor variabilidad) que las del modelo semiparamétrico.

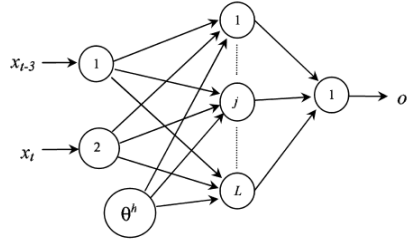


Figura 3.1: Estructura de la red neuronal diseñada para predecir los niveles de SO_2 .

3.3. Modelo con datos funcionales

En esta sección se va a presentar el método que fue propuesto en Fernández-de Castro et al. (2005) para abordar el problema medioambiental. En este artículo se cambia el enfoque con el que se afronta el mismo, considerando como objeto de estudio estadístico datos que, en lugar de números o vectores, ahora pasan a ser curvas. Para ello, la idea principal consiste en ver observaciones discretas en conjunto y considerarlas como una realización de una curva aleatoria. De este modo, el nuevo marco teórico en el que se desarrolla el modelo cambia sustancialmente, estando ahora ligado al contexto de datos funcionales. Algunas referencias que el lector podría consultar para profundizar en este ámbito son Ramsay y Silverman (2005), Horváth y Kokoszka (2012) o Ferraty (2006).

Dado que el horizonte de predicción es de 30 minutos, cada dato funcional muestral será una curva constituida por 6 datos consecutivos y se verá como una realización del proceso estocástico en tiempo continuo \mathcal{X} que modela la concentración de SO_2 . Para predecir valores futuros $\mathcal{X}(u)$, con $u \geq T$, se va

⁴ En general, estas redes neuronales tienen $ML + L + NL$ parámetros en total.

a emplear la información contenida en el infinito número de variables del pasado, $\mathcal{X}(u)$, $u \leq T^5$. En el caso del problema medioambiental, para predecir este proceso se restringe el análisis a la dependencia a retardo 1 porque, como las curvas van a ser observadas en tan solo 6 puntos, parece suficiente para cubrir la evolución del proceso⁶. Además, dado que el objetivo es poder predecir incrementos y descensos repentinos, no se va a suponer ninguna condición de suavidad sobre las curvas. Así pues, se consideran variables aleatorias funcionales con valores en el espacio de Hilbert dado por $H = \mathbb{L}_2[0, 6]^7$ de la forma $\mathcal{X}_n(u) = \mathcal{X}(6n + u)$, con $u \in [0, 6]$ y con $n = 1, 2, \dots$

3.3.1. Redefiniendo la matriz histórica

Debido al cambio de paradigma en los datos, se va a adaptar la idea propuesta en la Sección 1.5.1 para la construcción de la matriz histórica. La semilla va a ser construida mediante los aproximadamente 1500 datos relativos al año 2001 de la forma $(\mathcal{X}_n, \mathcal{X}_{n+1})$, donde cada dato \mathcal{X}_n está formado por 6 medidas consecutivas de la media bihoraria de SO_2 . La determinación de los rangos para la clasificación de los datos se realizó de 2 maneras diferentes.

El primer criterio consistió en llevar a cabo una clasificación ordinaria, al estilo de lo realizado en los otros modelos presentados. Así, se divide la matriz histórica funcional en 10 estratos⁸, cada uno asociado con un intervalo de los niveles de SO_2 del último valor de \mathcal{X}_{n+1} y con una longitud máxima de 200 vectores funcionales. Dado un nuevo vector funcional $(\mathcal{X}_n, \mathcal{X}_{n+1})$, se observa el último valor de la respuesta \mathcal{X}_{n+1} y se sitúa el vector en el estrato correspondiente. La matriz histórica así construida se denominó *matriz histórica de niveles*.

En cuanto al segundo criterio, se dividió la matriz histórica en 5 estratos asociados con una forma de curva diferente: «incrementos», «descensos», «mesetas», «cambios», y «todo lo demás». Para determinar a qué clase pertenece $\mathcal{X}_n = (X_n^1, \dots, X_n^6)^9$ se calculan las 5 diferencias $X_n^2 - X_n^1, \dots, X_n^6 - X_n^5$ y luego se cambia cada valor por «+», cuando este es mayor o igual que 5; «-», cuando es menor o igual que 5; o 0, cuando está entre -5 y 5. Mediante estos símbolos se definen los estratos: cinco «+», para un incremento; cinco «-», para un descenso; cinco «0» para una meseta; y al menos un «+» y un «-» y ningún «0» para un «cambio». Dado un nuevo vector $(\mathcal{X}_n, \mathcal{X}_{n+1})$ se calcula la forma asociada a \mathcal{X}_{n+1} y se sitúa en el estrato correspondiente, formando la denominada *matriz histórica de formas*.

3.3.2. Procedimientos bootstrap

En este apartado final, se van a realizar algunos comentarios acerca de cómo adaptar algunos de los procedimientos bootstrap que han sido incluidos en este trabajo para la construcción de intervalos de predicción bootstrap al contexto de variables aleatorias funcionales dependientes evaluadas sobre un espacio de Hilbert. Sean $(\mathcal{X}_i, \mathcal{X}_{i+1})$, con $i = 1, \dots, N$, los pares de curvas almacenados en la matriz histórica y sean \mathcal{Y}_i las curvas de las cuales se desea predecir \mathcal{Y}_{i+1} .

⁵ Debido a lo comentado en la Sección 3.1, en relación a la poca información que aportan las variables climatológicas disponibles, solamente se consideran valores pasados del proceso.

⁶ En el futuro se espera poder disponer de datos nuevos cada minuto, lo cual permitirá describir las curvas con mayor detalle y sacar mayor partido al enfoque funcional.

⁷ Esto hace referencia a variables aleatorias $\mathcal{X} : (\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \mathbb{L}_2[0, 6]$, donde $\mathcal{X}(\omega) : [0, 6] \rightarrow \mathbb{R}$ es una función (fijado el ω , ya es determinística) que cumple que $\int_0^6 [\mathcal{X}(\omega)(t)]^2 dt < \infty$.

⁸ Cada estrato se define adecuadamente para garantizar el almacenamiento de información en todo el rango de la variable.

⁹ Se tiene que X_n^1 es la evaluación de \mathcal{X}_n en el primero punto de discretización del intervalo.

Bootstrap para predicciones sin estructura específica de dependencia

Bajo el supuesto de que la serie de tiempo es un proceso de Markov, puede adaptarse el procedimiento presentado en la Sección 2.3.5. Así, para realizar predicciones a horizonte $k = 1$ a partir de \mathcal{Y}_m , basta con aplicar el Algoritmo 15 tomando bloques muestrales de tamaño 2, $B_j = (\mathcal{X}_j, \mathcal{X}_{j+1})$, con $j = 1, \dots, N$, a partir de la matriz histórica y construyendo las probabilidades

$$\hat{p}_i = \frac{K(\|\mathcal{X}_i - \mathcal{Y}_m\|/h)}{\sum_{j=1}^N K(\|\mathcal{X}_i - \mathcal{Y}_m\|/h)},$$

que pueden interpretarse como una medida de proximidad entre el dato \mathcal{Y}_m y los datos de la matriz histórica. Luego, como predicción bootstrap \mathcal{Y}_{m+1}^* , se tomaría el segundo elemento del bloque remuestreado. No obstante, así como en el caso de datos escalares era posible calcular mediante técnicas de Monte Carlo una aproximación empírica de la distribución bootstrap de las predicciones y, en base a ella, un intervalo de predicción, en el caso de datos funcionales esta idea no es válida.

Una alternativa es emplear el concepto de *profundidad* que definen en Fraiman y Muniz (2001). En un caso unidimensional, si F es la función de distribución de una variable aleatoria puede definirse una profundidad¹⁰ como $D(x) = \min\{F(x), 1 - F(x^-)\}$, la cual puede verse como una medida de proximidad a la mediana, donde se alcanza el máximo de esta función. Sea, ahora, un conjunto de datos funcionales $\mathcal{X}_1(t), \dots, \mathcal{X}_p(t)$, con $t \in [a, b]$, independientes y siguiendo la misma distribución. Para cada punto $t \in [a, b]$ considérese la profundidad empírica univariante del dato i -ésimo en t con respecto a $\{\mathcal{X}_i(t)\}_{i=1}^p$, es decir, $D_{p,t}(x) = \min\{F_{p,t}(x), 1 - F_{p,t}(x^-)\}$, donde $F_{p,t}(x) = p^{-1} \sum_{j=1}^p \mathbb{I}\{\mathcal{X}_j(t) \leq x\}$. La idea de Fraiman y Muniz (2001) consiste en definir la profundidad funcional del dato i -ésimo como un promedio de la profundidad univariante a lo largo de $[a, b]$, es decir,

$$FMD(\mathcal{X}_i) = \int_a^b D_{p,t}(\mathcal{X}_i(t)) dt,$$

donde se está midiendo la proximidad a la mediana empírica, la curva muestral que maximiza esa profundidad funcional. Con este criterio se ordenan las predicciones bootstrap de manera descendente y se toma la envolvente convexa de las $\lceil p(1 - \alpha) \rceil$ primeras como curvas límite para las predicciones.

Observación 3.7. Aunque no vayan a ser tratados con detalle, merece la pena mencionar los artículos Febrero et al. (2007) y Febrero et al. (2008), ya que en ellos se ilustra la utilidad de técnicas de datos funcionales basadas en la profundidad para la detección de outliers con datos relativos a la concentración de NO_x , los cuales ayudan a percibir niveles relevantes de esta sustancia.

Bootstrap para predicciones en un modelo ARH(1)

En este caso se va a suponer una estructura específica de dependencia sobre las variables aleatorias funcionales. En concreto, se va a suponer un modelo autorregresivo hilbertiano de orden 1, ARH(1):

$$\mathcal{X}_n = \rho(\mathcal{X}_{n-1}) + \mathcal{E}_n,$$

donde $\rho : H \rightarrow H$ es un operador que es necesario estimar¹¹ y donde $\{\mathcal{E}_n\}$ es un proceso hilbertiano de ruido blanco fuerte, es decir, una sucesión de variables aleatorias independientes e idénticamente distribuidas con valores en H y que satisfacen que $\mathbb{E}[\mathcal{E}_n] = 0$ y $0 < \mathbb{E}[\|\mathcal{E}_n\|_H^2] = \sigma^2 < \infty$ con $n \in \mathbb{Z}$.

¹⁰ La que se define se conoce como profundidad de Tukey pero existen muchas otras en la literatura. Para profundizar acerca de este concepto puede consultarse López-Pintado y Romo (2007) o López-Pintado y Romo (2009).

¹¹ Debido a la discretización de los puntos en los que se observan las variables funcionales, los estimadores de ρ van, en realidad, de \mathbb{R}^6 en \mathbb{R}^6 .

Observación 3.8. En cuanto a la estimación de $\rho(\cdot)$, se proponen dos métodos: uno basado en suponer que es un operador lineal acotado sobre H^{12} y otro construido mediante una extensión del estimador de Nadaraya-Watson en regresión al contexto de datos funcionales (véase Besse et al. (2000)).

En este contexto se buscará adaptar la metodología bootstrap de Cao et al. (1997), que fue introducida en la Sección 2.3.3 para construir intervalos de predicción bootstrap en modelos AR(p). La adaptación del esquema de remuestreo ha sido recogida en el Algoritmo 17, donde, una vez generadas las predicciones, se siguen los mismos pasos antes para obtener las curvas límite de las predicciones.

Por último, comentar que en los análisis que fueron realizados en Fernández-de Castro et al. (2005) se pudo ver que los picos de los episodios de contaminación se predijeron con mayor precisión empleando la matriz histórica de formas. Además, se compararon las predicciones obtenidas mediante esta nueva metodología con las del modelo semiparamétrico de la Sección 2.4.2 y las del modelo basado en redes neuronales de la Sección 3.2.2, resultando, por un lado que la aproximación mediante datos funcionales es competitiva y, por otro, que el bootstrap basado en el modelo ARH(1) parece conducir a mejores resultados que el método sin estructura específica de dependencia.

Algoritmo 17 Predicciones bootstrap en modelo ARH(1), Fernández-de Castro et al. 2005

- 1: **Estimar el operador** $\rho(\cdot)$: $\widehat{\rho}(\cdot)$.
- 2: **Construir los residuos hacia adelante**: $\widehat{\mathcal{E}}_i = \mathcal{X}_i - \widehat{\rho}(\mathcal{X}_{i-1})$, con $i = 2, \dots, n + 1$.
- 3: **Definir los residuos centrados**: $\widehat{\mathcal{E}}'_i = \widehat{\mathcal{E}}_i - \bar{\mathcal{E}}$, con $\bar{\mathcal{E}} = n^{-1} \sum_{i=2}^{n+1} \widehat{\mathcal{E}}_i$.
- 4: **Realizar un análisis de componentes principales**¹³ de los $\widehat{\mathcal{E}}'_i$ de la siguiente manera:

$$\widehat{\mathcal{E}}'_i = c_1^i \mathcal{V}_1 + \dots + c_{k_n}^i \mathcal{V}_{k_n}, \quad i = 1, \dots, n.$$

- 5: **Obtener la distribución empírica** de cada coordenada c_l : $\widehat{F}_n^{c_l}$, con $l = 1, \dots, k_n$.
 - 6: **Repetir** para $b = 1, \dots, B$:
 - 7: **Obtener una remuestra de las coordenadas**: c_l^{*b} a partir de $\widehat{F}_n^{c_l}$, con $l = 1, \dots, k_n$.
 - 8: **Construir el residuo bootstrap**: $\widehat{\mathcal{E}}^{*b} = c_1^{*b} \mathcal{V}_1 + \dots + c_{k_n}^{*b} \mathcal{V}_{k_n}$.
 - 9: **Generar una réplica bootstrap** mediante la ecuación del modelo: $\mathcal{Y}_{m+1}^{*b} = \widehat{\rho}(\mathcal{Y}_m) + \widehat{\mathcal{E}}^{*b}$.
-

3.4. Otras aproximaciones

En este apartado se enunciarán otros modelos desarrollados para atacar el problema medioambiental. Para no extender demasiado el trabajo, únicamente se destacarán las ideas principales.

3.4.1. Modelos aditivos generalizados

En Roca-Pardiñas et al. (2004) se propone otro enfoque para abordar el problema medioambiental. Dado que se dice que ocurre un episodio de contaminación cuando la serie de la concentración media bihoraria es mayor que un determinado nivel r , denotando por X_t al valor de dicha serie en el instante t y fijando un horizonte de predicción $h = 12$, el objetivo ahora es predecir $\mathbb{P}\{X_{t+12} > r \mid X_t, X_{t-1}, \dots\}$.

¹² Al algoritmo para la construcción del mismo puede consultarse en la página 215 de Fernández-de Castro et al. (2005).

¹³ El análisis de componentes principales con datos funcionales es una herramienta fundamental que surge de la generalización de un procedimiento análogo en el ámbito multivariante que permite observar la estructura de covarianzas de los datos de una manera «más informativa» que una examinación directa de la misma. Para profundizar sobre este método puede consultarse el Capítulo 8 de Ramsay y Silverman (2005).

No obstante, las predicciones de esto realizadas con el modelo semiparamétrico o con el parcialmente lineal no está garantizado que estén en el intervalo $[0, 1]$, como corresponde a una probabilidad. Es por ello que en el citado artículo consideran otros modelos: un modelo lineal generalizado (*GLM*, [McCullagh y Nelder 1989](#)), un modelo aditivo generalizado (*GAM*, [Hastie y Tibshirani 1990](#)), un modelo *single index* (*SIM*, [Horowitz 1998](#)) y un modelo GAM con función de link desconocida¹⁴ ([Horowitz 2001](#)).

Siguiendo el enfoque anterior, en [Roca-Pardiñas et al. \(2005\)](#) se propone un GAM con términos de interacción de segundo orden y se desarrolla un contraste de significación para los mismos que es calibrado mediante técnicas bootstrap que permiten aproximar la distribución del estadístico de test.

Por último, comentar que más recientemente, en [Sestelo et al. \(2014\)](#) se busca considerar también la influencia de algunas variables meteorológicas en la predicción de la probabilidad condicionada y para ello proponen un método basado en emplear un modelo aditivo y un algoritmo de selección de variables. No obstante, las conclusiones a las que llegan son, por un lado, que basta considerar solamente dos términos del pasado de la serie y, por otro, que la inclusión en el modelo de las variables meteorológicas consideradas no resulta significativa.

3.4.2. Modelos de regresión cuantil

En [Conde-Amboage et al. \(2017\)](#) el problema medioambiental fue abordado empleando técnicas de regresión cuantil¹⁵ para construir intervalos de predicción de NO_x . Estos métodos son más robustos en comparación con los modelos de regresión en media ante la presencia de datos atípicos y no requieren de suposiciones tan estrictas como la homocedasticidad y la normalidad de los errores.

En cuanto a los resultados observados en el estudio que llevan a cabo, se observa que la heterocedasticidad y la distribución no gaussiana de los errores es bastante habitual, lo cual se desvía de las suposiciones clásicas de los modelos de regresión en media y por ello explica el mejor desempeño de los métodos de regresión cuantil, obteniéndose errores de predicción más pequeños que los obtenidos con la regresión en media. Además, el método propuesto para construir intervalos de predicción basado en la estimación cuantil y la aproximación bootstrap de los errores de predicción tuvo un comportamiento significativamente mejor que los métodos clásicos en ese tipo de regresión.

Por último, cabe mencionar también el artículo [Martínez-Silva et al. \(2016\)](#), donde se propone la utilización de curvas cuantiles obtenidas mediante modelos aditivos, ya que de este modo se proporciona información no solo sobre la media sino sobre la distribución completa de los niveles de emisión.

3.4.3. Modelo semiparamétrico bivalente

En [González-Manteiga et al. \(2009\)](#) se lleva a cabo una generalización del modelo semiparamétrico que fue propuesto en la Sección 2.4.2. En concreto, se generaliza al caso bivalente para realizar predicciones conjuntas de los niveles de SO_2 y de NO_x . Estos modelos surgen en el contexto de la economía y es por ello que también se hace hincapié en el fenómeno de la cointegración, concepto desarrollado en [Granger \(1983\)](#) y [Engle y Granger \(1987\)](#) con diversas aplicaciones en el ámbito del análisis de datos econométricos.

En el contexto de las series de tiempo, imagínese que se tienen dos series X_t e Y_t no estacionarias. A pesar de que estas no lo son, podría tenerse una combinación lineal de estas que sí lo fuese, i.e., $Z_t = \alpha_1 X_t + \alpha_2 Y_t$. De existir, se dice que ambas series están cointegradas y se interpreta como que, a pesar de que el comportamiento de ambas pueda resultar «errático», en el sentido de que no sean

¹⁴ Contiene como casos particulares a los modelos anteriores.

¹⁵ Los primeros modelos de regresión cuantil fueron propuestos en [Koenker y Bassett Jr \(1978\)](#).

estables en términos de media y varianza, su relación a largo plazo sí que es estable. Es por ello que en [González-Manteiga et al. \(2009\)](#) se lleva a cabo un contraste para analizar si los errores del modelo propuesto están cointegrados y, en caso afirmativo, se emplea un modelo de corrección sobre ellos.

Por último, cabe mencionar que los resultados obtenidos con el modelo propuesto son similares al modelo no paramétrico y semiparamétrico cuando las componentes de la variable respuesta no están relacionadas pero en el contexto medioambiental, donde sí que hay dependencia entre ellas, este nuevo modelo obtiene mejores resultados que los otros.

3.4.4. Modelo de datos funcionales bivariante

En [Oviedo-de la Fuente et al. \(2020\)](#) se presenta un modelo funcional de localización y escala para predecir episodios de contaminación en los que están involucrados tanto SO_2 como NO_x . Más concretamente, se emplean modelos aditivos generalizados funcionales (*FGAM*, véase [McLean et al. 2014](#)) para estimar tanto las medias como la estructura de correlaciones entre ambas sustancias. La metodología que presentan incluye tanto predicciones puntuales conjuntas de las concentraciones de ambas como regiones de incertidumbre, con un nivel prefijado, donde estas deberían estar localizadas, alcanzándose buenos resultados en términos de cobertura y error de predicción.

Otra aproximación realizada en el caso de una respuesta bivariante es la llevada a cabo en [Roca-Pardiñas et al. \(2021\)](#), donde las medias y las desviaciones típicas son aproximadas mediante funciones de tipo núcleo en modelos aditivos, mientras que la matriz de covarianzas se obtiene a partir de los residuos de los modelos anteriores.

Observación 3.9. También podían haber sido aplicados al problema medioambiental una generalización de los modelos Box-Jenkins del Capítulo 1 para vectores aleatorios. Estos modelos se conocen como modelos autorregresivos (integrados) de medias móviles o, abreviadamente, VARMA (*VARIMA*). Algunos ejemplos de aplicación en el contexto del control de emisiones pueden ser [Hsu \(1992\)](#), [Kadiyala y Kumar \(2014\)](#) o [García-Nieto et al. \(2018\)](#).

3.4.5. Boosting en los modelos de redes neuronales y de datos funcionales

Los métodos *boosting* surgen en el ámbito del *machine learning*, donde la motivación de los mismos consiste en obtener una herramienta de clasificación poderosa como combinación de otras más débiles (véase el Capítulo 10 de [Hastie et al. 2009](#)). No obstante, estas técnicas, diseñadas para el contexto de los problemas de clasificación, se extendieron hacia otros ámbitos como el aprendizaje de algoritmos y la regresión. Es por ello que, conscientes de su potencial, en [Fernández-de Castro y González-Manteiga \(2008\)](#) se decidió aplicar estas técnicas al contexto de datos funcionales en los dos tipos de metodologías que fueron presentadas en la Sección 3.3 y también al contexto de datos escalares en el modelo de redes neuronales propuesto en la Sección 3.2.2.

Dada una variable respuesta y (escalar o funcional) y una variable explicativa x (vectorial o funcional) el interés está en la relación $y = f(x) + \varepsilon$, donde la estimación del operador f vendrá dada por una combinación lineal de una base de funciones conocida. Así, el proceso *boosting* busca ajustar un modelo aditivo sobre un conjunto de funciones de una base mediante un proceso iterativo que busca ir reduciendo un criterio de error en cada iteración.

En cuanto a los resultados, el procedimiento *boosting* permitió mejorar tanto las predicciones obtenidas con el modelo basado en redes neuronales como con los dos modelos basados en datos funcionales.

Capítulo 4

Test bootstrap para estructuras simples

En los modelos predictivos que han sido introducidos en los capítulos anteriores una de las premisas más importantes es la correcta especificación de los mismos. Es por ello que a lo largo de los últimos años se ha producido un auge en el desarrollo de contrastes de hipótesis para la especificación de modelos. Para ilustrar la naturaleza de estos contrastes, en el presente capítulo se introducirá el test de especificación bootstrap para estructuras simples en regresión no paramétrica de series de tiempo que fue propuesto en [Kreiss et al. \(2008\)](#).

Sea una muestra $\{(\mathbf{X}_t, Y_t)\}_{t=1}^T$ de un proceso estocástico en tiempo discreto estrictamente estacionario $\{\mathbf{X}_t, Y_t\}$, con $\mathbf{X}_t \in \mathbb{R}^d$ e $Y_t \in \mathbb{R}$. El objetivo que se persigue es realizar un test de hipótesis sobre si la función de regresión dada por $m(\mathbf{x}) = \mathbb{E}[Y_t | \mathbf{X}_t = \mathbf{x}]$ sigue algún tipo de «estructura simple», considerando como tal al modelo paramétrico, al modelo unidimensional y al modelo aditivo. Así pues, considérese el modelo de regresión en media general dado por¹

$$Y_t = m(\mathbf{X}_t) + \varepsilon_t, \quad t \geq 1,$$

donde $\mathbb{E}[\varepsilon_t | \mathcal{F}_t] = 0$ para todo t y \mathcal{F}_t es la σ -álgebra generada por $\{(\mathbf{X}_s, Y_{s-1}), s = t, t-1, \dots\}$. Se considerarán 3 tipos de hipótesis nulas acerca de la estructura de la esperanza condicionada:

$$\begin{cases} H_{0,\theta} : m(\cdot) \in \{m_\theta(\cdot) \mid \theta \in \Theta\}, \\ H_{0,m} : m(x_1, \dots, x_d) = m_0(x_1), \\ H_{0,a} : m(x_1, \dots, x_d) = m_1(x_1) + \dots + m_d(x_d). \end{cases}$$

Conviene mencionar que así como en la hipótesis nula $H_{0,m}$ se considera solamente un modelo unidimensional será posible testear también un modelo d_0 -dimensional para algún $d_0 < d$. En cuanto a la hipótesis nula relativa al modelo paramétrico, cabe destacar que bajo esta se encuentra una amplia variedad de modelos de regresión como por ejemplo:

- Modelo de regresión lineal:

$$m_\theta(x_1, \dots, x_d) = \sum_{i=1}^d \theta_i x_i.$$

¹ Nótese que no se establece ninguna condición de independencia sobre los errores ε_t , por lo que la heterocedasticidad condicional también estaría incluida en este modelo general. Además, si se toman como \mathbf{X}_t valores pasados de Y_t se estaría considerando el caso particular del modelo autorregresivo.

- Modelo ARCH²:

$$X_t = \sigma_\theta(X_{t-1}, \dots, X_{t-d}) e_t, \quad \mathbb{E}[e_t] = 0, \quad \mathbb{E}[e_t^2] = 1,$$

donde $\sigma_\theta(x_1, \dots, x_d) = \sqrt{\theta_0 + \sum_{i=1}^d \theta_i x_i^2}$. Para verlo como un caso particular de modelo de regresión paramétrico basta con tomar $Y_t = X_t^2$.

Previamente a introducir el estadístico de test es necesario considerar primero un estimador de $m(\cdot)$ adecuado la hipótesis nula que se esté considerando. Así, denotando de manera común a tal estimador como $\tilde{m}(x_1, \dots, x_d)$ se tendrá lo siguiente:

$$\tilde{m}(x_1, \dots, x_d) = \begin{cases} m_{\hat{\theta}}(x_1, \dots, x_d), & \text{si se considera } H_{0,\theta}, \\ \hat{m}_0(x_1), & \text{si se considera } H_{0,m}, \\ \hat{m}_1(x_1) + \dots + \hat{m}_d(x_d) & \text{si se considera } H_{0,a}. \end{cases}$$

Considerando este estimador de la función de regresión se tiene que el estadístico de test presentado en [Kreiss et al. \(2008\)](#) es el dado por

$$S_T = \int_{\mathbb{R}^d} \left(\frac{1}{T} \sum_{i=1}^T K_h(\mathbf{x} - \mathbf{X}_i) [Y_i - \tilde{m}(\mathbf{X}_i)] \right)^2 w(\mathbf{x}) d\mathbf{x},$$

donde $K_h(\cdot) = h^{-d} K(\cdot/h)$, $K(\cdot)$ es una función núcleo en \mathbb{R}^d , $h > 0$ es un parámetro ventana y $w(\cdot)$ es una función de pesos.

Observación 4.1. En $w(\cdot)$ se considera el cuadrado del estimador de la densidad estacionaria de \mathbf{X}_t , $\pi^2(\cdot)$, lo cual se interpreta como que solo se considera la diferencia en el soporte de $\pi(\cdot)$ y se presta menos atención a las diferencias donde los datos son más escasos.

Con cada hipótesis nula lo único que varía en el estadístico de test S_T es el estimador de la función de regresión. Así pues, en el caso de querer testear una hipótesis paramétrica del tipo $H_{0,\theta}$ se considerará un estimador \sqrt{T} -consistente del verdadero parámetro θ_0 , denotado por $\hat{\theta}$, para el cual se tenga que

$$m_{\hat{\theta}}(\cdot) - m_{\theta_0}(\cdot) = (\hat{\theta} - \theta_0)' \dot{m}_{\theta_0}(\cdot) + O_P\left(\|\cdot\|^2 \left(\sqrt{T} \log T\right)^{-1}\right).$$

donde $\dot{m}_\theta(\cdot)$ representa la derivada de $m_{\theta_0}(\cdot)$ con respecto de θ y $\|\cdot\|$ denota a la norma euclídea.

En cuanto al test para la hipótesis no paramétrica de unidimensionalidad, $H_{0,m}$, se considerará un estimador polinómico local de cierto orden p , con $[p/2] > 5d/16$ ³⁴. Luego, se estimará $m_0(x_1)$ por $\hat{m}_g(x_1) = \hat{a}$, donde

$$\left(\hat{a}, \hat{b}_1, \dots, \hat{b}_p\right) = \arg \min_{a, b_1, \dots, b_p} \sum_{t=1}^T [Y_t - a - b_1(x_1 - X_{t,1}) - \dots - b_p(x_1 - X_{t,1})^p]^2 W\left(\frac{x_1 - X_{t,1}}{g}\right),$$

donde W es un núcleo en \mathbb{R} , $g > 0$ es un parámetro ventana y $X_{t,1}$ es la primera componente de \mathbf{X}_t .

² Véase el Apéndice A.1.

³ $[p/2]$ denota a la parte entera de $p/2$.

⁴ Se toma un estimador polinómico local con un orden suficientemente grande en lugar de un estimador tipo núcleo convencional (local constante) para mantener el sesgo lo suficientemente pequeño. Nótese que tal y como se ha definido S_T está involucrada a mayores la suavización del estimador de $m_0(\cdot)$, lo cual conduce a que aumente aún más su sesgo bajo la hipótesis nula.

Por último, en cuanto al test para la hipótesis nula del modelo de regresión aditivo, $H_{0,a}$, se emplean estimadores de integración no paramétricos. En [Fan et al. \(1998\)](#) puede verse este tipo estimador basado en integración marginal, destacando el hecho de que se alcanza la tasa habitual de los estimadores unidimensionales de curvas no paramétricas, es decir, cada componente puede estimarse con el mismo sesgo y la misma varianza que el suavizador unidimensional, como si las demás componentes fuesen conocidas. Dicho de otro modo, el hecho de añadir más componentes en el modelo aditivo, si bien aumenta el número efectivo de parámetros, no añade ninguna dificultad extra en la estimación (al menos asintóticamente).

Observación 4.2. En virtud del comentario realizado al final del párrafo anterior, los resultados obtenidos para testear una hipótesis nula no paramétrica unidimensional se trasladan de manera inmediata al caso no paramétrico aditivo.

4.1. Comparación con el estadístico de Härdle y Mammen

En este apartado, con el fin de ilustrar mejor la naturaleza del estadístico S_T se va a introducir muy brevemente el estadístico de test presentado en [Hardle y Mammen \(1993\)](#) en el contexto de la bondad de ajuste en la regresión para luego comparar su estructura con la del estadístico de [Kreiss et al. \(2008\)](#) y comprender mejor la idea fundamental subyacente. Así pues, considérese el modelo de regresión general dado por

$$Y_t = m(\mathbf{X}_t) + \varepsilon_t, \quad t \geq 1, \quad \mathbb{E}[\varepsilon_t | \mathbf{X}_t] = 0.$$

Para estimar la función regresora $m(\cdot)$ puede utilizarse un estimador de tipo Nadaraya-Watson:

$$\hat{m}_h(\cdot) = \sum_{t=1}^T \frac{K_h(\cdot - X_t)}{\sum_{j=1}^T K_h(\cdot - X_j)} Y_t, \quad (4.1)$$

donde $K_h(\cdot) = h^{-d}K(\cdot/h)$, $K(\cdot)$ es una función núcleo en \mathbb{R}^d y $h > 0$ es un parámetro ventana. Considérese, además, un operador de suavizado (aleatorio), denotado por $\mathcal{K}_{h,T}$, y cuya definición es

$$\mathcal{K}_{h,T}g(\cdot) = \sum_{t=1}^T \frac{K_h(\cdot - X_t)}{\sum_{j=1}^T K_h(\cdot - X_j)} g(X_t).$$

A partir de la definición de este estimador resulta inmediato obtener que la esperanza condicionada del estimador de Nadaraya-Watson considerado en (4.1) no es la verdadera función de regresión sino una versión suavizada de la misma. Más concretamente,

$$\mathbb{E}[\hat{m}_h(\cdot) | X_1, \dots, X_T] = \mathcal{K}_{h,T}m(\cdot).$$

Conscientes de esto, en [Hardle y Mammen \(1993\)](#), con el fin de contrastar la hipótesis $H_{0,\theta}$ basada en observaciones independientes, se considera la siguiente distancia cuadrática ponderada entre \hat{m}_h y $m_{\hat{\theta}}$:

$$d(T, H_{0,\theta}) = \int (\hat{m}_h(x) - \mathcal{K}_{h,T}m_{\hat{\theta}}(x))^2 \pi(x) dx.$$

De este modo, lo que se hace a través de ese estadístico es comparar \hat{m}_h con la «estimación» paramétrica de la esperanza condicionada de \hat{m}_h , esto es, $\mathcal{K}_{h,T}m_{\hat{\theta}}(x)$. Así pues, la idea fundamental que se desprende del citado artículo consiste en emplear una distancia cuadrática modificada entre un

estimador paramétrico $(H_{0,\theta})$, $m_{\hat{\theta}}(\cdot)$, y uno no paramétrico, $\hat{m}_h(\cdot)$, que es bastante similar a lo que se hace en el estadístico propuesto en [Kreiss et al. \(2008\)](#):

$$S_T = \int_{\mathbb{R}^d} \left(\frac{1}{T} \sum_{i=1}^T K_h(\mathbf{x} - \mathbf{X}_t) [Y_t - \tilde{m}(\mathbf{X}_t)] \right)^2 w(\mathbf{x}) d\mathbf{x},$$

4.2. Calibrado del test

Antes de presentar las técnicas bootstrap para el calibrado del test conviene comenzar descomponiendo el estadístico S_T en varios sumandos. Así pues, se tiene que

$$\begin{aligned} Th^{d/2} S_T &= Th^{d/2} \bar{S}_T - \frac{2h^{d/2}}{T} \int \left[\sum_{t=1}^T K_h(\mathbf{x} - \mathbf{X}_t) \varepsilon_t \right] \left[\sum_{s=1}^T K_h(\mathbf{x} - \mathbf{X}_s) \{ \tilde{m}(\mathbf{X}_s) - m(\mathbf{X}_s) \} w(\mathbf{x}) \right] d\mathbf{x} \\ &\quad + \frac{h^{d/2}}{T} \int \left(\sum_{t=1}^T K_h(\mathbf{x} - \mathbf{X}_t) \{ \tilde{m}(\mathbf{X}_t) - m(\mathbf{X}_t) \} \right)^2 w(\mathbf{x}) d\mathbf{x}, \end{aligned}$$

donde

$$\bar{S}_T = \frac{1}{T^2} \int \left(\sum_{t=1}^T K_h(\mathbf{x} - \mathbf{X}_t) \varepsilon_t \right)^2 w(\mathbf{x}) d\mathbf{x},$$

lo cual es una forma cuadrática sobre las innovaciones $\{\varepsilon_t\}$. Además, este estadístico tiene como particularidad que, al contrario de lo que ocurría con S_T , es invariante bajo las tres hipótesis nulas, ya que no está involucrado el estimador de la función de regresión.

El siguiente resultado será de gran utilidad, ya que, además de afirmar que la distribución asintótica de $Th^{d/2} S_T$ es la misma que la de $Th^{d/2} \bar{S}_T$, establece que esta es gaussiana.

Teorema 4.1 ([Kreiss et al. 2008](#), p. 375). *Supóngase que se cumple una de las hipótesis nulas $H_{0,\theta}$, $H_{0,m}$ o $H_{0,a}$ y que el estadístico S_T antes introducido está definido en términos de uno de los estimadores presentados según la hipótesis nula considerada. Bajo ciertas condiciones de regularidad (véase [Kreiss et al. 2008](#), p. 374) se tiene que, cuando $T \rightarrow \infty$ se cumple lo siguiente:*

- 1) $S_T = \bar{S}_T + o_P(T^{-1}h^{-d/2})$.
- 2) $(Th^{d/2}) (\bar{S}_T - \mathbb{E}[\bar{S}_T]) \xrightarrow{d} N(0, V)$, donde

$$\begin{aligned} \mathbb{E}[\bar{S}_T] &= \frac{1}{Th^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} K^2(\mathbf{u}) w(\mathbf{x} + h\mathbf{u}) \pi(\mathbf{x}) \sigma^2(\mathbf{x}) d\mathbf{x} d\mathbf{u}, \\ V &= 2 \left[\int_{\mathbb{R}^d} \sigma^4(\mathbf{x}) \pi^2(\mathbf{x}) w(\mathbf{x}) d\mathbf{x} \right] \left[\int_{\mathbb{R}^{3d}} K(\mathbf{u}) K(\mathbf{v}) K(\mathbf{u} - \mathbf{z}) K(\mathbf{v} - \mathbf{z}) d\mathbf{u} d\mathbf{v} d\mathbf{z} \right]. \end{aligned}$$

Observación 4.3. Aunque la distribución de S_T depende de si la hipótesis nula correspondiente se cumple o no, esto no ocurre con la distribución de \bar{S}_T . Así pues, la aproximación bootstrap a la distribución de S_T bajo la hipótesis nula basada en \bar{S}_T es siempre válida, incluso cuando las observaciones $\{(Y_t, \mathbf{X}_t)\}$ provienen de una distribución poblacional en la cual no es cierta dicha hipótesis. La validez de esta aproximación es lo que garantiza una potencia razonable en el test bootstrap frente a desviaciones con respecto a la hipótesis nula considerada.

De este modo, en base al Teorema 4.1 y a la Observación 4.3, para imitar la distribución de S_T basta con realizar el bootstrap únicamente sobre la forma cuadrática \bar{S}_T . En cuanto al tipo de mecanismo de remuestreo, en [Hardle y Mammen \(1993\)](#) se estudian tres mecanismos bootstrap diferentes (Naïve, residual y wild) y concluyen que el más adecuado para testear la estructura del modelo de regresión es el wild bootstrap, el cual fue presentado en la Sección 2.3.4.

Ahora que se conoce que la distribución asintótica $Th^{d/2}S_T$ es la misma que la de $Th^{d/2}\bar{S}_T$ y que el mecanismo de remuestreo a emplear será el wild bootstrap, se va a definir el siguiente estadístico:

$$S_T^* = \frac{1}{T^2} \int \left(\sum_{t=1}^T K_h(\mathbf{x} - \mathbf{X}_t) \varepsilon_t^* \right)^2 w(\mathbf{x}) d\mathbf{x},$$

donde las innovaciones bootstrap $\{\varepsilon_t^*\}_{t=1}^T$ son condicionalmente independientes dadas las observaciones $\{(Y_t, \mathbf{X}_t)\}_{t=1}^T$ y, además cumplen que $\varepsilon_t^* = \hat{\varepsilon}_t \eta_t$, donde $\hat{\varepsilon}_t = Y_t - \hat{m}_h(\mathbf{X}_t)$ y $\{\eta_t\}_{t=1}^T$ es una sucesión de variables aleatorias independientes que cumplen que

$$\mathbb{E}^*[\eta_t] = 0, \quad \mathbb{E}^*[\eta_t^2] = 1, \quad \mathbb{E}^*[\eta_t^3] = 1.$$

Así pues, el esquema del mecanismo bootstrap que se empleará para el calibrado del test será el que se recoge en el Algoritmo 18.

Algoritmo 18 Test Bootstrap sobre el modelo de regresión, [Kreiss et al. \(2008\)](#)

- 1: **Construir los residuos** a partir del estimador no paramétrico de $m(x)$ tomando el parámetro ventana de partida h :

$$\hat{\varepsilon}_t = Y_t - \hat{m}_h(\mathbf{X}_t), \quad i = 1, 2, \dots, n.$$

- 2: **Generar errores bootstrap** $\varepsilon_t^* = \hat{\varepsilon}_t \eta_t$, $t = 1, 2, \dots, n$, condicionalmente a la muestra observada, donde $\{\eta_t\}_{t=1}^T$ es una sucesión de variables aleatorias independientes que cumplen que

$$\mathbb{E}^*[\eta_t] = 0, \quad \mathbb{E}^*[\eta_t^2] = 1, \quad \mathbb{E}^*[\eta_t^3] = 1.$$

- 3: **Generar el análogo bootstrap del estadístico \bar{S}_T :**

$$S_T^* = \frac{1}{T^2} \int \left(\sum_{t=1}^T K_h(\mathbf{x} - \mathbf{X}_t) \varepsilon_t^* \right)^2 w(\mathbf{x}) d\mathbf{x}$$

- 4: **Obtener el cuantil $t_{1-\alpha}^*$** , esto es, el punto tal que

$$\mathbb{P}^*\{S_T^* \leq t_{1-\alpha}^*\} = 1 - \alpha.$$

- 5: **Rechazar la hipótesis nula si $S_T > t_{1-\alpha}^*$.**
-

Observación 4.4. En el paso 4 del Algoritmo 18 para obtener el cuantil es necesario conocer la distribución en el remuestreo de S_T^* . No obstante, esta puede aproximarse mediante técnicas de Monte Carlo. Además, en ese caso puede obtenerse también un p -valor bootstrap de la manera siguiente:

$$p^* = \frac{1}{B} \# \{S_T^* \geq S_T\}.$$

4.3. Validez del bootstrap y ejemplo de aplicación

En este apartado final se comienza presentando un resultado teórico en el cual descansa la validez del test bootstrap para luego finalizar con la implementación del método en R y la aplicación del mismo a uno de los ejemplos de [Kreiss et al. \(2008\)](#).

Teorema 4.2 ([Kreiss et al. 2008](#), p. 375). *Bajo las mismas condiciones que en el Teorema 4.1 el estadístico bootstrap S_T^* , cuando $T \rightarrow \infty$ y condicionalmente a la muestra $\{(\mathbf{X}_t, Y_t)\}_{t=1}^T$ verifica que*

$$Th^{d/2} (S_T^* - \mathbb{E}^* [S_T^*]) \xrightarrow{d} N(0, V),$$

donde V es la misma que la del Teorema 4.1 y, además, se tiene que

$$Th^{d/2} (\mathbb{E} [\bar{S}_T] - \mathbb{E}^* [S_T^*]) \xrightarrow{p} 0.$$

Una consecuencia directa del Teorema 4.1 es que, en efecto, el nivel de significación del test bootstrap converge al valor nominal fijado a medida que aumenta el tamaño de la muestra. Dicho de otro modo, el test de Kreiss está bien calibrado con la metodología bootstrap presentada.

Corolario 4.3 ([Kreiss et al. 2008](#), p. 375). *En las condiciones del Teorema 4.1 y siendo $t_{1-\alpha}^*$ el cuantil $1 - \alpha$ de la distribución condicional de S_T^* dada la muestra $\{(\mathbf{X}_t, Y_t)\}_{t=1}^T$ se tiene que, bajo la correspondiente hipótesis nula, cuando $T \rightarrow \infty$ se cumple que*

$$\mathbb{P} \{S_T \leq t_{1-\alpha}^*\} \longrightarrow 1 - \alpha.$$

Antes de finalizar los comentarios relacionados con el aspecto metodológico es conveniente destacar antes que las diferentes versiones del test de Kreiss que han sido presentadas son test omnibus donde la alternativa es la regresión d -dimensional general dada por

$$Y_t = m(X_t) + \varepsilon_t, \quad t \geq 1,$$

donde $\mathbb{E}[\varepsilon_t | \mathcal{F}_t] = 0$. Ahora bien, como cabría esperar, la potencia de los test que han sido introducidos acostumbra a ser menor que la de aquellos otros que tienen una hipótesis alternativa más específica⁵.

Cabe mencionar que en el Apéndice C.1 se encuentra una implementación en R ([R Core Team 2024](#)) del test bootstrap que se ha presentado. Además, a título ilustrativo se ha incluido también la aplicación del mismo, considerando como hipótesis nula la de tipo paramétrico (linealidad), a uno de los modelos estacionarios simulados en [Kreiss et al. \(2008\)](#). Más concretamente, al modelo dado por

$$X_t = -0,9X_{t-1} + \varepsilon_t,$$

donde $\{\varepsilon_t\}$ es una muestra aleatoria simple de una normal estándar. Fijando una cobertura nominal del 95 % se realizan 500 simulaciones de dicho modelo con un tamaño de muestra de 100 y 200 réplicas bootstrap. Se toma como función de pesos $w(\cdot) \equiv 1$ y la función tipo núcleo de Epanechnikov. Además, para la selección del parámetro ventana en la estimación no paramétrica se emplea un criterio de validación cruzada. En cuanto al resultado obtenido, se tiene una proporción de rechazos del 6.4 %, razonablemente próxima al 5 % fijado.

Al final de [Kreiss et al. \(2008\)](#) puede verse la aplicación de este test a otros modelos, resultando que incluso cuando la distribución de las innovaciones es fuertemente asimétrica el test tiende a tomar la decisión en el sentido correcto⁶. Este hecho resulta muy llamativo, ya que algunos test clásicos rechazarían la hipótesis de linealidad ante la presencia de un modelo que, a pesar de ser lineal, posea unas innovaciones fuertemente asimétricas, es decir, en estos casos les resulta difícil distinguir la no normalidad de la no linealidad.

⁵ Véase [Fan y Jiang \(2007\)](#).

⁶ Aunque el comportamiento del mismo en esos casos es más bien conservador, es decir, rechaza la hipótesis nula una proporción de veces razonablemente menor que el valor nominal.

Capítulo 5

Ilustración con datos reales

En el presente capítulo se va a ilustrar someramente la implementación de algunos de los modelos presentados a lo largo del trabajo, empleando para ello datos reales de NO_x del ciclo combinado de la Unidad de Producción Térmica de As Pontes y buscando predecir un episodio no incluido en la matriz histórica.

Debido a que la empresa no ha sido consultada para la publicación de detalles relativos los datos, no se incluye el código relativo a la creación de matrices históricas, ya que contiene información acerca de la configuración interna de los archivos de datos, así como la fecha y localización de los episodios de contaminación.

No obstante, lo que sí cabe decir acerca de la matriz es que en este caso la conforman ternas de la forma (X_{t-5}, X_t, X_{t+30}) y que los estratos están contruidos en base a X_{t+30} . En cuanto a la naturaleza de los datos, se trata de medias horarias de inmisión de NO_x de frecuencia minutal de un determinado año. El código con el ajuste de los modelos puede verse en el Apéndice C.2.

5.1. Modelo paramétrico

En esta sección se va a emplear la metodología Box-Jenkins en el día del episodio que se quiere predecir. Para ello, se dividen los datos en una muestra de entrenamiento (hasta las 13:00h) y en una muestra de test (de 13:01 a 13:30). El objetivo consiste en ajustar un modelo $\text{ARIMA}(p, d, q)$ sobre la muestra de entrenamiento y luego realizar una predicción a un horizonte de 30 minutos, que es cuando se produce el pico del episodio.

En cuanto al ajuste del modelo, se ha empleado un procedimiento automático de selección de modelos basado en el criterio de información de Akaike corregido, obteniéndose como modelo sugerido un $\text{ARIMA}(1, 2, 2)$. En la Figura 5.1 puede verse que los residuos no parecen homocedásticos, ya que en la gráfica superior de la Figura 5.1 se observan instantes de mayor variabilidad que otros. Por otro lado, tanto en esa figura como en la Figura 5.2 se observa que los residuos, si bien parecen tener una distribución simétrica, la densidad estimada en el 0 es razonablemente mayor que lo esperado bajo un modelo gaussiano, hecho que se refrenda gráficamente en el *qq-plot* y de manera analítica mediante el Test de Normalidad de Shapiro-Wilk, donde se obtiene un p -valor menor que $2,2e - 16$. Así pues, los intervalos de predicción que se podrían construir en este caso no podrían realizarse empleando que las innovaciones son gaussianas, pudiendo ser aproximada su distribución mediante alguno de los métodos bootstrap comentados en el trabajo.

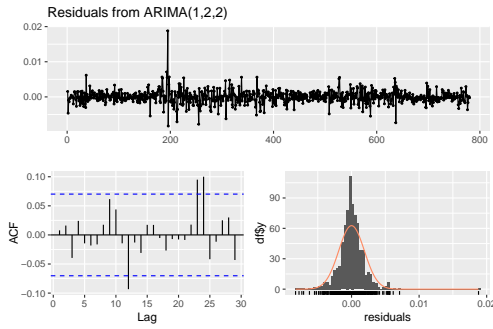


Figura 5.1: Análisis de los residuos del modelo ARIMA(1, 2, 2) ajustado.

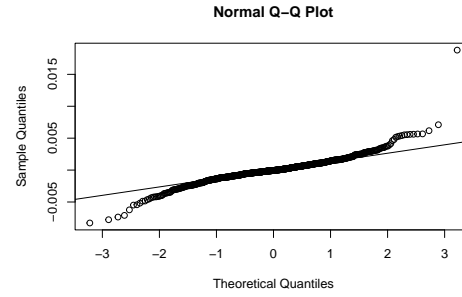


Figura 5.2: *qq-plot* de normalidad de los residuos del modelo ARIMA(1, 2, 2) ajustado.

No obstante, como el objetivo consiste en realizar predicciones puntuales, la ausencia de normalidad que ha sido comentada no supone un gran problema. Estas predicciones, realizadas sobre la muestra de test pueden verse en la Figura 5.3, donde se observa el bajo rendimiento de las predicciones obtenidas.

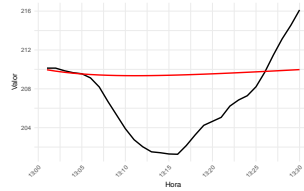


Figura 5.3: Predicciones (rojo) en la muestra de test (negro) bajo el modelo ARIMA(1, 2, 2) ajustado.

Observación 5.1. Se han realizado predicciones habiendo fijado los parámetros del modelo ARIMA en la muestra de entrenamiento. Otra posibilidad consistiría en actualizar los órdenes del modelo cada cierto tiempo, no necesariamente cada minuto, aunque esto añadiría mayor complejidad computacional al proceso predictivo, reduciendo así la capacidad de anticipación ante un episodio.

5.2. Modelo no paramétrico

En este apartado se va a ajustar un modelo no paramétrico de tipo Nadaraya-Watson, al estilo del presentado en la Sección 2.5.1. Más concretamente se va a estimar no paramétricamente la función de regresión dada por $\mathbb{E}[X_{t+30} | X_t, X_{t-5}]$ a partir de los datos recogidos en la matriz histórica. Luego, se realizan predicciones puntuales con dicho modelo sobre los datos relativos al episodio de contaminación.

Los resultados obtenidos tomando un núcleo gaussiano con un criterio de selección de ventana basado en validación cruzada pueden verse en la Figura 5.4. En primer lugar, antes de analizarlo conviene mencionar que en los cálculos de las predicciones ha habido que realizar una modificación (habitual en este tipo de modelos) al haber datos del episodio que caen fuera del soporte de las covariables con las que fue ajustada el modelo (matriz histórica). Esto puede verse de manera clara entre las 14:00h y las 15:30h, ya que en la matriz histórica no hay valores tan elevados, y por tanto es necesario aplicar la corrección mencionada, la cual se ha tomado como la predicción del valor más cercano (según una distancia euclídea) que sí esté en la matriz. Así pues, ha quedado reflejada la importancia de disponer de una matriz histórica con un número de datos razonable en cada uno de los estratos

En cuanto al rendimiento del modelo, cabe mencionar que en las zonas donde hay datos la aproximación es razonablemente buena, si bien se producen una serie de oscilaciones en la estimación que

aventuran, quizás, la necesidad de redefinir los estratos de la matriz histórica, de manera que el criterio de selección de ventana permita obtener predicciones más suaves¹.

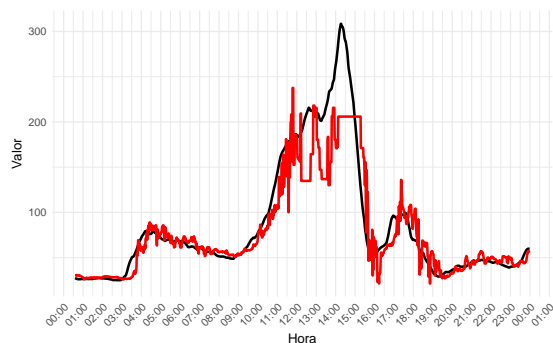


Figura 5.4: En rojo, las predicciones bajo el modelo no paramétrico ajustado; en negro, los datos reales.

Por último, cabe destacar el retraso en las predicciones del modelo, lo cual es debido a que se están tomando como variables explicativas los valores pasados de la serie hace 30 y 35 minutos, por lo que un dato recibido en un instante t no será empleado para predecir hasta el instante $t + 35$.

5.3. Modelo semiparamétrico

En este apartado se va a ajustar un modelo semiparamétrico, al estilo del propuesto en la Sección 2.5.2, solo que ahora el ajuste no paramétrico subyacente será el que ha sido comentado en la Sección 5.2. A partir de dicho ajuste se construye la serie de residuos asociados sobre la muestra de entrenamiento construida en la Sección 5.1 para luego definir como predicciones las predicciones no paramétricas en el conjunto de test más las obtenidas para los residuos bajo el modelo Box-Jenkins.

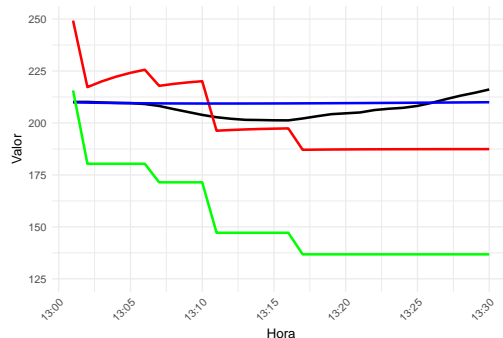


Figura 5.5: En rojo, las predicciones bajo el modelo semiparamétrico ajustado; en azul, las predicciones bajo el modelo Box-Jenkins de la Sección 5.1; en verde, las predicciones bajo el modelo no paramétrico de la Sección 5.2; en negro, los datos reales.

Los resultados obtenidos pueden verse en la Figura 5.5, donde se observa cierta mejora con respecto a los modelos de los dos apartados anteriores. No obstante, debido al problema que presentaba el modelo

¹ Nótese que se observa un mayor número de oscilaciones en valores elevados de la concentración de NO_x , lo cual es debido a que estos son mucho menos frecuentes que los valores bajos pero la ventana es común para todos los valores.

semiparamétrico con respecto a predecir sobre valores que están fuera del soporte de la matriz histórica, las predicciones todavía no se ajustan razonablemente bien a los datos reales.

5.4. Modelo de redes neuronales

Por último, en este apartado se va a ajustar el modelo basado en redes neuronales comentado en la Sección 3.2.2, con la salvedad de que los vectores de la matriz histórica que se va a emplear para el entrenamiento de la red son de la forma $(X_{t-35}, X_{t-30}, X_t)$. Las predicciones obtenidas con la red neuronal entrenada son los que se pueden ver en la Figura 5.6. Allí se observa que, en comparación con lo obtenido bajo el modelo no paramétrico, las predicciones son suaves y el problema que había entonces entre las 14:00 y las 15:30h no se observa, si bien debido a la falta de datos suficientes, no es capaz de predecir adecuadamente el mayor pico del episodio. A pesar de lo comentado, cabe decir que las predicciones imitan razonablemente bien la forma de los verdaderos valores, destacando también el retardo de 30 minutos en la imitación de esta forma.

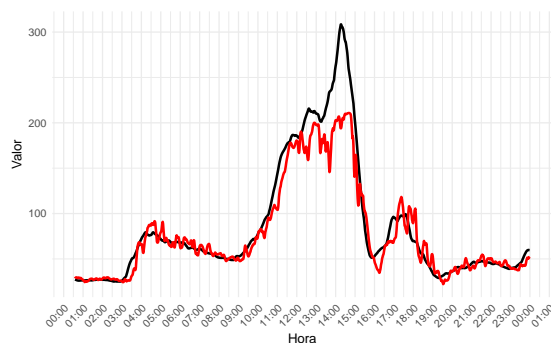


Figura 5.6: En rojo, las predicciones bajo el modelo basado en redes neuronales ajustado; en negro, los datos reales.

5.5. Conclusiones finales y *future work*

En cuanto a las predicciones sobre la muestra de test, estas pueden verse conjuntamente en la Figura 5.7, donde claramente el mejor ajuste viene dado por el modelo basado en redes neuronales, seguido por el modelo Box-Jenkins. No obstante, a la hora de realizar estas comparaciones es preciso tener en cuenta que el modelo paramétrico no ha sido entrenado mediante la matriz histórica y, en consecuencia, no es «justa» su comparación con los demás modelos.

En definitiva, a lo largo del trabajo se ha visto que la matriz histórica es una manera de construir muestras muy útil pero el diseño que se considera a la hora de construirla puede provocar que no se disponga de datos suficientes en algunos estratos, repercutiendo en el rendimiento de algunos modelos como el modelo no paramétrico y, por extensión, también el modelo semiparamétrico. Por lo tanto, quedaría pendiente realizar un estudio en profundidad del episodio que se ha considerado en este capítulo y analizar con qué configuraciones de la matriz histórica se obtienen mejores resultados. Quizás de esta manera podría vislumbrarse un procedimiento generalizable que ayudase a diseñar de antemano los estratos sin necesidad de llevar a cabo un estudio previo.

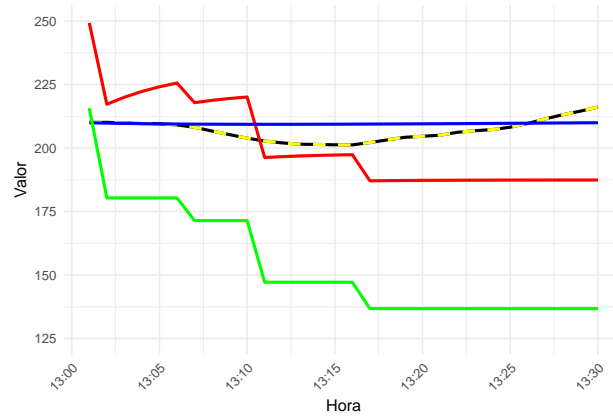


Figura 5.7: En amarillo discontinuo, las predicciones bajo el modelo basado en redes neuronales; en rojo, las predicciones bajo el modelo semiparamétrico; en azul, las predicciones bajo el modelo Box-Jenkins; en verde, las predicciones bajo el modelo no paramétrico ; en negro, los datos reales.

Apéndices

Apéndice A

Series de tiempo

A.1. Modelos de heterocedasticidad condicional

Notación. Para no estar introduciendo la notación en cada definición se establece la siguiente:

- $\{r_t\}$ denotará una serie de rendimientos.
- Estructura principal de un modelo de volatilidad:

$$\begin{aligned}r_t &= \mu_t + a_t, \\ a_t &= \sigma_t \varepsilon_t,\end{aligned}$$

donde se tiene que

- $\mu_t = \mathbb{E}[r_t | \mathcal{F}_{t-1}]$, con \mathcal{F}_{t-1} la σ -álgebra generada por $\{r_s, s = 1, 2, \dots, t-1\}$;
- $\sigma_t^2 = \text{Var}[r_t | \mathcal{F}_{t-1}]$;
- a_t , son los llamados shocks o innovaciones y
- ε_t , son las innovaciones estandarizadas, variables aleatorias i.i.d. de media 0 y varianza 1.

Observación A.1. Para la media condicionada, μ_t , pueden proponerse modelos autorregresivos. En cuanto a la volatilidad, σ_t^2 , los modelos son del estilo de los que se expondrán a continuación.

Definición A.1 (Modelo ARCH¹, Galeano (2023)). Con la Notación A.1, el modelo ARCH de orden m se define como aquel que admite una expresión de la forma

$$\sigma_t^2 = \omega + \alpha_1 a_{t-1}^2 + \dots + \alpha_m a_{t-m}^2,$$

donde, como la volatilidad debe ser positiva, se debe satisfacer que $\omega > 0$ y $\alpha_i \geq 0$ para $i = 1, 2, \dots, m$.

Definición A.2 (Modelo GARCH², Galeano (2023)). Con la Notación A.1, el modelo GARCH de órdenes m y s , denotado por GARCH(m, s), es aquel que admite una expresión de la forma

$$\sigma_t^2 = \omega + \sum_{i=1}^m \alpha_i a_{t-i}^2 + \sum_{j=1}^s \beta_j \sigma_{t-j}^2, \quad (\text{A.1})$$

donde $\{\varepsilon_t\}$ es una sucesión de variables aleatorias i.i.d. que tienen media 0 y varianza 1 y donde, además, los parámetros satisfacen que $\omega > 0$, $\alpha_i \geq 0$ para $i = 1, \dots, m$ y $\beta_j \geq 0$ para $j = 1, 2, \dots, s$.

Observación A.2. Un modelo ARCH(m) no es más que un modelo GARCH($m, 0$).

¹ Este modelo fue propuesto en Engle (1982).

² Este modelo fue introducido en Bollerslev (1986).

Apéndice B

Algunos conceptos de Teoría de la Probabilidad

En este apéndice se van a introducir una serie de conceptos de Teoría de la Probabilidad que resultan fundamentales en el desarrollo metodológico de los modelos para series de tiempo. Además, se enunciarán algunos de los resultados auxiliares que se han empleado en el trabajo.

B.1. El espacio de las variables aleatorias con momento de orden 2 finito

Definición B.1 (Politis y McElroy 2020, p. 97). Dado un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, visto como un espacio vectorial de dimensión infinita formado por variables aleatorias se define el espacio de las variables aleatorias con momento de orden 2 finito como

$$\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P}) = \{X : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathcal{B}) \text{ medible} \mid \mathbb{E}[X^2] < \infty\}.$$

Abreviadamente se acostumbra a escribir únicamente como \mathbb{L}_2 .

Definición B.2 (Producto interior en \mathbb{L}_2 , Politis y McElroy 2020, p. 97). El producto interior en el espacio $\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$ se define como

$$\langle X, Y \rangle = \mathbb{E}[XY], \quad X, Y \in \mathbb{L}_2.$$

Así, la norma de una variable aleatoria se define como

$$\|X\| = \sqrt{\langle X, X \rangle} = \sqrt{\mathbb{E}[X^2]}, \quad X \in \mathbb{L}_2.$$

Definición B.3 (Convergencia en $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$). Se dice que una sucesión de variables aleatorias $\{X_n\}_{n \in \mathbb{N}}$ contenida en el espacio $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$ converge a X_∞ , y se escribe $X_n \xrightarrow{\mathbb{L}_2} X_\infty$, si se cumple que

$$\|X_n - X_\infty\| \longrightarrow 0.$$

Definición B.4 (Sucesión de Cauchy y completitud). Una sucesión $\{x_n\}_{n \in \mathbb{N}}$ se dice que es *de Cauchy* si cumple que $\|x_n - x_m\| \longrightarrow 0$ cuando n y m tienden a infinito. Un espacio vectorial se dice que es *completo* si, y solo si, toda sucesión de Cauchy converge a un elemento del espacio.

Definición B.5 (Espacio de Hilbert). Un espacio dotado de producto interior que es completo recibe el nombre de *espacio de Hilbert*.

Teorema B.1 (Teorema de Riesz-Fischer sobre la completitud de $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$, [Shumway y Stoffer 2017](#), p. 474). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias en $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$. Se tiene que existe una variable aleatoria $X \in (\mathbb{L}_2, \langle \cdot, \cdot \rangle)$ tal que $\|X_n - X\| \rightarrow 0$ si, y solo si se cumple que

$$\lim_{m \rightarrow \infty} \left(\sup_{n \geq m} \|X_n - X_m\| \right) = 0.$$

Dicho de otro modo, el espacio $(\mathbb{L}_2, \langle \cdot, \cdot \rangle)$ es un espacio de Hilbert.

Proposición B.2. Sea X una variable aleatoria tal que $X \in \mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$. En ese caso se tiene que $\mathbb{E}[|X|] < \infty$.

Observación B.1. La Proposición [B.2](#) puede generalizarse al caso en que X sea una variable aleatoria tal que $\mathbb{E}[|X|^p] < \infty$. En tal caso, para $r \in [1, p]$ se tendría que $\mathbb{E}[|X|^r] < \infty$.

B.2. Convergencia de series de variables aleatorias

Teorema B.3 ([Billingsley 1995](#), p. 289). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes con $\mathbb{E}[X_n] = 0$ para todo $n \in \mathbb{N}$. Si se cumple que

$$\sum_{n=0}^{\infty} \text{Var}[X_n] < \infty$$

entonces la serie $\sum_{n=0}^{\infty} X_n$ converge con probabilidad 1.

B.3. Desigualdades

Proposición B.4 (Desigualdad de Jensen). Sea X una variable aleatoria y sea φ una función convexa tal que $\mathbb{E}[\varphi(X)] < \infty$. Entonces se tiene que

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

Proposición B.5 (Desigualdad de Cauchy-Schwarz). Sean X e Y dos variables aleatorias con momento de orden 2 finito. Se tiene que

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]} \sqrt{\mathbb{E}[Y^2]}.$$

Proposición B.6. Sean X e Y dos variables aleatorias. Se tiene que $\text{Cov}[X, Y]^2 \leq \text{Var}[X] \text{Var}[Y]$.

B.4. El concepto de esperanza condicionada

B.4.1. Definición, existencia y unicidad de la esperanza condicionada

Definición B.6 (Esperanza condicionada a una variable aleatoria, [Sánchez Sellero 2023](#)). Sean X e Y dos variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Se define la esperanza condicionada $\mathbb{E}[Y | X = x]$ como una función medible de x tal que

$$\int_{\{X \in B\}} Y d\mathbb{P} = \int_B \mathbb{E}[Y | X = x] d\mu_X(x), \quad \forall B \in \mathcal{B}.$$

Definición B.7 (Esperanza condicionada a una σ -álgebra, [Sánchez Sellero 2023](#)). Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad, Y una variable aleatoria medible respecto de la σ -álgebra \mathcal{A} y $\mathcal{F} \subset \mathcal{A}$ una σ -álgebra. Se define la esperanza condicionada de Y a \mathcal{F} , $\mathbb{E}[Y | \mathcal{F}]$ como la única variable aleatoria tal que

- $\mathbb{E}[Y | \mathcal{F}]$ es una función Borel medible respecto de \mathcal{F} , y
- $\int_F Y d\mathbb{P} = \int_F \mathbb{E}[Y | \mathcal{F}] d\mathbb{P}, \quad \forall F \in \mathcal{F}.$

Observación B.2. Se tiene que $\mathbb{E}[Y | X] = \mathbb{E}[Y | \sigma(X)]$. Además, si $Y = \mathbb{I}_A$ entonces se tiene que

$$\mathbb{E}[\mathbb{I}_A | \mathcal{F}] = \mathbb{P}\{A | \mathcal{F}\}.$$

Observación B.3. Como \mathcal{F} es una σ -álgebra contenida en \mathcal{A} la variable aleatoria Y no es, en general, una función \mathcal{F} -medible.

Observación B.4. El Teorema [B.16](#) garantiza la existencia y unicidad de la esperanza condicionada. Para ello, considérese un espacio de medida (Ω, \mathcal{F}) y la medida sobre él (bien definida si $Y \geq 0$ ¹) dada por

$$\nu(F) = \int_F Y d\mathbb{P}, \quad \forall F \in \mathcal{F}.$$

En ese caso la medida ν es absolutamente continua² con respecto a la medida \mathbb{P} restringida a \mathcal{F} . Aplicando el Teorema [B.16](#) se deduce la existencia y unicidad (de forma casi segura) de una función \mathcal{F} -medible, que se denota por $\mathbb{E}[Y | \mathcal{F}]$ y que verifica que

$$\nu(F) = \int_F \mathbb{E}[Y | \mathcal{F}] d\mathbb{P}, \quad \forall F \in \mathcal{F}.$$

Definición B.8 (Esperanza condicionada a una familia de variables aleatorias). Sea $X_t, t \in I$, una familia de variables aleatorias definidas sobre el mismo espacio de probabilidad, donde I puede ser un intervalo de la recta real o $I = \{1, 2, \dots\}$. Se define $\sigma(X_t, t \in I) = \sigma(\cup_t \sigma(X_t))$ como la menor σ -álgebra que hace medibles todas las variables aleatorias de la familia. Los sucesos de tal σ -álgebra se definen en términos de lo que ha ocurrido con las variables X_t .

Para finalizar la introducción del concepto de esperanza condicionada, cabe mencionar que, tal y como se afirma en la página 223 de [Durrett \(2019\)](#), de manera intuitiva se acostumbra a pensar en la σ -álgebra \mathcal{F} como una manera de describir la información que se tiene a disposición³. Así pues, la esperanza condicionada $\mathbb{E}[Y | \mathcal{F}]$ se puede ver como la «mejor suposición o conjetura» del valor de la variable Y , dada la información de la que se dispone mediante \mathcal{F} . En la sección siguiente se incluirán algunas de las principales propiedades de esta variable aleatoria para ayudar a reforzar esta idea.

B.4.2. Propiedades de la esperanza condicionada

A lo largo de toda esta sección se considerará un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, una variable aleatoria Y medible respecto de la σ -álgebra \mathcal{A} y $\mathcal{F} \subset \mathcal{A}$ una σ -álgebra.

Proposición B.7. Si Y es \mathcal{F} -medible entonces $\mathbb{E}[Y | \mathcal{F}] = Y$.

¹ Luego se probaría el caso general descomponiendo la variable aleatoria en su parte positiva y negativa.

² Véase la definición [B.10](#).

³ Para cada $F \in \mathcal{F}$, se tiene conocimiento de si ha ocurrido F o no.

Observación B.5. La proposición anterior puede interpretarse como que, en el caso de tener la «información perfecta» y saber lo que le ha ocurrido a la variable Y , entonces la mejor suposición que se puede hacer sobre Y es ella misma.

Proposición B.8. Si $\mathcal{F} = \{\emptyset, \Omega\}$ entonces $\mathbb{E}[Y | \mathcal{F}] = \mathbb{E}[Y]$.

Proposición B.9. Si $\mathcal{F}_1 \subset \mathcal{F}_2$ entonces se tiene que

$$\mathbb{E}[\mathbb{E}[Y | \mathcal{F}_2] | \mathcal{F}_1] = \mathbb{E}[Y | \mathcal{F}_1] = \mathbb{E}[\mathbb{E}[Y | \mathcal{F}_1] | \mathcal{F}_2]$$

Proposición B.10 (Shao 2006, pp. 27-28). Si X es \mathcal{F} -medible y se cumple que $\mathbb{E}[|XY|] < \infty$ y que $\mathbb{E}[|Y|] < \infty$ entonces

$$\mathbb{E}[XY | \mathcal{F}] = X \mathbb{E}[Y | \mathcal{F}].$$

Proposición B.11. Si Y es independiente de \mathcal{F}^4 entonces $\mathbb{E}[Y | \mathcal{F}] = \mathbb{E}[Y]$.

Observación B.6. La proposición anterior puede verse de manera opuesta a lo comentado en la Observación B.5, esto es, que en el caso de no tener información alguna acerca de la variable Y la mejor suposición que se puede hacer sobre ella es su media.

Por último, se van a enumerar algunas otras propiedades de la esperanza condicionada que serán de utilidad en apartados posteriores de este capítulo⁵.

Proposición B.12 (Durrett 2019, pp. 226-228). Sean X e Y dos variables aleatorias medibles en el espacio $(\Omega, \mathcal{A}, \mathbb{P})$ y sea $\mathcal{F} \subset \mathcal{A}$ una σ -álgebra.

1) (Monotonía) Si $Y \leq X$ entonces $\mathbb{E}[Y | \mathcal{F}] \leq \mathbb{E}[X | \mathcal{F}]$.

2) (Linealidad) Se tiene que

$$\mathbb{E}[aY + bX | \mathcal{F}] = a\mathbb{E}[Y | \mathcal{F}] + b\mathbb{E}[X | \mathcal{F}], \quad \forall a, b \in \mathbb{R}.$$

3) (Desigualdad de Jensen) Supóngase que $\mathbb{E}[|X|] < \infty$ y que $\mathbb{E}[|\varphi(X)|] < \infty$ y sea φ una función convexa. Entonces se tiene que

$$\varphi(\mathbb{E}[X | \mathcal{F}]) \leq \mathbb{E}[\varphi(X) | \mathcal{F}].$$

4) Si $\mathbb{E}[|Y|] < \infty$ entonces se tiene que

$$\mathbb{E}[\mathbb{E}[Y | \mathcal{F}]] = \mathbb{E}[Y].$$

B.4.3. Proyecciones de variables aleatorias

Definición B.9 (Proyección en \mathbb{L}_2 , Van der Vaart 2000, p. 153). Sean T y $\{S | S \in \mathcal{S}\}$ variables aleatorias en $\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$. Se dice que una variable aleatoria \hat{S} es una proyección de T en \mathcal{S} si cumple que $\hat{S} \in \mathcal{S}$ y que

$$\hat{S} = \arg \min_{S \in \mathcal{S}} \mathbb{E}[(T - S)^2].$$

⁴ Se dice que Y es independiente de \mathcal{F} si lo son las σ -álgebras $\sigma(Y)$ y \mathcal{F} .

⁵ Para profundizar en otras propiedades de la esperanza condicionada (e.g., el Lema de Fatou o el Teorema de la Convergencia dominada) pueden consultarse las páginas 39 y 40 de Shao (2003). Cabe destacar que la demostración de las mismas, si bien en dicha referencia se dejan al lector, pueden ser encontradas en su mayoría en los ejercicios del 35 al 39 que figuran en Shao (2006).

Teorema B.13 (Caracterización de las proyecciones en \mathbb{L}_2 sobre espacios lineales, [Van der Vaart 2000](#), p. 153). Sean T y $\{S \mid S \in \mathcal{S}\}$ variables aleatorias en $\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$ y sea \mathcal{S} un espacio lineal de variables aleatorias con momento de orden 2 finito. Se tiene que \widehat{S} es una proyección de T en \mathcal{S} si, y solo si, $\widehat{S} \in \mathcal{S}$ y

$$\mathbb{E} \left[(T - \widehat{S}) S \right] = 0, \quad \forall S \in \mathcal{S}. \quad (\text{B.1})$$

Además, dos proyecciones de T sobre \mathcal{S} son iguales, salvo un conjunto de probabilidad 0. Si, además, el espacio lineal \mathcal{S} contiene a las variables constantes entonces se tiene que

$$\mathbb{E}[T] = \mathbb{E}[\widehat{S}], \quad \text{Cov}[T - \widehat{S}, S] = 0, \quad \forall S \in \mathcal{S}.$$

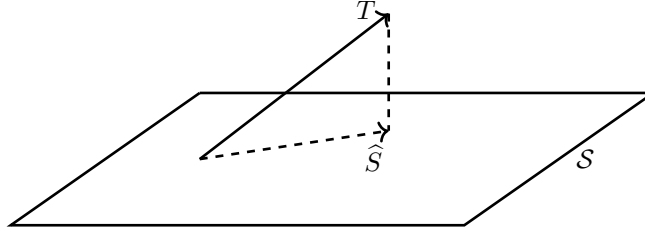


Figura B.1: Una variable aleatoria T y su proyección \widehat{S} sobre un espacio lineal \mathcal{S} . Basado en [Van der Vaart \(2000\)](#), p. 154.

Observación B.7. El Teorema B.13 no garantiza que la proyección exista, aunque una condición suficiente para la existencia es que \mathcal{S} sea cerrado con la norma de \mathbb{L}_2 . No obstante, hay muchas ocasiones en las que la existencia se establece más fácilmente comprobando que una variable \widehat{S} verifica la condición de ortogonalidad.

El siguiente resultado es el que otorga una mayor interpretación a $\mathbb{E}[X | Y]$, ya que, tal y como se comentaba de manera intuitiva al final de la Sección B.4.1, será la mejor predicción de X a partir de la información que da Y .

Teorema B.14 ([Shao 2003](#), p. 40). Sea X una variable aleatoria en el espacio $\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$ y sea Y una función medible de $(\Omega, \mathcal{A}, \mathbb{P})$ en (Ω', \mathcal{G}) . Sea, además, $g(Y)$ un predictor de la variable aleatoria X , con

$$g \in \mathfrak{N} = \left\{ g \text{ función Borel-medible} \mid \mathbb{E} \left[g(Y)^2 \right] < \infty \right\}.$$

Se tiene que la variable aleatoria $\mathbb{E}[X | Y]$ es el mejor predictor de X en el sentido de que

$$\mathbb{E}[X | Y] = \arg \min_{g \in \mathfrak{N}} \text{PMSE} [g(Y)] = \arg \min_{g \in \mathfrak{N}} \mathbb{E} \left[(X - g(Y))^2 \right].$$

B.4.4. Resultados auxiliares generales

NOTA (Sobre la definición de densidad de una medida, [Billingsley 1995](#), p. 213). Sea δ una función medible no negativa y sea ν una medida definida como

$$\nu(A) = \int_A \delta d\mu, \quad \forall A \in \mathcal{A}.$$

Nótese que no se ha asumido que δ sea integrable con respecto de μ . La medida ν que así ha sido definida se dice que tiene *densidad* δ con respecto de μ .

Teorema B.15 (Billingsley 1995, p. 214). Sean dos medidas μ y ν . Si ν tiene densidad δ con respecto de μ entonces se tiene que para cualquier f no negativa

$$\int f d\nu = \int f\delta d\mu.$$

Más aún, f (no necesariamente no negativa) es integrable con respecto de ν si, y solo si, $f\delta$ es integrable con respecto de μ , en cuyo caso se tiene que, además de la igualdad anterior, se cumple

$$\int_A f d\nu = \int_A f\delta d\mu, \quad \forall A \in \mathcal{A}.$$

Definición B.10 (Continuidad Absoluta, Sánchez Sellero 2023). Una medida ν se dice que es absolutamente continua con respecto de otra medida μ si para todo conjunto medible A tal que $\mu(A) = 0$ se tiene que $\nu(A) = 0$.

Observación B.8. Toda función de densidad induce una medida absolutamente continua con respecto a la medida de Lebesgue en \mathbb{R} . El Teorema de Radon-Nikodym garantiza que toda medida absolutamente continua tiene asociada una función de densidad respecto a la otra medida.

Teorema B.16 (Radon-Nikodym, Sánchez Sellero 2023). Sean dos medidas ν y μ tales que ν es absolutamente continua con respecto a μ . Entonces existe una función medible no negativa f , tal que

$$\nu(A) = \int_A f d\mu, \quad \forall A \in \mathcal{A}.$$

Luego ν tiene densidad f con respecto de μ . Además, en el caso de que existiesen dos funciones f y g cumpliendo tal condición se tendría que $\mu(f \neq g) = 0$.

Breve comentario de la demostración. Por el Teorema B.15 las integrales con respecto a ν de funciones medibles h pueden ser obtenidas mediante la fórmula

$$\int_A h d\nu = \int_A hf d\mu. \tag{B.2}$$

La densidad cuya existencia se pretende probar recibe el nombre de *derivada de Radon-Nikodym de ν con respecto de μ* y con frecuencia se denota por $d\nu/d\mu$. Nótese que la ecuación (B.2) puede reescribirse como

$$\int_A h d\nu = \int_A h \frac{d\nu}{d\mu} d\mu$$

■

Observación B.9. Luego si se consideran dos variables aleatorias se puede hablar de la continuidad absoluta de las medidas que inducen sus distribuciones y es posible calcular también la densidad de una respecto a la distribución de la otra, que se obtiene haciendo el cociente de sus densidades.

Apéndice C

Código de R empleado

C.1. Test de estructuras simples en regresión

```
#####  
# LIBRERÍAS NECESARIAS  
#####  
library(PLRModels) # Para calcular ventana por validación cruzada  
library(tseries)  
library(forecast)  
library(parallel) # Para la paralelización  
library(cubature) # Integración en varias variables  
#####  
# KERNEL DE EPANECHNIKOV  
#####  
kernel.ep <- function(u){prod(ifelse(abs(u)<=1,  
                                (3/4)*(1-u^2),  
                                0))}  
#####  
# ESTADÍSTICO DE TEST  
#####  
integrando <- function(x, h, Xt, errores, kernel){  
  d <- ncol(Xt)  
  n <- nrow(Xt)  
  integ <- sum(sapply(1:n,  
                    function(i){  
                      (h^(-d))*kernel((x-Xt[i, ])/h)*errores[i]  
                    })))^2  
  return(integ)}  
  
S.T <- function(h, covar, errores, kernel){  
  n <- nrow(covar)  
  integrando.fun <- function(x){integrando(x, h = h,  
                                           Xt = covar,  
                                           errores = errores,  
                                           kernel = kernel)}
```

```

res <- (1/n^2)*cubintegrate(Vectorize(integrando.fun),
                           lower = rep(-Inf, ncol(covar)),
                           upper = rep(Inf, ncol(covar)),
                           fDim = ncol(covar),
                           method = "pcubature",
                           relTol = 1e-02,
                           absTol = 1e-02)$integral

return(res)}
#####
# TEST BOOTSTRAP DE KREISS
#####
kreiss.test <- function(muestra, res.parametric, B = 500, alpha = 0.05, kernel){
  muestra <- as.matrix(muestra)
  p <- ncol(muestra)
  n <- nrow(muestra)

  # Ajuste tipo Nadaraya-Watson (No paramétrico) de la función de
  # regresión
  bw0 <- np::npregbw(xdat = muestra[, 2:p],
                    ydat = muestra[,1],
                    regtype = "lc",
                    ckertype = "epanechnikov")
  residuos <- np::npreg(bws = bw0,
                      residual = TRUE)$resid

  # Estadístico S_T sobre la muestra
  S.sample <- S.T(h = bw0$bw,
                covar = as.matrix(muestra[,2:p]),
                errores = res.parametric,
                kernel = kernel.ep)

  # Errores bootstrap - Cada fila es una réplica
  rwild <- matrix(sample(c((1 - sqrt(5))/2, (1 + sqrt(5))/2),
                       n*B,
                       replace = TRUE,
                       prob = c((5+sqrt(5))/10, 1-(5+sqrt(5))/10)),
                 ncol = n)

  # Réplicas bootstrap del estadístico S_T
  S.T.boot <- as.numeric(sapply(1:B,
                              function(b){
                                S.T(h = bw0$bw,
                                    covar = as.matrix(muestra[,2:p]),
                                    errores = residuos*rwild[b,],
                                    kernel = kernel.ep)
                              })))

  cuantil <- quantile(S.T.boot, 1-alpha)
  p.valor <- mean(S.T.boot>S.sample)
  decision <- ifelse(S.sample>cuantil, "Rechazar H0", "No rechazar H0")

```

```

    resultados <- list(S.T = S.sample,
                     Valor_critico = cuantil,
                     p.valor = p.valor,
                     Decision = decision)

    return(resultados)}
#####
# EJEMPLO DE SIMULACIÓN CON UN MODELO PARAMÉTRICO (HO CIERTA)
#####
# Generación de las N muestras artificiales del modelo mediante paralelización
modelo1 <- function(n){
  x1 <- numeric(n)
  x1[1] <- 0
  for (t in 2:n) {x1[t] <- -0.9 * x1[t - 1] + rnorm(1)}
  datos1 <- as.matrix(cbind(x1, c(0, x1[-n])))
  return(datos1)}

alpha <- 0.05
n <- 100
B <- 200

ncores <- detectCores()
cl <- makeCluster(ncores, type = "PSOCK") # Creamos un cluster
parallel::clusterSetRNGStream(cl)

# Cargar funciones y variables en todos los nodos del cluster
clusterExport(cl, c("kernel.ep", "S.T", "kreiss.test", "modelo1",
                   "B", "alpha", "n"))
clusterEvalQ(cl, {
  library(PLRModels)
  library(cubature)
  library(tseries)
  library(forecast)
})

N <- 500
system.time(res <- parLapply(cl, 1:N,
  function(isim){
    sample.i <- modelo1(n)
    ajuste1.i <- Arima(sample.i[, 1],
                      order=c(1,0,0),
                      lambda=NULL,
                      include.constant=FALSE)
    res.ajuste1.i <- ajuste1.i$residuals
    kreiss.test(muestra = sample.i,
                res.parametric = res.ajuste1.i,
                B = B,
                alpha = alpha,
                kernel = kernel.ep)
  })))

stopCluster(cl)

```

```

# Tiempo de ejecución con un equipo con 3 cores
# user system elapsed
# 1.69    1.11 6105.91

# Guardado de los resultados
# save(res, file = "Modelo1_boot_test.RData")

# Decisiones de los test y proporción de rechazos
decisiones <- do.call(rbind, lapply(res, function(x) x$Decision))

(rechazos <- apply(decisiones=="Rechazar H0", 2, mean))
# 95%
# 0.064

```

C.2. Modelos ajustados con datos reales

```

# MH de NOx de un año (salvo día episodio a predecir)
fichero <- paste("MH.dat", sep="")
abrir=open_file(fichero) # Lectura archivos con control errores (autoría no propia)
if(!abrir$error){
  aux=matrix(scan(abrir$con, skip=1, nlines=4000, quiet=T),
             4000, 8, byrow=TRUE)
  aux=matrix(aux[!is.na(aux)], ncol=8, byrow=F)
  colnames(aux) <- c("NOx_MH(t-5)", "NOx_MH(t)", "NOx_MH(t+30)",
                    "Año", "Mes", "Dia", "Hora", "Min")

  aux <- aux[aux[,1]!=-1&aux[,2]!=-1&aux[,3]!=-1,] # Elimina datos missing
  aux <- aux[aux[,5]!="10"& aux[,6]!="25", ]      # Elimina episodio a predecir
  close(abrir$con)}

# Muestra basada en la matriz histórica
orig_sample <- as.data.frame(aux[,1:3])
colnames(orig_sample) <- c("x1", "x2", "y")

# Datos del episodio a predecir
episodio <- leer_matriz(fecha = fecha_episodio, directorio=dir, skip = 2,
                      dim1=1440, dim2=5+17*15)
datos_episodio <- episodio$datos[,11,7] # Se toman los datos de NOx (11) de la
# estación que ocupa la columna 7
datos_episodio <- data.frame("Hora"=as.POSIXct(paste(fecha_episodio,
                                                    rownames(episodio$datos[, ,7])),
                              sep=" "),
                            format="%Y-%m-%d %H:%M"),
                          "Valor"=as.numeric(datos_episodio))

# Se definen X(t-5), X(t) en el episodio a predecir
load("datosano.RData") # Una lista: elemento 1 para datos durante un año de SO2,
#                       elemento 2, análogo para NOx.

```

```

saltos=matrix(c(5,30) ,nrow=2, ncol=length(c("SO2_MH","NOx_MH")),
              dimnames=list(NULL,c("SO2_MH","NOx_MH")))
fecha=as.POSIXlt(dimnames(datosano[[1]])[[1]])
fecha=matrix(c(format(fecha,"%Y"),
                 format(fecha,"%m"),
                 format(fecha,"%d"),
                 format(fecha,"%H"),
                 format(fecha,"%M")),
              ncol=5,
              dimnames=list(NULL,c("Ano","Mes","Dia","Hora","Min")))

ini_episodio <- which(rownames(datosano[[2]])== datos_episodio$Hora[1])
durac_episodio <- dim(datos_episodio)[1]

raw.data<-matrix(as.numeric(
c(datosano[[2]][ini_ep:(ini_ep+durac_ep-1-sum(saltos[,2])], 7),
  datosano[[2]][(ini_ep+saltos[1,2]):(ini_ep+durac_ep-1-saltos[2,2]), 7],
  datosano[[2]][(ini_ep+sum(saltos[,2])):(ini_ep+durac_ep-1), 7],
  fecha[(ini_ep+sum(saltos[,2]):(ini_ep+durac_ep-1),]),
  c(durac_ep-sum(saltos[,2]), dim(saltos)[1]+6))

# Se corrigen los valores missing
raw.data <- raw.data[raw.data[,1]!=-1&raw.data[,2]!=-1&raw.data[,3]!=-1,]
newdata <- data.frame("x1"=raw.data[,1], "x2"=raw.data[,2])

datos_episodio <- datos_episodio[datos_episodio$Valor!=-1,]
#####
# MODELO PARAMÉTRICO DE TIPO BOX - JENKINS
#####
library(fpp2)
train <- datos_episodio[1:(13*60+1),]
test <- datos_episodio[(13*60+2):(13*60+30+1),]

# ETAPA 1 - SELECCIÓN DE MODELOS
auto.arima(train$Valor, trace = T, stepwise=F, approximation = FALSE, seasonal = F)
# modelo sugerido: ARIMA(1,2,2), AICc=-817.9

# ETAPA 2 - AJUSTE DEL MODELO
ajuste <- Arima(train$Valor, order=c(1,2,2), lambda=0, include.constant=F)

# Significación de parámetros del modelo ajustado
abs(ajuste$coef) < 1.96*sqrt(diag(ajuste$var.coef)) # Todos significativos
# ar1 ma1 ma2
# FALSE FALSE FALSE

# ETAPA 3 - DIAGNOSIS DEL MODELO
checkresiduals(ajuste)

qqnorm(ajuste$residuals)
qqline(ajuste$residuals)
shapiro.test(ajuste$residuals)

```

```

# Shapiro-Wilk normality test
# data: ajuste$residuals
# W = 0.9006, p-value < 2.2e-16

# ETAPA 4 - PREDICCIONES Y REPRESENTACIÓN GRÁFICA
pred_BJ <- data.frame("Valor"=forecast(ajuste, h = 30, bootstrap=T)$mean,
                     "Hora"=test$Hora)

ggplot() +
  geom_line(data = test, aes(x=Hora, y=Valor), size=1) +
  geom_line(data = pred_BJ, aes(x=Hora, y=Valor), size=1, col = "red") +
  scale_x_datetime(date_labels = "%H:%M", date_breaks = "5 min") +
  labs(title="", x="Hora", y="Valor") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#####
# MODELO NO PARAMÉTRICO
#####
library(np)
library(sm)

# Criterio de ventana por validacion cruzada
bw <- np::npregbw(y~x1+x2, data=orig_sample, regtype="lc", ckertype="gaussian")
np_model <- np::npreg(bws = bw, residuals = T)

# PREDICCIONES Y REPRESENTACIÓN GRÁFICA
# Se evita el problema de que no haya datos para valores de X1 y X2 de newdata
# en orig_sample
pred_NP <- numeric(length = dim(newdata)[1])
for(i in 1:dim(newdata)[1]){
  vec <- as.numeric(newdata[i,1:2])
  mat <- as.matrix(orig_sample)[,1:2]
  distancias <- apply(mat, 1, function(row) {sqrt(sum((row - vec)^2))})
  pred_NP[i] <- np_model$mean[which.min(distancias)]}

pred_NP <- data.frame("Valor"=pred_NP,
                     "Hora"=as.POSIXct(paste(paste(raw.data[,4],
                                                    raw.data[,5],
                                                    raw.data[,6],
                                                    sep = "-"),
                                           paste(raw.data[,7],
                                                    raw.data[,8], sep = ":"),
                                           sep = " "),format="%Y-%m-%d %H:%M"))

ggplot() +
  geom_line(data = datos_episodio[datos_episodio$Hora %in% pred_NP$Hora,],
           aes(x=Hora, y=Valor), size=1) +
  geom_line(data = pred_NP, aes(x=Hora, y=Valor), size=1, col = "red") +
  scale_x_datetime(date_labels = "%H:%M", date_breaks = "1 hour") +
  labs(title="", x="Hora", y="Valor") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

```
#####
# MODELO SEMIPARAMÉTRICO
#####
# Residuos de la estimación no paramétrica en la muestra de entrenamiento
res_np <- datos_episodio[1:length(train$Valor),"Valor"]-
  pred_NP2$Valor[1:length(train$Valor)]

# ETAPA 1 - SELECCIÓN DE MODELOS
auto.arima(res_np, trace = T, stepwise=F, approximation = FALSE, seasonal = F)
# modelo sugerido: ARIMA(3, 1, 1), AICc=5485.8

# ETAPA 2 - AJUSTE DEL MODELO
ajuste <- Arima(res_np, order=c(3,1,1), lambda=0, include.constant=F)

# Significación de parámetros del modelo ajustado
abs(ajuste$coef) < 1.96*sqrt(diag(ajuste$var.coef))
# ar1 ar2 ar3 ma1
# FALSE TRUE TRUE FALSE

abs(ajuste$coef)/(1.96*sqrt(diag(ajuste$var.coef))) # Se fija a 0 el de menor valor

parametros.fijos <- rep(NA, 4)
parametros.fijos[3] <- 0
ajuste <- Arima(res_np, order=c(3,1,1), lambda=0, include.constant=F,
  fixed = parametros.fijos)
abs(ajuste$coef[ajuste$coef != 0])<(1.96*sqrt(diag(ajuste$var.coef)))
# ar1 ar2 ma1
# FALSE TRUE FALSE

parametros.fijos[2] <- 0
ajuste <- Arima(res_np, order=c(3,1,1), lambda=0, include.constant=F,
  fixed = parametros.fijos)
abs(ajuste$coef[ajuste$coef != 0])<(1.96*sqrt(diag(ajuste$var.coef))) #
# ar1 ma1
# FALSE FALSE

# ETAPA 3 - DIAGNOSIS DEL MODELO
checkresiduals(ajuste)
t.test(ajuste$residuals, mu=0)
# One Sample t-test #
# data: ajuste$residuals
# t = -0.26497, df = 238, p-value = 0.7913
# (...)

qqnorm(ajuste$residuals)
qqline(ajuste$residuals) # se ve la simetría
shapiro.test(ajuste$residuals)
# Shapiro-Wilk normality test #
# data: ajuste$residuals
# W = 0.77858, p-value < 2.2e-16
```

```

# ETAPA 4 - PREDICCIONES
pred_res_BJ <- data.frame("Valor"=forecast(ajuste, h = 30)$mean,
                        "Hora"=test$Hora)
t <- which(pred_NP2$Hora==head(test$Hora)[1])
pred_semipar <- data.frame("Valor"=pred_res_BJ$Valor+pred_NP2$Valor[t:(t+29)],
                        "Hora"=test$Hora)

ggplot() +
  geom_line(data = test, aes(x=Hora, y=Valor), size=1) +
  geom_line(data = pred_semipar, aes(x=Hora, y=Valor), size=1, col = "red") +
  geom_line(data = pred_BJ, aes(x=Hora, y=Valor), size=1, col = "blue") +
  geom_line(data = pred_NP2, aes(x=Hora, y=Valor), size=1, col = "green") +
  scale_x_datetime(date_labels = "%H:%M", date_breaks = "5 min",
                  limits = c(min(test$Hora), max(test$Hora))) +
  labs(title="", x="Hora", y="Valor") +
  ylim(125,250) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#####
# MODELO BASADO EN REDES NEURONALES
#####
library(nnet)

# Definición y entrenamiento de la red (linout=TRUE para una función de activación
#                                     igual a la identidad en la capa oculta)
set.seed(12345)
red <- nnet(y~x1+x2, data = orig_sample, size = 40, linout = TRUE, maxit = 1000,
           abstol = 1.0e-6)

# Predicciones basadas en la red neuronal
nnet_pred <- data.frame("Valor"=predict(red, newdata = newdata),
                      "Hora"=as.POSIXct(paste(paste(raw.data[,4],
                                                    raw.data[,5],
                                                    raw.data[,6], sep = "-"),
                                              paste(raw.data[,7],
                                                    raw.data[,8],
                                                    sep = ":"),
                                              sep = " "),
                      format="%Y-%m-%d %H:%M"))

ggplot() +
  geom_line(data = datos_episodio[datos_episodio$Hora %in% nnet_pred$Hora,],
           aes(x=Hora, y=Valor), size=1) +
  geom_line(data = nnet_pred, aes(x=Hora, y=Valor), size=1, col = "red") +
  scale_x_datetime(date_labels = "%H:%M", date_breaks = "1 hour") +
  labs(title="", x="Hora", y="Valor") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#####
# REPRESENTACIÓN CONJUNTA
#####

```



```
#####  
ggplot() +  
geom_line(data = test, aes(x=Hora, y=Valor), size=1) +  
geom_line(data = datos_episodio[datos_episodio$Hora %in% nnet_pred$Hora,],  
          aes(x=Hora, y=Valor), size=1, col = "yellow", linetype="dashed") +  
geom_line(data = pred_semipar, aes(x=Hora, y=Valor), size=1, col = "red") +  
geom_line(data = pred_BJ, aes(x=Hora, y=Valor), size=1, col = "blue") +  
geom_line(data = pred_NP2, aes(x=Hora, y=Valor), size=1, col = "green") +  
scale_x_datetime(date_labels = "%H:%M", date_breaks = "5 min",  
                 limits = c(min(test$Hora), max(test$Hora))) +  
ylim(125,250) +  
labs(title="", x="Hora", y="Valor") +  
theme_minimal() +  
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Apéndice D

Distribuciones notables

En este apéndice se recogen algunas de las características más notables de las distribuciones que se han mencionado a lo largo de todo el trabajo.

D.1. Distribución Normal univariante

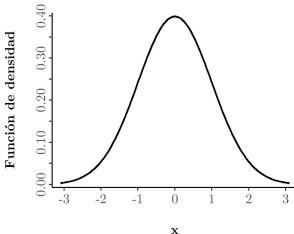
$X \in N(\mu, \sigma^2)$	
Normal(0, 1)	
	
Parámetros	$\mu \in \mathbb{R}, \sigma^2 > 0$
Función de densidad	$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
Función de distribución	$F(x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$
Media	μ
Varianza	σ^2

Tabla 2: La distribución normal univariante.

D.2. Distribución Normal multivariante

Definición D.1 (Normal multivariante no singular, [Tong 1990](#), p. 26). Un vector aleatorio \mathbf{X} d -dimensional con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ se dice que sigue una *distribución Normal multivariante no singular*, simbólicamente expresado como $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, si

- 1) $\boldsymbol{\Sigma}$ es definida positiva, y
- 2) la función de densidad de \mathbf{X} es de la forma

$$f(\mathbf{x}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad \mathbf{x} \in \mathbb{R}^d,$$

donde $|\boldsymbol{\Sigma}|$ representa al determinante de dicha matriz.

Definición D.2 (Normal multivariante singular, [Tong 1990](#), p. 28). Un vector aleatorio \mathbf{X} d -dimensional con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ se dice que sigue una *distribución Normal multivariante singular*, simbólicamente expresado como $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $|\boldsymbol{\Sigma}| = 0$, si

- 1) $\boldsymbol{\Sigma}$ es semidefinida positiva, y
- 2) para algún $r < d$ existe una matriz $\mathbf{C} \in \mathbb{R}^{d \times r}$ tal que \mathbf{X} y $\mathbf{C}\mathbf{Z}_r + \boldsymbol{\mu}$ tienen la misma distribución, donde $\mathbf{Z}_r \sim N_r(\mathbf{0}, \mathbf{I}_r)$.

Definición D.3 (Normal multivariante, [Tong 1990](#), p. 28). Un vector aleatorio \mathbf{X} d -dimensional con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ se dice que sigue una *distribución Normal multivariante*, simbólicamente expresado como $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ si \mathbf{X} sigue una distribución Normal univariante singular o una no singular.

Teorema D.1 (Caracterización de la normalidad multivariante, [Tong 1990](#), pp. 29-30). *Un vector aleatorio \mathbf{X} d -dimensional con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ sigue una distribución Normal multivariante si, y solo si*

$$\mathbf{c}'\mathbf{X} \sim N_1(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}), \quad \forall \mathbf{c} \in \mathbb{R}^d.$$

Ejemplo D.1 (Normal multivariante no singular estándar). Considérese el vector $\mathbf{Z} = (Z_1, Z_2, \dots, Z_d)'$ con vector de medias $\mathbf{0}$ y matriz de covarianzas $\boldsymbol{\Sigma} = \mathbf{I}_d$. En este caso, se puede escribir $\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$ con la función de densidad dada por

$$f(\mathbf{z}) = (2\pi)^{-d/2} \exp \left\{ -\frac{1}{2} \sum_{j=1}^d z_j^2 \right\}, \quad \mathbf{z} \in \mathbb{R}^d.$$

A partir de la función de densidad anterior es inmediato ver que

$$f(\mathbf{z}) = (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} z_1^2 \right\} \cdots (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} z_d^2 \right\} = f(z_1) \cdots f(z_d), \quad (\text{D.1})$$

donde f es la función de densidad de una $N(0, 1)$. Dicho de otro modo, la incorrelación de las variables Z_i , $i = 1, 2, \dots, d$, implica su independencia, lo cual se enunciará de manera más general a continuación, en el Teorema [D.4](#). Por otro lado, a partir de [\(D.1\)](#) puede verse que las distribuciones marginales 1-dimensionales son $N(0, 1)$, tal y como cabía esperar a partir del Teorema [D.1](#) tomando como vector \mathbf{c} aquel que tiene un 1 en una posición y el resto son 0.

Teorema D.2 (Normalidad bajo transformaciones afines, [Tong 1990](#), p. 32). Sea \mathbf{X} un vector aleatorio d -dimensional tal que $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y sea otro vector aleatorio \mathbf{Y} tal que $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, donde \mathbf{A} es una matriz $m \times d$ y $\mathbf{b} \in \mathbb{R}^m$. Entonces se tiene que $\mathbf{Y} \sim N_m(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$, $\boldsymbol{\Sigma}_Y > 0$, donde

$$\boldsymbol{\mu}_Y = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \quad \boldsymbol{\Sigma}_Y = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.$$

Teorema D.3 (Distribuciones marginales, [Tong 1990](#), p. 30). Sea \mathbf{X} un vector aleatorio d -dimensional tal que $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y para $k < d$ considérense las particiones de \mathbf{X} , $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ siguientes:

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)', \quad \boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)', \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad (\text{D.2})$$

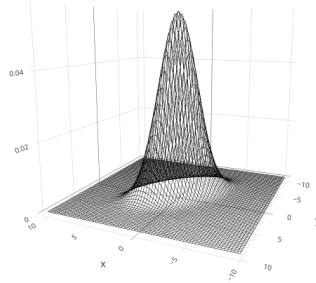
donde

$$\begin{aligned} \mathbf{X}_1 &= (X_1, X_2, \dots, X_k)', & \mathbf{X}_2 &= (X_{k+1}, X_{k+2}, \dots, X_d)', \\ \boldsymbol{\mu}_1 &= (\mu_1, \mu_2, \dots, \mu_k)', & \boldsymbol{\mu}_2 &= (\mu_{k+1}, \mu_{k+2}, \dots, \mu_d)', \end{aligned}$$

$\boldsymbol{\Sigma}_{12} = (\sigma_{ij})$, con $\sigma_{ij} = \text{Cov}[X_i, X_j]$, para $1 \leq i < j \leq d$, y $\boldsymbol{\Sigma}_{ii}$ es la matriz de covarianzas de \mathbf{X}_i , $i = 1, 2$. En tal caso se tiene que para cada k las distribuciones marginales de \mathbf{X}_1 y \mathbf{X}_2 son, respectivamente, $N_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ y $N_{d-k}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

Teorema D.4 ([Tong 1990](#), p. 31). Sea \mathbf{X} un vector aleatorio d -dimensional tal que $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y considérense las particiones definidas en (D.2). Se tiene que \mathbf{X}_1 y \mathbf{X}_2 son independientes si, y solo si, $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

$$X \in N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



Parámetros	$\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ semidefinida positiva
Función de densidad	$f(\mathbf{x}) = (2\pi)^{-d/2} \boldsymbol{\Sigma} ^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$, $\boldsymbol{\Sigma} > 0$
Vector de medias	$\boldsymbol{\mu}$
Matriz de covarianzas	$\boldsymbol{\Sigma}$

Tabla 3: La distribución normal multivariante.
En la figura, $\boldsymbol{\mu} = \mathbf{0}$ y $\boldsymbol{\Sigma} = ((4, 2)', (2, 3)')$.

D.3. Distribución χ^2

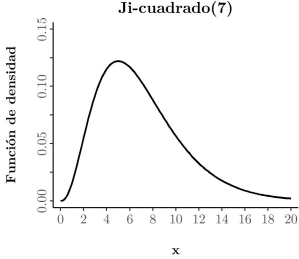
$X \in \chi^2(k)$	
 <p style="text-align: center;">Ji-cuadrado(7)</p>	
Parámetros	$k \in \mathbb{N} \setminus \{0\}$
Función de densidad	$f(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2}$
Función de distribución	$F(x) = \frac{1}{2^{k/2} \Gamma(k/2)} \int_0^x t^{k/2-1} e^{-t/2} dt$
Media	k
Varianza	$2k$

Tabla 4: La distribución χ^2 .

Bibliografía

- Aneiros, G. (2023). Apuntes de la asignatura Series de Tiempo del Máster en Técnicas Estadísticas.
- Bertail, P., Doukhan, P., y Soulier, P. (2006). *Dependence in probability and statistics*. Springer New York.
- Besse, P. C., Cardot, H., y Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27(4):673–687.
- Bhattacharya, R. N. y Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.*, 6(2):434–451.
- Billingsley, P. (1995). *Probability and measure*. John Wiley & Sons.
- Bobrowski, A. (2005). *Functional analysis for probability and stochastic processes: An introduction*. Cambridge University Press.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., y Ljung, G. M. (2008). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Boznar, M., Lesjak, M., y Mlakar, P. (1993). A neural network-based method for short-term predictions of ambient SO₂ concentrations in highly polluted industrial areas of complex terrain. *Atmospheric environment. Part B. urban atmosphere*, 27(2):221–230.
- Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107 – 144.
- Brockwell, P. J. y Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing.
- Bühlmann, P., Doukhan, P., y Nze, P. A. (2002). Weak dependence beyond mixing and asymptotics for nonparametric regression. *The Annals of Statistics*, 30(2):397–430.
- Cámara Oficial Minera de Galicia (28 de mayo de 2024a). Mina de As Pontes. En <https://patrimonio.camaraminera.org/gl/lugar/mina-de-pontes>.
- Cámara Oficial Minera de Galicia (28 de mayo de 2024b). Mina de Meirama. En <https://patrimonio.camaraminera.org/gl/lugar/mina-de-meirama>.
- Cao, R. (1994). Un estudio de simulación comparativo de técnicas no paramétricas, semiparamétricas y box-jenkins para la predicción con datos dependientes. *Estadística española*, 36(135):5–20.
- Cao, R. (1999). An overview of bootstrap methods for estimating and predicting in time series. *Test*, 8(1):95–116.

- Cao, R., Febrero-Bande, M., González-Manteiga, W., Prada-Sánchez, J., y García-Jurado, I. (1997). Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. *Communications in Statistics-Simulation and Computation*, 26(3):961–978.
- Cao, R. y Fernández, R. (2023). Apuntes de la asignatura Técnicas de Remuestreo del Máster en Técnicas Estadísticas.
- Ciarlet, P., Miara, B., y Thomas, J. (1989). *Introduction to numerical linear algebra and optimisation*. Cambridge Texts in Applied Mathematics. Cambridge University Press.
- Conde-Amboage, M., González-Manteiga, W., y Sánchez-Sellero, C. (2017). Predicting trace gas concentrations using quantile regression models. *Stochastic Environmental Research and Risk Assessment*, 31:1359–1370.
- Cramér, H. (1928a). On the composition of elementary errors: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.
- Cramér, H. (1928b). On the composition of elementary errors: Statistical applications. *Scandinavian Actuarial Journal*, 1928(1):141–180.
- Davison, A. C. y Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press.
- Durrett, R. (2019). *Probability: Theory and examples*, volume 49. Cambridge university press.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.
- Efron, B. y Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the econometric society*, pages 987–1007.
- Engle, R. F. y Granger, C. W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276.
- Fan, J., Härdle, W., y Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *The Annals of Statistics*, 26(3):943–971.
- Fan, J. y Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *Test*, 16:409–444.
- Febrero, M., Galeano, P., y González-Manteiga, W. (2007). A functional analysis of NOx levels: location and scale estimation and outlier detection. *Computational Statistics*, 22:411–427.
- Febrero, M., Galeano, P., y González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics: The official journal of the International Environmetrics Society*, 19(4):331–345.
- Fernández-de Castro, B. y González-Manteiga, W. (2008). Boosting for real and functional samples: An application to an environmental problem. *Stochastic Environmental Research and Risk Assessment*, 22:27–37.
- Fernández-de Castro, B., Guillas, S., y González-Manteiga, W. (2005). Functional samples and bootstrap for predicting sulfur dioxide levels. *Technometrics*, 47(2):212–222.

- Fernández-de Castro, B. M., Sánchez, J. M. P., Manteiga, W. G., Bande, M. F., Cela, J. L. B., y Fernández, J. J. H. (2003). Prediction of SO₂ levels using neural networks. *Journal of the Air & Waste Management Association*, 53(5):532–539.
- Ferraty, F. (2006). *Nonparametric functional data analysis*. Springer.
- Fraiman, R. y Muniz, G. (2001). Trimmed means for functional data. *Test*, 10:419–440.
- Franke, J., Kreiss, J.-P., y Mammen, E. (2002). Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli*, 8(1):1–37.
- Franke, J. y Wendel, M. (1992). A bootstrap approach for nonlinear autoregressions - some preliminary results. In *Bootstrapping and Related Techniques: Proceedings of an International Conference, Held in Trier, FRG, June 4–8, 1990*, pages 101–105. Springer.
- Galeano, P. (2023). Apuntes de la asignatura del Máster en Técnicas Estadísticas Ingeniería Financiera.
- García-Jurado, I., González-Manteiga, W., Prada-Sánchez, J., Febrero-Bande, M., y Cao, R. (1995). Predicting using Box-Jenkins, nonparametric, and bootstrap techniques. *Technometrics*, 37(3):303–310.
- García-Nieto, P. J., Sánchez-Lasheras, F., García-Gonzalo, E., y de Cos-Juez, F. (2018). Estimation of PM 10 concentration from air quality data in the vicinity of a major steelworks site in the metropolitan area of Avilés (Northern Spain) using machine learning techniques. *Stochastic environmental research and risk assessment*, 32:3287–3298.
- González-Manteiga, W., Piñeiro-Lamas, M., y Febrero-Bande, M. (2009). Multidimensional semiparametric prediction with cointegration in errors for pollution indicators.
- Granger, C. W. (1983). *Co-integrated variables and error-correcting models*. PhD thesis, Discussion Paper 83-13. Department of Economics, University of California at San Diego.
- Green, P., Jennison, C., y Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(2):299–315.
- Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.
- Hall, P., Horowitz, J. L., y Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.
- Hardle, W. y Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, pages 1926–1947.
- Härdle, W. y Vieu, P. (1992). Kernel regression smoothing of time series. *Journal of Time Series Analysis*, 13(3):209–232.
- Hastie, T. y Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall / CRC Monographs on Statistics & Applied Probability. Taylor & Francis.
- Hastie, T., Tibshirani, R., Friedman, J. H., y Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, volume 2. Springer.
- Horowitz, J. (1998). *Semiparametric methods in Econometrics*. Lecture Notes in Statistics. Springer New York.
- Horowitz, J. L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, 69(2):499–513.

- Horváth, L. y Kokoszka, P. (2012). *Inference for functional data with applications*. Springer Series in Statistics. Springer New York.
- Hsu, K.-J. (1992). Time series analysis of the interdependence among air pollutants. *Atmospheric Environment. Part B. Urban Atmosphere*, 26(4):491–503.
- Kadiyala, A. y Kumar, A. (2014). Vector time series models for prediction of air quality inside a public transportation bus using available software. *Environmental Progress & Sustainable Energy*, 33(4):1069–1073.
- Koenker, R. y Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.
- Kolassa, J. E. (2006). *Series approximation methods in statistics*, volume 88. Springer Science & Business Media.
- Kreiss, J.-P. (1985). A note on M-estimation in stationary ARMA processes. *Statistics & Risk Modeling*, 3(3-4):317–336.
- Kreiss, J.-P. y Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, 13(4):297–317.
- Kreiss, J.-P., Neumann, M. H., y Yao, Q. (2008). Bootstrap tests for simple structures in nonparametric time series regression. *Statistics and its Interface*, 1(2):367–380.
- Kreutzberger, E. (1993). *Bootstrap für nichtlineare AR (1)-Prozesse*. Universität Kaiserslautern.
- Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17:1217–1241.
- Liu, R. Y. y Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the limits of bootstrap*, pages 225–248.
- López-Pintado, S. y Romo, J. (2007). Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 51(10):4957–4968.
- López-Pintado, S. y Romo, J. (2009). On the concept of depth for functional data. *Journal of the American statistical Association*, 104(486):718–734.
- Mammen, E. (1992). *When does bootstrap work? Asymptotic results and simulations*. Lecture notes in statistics. Springer New York, NY.
- Martínez-Silva, I., Roca-Pardiñas, J., y Ordóñez, C. (2016). Forecasting SO2 pollution incidents by means of quantile curves based on additive models. *Environmetrics*, 27(3):147–157.
- Masry, E. (1996). Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101.
- McCullagh, P. y Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall / CRC.
- McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., y Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269.
- Monforte, C. (27 de diciembre de 2019). Endesa solicita oficialmente el cierre de sus centrales de Almería y As Pontes. *CincoDías (El País)*. https://cincodias.elpais.com/cincodias/2019/12/27/companias/1577459963_683477.html.
- Oviedo-de la Fuente, M., Ordóñez, C., y Roca-Pardiñas, J. (2020). Functional location-scale model to forecast bivariate pollution episodes. *Mathematics*, 8(6):941.

- Paparoditis, E. y Politis, D. N. (2000). The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics*, 52:139–159.
- Peña, D. (2005). *Análisis de series temporales*. Ciencias sociales. Alianza.
- Pérez, P., Trier, A., y Reyes, J. (2000). Prediction of PM_{2.5} concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment*, 34(8):1189–1196.
- Petrov, V. y Mordecki, E. (2008). *Teoría de la probabilidad*. DIRAC.
- Politis, D. N. y McElroy, T. S. (2020). *Time series: A first course with bootstrap starter*. CRC Press.
- Politis, D. N. y Romano, J. P. (1994a). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050.
- Politis, D. N. y Romano, J. P. (1994b). The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313.
- Prada-Sánchez, J., Febrero-Bande, M., Cotos-Yáñez, T., González-Manteiga, W., Bermúdez-Cela, J., y Lucas-Domínguez, T. (2000). Prediction of SO₂ pollution incidents near a power station using partially linear models and an historical matrix of predictor-response vectors. *Environmetrics: The official journal of the International Environmetrics Society*, 11(2):209–225.
- Prada-Sánchez, J. M. y Febrero-Bande, M. (1997). Parametric, non-parametric and mixed approaches to prediction of sparsely distributed pollution incidents: A case study. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(1):13–32.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ramsay, J. y Silverman, B. (2005). *Functional data analysis*. Springer Series in Statistics. Springer.
- Rio, E. (2017). *Asymptotic theory of weakly dependent random processes*, volume 80. Springer Berlin.
- Robinson, P. M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis*, 4(3):185–207.
- Roca-Pardiñas, J., Cadarso-Suárez, C., y González-Manteiga, W. (2005). Testing for interactions in generalized additive models: application to so₂ pollution data. *Statistics and Computing*, 15:289–299.
- Roca-Pardiñas, J., González-Manteiga, W., Febrero-Bande, M., Prada-Sánchez, J., y Cadarso-Suárez, C. (2004). Predicting binary time series of so₂ using generalized additive models with unknown link function. *Environmetrics*, 15(7):729–742.
- Roca-Pardiñas, J., Ordonez, C., y Lado-Baleato, O. (2021). Nonparametric location-scale model for the joint forecasting of SO₂ and NO_x pollution episodes. *Stochastic Environmental Research and Risk Assessment*, 35(2):231–244.
- Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the national Academy of Sciences*, 42(1):43–47.
- Sánchez Sello, C. (2023). Apuntes de la asignatura Procesos Estocásticos del Máster en Técnicas Estadísticas.
- Sestelo, M., Roca-Pardinas, J., y Ordóñez, C. (2014). Predicting SO₂ pollution incidents by means of additive models with optimum variable selection. *Atmospheric Environment*, 95:151–157.
- Shao, J. (2003). *Mathematical statistics*. Springer Science & Business Media.

- Shao, J. (2006). *Mathematical statistics: Exercises and solutions*. Springer Science & Business Media.
- Shumway, R. H. y Stoffer, D. S. (2017). *Time series analysis and its applications*. Springer.
- Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 50(3):413–436.
- Stine, R. A. (1982). *Prediction intervals for time series*. Princeton University.
- Stine, R. A. (1987). Estimating properties of autoregressive forecasts. *Journal of the American statistical association*, 82(400):1072–1078.
- Stute, W. (1995). Bootstrap of a linear model with AR-error structure. *Metrika*, 42(1):395–410.
- Thombs, L. A. y Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, 85(410):486–492.
- Tong, Y. (1990). *The multivariate normal distribution*. Research in Criminology. Springer New York.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.
- Van der Vaart, A. W. y Wellner, J. A. (1996). Weak convergence and empirical processes: With applications to statistics. pages 284–308.
- Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.
- Wikipedia (16 de enero de 2024). Central térmica de Puentes de García Rodríguez. En *Wikipedia*. https://es.wikipedia.org/w/index.php?title=Central_trmica_de_Puentes_de_García_Rodríguez&oldid=157278971.
- Wikipedia-PepedoCouto (2009). Central térmica de Puentes de García Rodríguez. Recuperada el 28 de mayo de 2024 de https://upload.wikimedia.org/wikipedia/commons/2/22/T%C3%A9rmica_das_Pontes.jpg.
- Wold, H. (1954). *A study in the analysis of stationary time series*. Almqvist & Wiksell, 2nd edition.
- Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295.
- Yakowitz, S. J. (1985). Nonparametric density estimation, prediction, and regression for Markov sequences. *Journal of the American Statistical Association*, 80(389):215–221.