Trabajo Fin de Máster

Estudio comparativo de modelos de espacio de estados para series temporales

Daniel Diz Castro

Máster en Técnicas Estadísticas

Curso 2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Estudo comparativo de modelos de espazo de estados para series temporais
Título en español: Estudio comparativo de modelos de espacio de estados para series temporales
English title: Comparative study of state-space models for time series
Modalidad: Modalidad B
Autor: Daniel Diz Castro, Universidad de Santiago de Compostela
Directora: Rosa María Crujeiras Casais, Universidad de Santiago de Compostela
Tutor: Andrés Padrones, SDG Consulting España, S.A.
Breve resumen del trabajo: Este trabajo presenta un estudio comparativo entre los modelos Orbit implementados en el paquete Orbit-ml para el lenguaje de programación Python y algunos modelos clásicos para series de tiempo, como los modelos Box-Jenkins o de suavización exponencial, bajo el marco común de los modelos de espacio de estados.

Doña Rosa María Crujeiras Casais, profesora titular del área de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela y don Andrés Padrones, de SDG Consulting España, S.A., informan que el Trabajo Fin de Máster titulado

Estudio comparativo de modelos de espacio de estados para series temporales

fue realizado bajo su dirección por don Daniel Diz Castro 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 26 de Enero de 2023.

CRUJEIRAS CASAIS ROSA MARIA - 52931780V

Firmado digitalmente por CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-52931780V, givenName=ROSA MARIA, sn=CRUJEIRAS CASAIS, cn=CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Fecha: 2023.01.26 18:54:22 +01'00'

La directora:
Doña Rosa María Crujeiras Casais

Firmado por PADRONES GARCIA ANDRES -

El tutor:
Don Andrés Padrones

El autor:
Don Daniel Diz Castro

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

Resumen en español

En el departamento de ciencias empresariales de *SDG Consulting España, S.A.* se encargan, entre muchas otras tareas, del análisis de series de tiempo de diversa índole y dedican mucho esfuerzo a tratar de mantenerse al día con las novedades que van surgiendo, con el paso de los años, en este ámbito.

El objetivo de este trabajo es tratar de determinar si los modelos DLT y LGT del paquete `Orbit-ml` para el lenguaje de programación Python resultan competitivos frente a los modelos clásicos de series temporales y, de ser el caso, cuales serían los escenarios en los que pueden ofrecer un buen rendimiento.

Con este fin, se revisarán las características que presentan, tanto los modelos candidatos como los modelos Box-Jenkins y de suavización exponencial, desde el punto de vista de los modelos de espacio de estados; familia en la que todos ellos se enmarcan. Posteriormente, se llevará a cabo una comparación con datos reales y simulados del rendimiento predictivo de distintos modelos.

English abstract

The business science department of *SDG Consulting España, S.A.* is in charge of, among many other tasks, the analysis of time series of different kinds and devotes a lot of effort to try to keep up to date with the new developments in this field as the years go by.

The aim of this work is to try to determine whether the DLT and LGT models of the `Orbit-ml` package for the Python programming language are competitive with respect to the classical time series models and, if so, in which scenarios they can offer good performance.

To this end, we will review the characteristics of the candidate models as well as the Box-Jenkins and exponential smoothing models from the point of view of state-space models, the family in which they all belong. Subsequently, a comparison of the predictive performance of different models will be carried out with real and simulated data.

Capítulo 1

Introducción

Uno de los objetivos más comunes del análisis estadístico de datos es la obtención de predicciones, para los valores de una determinada variable, en base a toda la información disponible hasta un determinado momento. Cuando buena parte de esa información se encuentra contenida en las mediciones a lo largo del tiempo para esa variable en cuestión, a través de una estructura de dependencia entre observaciones próximas en el tiempo, el análisis de series temporales juega un papel fundamental en la determinación de las características de esa relación de dependencia, que se emplearán como base para la obtención de predicciones para la variable en cuestión.

Con el paso de los años, a medida que la tecnología que limita la cantidad total de información que se puede almacenar y procesar, así como la velocidad con la que dicha información se recopila, han surgido algunas propuestas alternativas a los modelos para series de tiempo basados en la metodología Box-Jenkins o de suavización exponencial, predominantes en la segunda mitad del siglo XX d.C. Una de esas nuevas propuestas corre a cargo del equipo de ciencia de datos de la empresa *Uber*, que en el año 2020 liberó el packete `Orbit-ml` (Ng et al., 2020) con un par de modelos para series de tiempo basados en las ideas de suavización exponencial pero con algunas diferencias lo suficientemente importantes como para considerar a estos nuevos modelos como una alternativa claramente diferenciada.

1.1. Objetivos y metodología

El departamento de ciencias empresariales de *SDG Consulting España, S.A.*, la empresa colaboradora en el desarrollo de este trabajo, se ha hecho eco de los modelos Orbit y ha mostrado su interés en determinar si podría estar justificado incorporarlos al conjunto de herramientas que emplea en el análisis series de tiempo y, de ser el caso, en qué contextos concretos estos modelos pueden ofrecer algún tipo de ventaja con respecto a las metodologías clásicas.

Si bien el principal interés de los modelos de series de tiempo para *SDG Consulting España, S.A.* radica en la obtención de predicciones puntuales a diversos horizontes de predicción, comparten nuestra preocupación por entender los fundamentos teóricos detrás de los modelos de series de tiempo que se vayan a emplear. Tendremos, por tanto, un objetivo fundamental: buscar un modo de comparar los modelos Orbit con los modelos Box-Jenkins y de suavización exponencial, tanto desde un punto de vista teórico como aplicado, discriminando aquellos escenarios en los que unos métodos aparenten ofrecer mejores predicciones que los demás, de forma consistente.

Para abordar el problema de la comparabilidad teórica, vamos a recurrir a la teoría de modelos de espacio de estados, puesto que la gran familia de modelos que recibe este nombre incluirá, como veremos más adelante, tanto a los modelos clásicos como a los modelos Orbit. Por otro lado, llevaremos a cabo un amplio estudio de simulación en el que pondremos a prueba el rendimiento de los distintos modelos en términos de la precisión de sus predicciones, medida en base a varios criterios, así como la cobertura y longitud de sus intervalos de predicción, los tiempos computacionales necesarios para llevar

a cabo los procesos de ajuste y predicción, y otras cuestiones ligadas a distintos escenarios específicos del análisis de series de tiempo. Finalmente compararemos, siempre que sea pertinente, la precisión de las predicciones que ofrecerán algunos de los modelos estudiados sobre series de tiempo con datos reales, cedidas por la empresa colaboradora para tal fin.

1.2. Aspectos computacionales

Resulta evidente que, si pretendemos ajustar los modelos Orbit para alguna serie de tiempo, tendremos que emplear el único paquete en el que están implementados; esto es, el paquete `Orbit-ml` para el lenguaje de programación Python (Van Rossum, G. y Drake, F. L. (2009)).

Por otro lado, todos los gráficos presentes en las figuras de este trabajo, así como las series de tiempo simuladas de acuerdo con los procedimientos descritos en el Capítulo 5, y el cálculo de aquellas medidas que, en nuestra opinión, mejor representarán el rendimiento de los distintos modelos sobre el conjunto de series de tiempo reales y simuladas, se llevarán a cabo empleando, fundamentalmente, el paquete básico de R (R Core Team (2022)). Sin embargo, prestaremos especial atención a la librería `smooth` (Svetunkov, I. (2022)) de R, que permitirá simular y ajustar modelos ARIMA y de suavización exponencial como si de modelos de espacio de estados con una única fuente de error se tratase, una familia de modelos de la que hablaremos en detalle en el Capítulo 3.

El autor de este trabajo es el responsable del código de R y Python necesario para la generación de todas las figuras, simulaciones y ajustes presentes en este documento, al margen de las funciones extraídas de las librerías específicas mencionadas en los dos párrafos anteriores. Por otro lado, los cálculos habrán sido llevados a cabo con un ordenador portatil con 32 GB de memoria RAM, un procesador Intel Core i7-1165G7 y con sistema operativo Windows 11.

Por último, cabe mencionar un tercer lenguaje de programación relevante para este trabajo: Stan (Carpenter et al., 2017). Si bien no recurriremos a este lenguaje de forma directa en ningún caso, el paquete `Orbit-ml` depende de él para realizar los ajustes basados en inferencia bayesiana, permitiendo obtener estimaciones para los parámetros que caracterizan a determinados modelos paramétricos de acuerdo con la estimación máxima a posteriori o el de cadenas de Markov Monte Carlo. Discutiremos este tipo de estimaciones en la Sección 3.2.4.

1.3. Estructura del trabajo

Al margen de este primer capítulo introductorio en el que motivamos la necesidad de este proyecto e incluso introducimos algunas definiciones relevantes para el estudio de series de tiempo, el trabajo constará de otros cinco capítulos.

En el Capítulo 2 se revisará, brevemente, el enfoque clásico para el estudio de series de tiempo. A lo largo de la Sección 2.1 introduciremos la metodología asociada a los modelos Box-Jenkins, describiendo desde la formulación de los modelos más conocidos hasta las posibles extensiones de los mismos para incorporar alguna componente regresiva, pasando por los procedimientos de estimación, predicción o selección de modelos. En la Sección 2.2, por su parte, se llevará a cabo una exposición de los modelos de suavización exponencial más conocidos, explorando sus características más relevantes, y sus limitaciones.

El Capítulo 3 concentrará la mayor parte de la teoría de este trabajo, pues introduciremos la familia de modelos de espacio de estados con una única fuente de error, tratando de relacionarlos con los modelos Box-Jenkins y de suavización exponencial que, como podremos comprobar, son casos particulares de modelos pertenecientes a esta amplia familia. Se discutirán los mecanismos de estimación y predicción, tanto puntual como por intervalos, así como la propiedad de estabilidad, de gran relevancia.

La formulación de los modelos Orbit será descrita, haciendo uso de la limitada información disponible para ello, en el Capítulo 4. Dichos modelos serán relacionados con los modelos de espacio de estados y, por extensión, con los modelos clásicos para series temporales. También aludiremos, en el mismo capítulo, a la implementación de los modelos Orbit en el paquete `Orbit-ml`.

El Capítulo 5 será el más largo y contendrá un estudio comparativo con datos reales y simulados del rendimiento predictivo de los modelos Orbit así como algunos modelos Box-Jenkins y de suavización exponencial. El estudio de simulación estará dividido en dos partes. La primera consistirá en un estudio de control en el que se verifica si los modelos considerados se comportan adecuadamente al emplearse sobre series de tiempo con modelos generadores conocidos. La segunda parte pondrá a prueba los modelos en distintos escenarios que, si bien son habituales en el estudio de series de tiempo, se alejan del marco teórico sobre el que los modelos se construyen. Las series reales que se estudiarán habrán sido cedidas por *SDG Consulting España, S.A.* y se emplearán como comparación final entre los modelos que, en base a los resultados del estudio de simulación y las características de las propias series, parezcan ser los más adecuados a la hora de obtener predicciones certeras.

Por último, el Capítulo 6 contendrá las conclusiones a las que llegaremos, gracias a las descripciones de los modelos Orbit y las comparaciones basadas en series reales y simuladas con respecto a otros modelos de series de tiempo, sobre las cuestiones planteadas en la Sección 1.1. También se explorarán posibles extensiones o continuaciones de este trabajo.

1.4. Definiciones previas

Para que el trabajo sea lo más autocontenido posible, recordaremos en esta sección algunas de las cuestiones básicas ligadas al estudio de series de tiempo, entre ellas, qué entendemos por una serie de tiempo, al menos, para lo que a los objetivos de este trabajo se refiere.

Definición 1.1. Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad. Un proceso estocástico univariante, $\{Y_t\}_{t\in\mathcal{T}}$, es una colección de variables aleatorias, $\{Y_t, t \in \mathcal{T}\}$, definidas sobre el espacio de probabilidad común $(\Omega, \mathcal{F}, P)$, e indexadas por un parámetro, $t \in \mathcal{T}$.

La definición de proceso estocástico anterior es bastante general y, de hecho, nos centraremos en aquellos procesos estocásticos de variable real; esto es, aquellos procesos estocásticos (univariantes) en los que el espacio de medida en el que toman valores las variables aleatorias que lo componen, verifica:

- El espacio muestral es la recta real, $\mathbb{R}$.
- La σ-álgebra, de dicho espacio de medida, es la σ-álgebra de Borel, $\mathcal{B}$.

Además, consideraremos que el conjunto paramétrico que indexa a los procesos estocásticos de variable real, $\mathcal{T}$, es numerable y representa el tiempo. En concreto, asumiremos que $\mathcal{T} = \mathbb{Z}$, lo que da lugar a la denominación de "procesos estocásticos en tiempo discreto". Puesto que no habrá ambigüedad en el conjunto de índices, la notación reducida que emplearemos para un proceso estocástico real de variable real en tiempo discreto con término general Y_t será: $\{Y_t\}$.

A partir de esta noción de proceso estocástico de variable real en tiempo discreto (a los que nos referiremos simplemente como procesos estocásticos, de ahora en adelante), podemos ofrecer una definición bastante intuitiva para el concepto de serie de tiempo.

Definición 1.2. Sean $\omega \in \Omega$ y el proceso estocástico $\{Y_t\}$. A la colección de realizaciones de las variables aleatorias del proceso estocástico considerado, $\{Y_t(\omega)\}$, se le conoce como trayectoria del proceso $\{Y_t\}$ para el resultado, ω, del experimento aleatorio. Una serie de tiempo asociada al proceso $\{Y_t\}$ es una trayectoria parcial para dicho proceso; es decir, una colección de la forma $\{Y_t(\omega)\}_{\mathcal{S}}$ para cierto $\omega \in \Omega$ y $\mathcal{S} \subset \mathbb{Z}$.

En realidad, nos quedaremos con una versión simplificada de la definición de serie de tiempo anterior, para la que el conjunto de índices $\mathcal{S}$ es de la forma $\mathcal{S} = \{1, \cdots, n\}$ para cierto $n \in \mathbb{N}$, $n \geq 1$. De este modo, una serie de tiempo vendría a representar un conjunto finito de observaciones sucesivas de las variables aleatorias que conforman un proceso estocástico, identificando cada observación con un instante de tiempo.

De entre todos los procesos estocásticos, hay una familia destacada por sus buenas propiedades teóricas: la familia de procesos gaussianos.

Definición 1.3. Diremos que un proceso estocástico $\{Y_t\}$ es un proceso gaussiano si, para cualquier colección de índices $t_1, \cdots, t_n \in \mathbb{Z}$, el vector aleatorio $(Y_{t_1}, \cdots, Y_{t_n})$ tiene distribución normal multivariante.

Notemos que, de acuerdo con la Definición 1.3, el vector aleatorio que da lugar a una serie de tiempo asociada a un proceso estocástico gaussiano tendrá distribución normal multivariante. Como veremos más adelante, los procesos gaussianos juegan un papel muy importante en la teoría de los modelos Box-Jenkins para series de tiempo, entre otros.

Capítulo 2

Introducción a los modelos clásicos para series temporales

Tanto para comprender la evolución de los métodos de análisis y predicción de series temporales, como para medir la calidad de alternativas más recientes, es inevitable tomar como referencia los modelos para series de tiempo más relevantes del siglo XX.

En concreto, los modelos Box-Jenkins y el método de suavización exponencial comprenden dos metodologías distintas, aunque no por ello inconexas, que se siguen empleando ampliamente en la actualidad. El planteamiento en el que se basaron los modelos originales ha resultado ser lo suficientemente amplio como para que se hayan podido extender de varias formas, con el paso de los años, permitiendo dar respuesta a nuevas problemáticas a medida que estas surgían. Es por este motivo, junto con la experiencia derivada del análisis de sus resultados a lo largo del tiempo, que resultan particularmente interesantes para validar y comprobar el funcionamiento de nuevas propuestas.

Revisaremos brevemente los conceptos clave asociados a los modelos Box-Jenkins, basándonos para ello en Box et al. (2015) y en Brockwell y Davis (2016). Asimismo, introduciremos los métodos clásicos de suavización exponencial de la mano de Hyndman et al. (2008).

2.1. Modelos Box-Jenkins

Aunque se puedan enmarcar los modelos Box-Jenkins dentro de los modelos de espacio de estados, lo cierto es que, al margen de mejoras computacionales en los algoritmos de estimación, la metodología clásica no ha recibido grandes modificaciones con esta inclusión en un marco más amplio. En esta sección recordaremos los fundamentos detrás de estos modelos.

A grandes rasgos, podemos decir que los modelos Box-Jenkins pivotan en torno a la idea de estacionariedad.

Definición 2.1. Un proceso estocástico $\{Y_t\}$ es (débilmente) estacionario si:

1. $\mathbb{E}[Y_t] = \mu$ para todo t y cierta constante $\mu \in \mathbb{R}$.
2. Dados dos instantes t y s cualesquiera, la función de autocovarianzas solamente depende de la diferencia entre dichos instantes. Esto es, $\mathrm{Cov}(Y_t, Y_s) = \mathrm{Cov}(Y_{t-s}, Y_0)$.
3. $\mathbb{E}[Y_t^2] < \infty$ para todo t.

Notemos que, con la definición anterior, las variables aleatorias que conforman un proceso estocástico estacionario tendrán varianza finita y, además, será constante. La función de autocovarianzas también tomará valores finitos. De ahora en adelante, emplearemos μ para referirnos a la media de un proceso estacionario.

Ahora bien, ¿por qué resulta interesante que un proceso estocástico sea estacionario? La respuesta a esta pregunta se fundamenta en el teorema de descomposición de Wold. Antes de enunciarlo, introduzcamos alguna definición.

Definición 2.2. Dado un proceso estocástico estacionario $\{Y_t\}$, el mejor predictor lineal (en términos de error cuadrático medio) de Y_{t+h} con $h > 0$, dados Y_r con $-s \leq r \leq t$, es aquella variable de la forma $a + \sum_{j=-s}^{t} b_j Y_j$, que minimiza

$$\mathbb{E}\left[\left(Y_{t+h} - a - \sum_{j=-s}^{t} b_j Y_j\right)^2\right],$$

donde $a, b_{-s}, \cdots, b_t \in \mathbb{R}$. A su vez, el mejor predictor lineal de Y_{t+h} dados Y_r con $-\infty < r \leq t$, se define como el límite en media cuadrática cuando $s \to \infty$ del mejor predictor lineal dados Y_r con $-s \leq r \leq t$. Diremos que un proceso tal que, para todo t, Y_t coincida con su mejor predictor lineal, dados Y_r con $-\infty < r \leq t-1$, es un proceso determinista y, en caso contrario, no determinista.

Teorema 2.3 (Descomposición de Wold)**.** Todo proceso estocástico estacionario $\{Y_t\}$ no determinista admite la siguiente descomposición:

$$Y_t = d_t + e_t + \sum_{j=1}^{\infty} \psi_j e_{t-j}, \tag{2.1}$$

donde

- $\{e_t\}$ es un proceso compuesto por variables incorreladas, con media 0, y varianza constante, i.e, un proceso de ruido blanco,
- $\sum_{j=1}^{\infty} \psi_j^2 < \infty$,
- d_t es el mejor predictor lineal de si mismo dados Y_r con $-\infty < r \leq t$. Además, $\mathrm{Cov}[d_t, e_s] = 0$ para todo t, s. Si $d_t = 0$ para todo t, diremos que $\{Y_t\}$ es un proceso puramente no determinista.

El teorema de descomposición de Wold, asumiendo que tratamos con un proceso estacionario puramente no determinista (hipótesis no trivial, y pocas veces explicitada en la literatura), justificaría entonces la búsqueda de un modelo permita aproximar de algún modo la estructura de dependencia del proceso en cuestión respecto de otro proceso de ruido blanco.

2.1.1. Modelos para procesos estacionarios

El modelo más sencillo para aproximar la dependencia lineal de un proceso estacionario y un proceso de ruido blanco es el modelo de medias móviles de orden q, o MA(q), obtenido truncando la serie de (2.1).

Definición 2.4. Diremos que un proceso estocástico $\{Y_t\}$ sigue un modelo MA(q) si admite una representación del tipo:

$$Y_t - \mu = \sum_{j=1}^{q} \theta_j e_{t-j} + e_t. \tag{2.2}$$

Si consideramos el operador retardo, B, caracterizado por $BY_t = Y_{t-1}$, podemos reescribir la representación (2.2) a partir del operador de medias móviles de orden q,

$$\theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j.$$

De este modo, un proceso MA(q) cumplirá:

$$Y_t - \mu = \theta(B)e_t.$$

El modelo de medias móviles está cláramente limitado por el orden q. Un modelo alternativo, sugiere que cada variable del proceso dependa de las p variables anteriores: el modelo autoregresivo de orden p, o AR(p).

Definición 2.5. Diremos que un proceso estocástico $\{Y_t\}$ sigue un modelo AR(p) si es estacionario y admite una representación del tipo:

$$Y_t - \mu = \sum_{i=1}^{p} \phi_i (Y_{t-i} - \mu) + e_t. \tag{2.3}$$

Empleando el operador autoregresivo de orden p, $\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$, podemos reescribir (2.3) como

$$\phi(B)(Y_t - \mu) = e_t.$$

Nótese que no todo proceso estocástico con una representación como la de la Definición 2.5 va a ser estacionario. Una condición necesaria y suficiente para conseguir la estacionariedad, es que el polinomio, $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, no tenga raíces complejas de módulo unidad. Es importante mencionar que, al margen de una mayor facilidad de interpretación del modelo AR(p) frente al modelo MA(q) al hacer explícita la dependencia entre variables del mismo proceso en, lugar de entre un proceso dado y de ruido blanco, si $\phi(z) \neq 0$ para todo complejo z con $|z| \leq 1$, el modelo AR(p) también puede definir procesos puramente no deterministas, admitiendo una representación de la forma

$$Y_t - \mu = \sum_{j=1}^{\infty} \psi_j e_{t-j} + e_t, \tag{2.4}$$

con un número potencialmente infinito de coeficientes verificando $\sum_{j=1}^{\infty} |\psi_j| < \infty$. Se ofrece, por tanto, una representación parsimoniosa de un modelo que involucraría un número no necesariamente finito de coeficientes. Fijémonos en que la serie converge absolutamente, una condición más estricta que la que garantiza el teorema de Wold, y que otorga al proceso el apellido de "causal".

Combinando los modelos previos, surge el modelo ARMA, acrónimo de "autoregresive-moving average".

Definición 2.6. Diremos que un proceso estocástico $\{Y_t\}$ sigue un modelo ARMA(p,q) si es estacionario y admite una representación del tipo:

$$Y_t - \mu = \sum_{i=1}^{p} \phi_i (X_{t-i} - \mu) + \sum_{j=1}^{q} \theta_j e_{t-j} + e_t. \tag{2.5}$$

Reescribiendo (2.5) a partir de los operadores autoregresivo de orden p y de medias móviles de orden q, obtenemos la siguiente representación:

$$\phi(B)(Y_t - \mu) = \theta(B)e_t \iff \phi(B)Y_t = c + \theta(B)e_t, \text{ con } c = (1 - \phi_1 - \cdots - \phi_p)\mu.$$

Como se puede apreciar, en el modelo ARMA(p,q) se incluye una dependencia lineal de tanto las p variables previas, como de las variables del proceso de ruido blanco correspondientes con los q últimos instantes.

De nuevo, no todo proceso estocástico con una representación como la de la Definición 2.6 va a ser estacionario, aunque la condición necesaria y suficiente para conseguir la estacionariedad es la

misma; esto es, que $\phi(z)$ no tenga raíces complejas de módulo unidad. Asimismo, si $\phi(z) \neq 0$ para todo complejo z con $|z| \leq 1$, el proceso ARMA también será causal.

Por otro lado, si queremos que la representación asociada a un proceso ARMA(p,q) sea "única" necesitaremos añadir más condiciones.

Observación 2.7. Un mismo proceso estocástico estacionario podría dar lugar a infinitas representaciones como proceso ARMA(p,q) si no evitamos que los polinomios de la parte autoregresiva y de medias móviles tengan raíces en común. Por ejemplo, si $Y_t = e_t$, entonces $Y_t - kX_{t-1} = e_t - ke_{t-1}$ para todo k con $|k| < 1$ o, equivalentemente, $(1 - kB)Y_t = (1 - kB)e_t$. La primera representación se corresponde con un ARMA(0,0), y la segunda con un ARMA(1,1). Esto causaría un problema de identificabilidad del modelo.

Además, pueden existir representaciones ARMA distintas, asociadas a dos procesos causales distintos, que tengan la misma media y función de autocovarianzas. Aunque la igualdad de momentos hasta orden 2 no llega, en general, para caracterizar un proceso ARMA, sí lo hace bajo la hipótesis adicional de ruido blanco Gaussiano, que no solo es conveniente, si no que se utiliza por defecto a la hora de realizar estimaciones e inferencia sobre el modelo. Basta considerar $Y_t = e_t + \theta e_{t-1}$, pues si definimos $\hat{e}_t = \theta e_t$, tenemos un nuevo proceso de ruido blanco, de modo que el proceso $\{\hat{Y}_t\}$, dado por $\hat{Y}_t = \hat{e}_t + \frac{1}{\theta}\hat{e}_{t-1}$ tiene la misma media (0) y misma función de autocovarianzas que Y_t:

$$\mathrm{Cov}(Y_t, Y_{t+h}) = \mathrm{Cov}(\hat{Y}_t, \hat{Y}_{t+h}) = \begin{cases} (1+\theta^2)\mathrm{Var}(e_t), & \text{si } h = 0 \\ \theta \mathrm{Var}(e_t), & \text{si } |h| = 1 \\ 0, & \text{en otro caso.} \end{cases}$$

Para seleccionar un representante, dentro de los modelos ARMA(p,q) equivalentes en términos de media y autocovarianzas, se introduce concepto de invertibilidad.

Definición 2.8. Diremos que un proceso estocástico $\{Y_t\}$ es invertible cuando admite una representación del tipo

$$Y_t = c + e_t + \sum_{i=1}^{\infty} \pi_i X_{t-i},$$

donde c es una constante real, $\{e_t\}$ un proceso de ruido blanco, y $\sum_{i=1}^{\infty} |\pi_i| < \infty$. En el caso de un proceso estacionario invertible con media μ, se tiene que $c = \mu\left(1 - \sum_{i=1}^{\infty} \pi_i\right)$ y la representación anterior se puede reescribir como

$$\pi(B)(Y_t - \mu) = e_t,$$

con $\pi(z) = 1 - \sum_{i=1}^{\infty} \pi_i z^i$.

Una condición necesaria y suficiente para que un proceso ARMA(p,q) sea invertible, es que el polinomio de la parte de medias móviles, $\theta(z)$ tenga todas sus raíces con módulo mayor que 1.

La invertibilidad tiene especial interés en los procesos estacionarios, pues viene a decir que, obviando la media del proceso y el ruido blanco, cada variable depende linealmente de las anteriores. Esta propiedad parece deseable, si pretendemos emplear los datos del pasado de una serie de tiempo para tratar de predecir el futuro. Además, en el caso de los modelos ARMA causales, permite expresar el proceso de ruido blanco en función de los valores previos de $\{Y_t\}$, poniendo fin a la ambiguedad de la representación anteriormente comentada, en términos de media y autocovarianzas. De hecho, suponiendo que tenemos un proceso ARMA(p,q) invertible, causal, sin raíces en común entre $\phi(z)$ y $\theta(z)$,

$$\phi(B)(Y_t - \mu) = \theta(B)e_t \iff \frac{\phi(B)}{\theta(B)}(Y_t - \mu) = e_t = \pi(B)(Y_t - \mu),$$

estando $\pi(B)$ bien definido como la evaluación en B de un cociente de polinomios en variable conpleja con denominador que no se anula en el círculo unidad. Por este motivo, se suele restringir la búsqueda de modelos ARMA a versiones causales e invertibles.

Dentro de los procesos ARMA(p,q), hay una clase particular de modelos que tienen especial relevancia para modelar la estacionalidad.

Definición 2.9. Diremos que un proceso estocástico $\{Y_t\}$ sigue un modelo ARMA estacional de órdenes P y Q, con periodo estacional m, y lo denotaremos por $\mathrm{ARMA(P,Q)_m}$, si es estacionario y admite una representación del tipo

$$\Phi(B^m)(Y_t - \mu) = \Theta(B^m)e_t,$$

donde $\Phi(B^m) = 1 - \Phi_1 B^m - \cdots - \Phi_P B^{mP}$, $\Theta(B^m) = 1 + \Theta_1 B^m + \cdots + \Theta_Q B^{mQ}$, y $\Phi_P \neq 0 \neq \Theta_Q$.

Un proceso $\mathrm{ARMA(P,Q)_m}$, conocido no es más que un proceso ARMA(mP,mQ) con todos los coeficientes de sus polinomios autoregresivos y de medias móviles nulos, salvo los asociados a múltiplos del periodo estacional. Este modelo permite introducir de forma explícita una dependencia de las variables del proceso respecto de variables desplazadas por múltiplos del periodo estacional, tanto del propio proceso, como del proceso de ruido blanco.

Finalmente, podemos combinar los modelos ARMA y ARMA estacional para obtener otros que, con un número limitado de parámetros, puedan recoger una estructura de dependencia compleja, tanto de variables de instantes inmediatamente anteriores, como estacional.

Definición 2.10. Diremos que un proceso estocástico $\{Y_t\}$ sigue un modelo ARMA estacional multiplicativo de órdenes p, q, P y Q, con periodo estacional m, y lo denotaremos por $\mathrm{ARMA}(p,q)\times\mathrm{(P,Q)_m}$, si es estacionario y admite una representación del tipo

$$\phi(B)\Phi(B^m)(Y_t - \mu) = \theta(B)\Theta(B^m)e_t,$$

donde $\Phi(B^m) = 1 - \Phi_1 B^m - \cdots - \Phi_P B^{mP}$, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, $\Theta(B^m) = 1 + \Theta_1 B^m + \cdots + \Theta_Q B^{mQ}$, $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ y $\Phi_P \neq 0 \neq \Theta_Q$, $\phi_p \neq 0 \neq \theta_q$.

Es habitual encontrar referencias a los modelos ARMA estacionales multiplicativos como modelos SARMA, y no son más que modelos ARMA($mP + p$,$mQ + q$) con muchos parámetros nulos.

2.1.2. Modelos para procesos no estacionarios

Aunque existen varios factores que pueden provocar la falta de estacionariedad de un proceso estocástico, los más habituales en la práctica son:

1. Presencia de tendencia, esto es, media del proceso no constante, que se refleja en variaciones a largo plazo en las series de tiempo.

2. Cambios en la variabilidad, a menudo ligados a la presencia de tendencia.

3. Aparición de patrones que se repiten en intervalos regulares en las series de tiempo, conocidos como componente estacional de la serie.

Se han incluido en la Figura 2.1 tres series de tiempo generadas por procesos no estacionarios, representativas de las condiciones que acabamos de enumerar.

La filosofía de los modelos Box-Jenkins ante la falta de estacionalidad provocada por alguno de estos motivos será la transformación conveniente de los datos, con el objetivo de obtener series de tiempo que hayan podido ser generadas por un proceso estacionario.

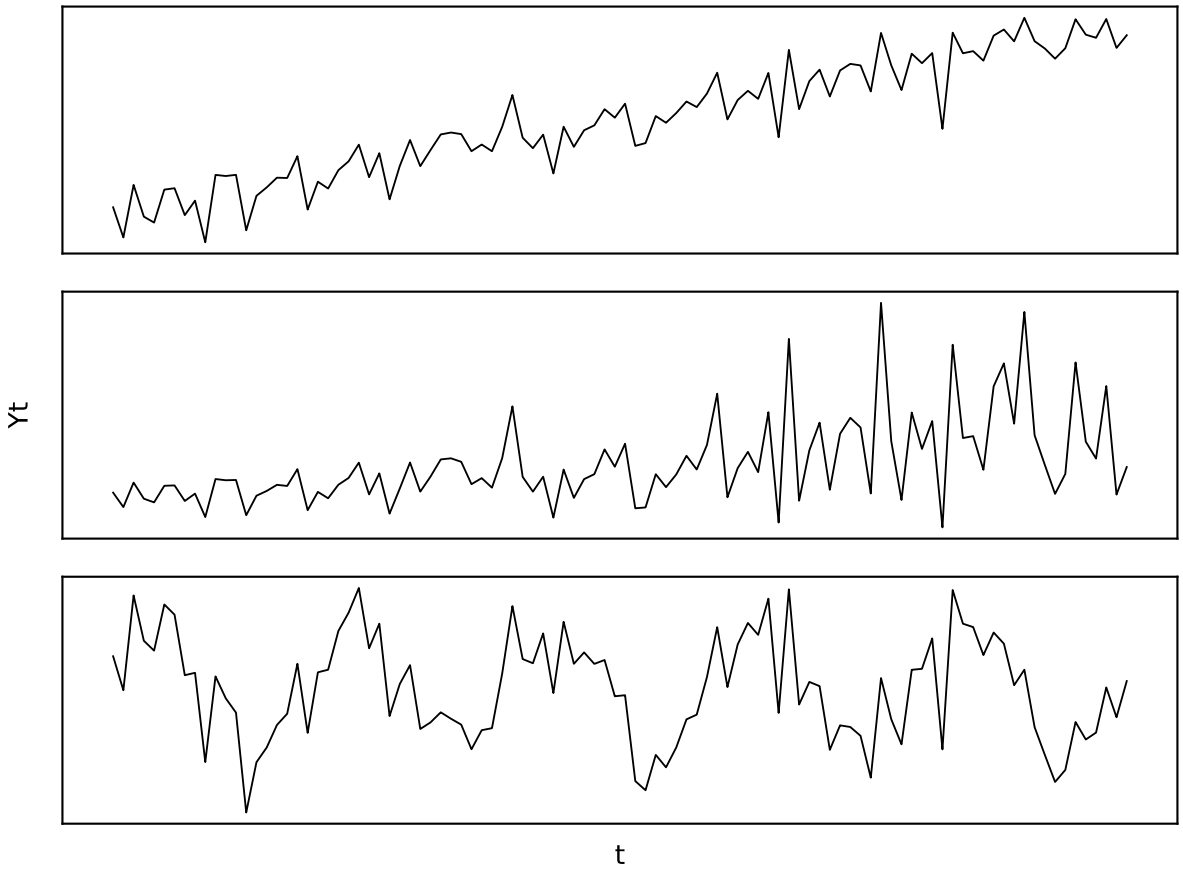

Figura 2.1: Representación de 3 series de tiempo generadas por procesos no estacionarios. Arriba: serie de tiempo con tendencia creciente. Medio: serie de tiempo con variabilidad incremental. Abajo: serie de tiempo con componente estacional

Por ejemplo, las transformaciones Box-Cox, que se emplean en muchos otros contextos, también pueden resultar útiles con el objetivo de estabilizar la varianza. Recordemos que se trata de la siguiente familia de transformaciones, caracterizada por un parámetro, λ:

$$f(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \text{si } \lambda \neq 0 \\ \log(x), & \text{si } \lambda = 0 \end{cases}$$

existiendo varios procedimientos para seleccionar de manéra automática un valor de λ adecuado, entre ellos el método de Guerrero y de máxima verosimilitud (Guerrero 1993).

Para lidiar con la presencia de tendencia o de componente estacional se introduce el concepto de diferenciación: regular, a través del operador $(1 - B)$, o estacional, a través del operador $(1 - B^m)$, siendo m el periodo estacional ligado a la componente correspondiente. La idea es convertir el proceso original en otro, aplicando el operador diferencial conveniente.

La aplicación de d diferenciaciones regulares permite eliminar una tendencia polinómica de grado hasta $d - 1$, convirtiendo una función de medias de tipo $\mathbb{E}(Y_t) = a_0 + a_1 t + \cdots + a_{d-1} t^{d-1}$ en $\mathbb{E}((1 - B)^d Y_t) = 0$. Análogamente, la aplicación de sucesivas diferenciaciones estacionales permitirá eliminar de la función de medias aquellas funciones periódicas de periodo igual al periodo estacional asociado a la

diferenciación. El modelo Box-Cox que incorpora la diferenciación regular y estacional simultáneamente es el modelo ARIMA estacional multiplicativo.

Definición 2.11. Diremos que un proceso estocástico $\{Y_t\}$ sigue un modelo ARIMA$(p,d,q)\times(P,D,Q)_m$; esto es, un modelo ARIMA estacional multiplicativo de órdenes p, d, q, P, D, Q, y periodo estacional m, cuando $\{(1-B)^d(1-B^m)^D Y_t\}$ sea un proceso ARMA$(p,q)\times(P,Q)_m$. Si $P=D=Q=0$, hablaremos de proceso ARIMA de órdenes p, d, q, denotado por ARIMA(p,d,q). Análogamente, si $p=d=q=0$, hablaremos de proceso ARIMA estacional de órdenes P, D, Q y periodo m, denotado por ARIMA$(P,Q)_m$.

2.1.3. Estimación, selección y validación de modelos

No entraremos en detalle en los procedimientos clásicos de estimación, puesto que los métodos implementados en los principales paquetes de los lenguajes de interés en este trabajo (R y Python) se basan en la representación de estos modelos como modelos de estado de espacios. Simplemente comentaremos que existen dos enfoques fundamentales: suma de residuos al cuadrado condicionales y máxima verosimilitud.

El primero de ellos consistiría en escoger aquellos parámetros que minimicen la suma de residuos al cuadrado asociados al modelo ajustado, asumiendo que las innovaciones previas al inicio de la serie tomaron el valor 0. El método de máxima verosimilitud buscará los valores de los parámetros que maximicen la función de verosimilitud del modelo, asumiendo que el proceso de ruido blanco es gaussiano (verosimilitud normal). Este último es el método empleado por defecto.

La pregunta natural, una vez tenemos modelos para procesos estocásticos, tanto estacionarios como no estacionarios, es la siguiente. ¿Cómo asocio una serie de tiempo con alguno de estos modelos?

En primer lugar debemos determinar si la serie se corresponde con un proceso estacionario, o no. Para ellos podemos fijarnos en su gráfico secuencial (presencia de tendencia, variabilidad constante o no, componente estacional y su periodo), o emplear procedimientos automatizados. En general, se transformaría la serie mediante una transformación Box-Cox adecuada, y se diferenciaría regular y estacionalmente la serie hasta que, por ejemplo, el test de Dickey-Fuller permita rechazar la hipótesis nula de no estacionariedad (Said y Dickey (1984)). Una vez conocidos d, D y m, un análisis de las autocorrelaciones simples y parciales muestrales, en base a resultados asintóticos de las mismas, podría permitir sugerir algún modelo sencillo. Sin embargo, el procedimiento más habitual es hacer uso de selectores automáticos de modelos, que emplean criterios de información, como los criterios AIC o BIC (definidos en la Sección 3.2.3) para escoger un modelo razonable. Cabe destacar que también existen tests que permiten elegir, sin necesidad de intervención humana, valores para d y D.

Asumiendo la correcta especificación del modelo ajustado por máxima verosimilitud, puede emplearse la teoría clásica correspondiente para realizar inferencia sobre los valores de los parámetros. En concreto, esto permite realizar contrastes de significación basados en la distribución asintóticamente normal de los estimadores.

Evidentemente, para ser rigurosos en el uso de estos modelos hay que tratar de validar las hipótesis subyacentes; esto es, el término de error en el modelo se debe corresponder con un proceso de ruido blanco de media 0. Además, es conveniente verificar la hipótesis de normalidad, que como ya hemos dicho, se asume implícitamente a la hora de ajustar el modelo, Como cabría esperar, la validación se realiza mediante un análisis de residuos. Los contrastes de incorrelación secuencial como el test de Ljung-Box y de media 0 como el T-test permiten verificar lo primero, mientras que los contrastes de normalidad como el test de Shapiro-Wilk, lo segundo.

La evidencia a favor del incumplimiento de las hipótesis de incorrelación y media 0 sugieren que el modelo no capta adecuadamente el comportamiento de la serie, e invitan a tratar de ajustar un modelo alternativo. La falta de normalidad es menos grave, pues los estimadores de máxima verosimilitud gaussiana seguirán siendo consistentes aunque las innovaciones no sean gaussianas, aunque se pierda eficiencia.

2.1.4. Valores atípicos y predicción

Un motivo que suele estar detrás de un análisis de residuos que invalide el modelo empleado, es la presencia de datos atípicos. Identificar y dar un tratamiento adecuado a aquellos datos que es alejen del comportamiento general de los demás es una constante en prácticamente cualquier modelo estadístico. De acuerdo con Box et al. (2015), podemos distinguir dos tipos de datos atípicos.

Definición 2.12. Si $\{\bar{Y}_t\}$ es un proceso ARIMA(p,d,q), y el proceso $\{Y_t\}$ generador de la serie de tiempo observada verifica

$$Y_t = k\mathbf{1}(t = r) + \tilde{Y}_t,$$

siendo $\mathbf{1}(t = r)$ la indicadora de que t toma el valor $r \geq 0$, entonces diremos que hay un atípico aditivo en $t = r$. Si, en cambio, $\{Y_t\}$ verifica

$$\phi(B)(1-B)^d Y_t = \theta(B)(k\mathbf{1}(t = r) + e_t) \iff Y_t = k\frac{\theta(B)}{\phi(B)(1-B)^d}\mathbf{1}(t = r) + \tilde{Y}_t,$$

diremos que hay un atípico innovativo en $t = r$.

El problema de identificación de un dato atípico en un instante $t = r$ se transforma en un test de significación del correspondiente parámetro k, habiéndolo incluido convenientemente en el modelo, de acuerdo con el típo de atípicos que estemos buscando. El propio valor ajustado para k nos permitiría recuperar el modelo, una vez sustraído el efecto del dato atípico.

Ahora bien, ¿cómo se realiza una búsqueda de datos atípicos sin conocer, a priori, ni cuántos hay ni en qué instante se encuentran? La idea sería estimar los valores para k considerando, una a una, cada una de las observaciones de la serie de tiempo como candidatas a ser observaciones atípicas, tanto aditivas como innovativas, efectuar los correspondientes contrastes de significación teniendo en cuenta un posible problema de comparaciones múltiples, y seleccionar como atípica aquella observación para la que el estadístico de contraste de significación tome un mayor valor, siempre y cuando el coeficiente estimado resulte estadísticamente significativo. Se podría aplicar este procedimiento de manera iterativa, incorporando en cada paso aquellas observaciones marcadas como atípicas, para obtener un modelo final con múltiples datos atípicos.

Una vez ajustado, validado y libre de datos atípicos, la finalidad fundamental del modelo de series de tiempo es ofrecer predicciones, tanto puntuales como por intervalos de predicción, de valores futuros de la serie. La predicción puntual habitual es la predicción en media, obtenida empleando los datos de la serie de tiempo y de acuerdo con la expresión del modelo ajustado, asumiendo el valor 0 (la media) para las futuras innovaciones. En cuanto a los intervalos de predicción, pueden contruirse en base a la normalidad de los predictores, en caso de no haber rechazado la hipótesis de normalidad de las innovaciones, o mediante el remuestreo de los residuos.

2.1.5. Regresión con errores ARIMA

Cuando disponemos de información externa a la propia serie, por ejemplo, en el caso de que se tenga constancia de algún suceso que haya podido provocar una perturbación en la recogida de datos, puede resultar útil emplear un modelo de regresión lineal con errores ARIMA. Este modelo de regresión es tanto una generalización del modelo de regresión lineal clásico, como una extensión del modelo ARIMA.

Supongamos, en el caso más general, que $\{X_t\}$ e $\{Y_t\}$ son dos procesos estocásticos, de modo que la serie de tiempo de interés se corresponde con una trayectoria parcial de $\{Y_t\}$, mientras que la información externa a dicha serie la aportaría una trayectoria parcial de $\{X_t\}$ (notemos que no tiene por qué existir una estructura de dependencia entre las variables del proceso $\{X_t\}$, si no que podría corresponderse con una serie de mediciones independientes a lo largo del tiempo). El modelo de regresión lineal con errores ARIMA(p,d,q) de $\{Y_t\}$ sobre $\{X_t\}$ se plantea como

$$\begin{cases} Y_t = \beta X_t + \eta_t, \\ \eta_t \sim \text{ARIMA}(p, d, q). \end{cases}$$

Es decir, asumimos que el proceso $\{Y_t - \beta X_t\}$ sigue un modelo ARIMA(p,d,q) o, equivalentemente, que el proceso diferenciado $\{(1-B)^d(Y_t - \beta X_t)\}$ es un proceso ARMA(p,q).

Este modelo de regresión puede emplearse, por ejemplo, para estimar el valor del coeficiente k asociado a un valor atípico aditivo, teniendo en cuenta la Definición 2.11. En este caso concreto, $\beta = k$ y $X_t = \mathbf{1}(t = r)$ para todo t, siendo r el instante de tiempo en el que se ha observado el dato atípico.

2.2. Suavización exponencial

A diferencia de los modelos Box-Jenkins, formulados para procesos estocásticos con la intención de asociar, posteriormente, las series de tiempo con algún proceso generador de la misma, los modelos de suavización exponencial surgieron como métodos de predicción cuyo único objetivo era la obtención de predicciones puntuales a partir de los datos recogidos en las series de tiempo. Por este motivo, un análisis clásico basado en suavización exponencial se limita a escoger un modelo adecuado y recoger las predicciones que ofrece. A día de hoy, como veremos más adelante, es posible un análisis estadístico más profundo de estos métodos al formar parte de la clase más amplia de modelos de espacio de estados, que se estudiarán en detalle en el Capítulo 3.

En esta sección describiremos los métodos más sencillos, que surgieron en la segunda mitad del siglo XX. El uso de unos u otros métodos estará motivado por la presencia de tendencia (comportamiento no constante a largo plazo) y/o componente estacional (patrones repetitivos en intervalos regulares) en la serie.

2.2.1. Suavización exponencial simple

Supongamos que queremos obtener predicciones para futuros valores de una serie de tiempo generada por un proceso estocástico, $\{Y_t\}$, que no presenta tendencia clara ni una componente estacional. Existen varios métodos sencillos para ofrecer predicciones a corto plazo.

El primero de ellos, consiste en tomar como predicción para valores futuros al último valor de la serie de tiempo, y es conocido como método "Naive"de predicción. Esto es, denotando por $\hat{Y}_{n+h|n}$ a la predicción a horizonte h a partir del instante n (el último instante para el que se observa la serie de tiempo), el método Naive consiste en tomar

$$\hat{Y}_{n+h|n} = Y_n,$$

para todo $h \geq 1$. Este método es muy simple pero presenta problemas evidentes. ¿Qué ocurre si, debido al ruido esperable para una serie de tiempo, aparece una tendencia estocástica localizada hacia el final de la serie de tiempo? El método Naive no es capaz de distinguir si el último valor de la serie de tiempo es representativo del comportamiento general de la misma.

Otro método, ligeramente más elaborado, se basa en predecir nuevos valores en base al promedio de los valores previos de la serie de tiempo. Esto es,

$$\hat{Y}_{n+h|n} = \frac{1}{n}\sum_{t=1}^{n} Y_t,$$

$\forall h \geq 1$. Si bien este método soluciona parcialmente el principal inconveniente del método Naive, al considerar el resto de valores de la serie, otorga el mismo peso a todas las variables: $1/n$. De nuevo, un método de predicción debería ser capaz de detectar tendencias espúreas provocadas por la aleatoriedad intrínseca a los datos, y ofrecer predicciones acorde a este principio. El método del promedio probablemente sea capaz de captar el nivel de la serie, en caso de mantenerse constante, pero no será capaz de ofrecer predicciones que se desvíen de dicho nivel cuando la serie oscile en una u otra dirección. Si el nivel de la serie cambiase a partir de un punto, dando un "salto", el promedio de todos los valores podría colocarse en un punto intermedio entre el nivel pre-salto y el nivel post-salto, dejando de ser representativo de la serie al completo.

Bajo este prisma surge el método de suavización exponencial simple, a medio camino entre el método Naive y la predicción en media, al considerar todas las variables de la serie, pero dando mayor importancia a los últimos valores de la misma. Podemos presentar este método a partir del siguiente algoritmo iterativo:

1. Fijar un valor para $\hat{Y}_{1|0} \equiv \hat{Y}_1$, por ejemplo, $\hat{Y}_1 = X_1$,
2. Tomar $\hat{Y}_{t+1|t} = \alpha^* Y_t + (1-\alpha^*)\hat{Y}_{t|t-1}$ para todo $t = 1, \cdots, n$, y con $\alpha \in [0,1]$,
3. Efectuar la predicción $\hat{Y}_{n+h|n} = \alpha^* \hat{Y}_{n+h-1|n} + (1-\alpha^*)\hat{Y}_{n+h-1|n} = \hat{Y}_{n+h-1|n} = \hat{Y}_{n+1|n}$ para todo $h \geq 2$.

Como vemos, se predice el mismo valor para todos los futuros horizontes, por lo que no se puede esperar un comportamiento excelente a largo plazo, en general. A horizonte 1, sin embargo, predice el nuevo valor como una media ponderada entre el valor previo y la predicción a horizonte 1 de dicho valor previo. Notemos que si $\alpha^* = 1$ recuperamos el método Naive, y si $\alpha^* = 0$ predecimos por el valor fijado inicialmente, $\hat{Y}_1$. El parámetro α^* se conoce como parámetro de suavizado, y regula la importancia que se le concede a los valores pasados de la serie. Si sustituimos progresivamente las expresiones de los $\hat{Y}_{t+1|t}$, con $t = 1, \cdots, n-1$, en la fórmula de la predicción a horizonte 1, $\hat{Y}_{n+1|n}$, se obtiene

$$\hat{Y}_{n+1|n} = \alpha^* Y_n + \alpha^*(1-\alpha^*)X_{n-1} + (1-\alpha^*)^2 \hat{Y}_{n-1|n-2} = \cdots = \sum_{t=1}^{n-1} \alpha^*(1-\alpha^*)^t X_{n-t} + (1-\alpha^*)^n \hat{Y}_1.$$

Por tanto, a excepción del valor inicial $\hat{Y}_1$, que podemos asumir que dependerá de la serie, la predicción a horizonte 1 es una media ponderada de los valores previos de la serie de tiempo, con unos pesos que decrecen (si $\alpha^* \in (0,1)$) exponencialmente al alejarnos del instante final, n. La velocidad con la que los pesos decrecen está determinada por el parámetro α^*, que "suaviza"la contribución del último valor de la serie a la predicción a horizonte 1. De ahí surge el nombre de "suavización exponencial".

En la Figura 2.2 aparece representada una serie de tiempo entre $t = 1$ y $t = 100$ generada por un proceso de ruido blanco gaussiano, con varianza 1, desplazada 5 unidades a partir del instante $t = 81$. También se incluyen a las predicciones a horizonte 20, a partir de $t = 100$, construidas con los métodos Naive, promedio y suavización exponencial simple, así como los valores que realmente terminó por tomar en esos instantes. Como la serie se mueve, en su mayor parte, entre -2 y 2, el método del promedio ofrece predicciones con valor inferior a 2, alejadas del comportamiento de la serie en el tramo final. La última observación muestral de la serie toma un valor relativamente bajo, por lo que la predicción que ofrece el método Naive tampoco capta el comportamiento de la serie a partir de $t = 81$. El método de suavización exponencial simple, con $\alpha^* = 0.5$ y $\hat{Y}_1 = Y_1$, es capaz de captar, aparentemente, el nivel en torno al que oscila la serie.

Con el fin de facilitar la comprensión de las extensiones de este método, vamos a ofrecer una representación alternativa para el método de suavización exponencial simple. Como ya hemos dado a entender a lo largo de esta subsección, el objetivo de este método es predecir el nivel que presenta la serie hacia el final de la misma. Dicha representación alternativa constará entonces de una ecuación de predicción, y una ecuación de actualización del nivel:

$$\begin{aligned} \text{Predicción}: \ & \hat{Y}_{t+h|t} = l_t, \\ \text{Nivel}: \ & l_t = \alpha^* Y_t + (1-\alpha^*) l_{t-1}, \end{aligned} \tag{2.6}$$

para todo $t \in [1,n]$, $h \geq 1$ y partiendo del valor inicial l_0.

2.2.2. Método lineal de Holt

Sobre la base del método de suavización exponencial simple, el método lineal de Holt (Hyndman et al., 2008, p.15), también conocido como suavización exponencial doble, introduce un término adicional

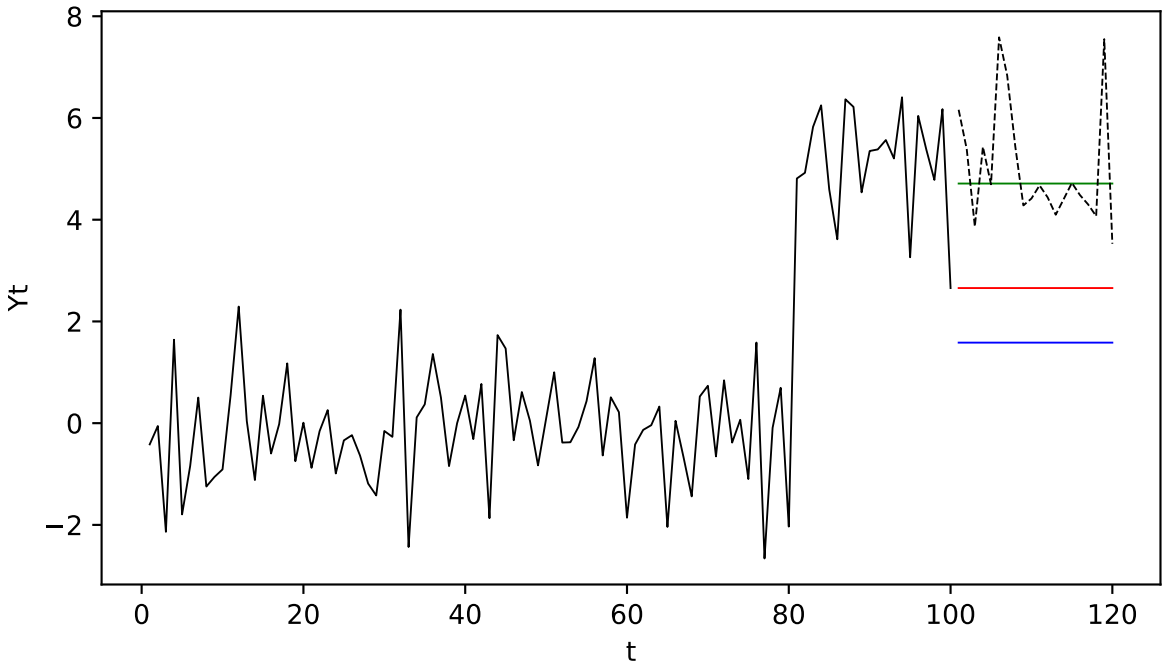

Figura 2.2: Serie de tiempo (en negro) representada en el intervalo (1,120), con un salto en $t = 80$, en linea discontinua a partir de $t = 101$. Representadas las predicciones a horizonte 20 desde $t = 100$ para los métodos Naive (en rojo), promedio (en azul) y suavización exponencial simple (en verde) con $\alpha^* = 0.5$.

en la representación (2.6) asociado a una tendencia lineal presente en la serie de tiempo, b_t, así como su correspondiente ecuación de actualización.

Se necesitarán dos valores iniciales, l_0 y b_0, que se corresponderán con el nivel y la pendiente de la tendencia inicial de la serie y, como podemos suponer al actualizar estos valores, el método trata de adaptarse a cambios de nivel y tendencia. Por ejemplo, podrá captar una ralentización en el crecimiento de una serie con tendencia lineal creciente (siempre que la nueva tendencia ralentizada sea lineal, aunque con menor pendiente).

A continuación, examinaremos las ecuaciones del método:

$$\begin{aligned} \text{Predicción}: &\quad \hat{Y}_{t+h|t} = l_t + hb_t, \\ \text{Nivel}: &\quad l_t = \alpha^* Y_t + (1-\alpha^*)(l_{t-1} + b_{t-1}), \\ \text{Tendencia}: &\quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}, \end{aligned} \tag{2.7}$$

para todo $t \in [1, n]$, $h \geq 1$ y partiendo de los valores iniciales l_0 y b_0.

En este caso, las predicciones a horizonte h dependen del valor de h, y se corresponden con lo que cabría esperar bajo la hipótesis de una tendencia lineal (local) de pendiente b_t. La ecuación de actualización de nivel de (2.7) se ve modificada en consonancia, pues $\hat{Y}_{t|t-1} = l_{t-1} + b_{t-1}$. El coeficiente de pendiente en el instante t, b_t, se construye como una media ponderada entre el coeficiente de pendiente en $t-1$ y la diferencia de niveles entre esos instantes, $(l_t - l_{t-1})$. El papel de $\beta^* \in [0, 1]$ sobre la tendencia de la serie es análogo al que posee α^* para el nivel de la misma, dando mayor importancia a los valores iniciales o finales de la serie, a la hora de cálcular el valor de b_t, cuanto más próximo a 0 o a 1 se encuentre β^*, respectivamente.

Se trata de un método de predicción enfocado a series de tiempo sin componente estacional, que presenten una tendencia que se pueda aproximar razonablemente bien de forma lineal, admitiendo

posibles puntos de cambio en la pendiente y/o el nivel de la serie. Como tal, tendrá una utilidad limitada cuando la tendencia de la serie diste de presentar un comportamiento lineal.

2.2.3. Método de tendencia amortiguada

El método de tendencia amortiguada (Hyndman et al., 2008, p.16) surge como una modificación del método lineal de Holt, con el objetivo de ofrecer mejores predicciones cuando no resulta realista asumir que se va a mantener, de manera sostenida, una tendencia lineal. En su lugar ofrecerá predicciones que, si bien a corto plazo estarán próximas a la linealidad, verán la contribución lineal de h "amortiguada"hasta el punto de que una predicción a cierto horizonte y la siguiente sean prácticamente indistinguibles.

Este método también presenta 3 ecuaciones: una de predicción y otras dos de actualización de nivel y tendencia. Sin embargo, añade otro parámetro de amortiguamiento : $\nu \in [0,1]$. Las ecuaciones en cuestión son las siguientes:

$$\begin{aligned} \text{Predicción}: \quad & \hat{Y}_{t+h|t} = l_t + (\nu + \nu^2 + \cdots \nu^h) b_t, \\ \text{Nivel}: \quad & l_t = \alpha^* Y_t + (1-\alpha^*)(l_{t-1} + \nu b_{t-1}), \\ \text{Tendencia}: \quad & b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)\nu b_{t-1}, \end{aligned} \tag{2.8}$$

para todo $t \in [1,n]$, $h \geq 1$ y partiendo de los valores iniciales l_0 y b_0. Como vemos, las ecuaciones de (2.8) con $h = 1$ son iguales a las de (2.7) salvo por el parámetro ν, que aparece multiplicando los coeficientes de las pendientes. Al predecir a horizonte $h \geq 2$, se considera $(\nu + \nu^2 + \cdots \nu^h) b_t$ en lugar de $h b_t$, como pendiente amortiguada h veces. Fijémonos en que, si $\nu = 0$ recuperamos el método de suavización exponencial simple, mientras que con $\nu = 1$ tenemos el método lineal de Holt.

Si hacemos tender h a infinito, las predicciones a horizonte h a partir del instante n se aproximan asintóticamente al valor $l_n + \frac{\nu}{1-\nu} b_n$, si $\nu \in (0,1)$. Así pues, las predicciones a corto plazo serán aproximadamente lineales, mientras que a largo plazo serán prácticamente constantes.

En la Figura 2.3, podemos comprobar las predicciones que efectuarían el método lineal de Holt y el método de tendencia amortiguada ante una serie con un crecimiento de tipo logístico :

$$Y_t = \frac{10}{1 + \exp(-(t-70)/10)} + e_t, \text{ con } e_t \text{ i.i.d } N(0,1), \ \forall t \in [1,120].$$

Para ambos métodos, los parámetros de suavizado son $\alpha^* = 0.5$ y $\beta^* = 0.2$, el parámetro de amortiguamiento toma el valor $\nu = 0.9$ y las condiciones iniciales son $l_0 = Y_1$ y $b_0 = 0.5$. La predicción del método de tendencia amortiguada se corresponde mejor con lo esperable bajo un modelo logístico.

2.2.4. Método de Holt-Winters

El último de los métodos clásicos de predicción puntual que vamos a ver, y el más complejo de todos ellos es el método de Holt-Winters, o método de suavización exponencial triple (Hyndman et al., 2008, p.16-18). Es un método diseñado para tratar con la estacionalidad de las series de tiempo aunque, de hecho, se distingan dos tipos de estacionalidad, dando lugar a dos versiones del método.

Si las variaciones debidas a la componente estacional de la serie parecen más o menos constantes a lo largo de la misma, entonces se dice que es una componente estacional de tipo aditivo. Si, en cambio, la magnitud de las variaciones estacionales parece estar ligada al nivel de la serie (escenario muy común en el caso de series de tiempo financieras), se dice que la componente estacional es de tipo multiplicativo.

El método de Holt-Winters aditivo consta de las siguientes ecuaciones:

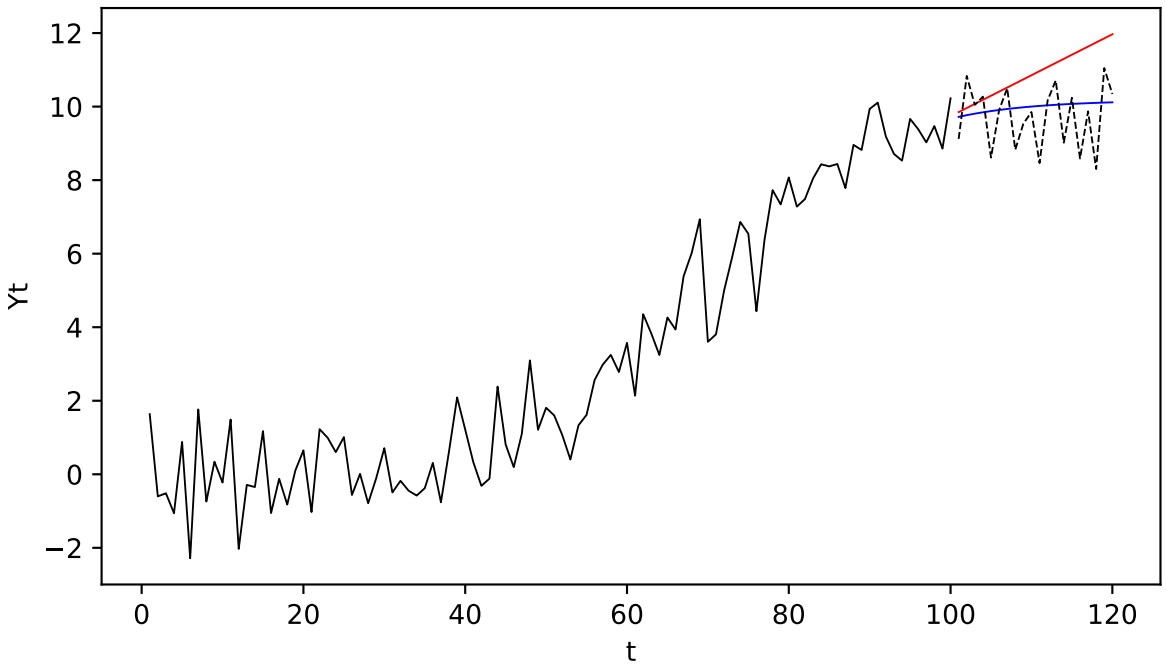

Figura 2.3: Serie de tiempo logística (en negro) representada en el intervalo (1,120), en linea discontinua a partir de $t = 101$ junto con las predicciones a horizonte 20 desde $t = 100$, para los métodos lineal de Holt (en rojo) y de tendencia amortiguada (en azul) con $\alpha^* = 0.5$, $\beta^* = 0.1$ y $\nu = 0.9$.

$$\begin{aligned} \text{Predicción}: &\quad \hat{Y}_{t+h|t} = l_t + hb_t + s_{t-m+h_m^+}, \\ \text{Nivel}: &\quad l_t = \alpha^*(Y_t - s_{t-m}) + (1-\alpha^*)(l_{t-1} + b_{t-1}), \\ \text{Tendencia}: &\quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}, \\ \text{Estacionalidad}: &\quad s_t = \gamma^*(Y_t - l_{t-1} - b_{t-1}) + (1-\gamma^*)s_{t-m}, \end{aligned} \tag{2.9}$$

para todo $t \in [1, n]$, $h \geq 1$ y partiendo de los valores iniciales l_0, b_0 y $s_0, s_{-1}, \cdots, s_{-m+1}$, siendo $h_m^+ = [(h-1) \text{mod } m] + 1 \in \{1, \cdots, m\}$.

La predicción a horizonte h derivada de la ecuación de predicción en (2.9) es la suma de nivel propuesto para el instante t, h veces la pendiente asociada a la tendencia lineal en dicho instante, y una componente que tomará el mismo valor para aquellos horizontes cuyo resto módulo m coincida.

Las ecuaciones de actualización de nivel y tendencia son esencialmente las mismas que las que aparecían en (2.7), salvo por la resta $Y_t - s_{t-m}$, que pretende recoger la parte no estacional de la serie en t, de este modo, la actualización de l_t y b_t se realizaría "al margen"de la componente estacional de la serie. Por su parte, la ecuación de actualización de la parte estacional, s_t, es una media ponderada por el parámetro de suavizado $\gamma^* \in [0, 1]$ entre el valor de dicha componente m instantes atrás, s_{t-m}, y lo que resta al retirar del valor observado en la serie, Y_t, la parte lineal de $X_{t|t-1}$: $l_{t-1} + b_{t-1}$.

Por otro lado, las ecuaciones del método de Holt-Winters multiplicativo son:

$$\begin{aligned}
\text{Predicción :} \quad & \hat{Y}_{t+h|t} = (l_t + hb_t)s_{t-m+h_m^+}, \\
\text{Nivel :} \quad & l_t = \alpha^* \frac{Y_t}{s_{t-m}} + (1-\alpha^*)(l_{t-1} + b_{t-1}), \\
\text{Tendencia :} \quad & b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}, \\
\text{Estacionalidad :} \quad & s_t = \gamma^* \frac{Y_t}{l_{t-1} + b_{t-1}} + (1-\gamma^*)s_{t-m},
\end{aligned} \tag{2.10}$$

para todo $t \in [1, n]$, $h \geq 1$ y partiendo de los valores iniciales l_0, b_0 y $s_0, s_{-1}, \cdots, s_{-m+1}$. Como vemos, las ecuaciones de (2.10) y (2.9) son análogas, cambiando convenientemente sumas y restas por productos y divisiones allá donde intervenga la parte estacional.

En ambos métodos, el parámetro γ^* regula la importancia de nuevos valores de la serie a la hora de actualizar el valor de la parte estacional. Si $\gamma^* = 0$, entonces se repiten constantemente los valores iniciales $s_0, s_{-1}, \cdots, s_{-m+1}$.

Aunque el método se ha definido considerando una tendencia lineal, se podría adaptar fácilmente a una tendencia amortiguada, añadiendo el parámetro ν como factor de amortiguación que multiplique a los coeficientes b_t en las ecuaciones de actualización, y sustituyendo hb_t por $(\nu + \nu^2 + \cdots \nu^h)b_t$.

2.2.5. Estimación de parámetros y selección de modelos

Inicialmente, de acuerdo con Hyndman et al. (2008, p.23-28), los métodos de suavización exponencial, planteados exclusivamente como métodos de predicción para series de tiempo, se ajustaban introduciendo a mano valores de parámetros que, de acuerdo con la experiencia previa del usuario, podrían considerarse como aceptables. También se desarrollaron algunos criterios heurísticos para obtener valores iniciales para $l_0, b_0, s_0, \cdots, s_{-m+1}$. Entre ellos, ajustar una recta de regresión para los 10 primeros valores muestrales y tomar como nivel y tendencia iniciales el intercepto y la pendiente de dicha recta respectivamente. Una alternativa más razonable pasaba por tomar aquellos valores, tanto para parámetros como inicializadores de las ecuaciones de actualización, que diesen lugar a modelos que minimizasen algún criterio de error basado en las predicciones. Por ejemplo, el error cuadrático medio de predicción a horizonte 1,

$$\text{MSE} = \sum_{t=1}^{n} (Y_t - \hat{Y}_{t|t-1})^2.$$

Ahora bien, si denotamos por e_t al error de predicción a horizonte 1 para el instante t,

$$e_t = Y_t - \hat{Y}_{t|t-1},$$

se pueden reescribir los métodos que hemos presentado anteriormente, en forma de modelos estadísticos con términos de error que representan la aleatoriedad subyacente a los datos y que, del mismo modo que en el contexto de los modelos Box-Jenkins, suelen denominarse como innovaciones del modelo. Ilustraremos esta idea con el método lineal de Holt (Hyndman et al., 2008, p.19-20), empezando por reescribir la ecuación de predicción a horizonte 1:

$$\hat{Y}_{t|t-1} = l_{t-1} + b_{t-1} \iff Y_t = l_{t-1} + b_{t-1} + e_t. \tag{2.11}$$

Si ahora sustituimos (2.11) en (2.7), obtenemos el sistema (2.12),

$$\begin{aligned}
Y_t &= l_{t-1} + b_{t-1} + e_t, \\
l_t &= l_{t-1} + b_{t-1} + \alpha^* e_t, \\
b_t &= b_{t-1} + \beta^* \alpha^* e_t,
\end{aligned} \tag{2.12}$$

que podemos representar, de manera equivalente, en formato matricial:

$$\begin{aligned} Y_t &= \begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{X}_{t-1} + e_t, \\ \mathbf{X}_t &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{X}_{t-1} + \begin{bmatrix} \alpha^* \\ \alpha^*\beta^* \end{bmatrix} e_t, \end{aligned} \tag{2.13}$$

con $\mathbf{X}_t^{'} = \begin{bmatrix} l_t & b_t \end{bmatrix}$. Imponiendo una determinada distribución de probabilidad para las innovaciones e_t con $t \geq 1$, asumiendo por ejemplo que son independientes e idénticamente distribuidas según una distribución $N(0, \sigma^2)$ para cierto σ^2, tendremos un modelo estadístico correctamente especificado. En efecto, en virtud de (2.13) y asumiendo conocido el valor de $\mathbf{X}_0$, se tiene que la aleatoriedad del modelo proviene únicamente de las innovaciones. Denotando por f_{α^*,β^*} a la función de densidad conjunta del vector aleatorio $\mathbf{Y} = (Y_1, \cdots, Y_n)$ (y cometiendo un abuso de notación al denotar del mismo modo a las funciones de densidad de los correspondientes subvectores),

$$f_{\alpha^*,\beta^*,\mathbf{X}_0}(y_1, \cdots, y_n) = f_{\alpha^*,\beta^*,\mathbf{X}_0}(y_1) \prod_{t=2}^{n} f_{\alpha^*,\beta^*,\mathbf{X}_0}(y_t|y_1, \cdots, y_{t-1}) = f_{\alpha^*,\beta^*,\mathbf{X}_0}(y_1) \prod_{t=2}^{n} f_{\alpha^*,\beta^*,\mathbf{X}_0}(y_t|\mathbf{x}_{t-1}).$$

La segunda igualdad es consecuencia de la representación (2.7) del método lineal de Holt, puesto que se emplean las observaciones Y_1 hasta Y_{t-1} para construir los valores de l_{t-1} y b_{t-1}, dando lugar a $\mathbf{X}_{t-1}$. Puesto que, condicionando por $\mathbf{X}_{t-1}$, la distribución de Y_t coincide con la de e_t (salvo por un cambio de localización), la distribución de los innovaciones caracteriza el modelo y, en caso de ser normales, la verosimilitud asociada al modelo será gaussiana.

Este resultado es análogo para cualquiera de los métodos que hemos presentado en esta sección, si los extendendemos a modelos estadísticos a partir de la definición que hemos empleado para las innovaciones del modelo, conocidas como innovaciones aditivas. Un forma alternativa de introducir dichas innovaciones sería tomando $e_t = \frac{Y_t - \hat{Y}_{t|t-1}}{\hat{Y}_{t|t-1}} = \frac{Y_t}{\hat{Y}_{t|t-1}} - 1$, de modo que $Y_t = \hat{Y}_{t|t-1}(1 + e_t)$. Las innovaciones en este caso se conocen como innovaciones multiplicativas, de manera análoga a la componente estacionaria multiplicativa que vimos en la Sección 2.2.4. También se pueden construir modelos con una tendencia multiplicativa, en lugar de aditiva; esto es, allá donde apareciese una suma con el coeficiente b_t, se sustituye por un producto.

La distinción entre presencia de tendencia (aditiva o multiplicativa) que podrá estar amortiguada o no, componente estacional (aditiva o multiplicativa) y el tipo de innovaciones (aditivas o multiplicativas), dan lugar a una conocida clasificación de los modelos de suavización exponencial a partir de tripletes de la forma (E,T,S), haciendo referencia al tipo de error, tendencia y estacionalidad respectivamente, mediante las siguientes abreviaturas: A para aditivo, A_d para aditivo amortiguado, M para multiplicativo y N para ninguno. Así, al modelo lineal de Holt con innovaciones aditivas que hemos descrito en esta Sección es conocido como ETS(A, A, N).

La introducción de una base estadística permite realizar tareas de inferencia, así como construir intervalos de predicción, e incluso introducir covariables (en particular variables indicadoras) para añadir información externa a la propia serie de tiempo o detectar valores atípicos, siguiendo una idea similar a la que comentamos en la Sección 2.1.4. Estas ventajas vienen con un coste añadido, pues habrá que validar las hipótesis distribucionales efectuadas sobre las innovaciones del modelo, mediante la realización de tests adecuados, empleando para ello los residuos obtenidos al ajustar el modelo sobre la serie de tiempo.

En cuanto a la selección entre los distintos métodos disponibles, un análisis exploratorio de los datos podría ayudar a identificar las características distintivas de las series que motivaron la construcción de los métodos presentados. Así, estudiando los gráficos secuenciales y las autocorrelaciones simples y parciales puede detectarse la presencia o no de tendencia, si esta tendencia permitiría efectuar predicciones

que siguiesen un ritmo de crecimiento lineal, o si por el contrario sería conveniente amortiguarlas, si existe una componente estacional, y si puede enmarcarse en el tipo aditivo o multiplicativo. En caso de querer prescindir de la intervención humana en la selección del modelo, podrían escogerse aquellos modelos que minimizasen algún criterio de información, como el AIC o BIC.

Capítulo 3

Introducción a los modelos de espacio de estados

Los modelos de espacio de estados conforman una gran familia de modelos destinados a analizar sistemas dinámicos (que evolucionen en el tiempo) con una componente estocástica y que poseen una estructura común caracterizada por dos grupos de ecuaciones: una ecuación de observación o medida y una ecuación de estados. La ecuación de observación, como su propio nombre indica, describe la evolución de la parte observable del sistema (en nuestro caso, de la serie de tiempo) en función de componentes no observables del mismo, llamados estados, y un término de error. La ecuación de estados determinará la evolución de los estados del sistema en función de sus valores previos, incluyendo también la posibilidad de perturbaciones aleatorias. Así pues, manteniendo la misma notación para las series de tiempo que en las secciones previas, los modelos de espacio de estados se pueden formular como sigue:

$$\begin{aligned} Y_t &= w_t(\mathbf{X}_{t-1}) + u_t(\mathbf{X}_{t-1})e_t, \\ \mathbf{X}_t &= \mathbf{M}_t(\mathbf{X}_{t-1}) + \hat{\mathbf{v}}_t(\mathbf{X}_{t-1})\nu_t, \end{aligned} \tag{3.1}$$

donde, para cada instante t, $\mathbf{X}_t$ será el vector conformado por los estados del sistema, w_t y $u_t \neq 0$ son funciones que toman valores en la recta real, mientras que $\mathbf{M}_t$ y $\hat{\mathbf{v}}_t$ son funciones que devuelven vectores. Los errores en la ecuación de observación de (3.1), e_t, se asumen i.i.d con media 0, y lo mismo se supone para los errores en la ecuación de estados, ν_t. En función de la relación entre los procesos $\{e_t\}$ y $\{\nu_t\}$ se distinguirán dos enfoques claramente diferenciados en el estudio de este tipo de modelos. Si se asume que los procesos son independientes, entonces se habla de modelos de espacio de estados con múltiples fuentes de error (o modelos MSOE, acrónimo de "Multiple Source of Error"). Si, en cambio, suponemos que $\hat{\mathbf{v}}_t(\mathbf{X}_{t-1})\nu_t = \mathbf{v}_t(\mathbf{X}_{t-1})e_t$, entonces la aleatoriedad provendrá únicamente del proceso $\{e_t\}$, llamado proceso de innovaciones, dando lugar a los modelos de espacio de estados con una fuente de error (o modelos SSOE, acrónimo de "Single Source of Error").

En base a la finalidad última de este trabajo, de relacionar la metodología clásica analizada en el Capítulo 2 con determinados enfoques novedosos (como por ejemplo Orbit), nos restringiremos al estudio de los modelos con única fuente de error, guiándonos por Hyndman et al. (2008). Un estudio detallado de los modelos MSOE se puede encontrar en Harvey (1990). De todos modos, al menos cuando las funciones implicadas en (3.1) son lineales, y de acuerdo con Hyndman et al. (2008, p.219-220), todo modelo MSOE lineal admite una formulación equivalente en forma de modelo SSOE lineal, mientras que no todo modelo SSOE lineal puede expresarse como modelo MSOE lineal. Podemos esperar, por tanto, que la familia de modelos SSOE sea extraordinariamente amplia.

3.1. Modelos SSOE lineales

Los modelos SSOE lineales son un caso particular de modelos de espacio de estados con única fuente de error, que permiten relacionar la evolución de una serie de tiempo con la evolución de una combinación lineal de estados, que de algún modo reflejan características de la misma, como pueden ser el nivel de la serie en cada instante y una componente estacional. Estos modelos comparten la estructura

$$\begin{aligned} Y_t &= \mathbf{w}'\mathbf{X}_{t-1} + e_t, \\ \mathbf{X}_t &= \mathbf{M}\mathbf{X}_{t-1} + \mathbf{v}e_t, \end{aligned} \tag{3.2}$$

donde $\mathbf{w}$ y $\mathbf{v}$ son vectores que regulan el efecto del pasado sobre Y_t y del ruido, e_t, sobre $\mathbf{X}_t$, respectivamente. La matriz $\mathbf{M}$, conocida como matriz de transición, determina la componente determinista en la evolución del vector de estados $\mathbf{X}_t$. Se asume además que $\{e_t\}$ es un proceso innovativo con variables i.i.d de media 0 y con varianza σ^2, aunque es habitual suponer además distribución gaussiana.

Esta formulación permite incorporar todos aquellos modelos de suavización exponencial que no tengan ninguna componente multiplicativa.

Ejemplo 3.1 (Modelo de Holt como modelo SSOE lineal)**.** De acuerdo con el desarrollo expuesto en la Sección 2.2.5, el modelo lineal de Holt se puede expresar como

$$\begin{aligned} Y_t &= \begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{X}_{t-1} + e_t, \\ \mathbf{X}_t &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{X}_{t-1} + \begin{bmatrix} \alpha^* \\ \alpha^*\beta^* \end{bmatrix} e_t, \end{aligned}$$

con $\mathbf{X}_t^{'} = \begin{bmatrix} l_t & b_t \end{bmatrix}$, que se corresponde con la formulación de un modelo SSOE lineal con

$$\mathbf{w}' = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \mathbf{v}' = \begin{bmatrix} \alpha^* & \alpha^*\beta^* \end{bmatrix} \quad \text{y} \quad \mathbf{M} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

La formulación (3.2) para los modelos SSOE lineales también permite incluir a todos los modelos ARIMA, por sorprendente que pueda parecer esta afirmación.

Ejemplo 3.2 (Modelos ARIMA como casos particulares de los modelos SSOE lineales)**.** Nos basáremos en el desarrollo propuesto por Svetunkov y Boylan (2019) (se trata de una ligera modificación de la formulación de Hyndman et al. (2008) que permite incluir la constante c del modelo ARIMA). Recordemos la formulación clásica de un modelo ARIMA(p, d, q), desarrollando los polinomios de la parte autoregresiva y de medias móviles:

$$\phi(B)(1-B)^d Y_t = c + \theta(B)e_t \iff (1 - \phi_1 B - \cdots - \phi_p B^p)(1-B)^d Y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)e_t.$$

Denotemos por K al mayor exponente acompañando al operador retardo, B, en la expresión anterior; esto es, $K = \text{máx}\{p + d, q\}$. La representación de un modelo ARIMA(p, d, q) se puede reescribir como

$$Y_t = c + \sum_{i=1}^{K} \varphi_i Y_{t-i} + \sum_{i=1}^{K} \vartheta_i e_{t-i} + e_t, \tag{3.3}$$

donde los coeficientes φ_i y ϑ_i dependen de los coeficientes ϕ_i y θ_i respectivamente, y algunos tomarán el valor 0. Definiendo $X_{t-j,j} = \sum_{i=j}^{K}(\varphi_i Y_{t-i} + \vartheta_i e_{t-i})$ con $1 < j \leq K$, $X_{t-j,j} = 0$ cuando $j > k$, $X_{t-1,1} = c + \sum_{i=1}^{K}(\varphi_i Y_{t-i} + \vartheta_i e_{t-i})$ y sustituyendo en (3.3) obtenemos

$$Y_t = X_{t-1,1} + e_t. \tag{3.4}$$

Por otro lado, $X_{t-1,1} = c + X_{t-2,2} + \varphi_1 Y_{t-1} + \vartheta_1 e_{t-1}$, y $X_{t-j,j} = X_{t-(j+1),j+1} + \varphi_j Y_{t-j} + \vartheta_j e_{t-j}$ si $j > 1$. Además, $Y_{t-j} = X_{t-(j+1),1} + e_{t-j}$, de modo que

$$\begin{aligned} X_{t-j,j} &= \varphi_j X_{t-(j+1),1} + (\varphi_j + \vartheta_j)e_{t-j} + X_{t-(j+1),j+1}, \quad 1 < j < K \\ X_{t-K,K} &= \varphi_K X_{t-(K+1),1} + (\varphi_K + \vartheta_K)e_{t-K}, \\ X_{t-1,1} &= \varphi_1 X_{t-2,1} + (\varphi_1 + \vartheta_1)e_{t-1} + X_{t-2,2} + c. \end{aligned} \tag{3.5}$$

Sustituyendo en (3.5) $t - j$ por t para $1 \leq j \leq K$,

$$\begin{aligned} X_{t,j} &= \varphi_j X_{t-1,1} + (\varphi_j + \vartheta_j)e_t + X_{t-1,j+1}, \quad 1 < j \leq K \\ X_{t,K} &= \varphi_K X_{t-1,1} + (\varphi_K + \vartheta_K)e_t, \\ X_{t,1} &= \varphi_1 X_{t-1,1} + (\varphi_1 + \vartheta_1)e_t + X_{t-1,2} + c. \end{aligned} \tag{3.6}$$

Finalmente, podemos combinar (3.4) y (3.6) de manera que, si denotamos por $\mathbf{X}_t$ al vector de componente i-ésima $X_{t,i}$, y añadiendo $X_{t,K+1} = c$ para todo t, tenemos la representación

$$\begin{aligned} Y_t &= \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \mathbf{X}_{t-1} + e_t, \\ \mathbf{X}_t &= \begin{bmatrix} \varphi_1 & 1 & 0 & \cdots & 0 & 0 & 1 \\ \vdots & & \vdots & & \vdots & & \\ \varphi_{K-1} & 0 & 0 & \cdots & 0 & 1 & 0 \\ \varphi_K & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \mathbf{X}_{t-1} + \begin{bmatrix} \varphi_1 + \vartheta_1 \\ \cdots \\ \varphi_{K-1} + \vartheta_{K-1} \\ \varphi_K + \vartheta_K \\ 0 \end{bmatrix} e_t. \end{aligned} \tag{3.7}$$

Notemos que los estados del vector $\mathbf{X}_t$ no tienen una interpretación evidente, salvo por el primero ($X_{t,1} = Y_t - e_t$) y el último ($X_{t,K+1} = c$).

Observación 3.3. De acuerdo con el Ejemplo 3.2, todo modelo ARIMA puede reescribirse como un modelo SSOE lineal gaussiano. Lo cierto es que el recíproco también se cumple, bajo ciertas condiciones. Para demostrarlo, reescribamos la ecuación de estados de (3.2) empleando el operador retardo introducido en la Sección 2.1:

$$\mathbf{X}_t = \mathbf{M}\mathbf{X}_{t-1} + \mathbf{v}e_t \iff (\mathbf{I} - \mathbf{M}B)\mathbf{X}_t = \mathbf{v}e_t. \tag{3.8}$$

Aunque $(\mathbf{I} - \mathbf{M}B)$ puede no ser invertible, podemos multiplicar por el operador construido a partir de la matriz adjunta (que denotaremos por $\text{adj}(\mathbf{I} - \mathbf{M}B)$) a ambos lados de la ecuación (3.8) para obtener

$$\det(\mathbf{I} - \mathbf{M}B)\mathbf{X}_t = \text{adj}(\mathbf{I} - \mathbf{M}B)\mathbf{v}e_t, \tag{3.9}$$

y podemos sustituir (3.9) en la ecuación de observación de (3.2), una vez aplicado el operador $\det(\mathbf{I} - \mathbf{M}B)$ a ambos lados de dicha ecuación, para obtener

$$\det(\mathbf{I} - \mathbf{M}B)\mathbf{Y}_t = \mathbf{w}' \det(\mathbf{I} - \mathbf{M}B)\mathbf{X}_{t-1} + \det(\mathbf{I} - \mathbf{M}B)e_t = \mathbf{w}'\text{adj}(\mathbf{I} - \mathbf{M}B)\mathbf{v}e_{t-1} + \det(\mathbf{I} - \mathbf{M}B)e_t.$$

Así pues, todo modelo SSOE lineal gaussiano (la hipótesis de normalidad aquí solo es necesaria en la medida en que se requiera para definir los modelos ARIMA) se puede "reducir" a un modelo ARIMA con $\theta(B) = \mathbf{w}'\text{adj}(\mathbf{I} - \mathbf{M}B)\mathbf{v}B + \det(\mathbf{I} - \mathbf{M}B)$ y $\pi(B)\delta(B) = \det(\mathbf{I} - \mathbf{M}B)$, donde $\delta(x)$ será un polinomio con todas las raíces de módulo unidad de $\det(\mathbf{I} - \mathbf{M}x)$. De este modo, $\{\delta(B)Y_t\}$ será un proceso estocástico asociado a un modelo ARMA y si $\pi(x)$ y $\theta(x)$ tienen todas sus raíces con módulo mayor que la unidad, se verificarán las condiciones de invertibilidad y causalidad. De acuerdo con Hyndman et al. (2008, p.171-173), si los autovalores de $\mathbf{M}$ tienen todos módulo menor que 1, $\{Y_t\}$ será causal, mientras que si todos los autovalores de $\mathbf{M} - \mathbf{v}\mathbf{w}'$ tienen módulo menor que 1, el proceso será invertible.

Por lo que hemos visto en el ejemplo 3.2 y la Observación 3.3, la relación entre los modelos SSOE lineales y los modelos ARIMA es estrecha, aunque los primeros podrán generalizarse de manera bastante intuitiva, como veremos más adelante, para incorporar más tipos de modelos dentro de una formulación común.

Volviendo a la formulación (3.2), va a resultar crucial distinguir entre dos escenarios alternativos. El primero de ellos consiste en asumir que existe un vector de estados inicial $\mathbf{X}_0$ no aleatorio, cuyas componentes serán parámetros que pasarán a formar parte del modelo, y que habrá que estimar de alguna manera. Este es el escenario habitual asociado a los modelos de suavización exponencial, y será adecuado cuando no exista un pasado previo a la primera observación de la serie de tiempo para la que intentamos ajustar un modelo. El segundo escenario consiste en asumir que no existe tal vector de estados inicial, o equivalentemente, que dicho vector de estados es una variable aleatoria, del mismo modo que lo es $\mathbf{X}_t$ con $t > 0$, de acuerdo con la ecuación de estados de (3.2). Este sería el enfoque preferible cuando no se puede asumir que la serie de tiempo haya comenzado en el mismo instante que el proceso estocástico subyacente.

3.1.1. Estimación y predicción

Comenzaremos asumiendo que $\mathbf{X}_0$ está fijado. Mediante un razonamiento análogo al de la Sección 2.2.6, la densidad conjunta asociada a una realización muestral concreta $y_1, \cdots, y_n$ bajo el modelo (3.2) se relaciona directamente con la densidad conjunta de las innovaciones e_t. En efecto, tengamos que cuenta que, si despejamos el término de error de la ecuación de observación y lo introducimos en la ecuación de estados de (3.2), obtenemos

$$\mathbf{X}_t = \mathbf{M}\mathbf{X}_{t-1} + \mathbf{v}(Y_t - \mathbf{w}'\mathbf{X}_{t-1}) = \tilde{\mathbf{M}}\mathbf{X}_{t-1} + \mathbf{v}Y_t = \cdots = \tilde{\mathbf{M}}^t\mathbf{X}_0 + \sum_{i=0}^{t-1}\tilde{\mathbf{M}}^i\mathbf{v}Y_{t-i}, \tag{3.10}$$

con $\tilde{\mathbf{M}} = \mathbf{M} - \mathbf{v}\mathbf{w}'$. Así pues, conociendo $y_1, \cdots, y_{t-1}$, para cierto instante $t > 0$, así como $\mathbf{X}_0$ y $\tilde{\mathbf{M}}$, quedan completamente determinados $\mathbf{x}_1, \cdots, \mathbf{x}_{t-1}$. Dicho de otra manera, la información que aportan simultáneamente $y_1, \cdots, y_{t-1}, \mathbf{X}_0$ y $\tilde{\mathbf{M}}$ es, como mínimo, la misma información que aportan $\mathbf{x}_1, \cdots, \mathbf{x}_{t-1}$, $\mathbf{X}_0$ y $\tilde{\mathbf{M}}$. Por otro lado, e_t es independiente de $(Y_1, \cdots, Y_{t-1})$ puesto que dicho vector se obtiene, en última instancia, a partir de transformaciones lineales de $(\mathbf{X}_0, e_1, \cdots, e_{t-1})$, de modo que la distribución de $Y_t|_{Y_{t-1},\cdots,Y_1,\mathbf{X}_0}$ coincidirá con la distribución de $Y_t|_{\mathbf{x}_{t-1},\cdots,\mathbf{x}_1,\mathbf{x}_0}$.

Ahora bien, $\mathbf{M}$, $\mathbf{v}$ y $\mathbf{w}$ (y por extensión $\tilde{\mathbf{M}}$) tendrán una estructura que dependerá, a lo sumo, de un conjunto de parámetros Υ, que junto al vector inicial $\mathbf{X}_0$ caracterizarán el modelo. A partir de esta idea, se puede deducir inmediatamente que

$$\begin{aligned} f(y_1, \cdots, y_n|\mathbf{X}_0, \Upsilon, \sigma^2) &= f(y_1|\mathbf{X}_0, \Upsilon, \sigma^2)\prod_{t=2}^{n} f(y_t|y_{t-1}, \cdots, y_1, \mathbf{X}_0, \Upsilon, \sigma^2) \\ &= f(y_1|\mathbf{X}_0, \Upsilon, \sigma^2)\prod_{t=2}^{n} f(y_t|\mathbf{x}_{t-1}, \cdots, \mathbf{x}_1, \mathbf{X}_0, \Upsilon, \sigma^2), \end{aligned}$$

siendo $f(y_1, \cdots, y_n|\mathbf{X}_0, \Upsilon, \sigma^2)$ la densidad conjunta asociada a la serie de tiempo observada. Naturalmente y en virtud de (3.2), conociendo $\mathbf{x}_{t-1}$ se tiene que $y_t = \mathbf{w}'\mathbf{x}_{t-1} + e_t$. En términos de densidades condicionadas,

$$\begin{aligned} f(y_1, \cdots, y_n|\mathbf{X}_0, \Upsilon, \sigma^2) &= f(y_1|\mathbf{X}_0, \Upsilon, \sigma^2) \prod_{t=2}^{n} f(y_t|\mathbf{x}_{t-1}, \mathbf{X}_0, \Upsilon, \sigma^2) \\ &= f(e_1 + \mathbf{w}'\mathbf{X}_0|\mathbf{X}_0, \Upsilon, \sigma^2) \prod_{t=2}^{n} f(e_t + \mathbf{w}'\mathbf{x}_{t-1}|\mathbf{x}_{t-1}, \mathbf{X}_0, \Upsilon, \sigma^2) \\ &= \prod_{t=1}^{n} f(e_t|\sigma^2), \end{aligned}$$

donde la última igualdad es consecuencia de la independencia entre e_t y X_{t-1}, derivada de la independencia intrínseca a las innovaciones del modelo, y de que la distribución de las innovaciones no depende ni del vector de estados inicial ni de los parámetros recogidos en Υ.

Veamos como podemos estimar, de acuerdo con un enfoque frecuentista, los parámetros que caracterizan el modelo. Como es habitual, la idea será tratar de obtener los estimadores de máxima verosimilitud, aunque al asumir que $\mathbf{X}_0$ está fijado se interpretará como máxima verosimilitud condicionada (por $\mathbf{X}_0$). Se escogerán los valores de los parámetros que maximicen

$$\mathcal{L}(\mathbf{X}_0, \Upsilon, \sigma^2|y_1, \cdots, y_n) = f(y_1, \cdots, y_n|\mathbf{X}_0, \Upsilon, \sigma^2) = \prod_{t=1}^{n} f(e_t|\sigma^2). \tag{3.11}$$

Observación 3.4. Es posible que no todos los parámetros tengan la misma importancia dentro del modelo, una vez especificada la distribución de las innovaciones. Si asumimos que el proceso de innovaciones es un proceso de ruido blanco gaussiano con varianza σ^2, la función de verosimilitud depende de esta varianza y generalmente no es necesario más que un estimador consistente para ella (se suele denominar como parámetro "nuisance"), pues no se busca realizar inferencia sobre un estimador de la varianza, a pesar de que este sea necesario para realizar inferencia asintótica para el resto de estimadores de máxima verosimilitud.

Podemos simplificar la tarea de estimación en presencia de algún parámetro "nuissance" al pasar al cálculo de una verosimilitud condicionada perfil, reduciendo así la carga computacional al tener que estimar un parámetro menos mediante algoritmos de optimización numérica, algo que sin duda es un objetivo razonable teniendo en cuenta que el vector de estados iniciales podría tener un tamaño considerable (en caso de los modelos ARIMA, tiene hasta $K+1$ componentes siendo $K = \text{máx}\{p+d, q\}$). En el caso de innovaciones gaussianas, asumiendo conocidos $\mathbf{X}_0$ y Υ, la verosimilitud

$$\prod_{t=1}^{n} f(e_t|\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{n} e_t^2/\sigma^2\right),$$

se maximiza, en términos de σ^2, tomando $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2$. Sustituyendo σ^2 por $\hat{\sigma}^2$ en (3.11) se llega a la función de verosimilitud perfil

$$\mathcal{L}(\mathbf{X}_0, \Upsilon|y_1, \cdots, y_n) = \prod_{t=1}^{n} f(e_t|\hat{\sigma}^2) = (2\mathrm{e}\pi\hat{\sigma}^2)^{-\frac{n}{2}}, \tag{3.12}$$

que se maximiza al minimizar $\hat{\sigma}^2$. Estamos entonces en el contexto clásico de estimación minimizando la suma de residuos al cuadrado del modelo, puesto que $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2$ y además $e_t = y_t - \mathbf{w}_\Upsilon'\mathbf{x}_{t-1} = y_t - \mathbf{w}_\Upsilon'\left(\tilde{\mathbf{M}}_\Upsilon^{t-1}\mathbf{X}_0 + \sum_{i=0}^{t-2}\tilde{\mathbf{M}}_\Upsilon^{i}\mathbf{v}_\Upsilon y_{t-1-i}\right)$, es la diferencia entre el valor observado en el instante t y el

correspondiente ajuste propuesto por el modelo SSOE lineal gaussiano (hemos hecho explícita la dependencia de los vectores y matrices que determinan las ecuaciones de observación y estados del vector de parámetros Υ). En caso de que las innovaciones no sean gaussianas, la estimación de un modelo lineal de este tipo sería completamente análoga, modificando solamente la función de verosimilitud de acuerdo con la distribución que se considere para las innovaciones.

Pasemos ahora al segundo escenario, en el que no consideramos fijado $\mathbf{X}_0$ si no que será tratado como una variable aleatoria, asumiendo que ha habido valores previos de la serie de tiempo, no observados. Iterando la ecuación de estados de (3.2) tenemos que

$$\mathbf{X}_t = \mathbf{M}\mathbf{X}_{t-1} + \mathbf{v}e_t = \mathbf{M}(\mathbf{M}\mathbf{X}_{t-2} + \mathbf{v}e_{t-1}) + \mathbf{v}e_t = \cdots = \sum_{j=0}^{\infty} \mathbf{M}^j \mathbf{v} e_{t-j}.$$

Habrá que distinguir dos casos: el proceso estocástico subyacente a los vectores de estados es un proceso estacionario o habrá algún estado que no sigue un proceso estacionario. En el primero de los casos, de la expresión anterior se deduce que $E[\mathbf{X}_t] = 0$, $\mathrm{Var}(\mathbf{X}_t) = \mathrm{Var}(e_t)\sum_{j=0}^{\infty} \mathbf{M}^j \mathbf{v}\mathbf{v}'(\mathbf{M}^j)' < \infty$, pero en el segundo de los casos los estados pueden tener varianza no finita. En ambos escenarios, la función de verosimilitud incondicional es:

$$\mathcal{L}(\Upsilon, \sigma^2 | y_1, \cdots, y_n) = f(y_1, \cdots, y_n | \Upsilon, \sigma^2) = f(y_1 | \Upsilon, \sigma^2) \prod_{t=2}^{n} f(y_t | y_{t-1}, \cdots, y_1, \Upsilon, \sigma^2).$$

Asumiendo estacionariedad y normalidad multivariante para $\mathbf{X}_0$, teniendo en cuenta la ecuación de observación en (3.2), la distribución de $\mathbf{Y}_t$ condicionada por $\mathbf{Y}_{t-1}, \cdots, \mathbf{Y}_1$ también será normal. En efecto, recordemos que

$$Y_t = \mathbf{w}'_\Upsilon \left(\tilde{\mathbf{M}}_\Upsilon^{t-1} \mathbf{X}_0 + \sum_{i=0}^{t-2} \tilde{\mathbf{M}}_\Upsilon^{i} \mathbf{v}_\Upsilon Y_{t-1-i} \right) + e_t,$$

y que e_t es independiente de $(Y_1, \cdots, Y_{t-1})$. Así pues, basta con demostrar que la distribución de $\mathbf{X}_0|_{Y_1, \cdots, Y_{t-1}}$ es normal. Por el teorema de Bayes,

$$f(\mathbf{X}_0 | Y_{t-1}, \cdots, Y_1, \Upsilon, \sigma^2) \propto f(Y_{t-1}, \cdots, Y_1 | \mathbf{X}_0, \Upsilon, \sigma^2) f(\mathbf{X}_0 | \Upsilon, \sigma^2),$$

y sabemos por la discusión bajo la hipótesis de que $\mathbf{X}_0$ está fijado que $f(Y_{t-1}, \cdots, Y_1 | \mathbf{X}_0, \Upsilon, \sigma^2)$ es una densidad asociada a una distribución gaussiana. Como el producto de densidades gaussianas se corresponde con otra densidad gaussiana (tras la normalización conveniente), se tiene el resultado.

La verosimilitud gaussiana asociada al modelo será

$$\mathcal{L}(\Upsilon, \sigma^2 | y_1, \cdots, y_n) = (2\pi \mathrm{Var}[Y_t | y_{t-1}, \cdots, y_1])^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \sum_{t=1}^{n} \frac{y_t - \mathbb{E}[Y_t | y_{t-1}, \cdots, y_1]}{\mathrm{Var}[Y_t | y_{t-1}, \cdots, y_1]} \right),$$

y el problema de estimación radicará en el cálculo, en cada instante, de la media y varianza condicionadas de Y_t. Para ello, se pueden emplear algoritmos conocidos como filtros y, entre ellos, el más conocido es el filtro de Kalman (Hyndman et al., 2008, p.197-200).

En el caso en que alguno de los estados esté asociado a un proceso no estacionario, el problema de estimación se vuelve aún más complejo, pues algunos filtros (entre ellos el de Kalman) no se pueden definir adecuadamente en ese contexto. Hay varias soluciones, entre las que se encuentran transformar la serie hasta obtener una serie que se corresponda con un proceso estacionario, y aplicar lo comentado anteriormente, o recurrir a filtros más complejos. No profundizaremos más en estas ideas pues no existen a día de hoy (que nosotros conozcamos) librerías en los lenguajes de programación relevantes para este trabajo (R, Python) que implementen simultáneamente modelos SSOE con vector de estados iniciales aleatorio y estimación frecuentista basada en filtros. De todos modos, podemos comentar que, en base a lo expuesto por Hyndman et al. (2008), aunque los estimadores obtenidos por procedimientos de máxima verosimilitud condicional pierdan en términos de eficiencia frente a los obtenidos por máxima

verosimilitud, el hecho de que no sea necesario preocuparse por la estacionalidad de la serie para obtener estas estimaciones, ligado a la mayor flexibilidad que ofrece el primer método de estimación (pues pueden implementarse fácilmente distribuciones alternativas para las innovaciones, a diferencia de lo que ocurre si empleamos un filtro) vuelven tomar $\mathbf{X}_0$ como fijo una opción razonable. En toda la discusión posterior, asumiremos que $\mathbf{X}_0$ está fijado, de modo que se recurre a estimaciones de máxima verosimilitud condicional.

A la hora de predecir, estaremos interesados tanto en obtener predicciones puntuales, como en obtener intervalos de predicción. Para ello, con el enfoque frecuentista se recurre a la distribución condicionada de un futuro valor de la serie dada la información de la que se dispone, asumiendo que el modelo ha sido correctamente especificado, y los parámetros estimados coinciden con los verdaderos parámetros que pretenden estimar. El no incorporar la incertidumbre asociada a la estimación de los parámetros es precísamente una de las críticas que se hacen al enfoque clásico frecuentista desde el bayesianismo, que trata de modelarla a través de las distribuciones a posteriori de los parámetros.

Debemos determinar la distribución de $Y_{n+h}|_{\mathbf{X}_n}$. Notemos que, como ya hemos comentado con anterioridad, conociendo $\mathbf{X}_0$ y el resto de parámetros que caracteriza el modelo, así como $Y_1, \cdots, Y_{t-1}$, se puede determinar el valor de $\mathbf{X}_n$. Para ello, basta con establecer un paralelismo con el problema de estimación de los parámetros del modelo para darnos cuenta de que la distribución de $Y_{n+h}|_{\mathbf{X}_n}$ será la misma que la de $Y_h|_{\mathbf{X}_0}$ tomando $\mathbf{X}_0 = \mathbf{X}_n$; es decir, tendremos distribuciones gaussianas. Por tanto, resulta de lo más razonable ofrecer predicciones en media, así como intervalos de predicción basados en la varianza de la distribución normal en cuestión. ¿Cómo las obtenemos?

Cuando $h = 1$, la media y varianza se derivan de forma inmediata de la ecuación de observación de (3.2). En concreto, $\mathbb{E}[Y_{n+1}|\mathbf{X}_n] = \mathbf{w}'\mathbf{X}_n$ y $\mathrm{Var}(Y_{n+1}|\mathbf{X}_n) = \mathrm{Var}(e_n) = \sigma^2$. Si $h > 1$, en virtud de la ecuación de estados de (3.2) tenemos que

$$\begin{aligned}
\mathbb{E}[\mathbf{X}_{n+h-1}|\mathbf{X}_n] &= \mathbf{M}\mathbb{E}[\mathbf{X}_{n+h-2}|\mathbf{X}_n] = \cdots = \mathbf{M}^{h-1}\mathbf{X}_n \implies \mathbb{E}[Y_{n+h}|\mathbf{X}_n] = \mathbf{w}'\mathbf{M}^{h-1}\mathbf{X}_n, \\
\mathrm{Var}[\mathbf{X}_{n+h-1}|\mathbf{X}_n] &= \mathbf{M}\mathrm{Var}[\mathbf{X}_{n+h-2}|\mathbf{X}_n]\mathbf{M}' + \mathbf{v}\mathbf{v}'\sigma^2 = \cdots = \sigma^2 \sum_{j=0}^{h-2} \mathbf{M}^j \mathbf{v}\mathbf{v}'(\mathbf{M}^j)' \\
\mathrm{Var}[Y_{n+h}|\mathbf{X}_n] &= \mathbf{w}'\mathrm{Var}[\mathbf{X}_{n+h-1}|\mathbf{X}_n]\mathbf{w} + \sigma^2 = \sigma^2 \left(1 + \sum_{j=0}^{h-2} \mathbf{w}'\mathbf{M}^j \mathbf{v}\mathbf{v}'(\mathbf{M}^j)'\mathbf{w}\right).
\end{aligned} \tag{3.13}$$

Así pues, la predicción a horizonte h dada la serie de tiempo hasta el instante n sería

$$\hat{Y}_{n+h|n} = \mathbb{E}[Y_{n+h}|\mathbf{X}_n] = \mathbf{w}'\mathbf{M}^{h-1}\mathbf{X}_n,$$

mientras que el correspondiente intervalo de predicción de nivel α tendría la siguiente expresión:

$$(\mathbb{E}[Y_{n+h}|\mathbf{X}_n] \pm Z_{1-\alpha/2}\sigma\sqrt{\mathrm{Var}[Y_{n+h}|\mathbf{X}_n]}) = \left(\mathbf{w}'\mathbf{M}^{h-1}\mathbf{X}_n \pm Z_{1-\alpha/2}\sigma\sqrt{1 + \sum_{j=0}^{h-2} \mathbf{w}'\mathbf{M}^j \mathbf{v}\mathbf{v}'(\mathbf{M}^j)'\mathbf{w}}\right),$$

donde $Z_{1-\alpha/2}$ denota el cuantil de orden $1-\alpha/2$ de la distribución normal estándar. Naturalmente, se tendrá que sustituír en las expresiones anteriores los verdaderos parámetros por sus estimaciones de máxima verosimilitud condicionada. Se trata por tanto de predicciones consistentes e intervalos asintóticos, que podrán ser inadecuados para series temporales de tamaño reducido.

3.1.2. Dimensión mínima y estabilidad

Del mismo modo que la representación de un proceso ARIMA no es única si no se imponen una serie de restricciones, como comentamos en la Sección 2.1.1 al poder existir raíces en común entre el polinomio de la parte autoregresiva y de medias móviles, un mismo proceso estocástico puede

admitir varias representaciones de modelos SSOE lineales (la demostración de este resultado es trivial, considerando la relación entre ARIMA y SSOE analizada en la Sección 3.1). Con el objetivo de reducir la carga computacional a la hora de estimar los estados iniciales, interesará conocer la representación que involucre al menor número posible de dichos estados. Esta representación se dice que tiene dimensión mínima, y el siguiente resultado de Hyndman et al. (2008, p.150) permite caracterizar esta propiedad en función de los vectores y matrices que definen las ecuaciones de observación y estados de un modelo SSOE lineal.

Teorema 3.5 (Dimensión mínima, Hyndman et al., 2008)**.** Un modelo SSOE lineal tiene dimensión mínima si y solo si el rango de las dos siguientes matrices

$$\left[\mathbf{w} \mid \mathbf{M}'\mathbf{w} \mid \cdots \mid (\mathbf{M}')^{\dim(\mathbf{X}_0)-1}\mathbf{w}\right], \quad \left[\mathbf{v} \mid \mathbf{M}\mathbf{v} \mid \cdots \mid \mathbf{M}^{\mathbf{X}_0-1}\mathbf{v}\right],$$

es máximo ($\dim(\mathbf{X}_0)$).

Ejemplo 3.6 (Dimensión mínima para el modelo de Holt)**.** En el caso del modelo lineal de Holt, que reescribimos como modelo SSOE lineal en la Sección 3.1 y para el que $\dim(\mathbf{X}_0) = 2$, se verifica

$$\left[\mathbf{w} \mid \mathbf{M}'\mathbf{w}\right] = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \left[\mathbf{v} \mid \mathbf{M}\mathbf{v}\right] = \begin{bmatrix} \alpha^* & \alpha^*(1+\beta^*) \\ \alpha^*\beta^* & \alpha^*\beta^* \end{bmatrix}.$$

La primera matriz es de rango máximo, y la segunda también siempre que α^* y β^* sean distintos de 0. Bajo esta última restricción, el modelo lineal de Holt (ETS(A, A, N)) tiene dimensión mínima. Dicha restricción resulta de lo más razonable, pues si $\beta^* = 0$ entonces el término asociado a la tendencia en (2.12), b_t, es constante en el tiempo, de manera que podría eliminarse la correspondiente ecuación de actualización de estados manteniéndose la constante b_0 en las otras dos ecuaciones. Por su parte, si $\alpha^* = 0$ entonces (2.12) puede reducirse a la expresión

$$Y_t = l_0 + b_0 t + e_t,$$

esto es, se propone un modelo de regresión lineal simple tomando el tiempo como covariable.

De forma análoga al desarrollo del Ejemplo 3.6, puede demostrarse que los modelos ETS(A, N, N) y ETS(A, A_d, N) (los otros dos modelos de suavización exponencial lineales no estacionales) son también de dimensión mínima bajo la condición de que los parámetros que los definen sean no nulos. No obstante, los modelos estacionales no poseen, en ningún caso, la mínima dimensionalidad.

Ejemplo 3.7 (Dimensión mínima del modelo ETS(A, N, A))**.** Consideremos ahora el modelo ETS(A, N, A), es decir, el modelo de Holt-Winters aditivo con $\beta^* = b_0 = 0$ e innovaciones gaussianas:

$$\begin{array}{rlrl} \hat{Y}_{t|t-1} &= l_{t-1} + s_{t-m}, & Y_t &= l_{t-1} + s_{t-m} + e_t, \\ l_t &= \alpha^*(Y_t - s_{t-m}) + (1-\alpha^*)l_{t-1}, & \Longleftrightarrow \quad l_t &= l_{t-1} + \alpha^* e_t, \\ s_t &= \gamma^*(Y_t - l_{t-1}) + (1-\gamma^*)s_{t-m}. & s_t &= s_{t-m} + \gamma^* e_t. \end{array} \tag{3.14}$$

En este caso, $\dim(\mathbf{X}_0)$ pasa a ser $1 + m$, con

$$\mathbf{w} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} \alpha^* \\ \gamma^* \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \mathbf{X}_t = \begin{bmatrix} l_t \\ s_t \\ \cdots \\ s_{t-m+1} \end{bmatrix},$$

de modo que no tiene dimensión mínima, pues $(\mathbf{M}')^m\mathbf{w} = \mathbf{I}_{m+1}\mathbf{w} = \mathbf{w}$ y $\mathbf{M}^m\mathbf{v} = \mathbf{I}_{m+1}\mathbf{v} = \mathbf{v}$. Notemos que tal y como está definida esta componente según (3.14) no tiene por qué oscilar en torno a 0, una condición que parece razonable, pues lo ideal sería que cualquier variación del nivel de la serie estuviese recogida en el correspondiente término l_t. Así pues, existe una redundancia en la formulación del modelo. Se tiene que

$$s_t = s_{t-m} + \gamma^* e_t \iff (1-B^m)s_t = \gamma^* e_t \iff s_t = \frac{\gamma^* e_t}{(1-B^m)}.$$

Denotando por $S_1(B) = 1 + B + \cdots + B^{m-1}$, $s_t^* = \frac{1-S_1(B)/m}{1-B^m}\gamma^* e_t$ y $l_t^* = \frac{1}{1-B}\frac{1}{m}e_t$ se deduce inmediatamente que $s_t = s_t^* + l_t^*$. En efecto,

$$s_t = \frac{(1-S_1(B)/m) + S_1(B)/m}{(1-B^m)}\gamma^* e_t = s_t^* + \frac{S_1(B)}{1-B^m}\frac{1}{m}\gamma^* e_t = s_t^* + l_t^*.$$

Si denotamos por $S_2(B) = \frac{1}{m}((m-1)+(m-2)B+\cdots+2B^{m-3}+B^{m-2})$, se tiene que $(1-B)S_2(B) = (1-S_1(B)/m)$. En efecto,

$$\begin{aligned}(1-B)S_2(B) &= \frac{1}{m}((m-1)(1-B)+(m-2)B(1-B)+\cdots+2B^{m-3}(1-B)+B^{m-2}(1-B))\\ &= \frac{1}{m}((m-1)-B-B^2-\cdots-B^{m-1}) = \frac{1}{m}(m-(1+B+B^2+\cdots+B^{m-1})).\end{aligned}$$

Entonces, $\frac{1-S_1(B)/m}{1-B^m} = \frac{S_2(B)}{S_1(B)}$, de modo que l_t^* y s_t^* verifican

$$\begin{aligned} l_t^* = \frac{1}{1-B}\frac{1}{m}e_t \quad &\iff l_t^* = l_{t-1}^* + \frac{1}{m}\gamma^* e_t \\ s_t^* = \frac{S_2(B)}{S_1(B)}\gamma^* e_t \quad &\iff s_t^* = -\sum_{i=1}^{m-1} s_{t-i}^* + \frac{m-1}{m}\gamma^* e_t + \frac{m-2}{m}\gamma^* e_{t-1} + \cdots + \frac{1}{m}\gamma^* e_{t-m+2}.\end{aligned} \tag{3.15}$$

De acuerdo con las ecuaciones al lado derecho de las equivalencias en (3.15), la componente estacional del modelo se puede descomponer en un paseo aleatorio, l_t^*, y otra componente que depende de los valores que ha tomado en $m-1$ instantes anteriores, así como de los errores en dichos instantes, s_t^*.

Para conseguir dimensión mínima en el caso de los modelos de suavización exponencial con componente estacional, se "normaliza" esta componente, forzando que la suma de los últimos m valores sea siempre 0: $\sum_{i=0}^{m-1} s_{t-i} = 0$. No entraremos en detalle de la deducción del método de normalización, pero de acuerdo con Hyndman et al. (2008, p. 152) se consiguen métodos de dimensión mínima mediante este procedimiento. La intuición es muy sencilla: si los últimos m valores de la componente estacional suman 0, se puede despejar uno de ellos en función de los demás, permitiendo así reducir en 1 la dimensión del vector de estados. Notemos que la ecuación asociada a l_t^* en (3.15) será absorbida por la ecuación asociada al nivel l_t en el método normalizado, de modo que la estimación de α^* se verá modificada tras normalizar la componente estacional mediante la adición de $\frac{\gamma^*}{m}$.

Por otro lado, en la Sección 3.1.1 hemos determinado que asumir fijado el vector de estados iniciales $\mathbf{X}_0$ parece razonable. Sin embargo, no nos hemos preocupado de verificar que la dependencia de los estados iniciales del modelo desaparece a medida que aumenta el tamaño de la serie de tiempo.

Recordemos que podemos expresar $\mathbf{X}_t$ (y por ello también las predicciones para Y_t condicionadas por la información previa) en función del vector de estados iniciales y las observaciones pasadas de la serie de tiempo como

$$\mathbf{X}_t = \tilde{\mathbf{M}}^t\mathbf{X}_0 + \sum_{i=0}^{t-1}\tilde{\mathbf{M}}^i\mathbf{v}Y_{t-i},$$

siendo $\tilde{\mathbf{M}} = \mathbf{M} - \mathbf{v}\mathbf{w}'$. Por tanto, para obtener un modelo en el que $\tilde{\mathbf{M}}^t \mathbf{X}_0 \to 0$ cuando $t \to \infty$, condición a la que nos referiremos como "estabilidad"del modelo SSOE lineal (Hyndman et al., 2008, p.158), tendremos que imponer ciertas restricciones sobre los parámetros involucrados en la construcción de $\tilde{\mathbf{M}}$.

Proposición 3.8 (Hyndman et al., 2008)**.** *Un modelo SSOE lineal será estable si y solo si los autovalores de $\boldsymbol{M} - \boldsymbol{v}\boldsymbol{w}'$ tienen módulo menor que 1. Equivalentemente, el modelo será estable si el proceso ARMA reducido asociado, $\{\delta(B)Y_t\}$, es invertible.*

De acuerdo con la Proposición 3.5, se puede caracterizar la estabilidad de un modelo SSOE lineal mediante los autovalores de la matriz $\tilde{\mathbf{M}}$. En el caso de los modelos ARIMA, la condición de estabilidad se reduce a que el modelo ARMA asociado sea invertible, una de las dos propiedades, junto a la causalidad, que se suelen imponer para evitar problemas de falta de identificabilidad, tal y como mencionamos en la Sección 2.1.1. Para los modelos de suavización exponencial aditivos sin componente estacional, las condiciones necesarias para garantizar la estabilidad se pueden consultar en Hyndman et al. (2008, p.155).

Ejemplo 3.9 (Estabilidad del modelo ETS(A, N, N))**.** Vamos a analizar la estabilidad del método de suavización exponencial simple, o modelo ETS(A, N, N), definido por las ecuaciones

$$\begin{aligned} Y_t &= l_{t-1} + e_t, \\ l_t &= l_{t-1} + \alpha^* e_t \end{aligned}$$

donde $\mathbf{X}_t = l_t$. En este caso, $\mathbf{M} = [1]$, $\mathbf{w} = [1]$ y $\mathbf{v} = [\alpha^*]$, por lo que $\tilde{\mathbf{M}} = 1 - \alpha^*$. El único autovalor de $\tilde{\mathbf{M}}$ es $1 - \alpha^*$ que tiene módulo menor que 1 cuando $0 < \alpha^* < 2$. Así pues, la región de estabilidad del parámetro α^* es más amplia que la región clásica bajo la que se definía el método, en la Sección 2.2.1, donde α^* tomaba valores en $[0, 1]$, exceptuando el valor 0. Como detalle adicional, notemos que

$$l_{t-1} - l_{t-2} = \alpha^* e_{t-1} \implies Y_t - Y_{t-1} = l_{t-1} - l_{t-2} + e_t - e_{t-1} = e_t + (\alpha^* - 1)e_{t-1},$$

por lo que ETS(A, N, N) se reduce a un modelo ARIMA$(0, 1, 1)$ con $\theta_1 = \alpha^* - 1$, que efectivamente verificará las condiciones de invertibilidad del modelo ARMA asociado si $0 < \alpha^* < 2 \iff |\theta_1| < 1$.

Cabe destacar, que ningún modelo de suavización exponencial con componente estacional es estable (Hyndman et al. 2008, p.156). No obstante, al menos para los modelos aditivos, tras normalizar dicha componente se puede garantizar la estabilidad mediante limitaciones a los posibles valores que pueden tomar α^*, β^*, γ^* y ν.

En ausencia de estabilidad, la estimación del vector de estados iniciales tendrá repercusión en los vectores de estados para cualquier instante posterior. Como consecuencia, las predicciones que ofrecerá un modelo SSOE inestable, por muy elevado que sea el tamaño de la serie empleada para su ajuste, dependerán de la estimación de los estados iniciales del modelo, algo que va en contra del comportamiento aparente de las series de tiempo con datos reales, para las que se asume que la dependencia entre observaciones se diluye a medida que aumenta el tiempo transcurrido entre ellas.

3.2. Modelo SSOE general

En esta sección, revisaremos brevemente la extensión del modelo SSOE lineal al caso en que las ecuaciones de observación y estados incorporen funciones no lineales. La estructura común a estos modelos se presenta a continuación:

$$\begin{aligned} Y_t &= w_t(\mathbf{X}_{t-1}) + u_t(\mathbf{X}_{t-1})e_t, \\ \mathbf{X}_t &= \mathbf{M}_t(\mathbf{X}_{t-1}) + \mathbf{v}_t(\mathbf{X}_{t-1})e_t, \end{aligned} \tag{3.16}$$

donde w_t y $u_t \neq 0$ son funciones que toman valores en la recta real, mientras que $\mathbf{M}_t$ y $\mathbf{v}_t$ son funciones que devuelven vectores. Las innovaciones, e_t, se asumen i.i.d con media 0 y varianza σ^2.

En particular, todos los modelos ETS (con todas las posibles combinaciones de tripletes) admiten la formulación (3.16). Entre ellos destacan los modelos ETS con componente estacional multiplicativa, que será necesaria en aquellos casos en los que la componente estacional de la serie de tiempo aparente aumentar acorde con el nivel de la misma, o los modelos con errores multiplicativos, que permitirán recoger en el modelo heterocedasticidad de las innovaciones ligada al nivel de la serie. En la Figura 3.1 (derecha) podemos ver ejemplo de serie de tiempo con componente estacional aparentemente multiplicativa y una serie de tiempo con variabilidad aumentando acorde con el nivel (izquierda).

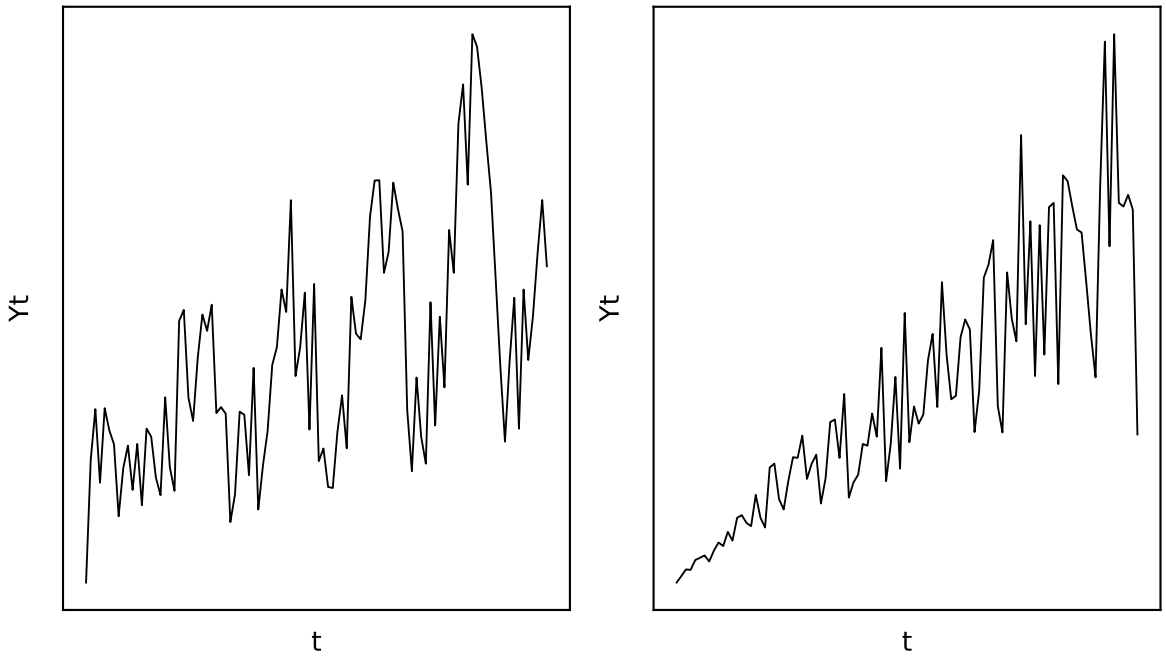

Figura 3.1: Izquierda: serie de tiempo con componente estacional de tipo multiplicativo. Derecha: serie de tiempo con innovaciones heterocedásticas, cuya variabilidad aumenta con el nivel.

3.2.1. Estimación, estabilidad y predicción

Podemos despejar el error en la ecuación de observación de (3.16) e introducir la igualdad obtenida en la ecuación de estados

$$e_t = \frac{Y_t - w_t(\mathbf{X}_{t-1})}{u_t(\mathbf{X}_{t-1})} \implies \mathbf{X}_t = \mathbf{M}_t(\mathbf{X}_{t-1}) + \mathbf{v}_t(\mathbf{X}_{t-1})\frac{Y_t - w_t(\mathbf{X}_{t-1})}{u_t(\mathbf{X}_{t-1})} = \tilde{\mathbf{M}}_t(\mathbf{X}_{t-1}) + \mathbf{v}_t(\mathbf{X}_{t-1})\frac{Y_t}{u_t(\mathbf{X}_{t-1})}$$

con $\tilde{\mathbf{M}}_t(\mathbf{X}_{t-1}) = \mathbf{M}_t(\mathbf{X}_{t-1}) - \mathbf{v}_t(\mathbf{X}_{t-1})\frac{w_t(\mathbf{X}_{t-1})}{u_t(\mathbf{X}_{t-1})}$. Esta ecuación recursiva es importante, pues aun que no sea tan explícita como (3.10), muestra que, fijado el vector de estados inicial $\mathbf{X}_0$ y condicionando por los valores pasados de la serie de tiempo, $Y_1, \cdots, Y_{t-1}$, el vector de estados $\mathbf{X}_{t-1}$ está completamente determinado (asumiendo conocidas $w_t, \mathbf{v}_t, u_t$ y $\mathbf{M}_t$). Así pues, la distribución de $Y_t|_{Y_1,\cdots,Y_{t-1}}$ coincide con la de $Y_t|_{Y_1,\cdots,Y_{t-1},\mathbf{X}_{t-1}}$ (conocer $\mathbf{X}_{t-1}$ no aporta información adicional). Ahora bien, por la ecuación de observación de (3.16), y en virtud de la independencia de las innovaciones, que implica

la independencia entre e_t y $(Y_1, \cdots, Y_{t-1})$, se tiene que, dadas las realizaciones muestrales $y_1, \cdots, y_n$,

$$f(y_1, \cdots, y_n|\mathbf{X}_0) = f(y_1|\mathbf{X}_0)\prod_{t=2}^{n} f(y_t|y_{t-1}, \cdots, y_1, \mathbf{X}_0) = f(y_1|\mathbf{X}_0)\prod_{t=2}^{n} f(y_t|y_{t-1}, \cdots, y_1, \mathbf{x}_{t-1}, \mathbf{X}_0)$$
$$= f(y_1|\mathbf{X}_0)\prod_{t=2}^{n} f(y_t|\mathbf{x}_{t-1}, \mathbf{X}_0).$$

Ahora, como $\mathbb{P}(Y_t \leq y_t|\mathbf{x}_{t-1}, \mathbf{X}_0) = \mathbb{P}\left(e_t \leq \frac{y_t - w_t(\mathbf{x}_{t-1})}{u_t(\mathbf{x}_{t-1})}\right)$, cuando $u_t(\mathbf{x}_{t-1}) > 0$, mientras que si $u_t(\mathbf{x}_{t-1}) < 0$ entonces $\mathbb{P}(Y_t \leq y_t|\mathbf{x}_{t-1}, \mathbf{X}_0) = 1 - \mathbb{P}\left(e_t \leq \frac{y_t - w_t(\mathbf{x}_{t-1})}{u_t(\mathbf{x}_{t-1})}\right)$, podemos deducir que las densidades cumplen $f(y_t|\mathbf{x}_{t-1}, \mathbf{X}_0) = \frac{f(e_t)}{|u_t(\mathbf{x}_{t-1})|}$. Es decir, a la hora de estimar los parámetros Υ y σ^2 de los que depende el modelo, la función de verosimilitud condicionada por $\mathbf{X}_0$ adopta la forma:

$$\mathcal{L}(\Upsilon, \sigma^2, \mathbf{X}_0|y_1, \cdots, y_n) = f(y_1, \cdots, y_n|\Upsilon, \sigma^2, \mathbf{X}_0) = \prod_{t=1}^{n} \frac{f(e_t)}{|u_t(\mathbf{x}_{t-1})|}, \tag{3.17}$$

donde $\mathbf{x}_0 \equiv \mathbf{X}_0$. Se trata de una generalización sencilla de la verosimilitud del caso lineal, (3.11).

Observación 3.10. Si asumimos que los errores son gaussianos, por ejemplo, podemos concentrar la verosimilitud al margen de σ^2, buscando la máxima verosimilitud perfil, de manera que el estimador de σ^2 siga siendo

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2,$$

pero la función de verosimilitud perfil pase a ser

$$\mathcal{L}(\mathbf{X}_0, \Upsilon|y_1, \cdots, y_n) = (2\mathrm{e}\pi\hat{\sigma}^2)^{-\frac{n}{2}} \left|\prod_{t=1}^{n} u_t(\mathbf{x}_{t-1})\right|^{-1},$$

que se maximizará minimizando la siguiente función de los parámetros

$$\left|\prod_{t=1}^{n} u_t(\mathbf{x}_{t-1})\right|^{\frac{2}{n}} \frac{1}{n}\sum_{t=1}^{n} e_t^2,$$

con $e_t = \frac{y_t - w_{t,\Upsilon}(\mathbf{x}_{t-1})}{u_{t,\Upsilon}(\mathbf{x}_{t-1})}$, haciendo explícita la dependencia de los parámetros del modelo.

Una vez que sabemos estimar el modelo, podemos preguntarnos para que valores de sus parámetros el comportamiento del mismo será el que esperamos, haciendo una analogía con el caso lineal. Desafortunadamente, no es sencillo extender el concepto de mínima dimensionalidad decrito en la Sección 3.1.2 al caso de los modelos SSOE no lineales. No obstante, podemos pensar en buscar las condiciones que garantizan la estabilidad del modelo, entendida como la desaparición del efecto que tienen los estados iniciales sobre sucesivos vectores de estados, a medida que avanza el tiempo. No existe una solución genérica para este problema, aunque sí es posible analizar la estabilidad para un subgrupo de modelos de este tipo. En concreto, supongamos que $\tilde{\mathbf{M}}_t$ es lineal y que $\mathbf{v}_t/u_t$ es constante. En ese caso,

$$\mathbf{X}_t = \tilde{\mathbf{M}}_t(\mathbf{X}_{t-1}) + \mathbf{v}_t(\mathbf{X}_{t-1})\frac{Y_t}{u_t(\mathbf{X}_{t-1})} = \tilde{\mathbf{M}}_t\mathbf{X}_{t-1} + \frac{\mathbf{v}_t}{u_t}Y_t = \cdots = \left(\prod_{j=1}^{t}\tilde{\mathbf{M}}_j\right)\mathbf{X}_0 + \sum_{i=0}^{t-1}\left(\prod_{j=0}^{i}\tilde{\mathbf{M}}_j\right)\frac{\mathbf{v}_{t-i}}{u_{t-i}}Y_{t-i},$$

con $\tilde{\mathbf{M}}_0$ la matriz identidad. Si además, $\tilde{\mathbf{M}}_t \equiv \tilde{\mathbf{M}}$ para todo $t > 0$, el sistema será estable, en virtud del Teorema 3.2, si los autovalores de $\tilde{\mathbf{M}}$ tienen módulo menor a la unidad. Se puede demostrar que todos los modelos ETS con componentes no multiplicativas y errores multiplicativos se encuentran en esta situación.

Ejemplo 3.11 (Estabilidad del modelo ETS(M, A, N))**.** El modelo de Holt con errores multiplicativos, también conocido como ETS(M, A, N), caracterizado por el siguiente sistema:

$$\begin{aligned} Y_t &= (l_{t-1} + b_{t-1})(1 + e_t), \\ l_t &= l_{t-1} + b_{t-1} + \alpha^*(l_{t-1} + b_{t-1})e_t, \\ b_t &= b_{t-1} + \beta^*(l_{t-1} + b_{t-1})e_t, \end{aligned}$$

Cláramente $\mathbf{v}_t/u_t = [1 \quad 1]'$ y, como $w_t = u_t$, tenemos que $\tilde{\mathbf{M}}_t(\mathbf{X}_{t-1}) = \mathbf{M}_t(\mathbf{X}_{t-1}) - \mathbf{v}_t(\mathbf{X}_{t-1})$, no depende de t, siendo tanto $\mathbf{M}_t$ como $\mathbf{v}_t$ lineales. Además,

$$\tilde{\mathbf{M}} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \alpha^* & \alpha^* \\ \beta^* & \beta^* \end{bmatrix} = \begin{bmatrix} 1-\alpha^* & 1-\alpha^* \\ -\beta^* & 1-\beta^* \end{bmatrix},$$

que curiosamente (Hyndman et al. p.154) coindice con la matriz de mismo nombre para el modelo lineal de Holt con arrores aditivos, ETS(A, A, N). Así pues, las restricciones sobre los parámetros para garantizar la estabilidad coinciden para el modelo de Holt con ambos tipos de error.

Pasemos a hablar de predicciones en el contexto del modelo SSOE general. En base a las ecuaciones de (3.16), la media de la distribución de $Y_{n+1}|_{Y_1,\cdots,Y_n} \equiv Y_{n+1}|_{\mathbf{X}_n}$ será $\mathbb{E}[w_{n+1}(\mathbf{X}_n)|\mathbf{X}_n] = w_{n+1}(\mathbf{X}_n)$, y recordemos que se pueden construir recursivamente los valores de $\mathbf{X}_t$ con $1 \leq t \leq n$ a partir de la serie de tiempo y del valor inicial $\mathbf{X}_0$, una vez fijados los parámetros del modelo. Por su parte, la varianza de la distribución condicionada tomará el valor $\mathrm{Var}[u_{n+1}(\mathbf{X}_n)e_t|\mathbf{X}_n] = u_{n+1}(\mathbf{X}_n)^2\sigma^2$. Ahora bien, esta varianza no tendrá la misma utilidad que en el caso lineal gaussiano (construcción de intervalos de predicción) si no asumimos que la distribución de e_t es simétrica en torno al 0, y que los cuantiles de dicha distribución son conocidos.

Si tratamos de predecir en media a horizonte $h > 1$, encontraremos más dificultades, puesto que

$$\begin{aligned} \mathbb{E}[Y_{n+h}|\mathbf{X}_n] &= \mathbb{E}[w_{n+h}(\mathbf{X}_{n+h-1})|\mathbf{X}_n] \neq w_{n+h}(\mathbb{E}[\mathbf{X}_{n+h-1}|\mathbf{X}_n]) \\ &= w_{n+h}(\mathbb{E}[\mathbf{M}_{n+h-1}(\mathbf{X}_{n+h-2})|\mathbf{X}_n]) \neq w_{n+h}(\mathbf{M}_{n+h-1}(\mathbb{E}[\mathbf{X}_{n+h-2}|\mathbf{X}_n])), \end{aligned}$$

donde las desigualdades son en general. Solamente se podrá predecir en media cuando, o bien w_t y $\mathbf{M}_t$ sean funciones lineales en los vectores de estados, o bien se conozcan las expresiones de $\mathbb{E}[w_t(\mathbf{X})]$ y $\mathbb{E}[\mathbf{M}_t(\mathbf{X})]$ a partir de $\mathbb{E}[\mathbf{X}]$. Con la varianza pasa algo parecido, sumado a la dificultad ya comentada de necesitar de una distribución simétrica en torno al 0 para que estimar la varianza condicionada permita construir intervalos simétricos en torno a la media condicionada. En Hyndman et al. (2008, Ch.6) se pueden encontrar fórmulas explícitas para los intervalos de confianza de la mayoría de modelos ETS, aunque no es posible derivar fórmulas explícitas para modelos con componente estacional multiplicativa, por ejemplo.

Una solución general para obtener predicciones en media e intervalos de confianza a horizonte h se basa en emplear algún método de simulación. La idea es generar múltiples trayectorias para la serie de tiempo partiendo de $\mathbf{X}_n$, de acuerdo con el modelo SSOE estimado, simulando valores de $e_{n+1}, \cdots, e_{n+h}$ a partir de la distribución de las innovaciones, en caso de ser conocida, o mediante el remuestreo de los residuos (cuya distribución se asume "similar" a la de las innovaciones) del modelo, en caso de no ser conocida la distribución de las innovaciones. Este procedimiento resulta particularmente útil en caso de que se asuma una distribución distinta a la gaussiana para las innovaciones del modelo y la estimación por máxima verosimilitud asociada de lugar a problemas en los algoritmos de optimación numérica. Para algunos modelos SSOE, los estimadores de máxima verosimilitud gaussiana serán consistentes aun cuando la distribución subyacente al modelo (en este caso a las innovaciones) no sea gaussiana. Por ejemplo, en Francq et al. (2004) se pueden encontar condiciones que garantizan la consistencia y normalidad asintótica de estos estimadores de quasi máxima verosimilitud para modelos ARMA-GARCH (en particular para modelos ARMA con condiciones de invertibilidad y causalidad).

Dichos estimadores, sin embargo, no serán eficientes en general. Así pues, si el tamaño muestral es suficientemente grande, podría recurrirse a la estimación basada en verosimilitud gaussiana para obtener estimaciones de sus parámetros, y posteriormente, recurrir a simulación, empleando la verdadera distribución de los errores (asumiendo que está completamente especificada o hay algún mecanismo de estimación de los parámetros de los que depende) o su aproximación a través de la distribución empírica de los residuos.

3.2.2. Regresión en modelos SSOE

Del mismo modo que en la Sección 2.1.5 introdujimos el modelo de regresión con errores ARIMA, ahora haremos lo propio para incorporar en el modelo SSOE general información externa a la propia serie de tiempo que pueda ayudar a explicar el comportamiento de la misma. En concreto, introduciremos una componente de regresión lineal de las variables de la serie sobre ciertas covariables, que podrían representar otra serie de tiempo, o simplemente una serie de factores con influencia en múltiples instantes de tiempo sobre la serie de tiempo de interés. Denotemos por $\{\mathbf{Z}_t\}$ al proceso estocástico (multivariante) generador de esas covariables (que no tienen por qué tener relación entre si, ni entre sus valores para distintos instantes de tiempo). Consideraremos modelos con una estructura como la siguiente

$$\begin{aligned} Y_t &= w_t(\mathbf{X}_{t-1}, \boldsymbol{\beta}\mathbf{Z}_t) + u_t(\mathbf{X}_{t-1}, \boldsymbol{\beta}\mathbf{Z}_t)e_t, \\ \mathbf{X}_t &= \mathbf{M}_t(\mathbf{X}_{t-1}) + \mathbf{v}_t(\mathbf{X}_{t-1})e_t, \end{aligned} \tag{3.18}$$

que incorpora el término $\boldsymbol{\beta}\mathbf{Z}_t$, donde $\boldsymbol{\beta}$ es un vector de coeficientes, en la ecuación de observación de (3.16). Se permite que u_t dependa de $\boldsymbol{\beta}\mathbf{Z}_t$ para poder considerar modelos con errores heterocedásticos (en particular errores de tipo multiplicativo). Ahora bien, si extendemos el vector de estados para contener el vector de parámetros, denotando por $\tilde{\mathbf{X}}_t = [\mathbf{X}_t \quad \boldsymbol{\beta}]'$, podemos reescribir (3.18) como un caso particular del modelo SSOE general verificando

$$\begin{aligned} Y_t &= w_t^*(\tilde{\mathbf{X}}_{t-1}) + u_t^*(\tilde{\mathbf{X}}_{t-1})e_t, \\ \tilde{\mathbf{X}}_t &= \mathbf{M}_t^*(\tilde{\mathbf{X}}_{t-1}) + \mathbf{v}_t^*(\tilde{\mathbf{X}}_{t-1})e_t, \end{aligned} \tag{3.19}$$

de manera que

$$\begin{aligned} w_t^*(\tilde{\mathbf{X}}_{t-1}) &= w_t(\mathbf{X}_{t-1}, \boldsymbol{\beta}\mathbf{Z}_t), & u_t^*(\tilde{\mathbf{X}}_{t-1}) &= u_t(\mathbf{X}_{t-1}, \boldsymbol{\beta}\mathbf{Z}_t), \\ \mathbf{M}_t^*(\tilde{\mathbf{X}}_{t-1}) &= [\mathbf{M}_t(\mathbf{X}_{t-1})' \quad \boldsymbol{\beta}]', & \mathbf{v}_t^*(\tilde{\mathbf{X}}_{t-1}) &= [\mathbf{v}_t(\mathbf{X}_{t-1})' \quad 0]'. \end{aligned}$$

En virtud de la representación (3.19), podemos proceder con la estimación e inferencia para el modelo tal y como hemos indicado en la Sección 3.2.1. Notemos que al pasar a formar parte del vector de estados, las estimaciones de los coeficientes de regresión se obtienen como parte del vector $\tilde{\mathbf{X}}_0$.

La introducción de esta componente regresiva, no viene sin un precio, pues los modelos pasarán a ser inestables. Para argumentar por qué, consideremos el caso simplificado de un modelo SSOE lineal. El sistema (3.19) se particulariza en

$$\begin{aligned} Y_t &= (\mathbf{w}_t^*)'\tilde{\mathbf{X}}_{t-1} + e_t, \\ \tilde{\mathbf{X}}_t &= \mathbf{M}^*\tilde{\mathbf{X}}_{t-1} + \mathbf{v}^* e_t, \end{aligned} \tag{3.20}$$

con vectores y matriz definidos a partir de las siguientes ecuaciones.

$$\mathbf{w}_t^* = [\mathbf{w}' \quad \mathbf{Z}_t']', \quad \mathbf{M}^* = \begin{bmatrix} \mathbf{M} & \mathbf{0}_{\dim(\mathbf{M})\times\dim(\boldsymbol{\beta})} \\ \mathbf{0}_{\dim(\boldsymbol{\beta})\times\dim(\mathbf{M})} & \mathbf{I}_{\dim(\boldsymbol{\beta})\times\dim(\boldsymbol{\beta})} \end{bmatrix}, \quad \mathbf{v}^* = [\mathbf{v}' \quad 0]'.$$

Se cumplirá, por tanto,

$$\tilde{\mathbf{X}}_t = \left(\prod_{j=1}^{t} \tilde{\mathbf{M}}_j^*\right) \mathbf{X}_0 + \sum_{i=0}^{t-1} \left(\prod_{j=0}^{i} \tilde{\mathbf{M}}_j^*\right) \mathbf{v}^* Y_{t-i}, \quad \tilde{\mathbf{M}}_t^* = \mathbf{M}^* - \mathbf{v}^*(\mathbf{w}_t^*)' = \begin{bmatrix} \mathbf{M} - \mathbf{v}\mathbf{w}' & -\mathbf{v}\mathbf{Z}_t' \\ \mathbf{0}_{\dim(\boldsymbol{\beta})\times\dim(\mathbf{M})} & \mathbf{I}_{\dim(\boldsymbol{\beta})\times\dim(\boldsymbol{\beta})} \end{bmatrix},$$

con $\tilde{\mathbf{M}}_0^* = \mathbf{I}$. Empleando las reglas del producto de matrices por bloques, se puede deducir la igualdad

$$\left(\prod_{j=1}^{t} \tilde{\mathbf{M}}_j^*\right) = \begin{bmatrix} (\mathbf{M} - \mathbf{v}\mathbf{w}')^t & -\sum_{i=0}^{t} (\mathbf{M} - \mathbf{v}\mathbf{w}')^i \mathbf{v}\mathbf{Z}_{t-i}' \\ \mathbf{0}_{\dim(\boldsymbol{\beta})\times\dim(\mathbf{M})} & \mathbf{I}_{\dim(\boldsymbol{\beta})\times\dim(\boldsymbol{\beta})}, \end{bmatrix}$$

de modo que, aunque el modelo SSOE lineal sin componente regresiva sea estable, al añadirle esta componente dicha estabilidad se pierde, por la constancia del vector de coeficientes de regresión dentro del vector de estados. Esto no deja de ser un tecnicismo, puesto que el verdadero problema radica en el bloque $-\sum_{i=0}^{t} (\mathbf{M} - \mathbf{v}\mathbf{w}')^i \mathbf{v}\mathbf{Z}_{t-i}' \xrightarrow{t\to\infty} -\sum_{i=0}^{\infty} (\mathbf{M} - \mathbf{v}\mathbf{w}')^i \mathbf{v}\mathbf{Z}_{t-i}'$, que recoge el efecto que tendrán los coeficientes de regresión sobre el resto de estados del modelo. En el mejor de los casos, en el que la serie anterior converja, existirá un efecto que tiende a ser constante del vector de coeficientes sobre los estados del modelo, mientras que en el peor, algunas de las componentes del vector de estados tendrán un comportamiento explosivo causado por esta componente regresiva. De todos modos, poder incluir covariables permite realizar tareas básicas de inferencia con series de tiempo como la detección de datos atípicos y el análisis de intervención.

Con respecto a la detección de atípicos, en realidad podremos detectar atípicos aditivos, definidos para el modelo SSOE general de manera análoga a la Definición 2.11, como aquellos valores para los que el coeficiente de regresión asociado a la covariable $Z_t = \mathbf{1}(t = r)$, $1 \leq r \leq n$, es significativamente distinto de 0. Para detectar atípicos innovativos, para los que la perturbación se produce sobre alguna innovación, se podría (por ejemplo) modificar la función de verosimilitud añadiendo un coeficiente extra que recoja el desplazamiento de la innovación afectada. Esto es, si existe un atípico innovativo en el instante r, la innovación asociada pasa a tomar el valor $\tilde{e}_r = k + e_r$ (siendo k el desplazamiento). La densidad de la innovación cumplirá $f_{\tilde{e}_r}(x) = f_{e_r}(x - k)$.

Por otro lado, en lugar de emplear variables indicadoras para detectar perturbaciones puntuales, podrían emplearse covariables que recojan un cambio permanente en la serie a partir de un instante, como $Z_t = \mathbf{1}(t \geq r)$, u otro tipo de variaciones. Los coeficientes de regresión estimados determinarían si esos cambios son significativamente distintos de 0, o no lo son.

3.2.3. Validación y selección de modelos SSOE

Los modelos SSOE se basan una serie de hipótesis sobre los procesos estocásticos generadores de las series de tiempo que es pertinente verificar. De lo contrario, nos arriesgamos a ajustar modelos subóptimos e incluso modelos que den lugar a predicciones alejadas ya no de la realidad, pues resultará prácticamente imposible encontrar un modelo que refleje por completo la evolución de una auténtica serie de tiempo, si no de cualquier comportamiento esperable a la vista de los datos disponibles. Por ejemplo, los modelos de suavización exponencial forman parte de un grupo de modelos llamados "modelos de descomposición", pues asumen que pueden descomponer el comportamiento en media condicionada de la serie en una serie de términos, que verificarán ciertas ecuaciones de recurrencia con respecto a valores pasados. Al tratar de ajustar un modelo con componente estacional a una serie que no refleje dicha componente, se podrían identificar perturbaciones aleatorias como un patrón recurrente, que será reflejado hacia predicciones futuras. Por otro lado, podríamos estar quedándonos cortos si no incluimos una componente que recoja una tendencia a largo plazo del modelo y en su lugar contamos solamente con una componente que trate de actualizar el nivel de la misma. Si la serie presentase

una componente estacional que incrementase su peso con el nivel de la serie, en lugar de mantenerse más o menos constante a lo largo del tiempo, deberemos incorporar esta información en el modelo que propongamos. Por ejemplo, empleando modelos de la forma ETS$(\cdot,\cdot,M)$ o transformando los datos para eliminar este comportamiento, como se sugiere para los modelos ARIMA. Lo mismo se puede decir para el caso de modelos heterocedásticos, pues de lo contrario obtendremos estimaciones de la variabilidad de la serie, e intervalos de predicción, incorrectos. Para poder identificar estas características previo ajuste de un modelo, será necesario recurrir al gráfico secuencial de la serie de tiempo. Sin embargo, la forma más efectiva de determinar si los datos se ajustan a las hipótesis del modelo propuesto radica en el análisis de los residuos. En este sentido, las ideas coinciden con las expuestas en la sección 2.1.3 para modelos Box-Jenkins. Notemos que el incumplimiento de ciertas hipótesis no tiene por qué resultar en la invalidación del modelo, si no que pueden servir como guía para determinar las carencias o los excesos del modelo inicialmente propuesto. Si los residuos no están centrados en 0 puede ser necesario añadir una constante; si no parecen gaussianos quizá sea conveniente emplear otra hipótesis distribucional o, en caso de que el tamaño muestral sea suficientemente grande, asumir que el error de los estimadores será razonablemente pequeño y recurrir a intervalos de predicción basados en simulación; si no se pueden asumir como incorrelados, quiere decir que el modelo no captura adecuadamente la estructura de dependencia de los datos, y habrá que proponer algún otro en consonancia.

La selección, de igual modo que con los modelos Box-Jenkins, se puede realizar en base a algun criterio de error. Algunos de los más conocidos (Hyndman et al., 2008) son el SMAPE ("Simmetric Mean Absolute Percentage Error"), definido como

$$\text{SMAPE} = 100\frac{1}{n}\sum_{i=1}^{n}\frac{|e_t|}{|y_t| + |\hat{y}_{t|t-1}|},$$

que mide la desviación, en promedio, del error relativo entre los valores ajustados y los observados con respecto a la suma de los valores absolutos de los ajustes y las observaciones, pudiendo tomar valores entre 0 y 100, o el MSE ("Mean Squared Error"),

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_t^2,$$

que no es más que la suma de residuos al cuadrado usual. El MSE devuelve una medida del error en términos absolutos, en escala cuadrática respecto a las unidades de la serie, de modo que será útil como criterio comparativo entre modelos aplicados a una misma serie, pero no como medida aislada. En cambio, el SMAPE es una medida del error relativa, en porcentaje, que permite comparar ajustes para series distintas y también ofrece una idea de la bondad del ajuste del modelo; no obstante, cuando los valores observados y sus ajustes están muy próximos a 0, aparecerán problemas numéricos (al encontrarse dichas cantidades en el denominador). Es habitual basarse no en uno, sino en varios de estos criterios simultáneamente, para compensar sus carencias.

En el caso particular de los modelos SSOE ajustados por máxima verosimilitud condicionada (por el vector de estados inicial), se puede emplear algún criterio de información basado en dicha verosimilitud, para escoger como modelo más adecuado aquel que minimice el criterio en cuestión. Esta es una ventaja adicional frente al ajuste por máxima verosimilitud no condicionada, pues el empleo de determinados filtros vuelven las verosimilitudes no comparables. Esto ocurre por ejemplo en caso de los modelos ARIMA con distintos grados de diferenciación que empleen el filtro de Kalman (Hyndman et al. 2008, p.195) para la estimación máximo verosimil. Algunos de los criterios de información más empleados son el AICc ("Akaike information criterion bias corrected") definido como

$$\text{AICc} = -2\log(\mathcal{L}(\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0|y_1, \cdots, y_n)) + \frac{2nN_{\text{param}}}{n - N_{\text{param}} - 1}$$

donde N_{param} es el número de parámetros estimados, que se trata de una versión corregida por sesgo del criterio AIC ("Akaike information criteria"), que minimiza asintóticamente el error de predicción

cometido; y el criterio BIC ("Bayesian information criterion"), definido por

$$\text{BIC} = -2\log(\mathcal{L}(\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0|y_1, \cdots, y_n)) + N_{\text{param}}\log(n),$$

que seleccionará, cuando el tamaño muestral es suficientemente elevado, modelos con el número adecuado de parámetros (asumiendo que existe algún modelo correcto). En este sentido, el criterio BIC tiende a escoger modelos más parsimoniosos, mientras que el criterio AIC tiende a obtener mejores ajustes (sobreajustando en algunas ocasiones).

3.2.4. Enfoque bayesiano para series de tiempo

En las secciones previas hemos tratado el problema de estimación y predicción de los modelos SSOE desde un punto de vista frecuentista. Hemos considerado que los parámetros involucrados en el modelo, aunque desconocidos, eran fijos. Además, hemos asumido que existía un vector de estados inicial, $\mathbf{X}_0$, que también hemos considerado como fijo, por convenciencia. En esta sección veremos como abordar estas tareas desde el punto de vista bayesiano, en el que se trata tanto a los parámetros del modelo como al vector de estados iniciales como aleatorios, y discutiremos las repercusiones de este punto de vista alternativo sobre el modelo. A pesar del cambio de perspectiva, seguiremos hablando de "parámetros"para no perder de vista el contexto, aunque emplearemos la coletilla de "aleatorios". Además, para no repetirnos constantemente, incluiremos el vector de estados iniciales dentro de esa categoría de "parámetros aleatorios".

La estadística bayesiana se fundamenta en el teorema de Bayes, que ya hemos mencionado en secciones previas. Este teorema permite relacionar una distribución incondicionada, conocida como distribución "a priori", con otra distribución condicionada por los datos observados, llamada distribución "a posteriori". En el caso particular que nos ocupa, dada la serie de tiempo observada $(y_1, \cdots, y_n)$, los parámetros $(\Upsilon, \sigma^2, \mathbf{X}_0)$ y una posible muestra para ellos $(\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0)$, el teorema de Bayes establece que

$$f_{\Upsilon,\sigma^2,\mathbf{X}_0}(\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0|y_1, \cdots, y_n) = \frac{f_{Y_1,\cdots,Y_n}(y_1, \cdots, y_n|\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0) f_{\Upsilon,\sigma^2,\mathbf{X}_0}(\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0)}{f_{Y_1,\cdots,Y_n}(y_1, \cdots, y_n)}, \tag{3.21}$$

aunque si restringimos la atención a la distribución de los parámetros y vector de estados iniciales, el denominador en (3.21) no deja de ser una constante normalizadora que asegura que el numerador es, efectivamente, una función de densidad. Así pues, es habitual hacer referencia al teorema de Bayes en términos de proporcionalidad entre la distribución a posteriori (izquierda de la igualdad en (3.21)) y el producto entre la densidad condicionada de la muestra y la distribución a priori (derecha de la igualdad en (3.21)), tal y como escribimos a continuación:

$$f_{\Upsilon,\sigma^2,\mathbf{X}_0}(\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0|y_1, \cdots, y_n) \propto f_{Y_1,\cdots,Y_n}(y_1, \cdots, y_n|\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0) f_{\Upsilon,\sigma^2,\mathbf{X}_0}(\hat{\Upsilon}, \hat{\sigma}^2, \hat{\mathbf{X}}_0). \tag{3.22}$$

Fijémonos en los términos de la ecuación en (3.22). La densidad a priori representará las ideas preconcebidas sobre la distribución de los parámetros que, una vez observada la serie de tiempo que caracterizan bajo el modelo que se haya especificado, se actualizará dando lugar a la densidad a posteriori. Notemos además que la densidad condicionada de la serie de tiempo dados los parámetros del modelo, no es otra cosa que la función de verosimilitud asociada al modelo. Por tanto, teniendo en cuenta (3.22), la estimación máximo verosimil coincide con la estimación que maximiza el ratio entre la distribución a posteriori y la distribución a priori.

Observación 3.12. Si imponemos que la densidad a priori tome el valor constantemente igual a 1, que equivaldría a decir que no se posee ninguna información sobre la distribución de los parámetros, la densidad a posteriori coincidirá con la función de verosimilitud, de modo que el estimador máximo verosimil será la moda de la distribución a posteriori (estimador máximo a posteriori o MAP). Técnicamente, si el espacio paramétrico no fuese acotado, una función constante no sería una densidad, aunque

en la práctica no resulte un problema. Ese tipo de densidades a priori que realmente no sean tales, pero que dada la constante normalizadora de (3.21) se pueden introducir en la fórmula de Bayes dando lugar a densidades a posteriori bien definidas, se denominan impropias y son ampliamente empleadas.

¿Cómo podemos asegurar que la distribución a posteriori, que depende de la subjetividad introducida a través de la distribución a priori, puede considerarse una alternativa razonable a la estimación frecuentista por máxima verosimilitud? Una extensión del teorema de Bernstein-von Mises para procesos estocásticos (Basawa et al., 1980, Ch.10) establece que, bajo condiciones de regularidad análogas a las requeridas para garantizar la normalidad asintótica del estimador de máxima verosimilitud, junto a alguna restricción sobre la distribución a priori, la distribución a posteriori se comportará, asintóticamente, como una distribución normal centrada en el estimador máximo verosimil con matriz de covarianzas la inversa de la matriz de información de Fisher, ponderada por el tamaño muestral (esto implica que la distribución a posteriori converge a una distribución degenerada en el verdadero valor del parámetro). Nos limitaremos a enunciar, informalmente, este resultado. Para más detalle, véase Basawa et al. (1980).

Teorema 3.13 (Bernstein von-Mises, Basawa et al., 1980)**.** Dada la serie de tiempo $Y_1, \cdots, Y_n$, y bajo ciertas condiciones de regularidad, la distribución a posteriori, $P_{\Upsilon,\sigma^2,\mathbf{X}_0|Y_1,\cdots,Y_n}$, verificará

$$P_{\Upsilon,\sigma^2,\mathbf{X}_0|Y_1,\cdots,Y_n} \approx N\left((\hat{\Upsilon},\hat{\sigma}^2,\hat{\mathbf{X}}_0), \frac{1}{n} I^{-1}(\hat{\Upsilon},\hat{\sigma}^2,\hat{\mathbf{X}}_0)\right),$$

siendo $(\hat{\Upsilon},\hat{\sigma}^2,\hat{\mathbf{X}}_0)$ el estimador de máxima verosimilitud de $(\Upsilon,\sigma^2,\mathbf{X}_0)$ y $I^{-1}(\hat{\Upsilon},\hat{\sigma}^2,\hat{\mathbf{X}}_0)$ la matriz de información de Fisher evaluada en dicho estimador.

Hemos establecido un claro enlace entre los enfoques frecuentistas y bayesianos, que sugiere el uso de la distribución a posteriori como mecanismo de estimación para los modelos. En concreto, teniendo en cuenta la descomposición de la función de verosimilitud propuesta en (3.17), la distribución a posteriori queda completamente especificada fijando una distribución a priori. Esto no significa que vayamos a obtener una expresión explícita para la densidad a posteriori, algo que solamente sería posible para distribuciones a priori y verosimilitudes muy concretas. Por ejemplo, en Forbes et al. (2000) se trata el caso particular de los modelo SSOE lineales gaussianos en el que se impone que $f_{\Upsilon,\sigma^2,\mathbf{X}_0}(\hat{\Upsilon},\hat{\sigma}^2,\hat{\mathbf{X}}_0) \propto \sigma^{-2} f_\Upsilon(\hat{\Upsilon})$. Se puede demostrar, en base a la descomposición

$$f_{\Upsilon,\sigma^2,\mathbf{X}_0}(\hat{\Upsilon},\hat{\sigma}^2,\hat{\mathbf{X}}_0|y_1,\cdots,y_n) = f_{\mathbf{X}_0}(\hat{\mathbf{X}}_0|y_1,\cdots,y_n,\hat{\Upsilon},\hat{\sigma}^2) f_{\sigma^2}(\hat{\sigma}^2|y_1,\cdots,y_n,\hat{\Upsilon}) f_\Upsilon(\hat{\Upsilon}|y_1,\cdots,y_n),$$

que $f_{\mathbf{X}_0}(\hat{\mathbf{X}}_0|y_1,\cdots,y_n,\hat{\Upsilon},\hat{\sigma}^2)$ es la densidad de una distribución normal multivariante, mientras que $f_{\sigma^2}(\hat{\sigma}^2|y_1,\cdots,y_n,\hat{\Upsilon})$ es la densidad de una distribución gamma inversa. Por su parte, la densidad $f_\Upsilon(\hat{\Upsilon}|y_1,\cdots,y_n)$ será una transformación sin especificar de $f_\Upsilon(\hat{\Upsilon})$.

Como obtener epxresiones explícitas para la distribución a posteriori es desde complicado hasta imposible, lo habitual es recurrir a algoritmos numéricos que permitan, o bien calcular la moda de dicha distribución, o bien obtener "muestras"para esa distribución a posteriori. El primero de estos enfoques nos lleva a la estimación máxima a posteriori (MAP), que no deja de ser una estimación puntual para los parámetros del modelo. El segundo, permite construir una aproximación empírica de la distribución a posteriori que, si bien puede emplearse para construir funcionales de la misma, como la media (estimador de Bayes), también es compatible con la obtención de intervalos de predicción basados en simulación tal y como hemos mencionado en la Sección 3.2.1, pero permitiendo incorporar la variabilidad asociada a la incertidumbre de los parámetros del modelo, ya que

$$f_{(Y_{n+h},\Upsilon,\sigma^2,\mathbf{X}_0)|Y_1,\cdots,Y_n} = f_{Y_{n+h}|Y_1,\cdots,Y_n,\Upsilon,\sigma^2,\mathbf{X}_0} f_{(\Upsilon,\sigma^2,\mathbf{X}_0)|Y_1,\cdots,Y_n}.$$

Así pues, simular de la distribución a posteriori y encadenar dicha simulación con otra del modelo específicado con los valores simulados de los parámetros es equivalente a simular de la distribución conjunta entre la predicción y los parámetros condicionada por la información de la serie de tiempo. En particular, quedándose con la distribución marginal asociada a la predicción tendremos la información que buscamos, habiendo incorporado la incertidumbre asociada a la estimación.

Observación 3.14. Teniendo en cuenta el Teorema 3.3, que nos garantiza que cualquier distribución a posteriori, bajo ciertas condiciones, se comporta, asintóticamente, como una distribución normal centrada en el estimador de máxima verosimilitud, podría resultar razonable exportar la idea de simulación compuesta al contexto frecuentista, para incorporar la variabilidad asociada al estimador de máxima verosimilitud en la construcción de intervalos de confianza. En concreto, se podrían simular valores para el parámetro a partir de la distribución normal asintótica de este estimador, y con ellos obtener simulaciones de las predicciones.

Algunos ejemplos de algoritmos numéricos destinados a la estimación máxima a posteriori, que maximizarán el producto entre la verosimilitud del modelo y la densidad a priori, son el método de Newton o el algoritmo L-BFGS ("Limited memory- BFGS", empleado por el lenguage de programación Stan, en el que se basan las implementaciones de la estimación máxima a posteriori para los modelos Orbit en el paquete `Orbit-ml`). En Nocedal y Stephen (2006) puede encontrarse información detallada acerca de ambos métodos.

Por otro lado, algunos de los algoritmos que permiten obtener simulaciones de muestras distribuidas según la distribución a posteriori, fijada una distribución a priori y el modelo subyacente a la serie de tiempo, se basan en el método de simulación por cadenas de Markov Monte Carlo ("MCMC"). Este tipo de métodos buscan simular secuencias de valores que se puedan asumir como trayectorias parciales de procesos estocásticos que tengan la propiedad de Markov.

Definición 3.15. Un proceso estocástico $\{a_t\}_{t\geq 0}$ posee la propiedad de Markov si para cualquier colección de instantes $0 \leq s_1 \leq s_2 \leq \cdots \leq s_k \leq t$, con $k \geq 1$, se verifica

$$F_{a_t|a_{s_1},\cdots,a_{s_k}} = F_{a_t|a_{s_k}},$$

donde $F_{X|Y}$ denota la función de distribución de X condicionada por Y.

De hecho, se busca que la "cadena" de Markov (realmente el término cadena estaría restringido a procesos de Markov cuyas variables pueden tomar un conjunto numerable de valores, mientras que en este contexto se pretende simular de distribuciones que pueden ser continuas, aunque esta distinción no sea relevante desde el momento en que empleamos un ordenador para la simulación, discretizando el conjunto de posibles valores que se pueden tomar) sea estacionaria en el tiempo y tenga una distribución estacionaria.

Definición 3.16. Una cadena de Markov, $\{a_t\}_{t\geq 0}$, es estacionaria en el tiempo si para cualesquiera instantes $0 \leq s \leq t$,

$$F_{a_t|a_s} = F_{a_{t-s}|a_0}.$$

Una cadena posee una distribución estacionaria (o distribución de equilibrio) F_{eq} si

$$F_{a_0} = F_{eq} \implies F_{a_t} = F_{eq} \quad \forall t \geq 0.$$

En el caso del lenguaje de programación Stan (y por extensión, el caso de los modelos Orbit implementados en `Orbit-ml`), se emplea el conocido como método de Monte Carlo hamiltoniano ('HMC'), algoritmo particular dentro de la familia de métodos basados en cadenas de Markov Monte Carlo. La idea fundamental detrás de este método consiste en simular no de la distribución a posteriori de interés, con densidad $f_{\Upsilon,\sigma^2,\mathbf{X}_0}(\cdot|y_1,\cdots,y_n)$, si no de la distribución conjunta del vector ampliado $(\Upsilon,\sigma^2,\mathbf{X}_0,\zeta)|_{Y_1,\cdots,Y_n}$, teniendo ζ la misma dimensión que el vector de parámetros original y siendo independiente de estos parámetros y de la serie de tiempo, de tal forma que tengamos $f_{\Upsilon,\sigma^2,\mathbf{X}_0,\zeta|Y_1,\cdots,Y_n} = f_{\Upsilon,\sigma^2,\mathbf{X}_0|Y_1,\cdots,Y_n} f_\zeta$. Como distribución de ζ se elige habitualmente la distribución normal multivariante, con media 0 y matriz de covarianzas Σ_ζ que, cuanto más parecida sea a la inversa de la matriz de covarianzas de la distribución a posteriori que perseguimos, más eficiente será el método HMC. Además, se pretende que la distribución estacionaria para las cadenas de Markov construidas sea precisamente la distribución del vector ampliado $(\Upsilon,\sigma^2,\mathbf{X}_0,\zeta)|_{Y_1,\cdots,Y_n}$, de modo que una cadena de Markov de tamaño N generada con dicha distribución estacionara dará lugar a una

muestra de tamaño N de $(\Upsilon, \sigma^2, \mathbf{X}_0, \zeta)|_{Y_1,\cdots,Y_n}$ y, en particular, una muestra de ese tamaño para la distribución a posteriori de los parámetros. En Brooks et al. (2011) puede consultarse los detalles de este y otros métodos similares destinados al muestreo para la dsitribución a posteriori basados en cadenas de Markov.

Si bien no explicaremos por qué este tipo de métodos permiten obtener simulaciones que puedan ser consideradas como muestras asociadas a la distribución a posteriori de interés, es importante destacar que los algoritmos necesitan de un periodo de "calentamiento"(warmup) hasta que se logra alcanzar (aproximadamente) una distribución estacionaria para la cadena que se va construyendo. A partir de dicho periodo, todos los valores obtenidos para la cadena pasan a ser considerados como simulaciones para la distribución a posteriori. Para tener una idea de si, efectivamente, se alcanza una distribución estacionaria con las propiedades deseadas, es habitual construir no una, si no un determinado número, n_c (habitualmente $n_c = 4$), de cadenas de Markov, y calcular algún estadístico que ayudará a determinar el estado de la cadena.

El estadístico $\hat{R}$, definido para cada parámetro $\upsilon \in (\Upsilon, \sigma^2, \mathbf{X}_0)$, comparará estimaciones de varianzas en una misma cadena y varianzas entre distintas cadenas. Se dividen las cadenas de Markov (tras el calentamiento) en 2 trozos distintos dando lugar a $2n_c$ trozos. Para cada trozo de cadena, de tamaño $N/2$, se computan, a partir de las simulaciones υ_{ij} con $i = 1, \cdots, N/2$, $j = 1, \cdots, 2n_c$:

$$\sigma^2_{bet} = \frac{N/2}{2n_c - 1}\sum_{j=1}^{2n_c}(\bar{\upsilon}_{\cdot j} - \bar{\upsilon}_{\cdot\cdot})^2, \quad \text{donde} \quad \bar{\upsilon}_{\cdot j} = \frac{1}{N/2}\sum_{i=1}^{N/2}\upsilon_{ij}, \quad \bar{\upsilon}_{\cdot\cdot} = \frac{1}{2n_c}\sum_{j=1}^{2n_c}\bar{\upsilon}_{\cdot j},$$

$$\sigma^2_{in} = \frac{1}{2n_c}\frac{1}{N/2 - 1}\sum_{j=1}^{2n_c}\sum_{i=1}^{N/2}(\upsilon_{ij} - \bar{\upsilon}_{\cdot j})^2.$$

Se define $\hat{R}$ como

$$\hat{R} = \sqrt{\frac{\frac{2n_c-1}{2n_c}\sigma^2_{in} + \frac{1}{2n_c}\sigma^2_{bet}}{\sigma^2_{in}}},$$

que tenderá a 1 (Brooks et al., 2011) cuando $N \to \infty$ bajo ciertas condiciones, y se interpreta como un factor de reducción de la varianza de la distribución actual de los parámetros en la cadena a medida que N tiende a infinito. Se suele tomar como umbral $\hat{R} < 1.1$ para decidir que ha habido convergencia. Este factor será más fiable a medida que la distribuciones marginales a posteriori de los parámetros se aproximen a una distribución normal.

Para validar un modelo basado en la estimación bayesiana, la manera más simple de hacerlo, por analogía con el enfoque frecuentista, es fijar valores para los parámetros (por ejemplo el estimador Bayes o el MAP) y seguir con un análisis de residuos usual. Del mismo modo, a partir de esos valores fijados podemos calcular la verosimilitud del modelo y con ella criterios de información, que podremos emplear para seleccionar entre diversos modelos, siempre que las verosimilitudes sean comparables. También podremos basarnos en criterios de error como los expuestos en la Seccción 3.2.3. Esta metodología sin duda sería criticada por adeptos al bayesianismo al recurrir a predicciones puntuales de los parámetros en lugar de emplear toda la distribución a posteriori.

Una alternativa completamente bayesiana pasa por considerar el factor de Bayes, utilizado como herramienta bayesiana de contraste de hipótesis, profundamente ligada al criterio BIC. Supongamos que queremos elegir entre dos modelos alternativos caracterizados por ciertos parámetros: el modelo H_0 con parámetros τ_0 y el modelo H_1 con parámetros τ_1. El factor de Bayes se define como

$$B_{01} = \frac{f_{Y_1,\cdots,Y_n|H_0}}{f_{Y_1,\cdots,Y_n|H_1}} = \frac{\int f_{Y_1,\cdots,Y_n|\tau_0,H_0} f_{\tau_0|H_0} d\tau_0}{\int f_{Y_1,\cdots,Y_n|\tau_1,H_1} f_{\tau_1|H_1} d\tau_1} = \frac{\mathbb{E}_{\tau_0|H_0}[f_{Y_1,\cdots,Y_n|\tau_0,H_0}]}{\mathbb{E}_{\tau_1|H_1}[f_{Y_1,\cdots,Y_n|\tau_1,H_1}]},$$

donde $f_{Y_1,\cdots,Y_n|\tau_i,H_i}$ representa la densidad a posteriori en el modelo i y $f_{\tau_i|H_i}$ la densidad a priori para dicho modelo, $1 \leq i \leq 2$. El factor de Bayes no es más que un ratio de verosimilitudes marginal, habiendo integrado la verosimilitud del modelo sobre el espacio paramétrico, lo que le confiere un

carácter de robustez ante la dimensión de dicho espacio. Un factor mayor que uno sugiere que H_0 es más plausible que H_1 y viceversa. S suele fijar un umbral en torno a 3 (equivalentemente $\frac{1}{3}$) para descartar evidencia anecdótica (Kass et al., 1995), aunque no hay un apoyo formal (no existe una distribución con la que comparar el valor del factor) para este criterio. Cabe destacar que es una herramienta comparativa simétrica, en la que se trata ambas hipótesis con el mismo peso. Además, se puede demostrar que, asintóticamente,

$$-2\log(B_{01}) \approx \mathrm{BIC}_0 - \mathrm{BIC}_1$$

de modo que un factor de Bayes $B_{01} > 1$ es asintóticamente equivalente a que el criterio de información sea menor para el modelo 0 que para el modelo 1 y, por tanto, que el modelo 0 sea preferible en términos del criterio BIC. Basándonos en esta discusión, nos centraremos en el criterio BIC como criterio para la selección de modelos basados en estimación bayesiana.

Observación 3.17. El factor de Bayes puede ser aproximado computacionalmente, al disponer de una muestra para la distribución a posteriori, por Monte Carlo:

$$\mathbb{E}_{\tau_0|H_0}[f_{Y_1,\cdots,Y_n|\tau_0,H_0}] \approx \frac{1}{Nn_c}\sum_{i=1}^{Nn_c} f_{Y_1,\cdots,Y_n|\tau_{0\,i},H_0}.$$

También podrían emplearse métodos de integración más sofisticados (basados en muestreo por importancia, por ejemplo).

Capítulo 4

Modelos Orbit para series de tiempo

En este capítulo revisaremos los dos modelos propuestos por Ng et al. (2020) como parte del paquete `Orbit-ml`, para el lenguaje de programación Python. Esta es la única referencia, junto con la documentación del propio paquete (https://orbit-ml.readthedocs.io/en/latest/index.html) que está disponible al respecto. Dichos modelos van a poder ser expresados como modelos de estado de espacios con única fuente de error y, por tanto, se les podría aplicar toda la metodología revisada en el Capítulo 3.

Nos centraremos en reescribir los modelos, a partir de su formulación original, como modelos SSOE para estudiar sus similitudes teóricas con algún modelo de suavización exponencial bien conocido, tratando de sacar conclusiones sobre la estabilidad de los métodos. Posteriormente abordaremos la implementación particular que se ha llevado a cabo para estos modelos Orbit. Como ya he señalado la información disponible acerca de este último aspecto es limitada, por lo que hemos tenido que profundizar en el propio código del paquete en busca de aquellos detalles ausentes en las referencias propuestas por los autores.

4.1. Modelo de tendencia local y global

El primer modelo que vamos a discutir recibe el nombre de "modelo de tendencia local y global" (LGT), y consiste en una modificación del modelo de Holt-Winters aditivo. Manteniendo la notación de Y_t para referirnos a la variable aleatoria asociada al proceso generador de una serie de tiempo en el instante t, la formulación original para el modelo LGT, de acuerdo con Ng et al. (2020), es la siguiente:

$$\begin{aligned} Y_t &= l_{t-1} + \xi_1 b_{t-1} + \xi_2 l_{t-1}^{\lambda} + s_t^* + e_t, \\ l_t &= \alpha^*(Y_t - s_t^*) + (1-\alpha^*) l_{t-1}, \\ b_t &= \beta^*(l_t - l_{t-1}) + (1-\beta^*) b_{t-1}, \\ s_{t+m}^* &= \gamma^*(Y_t - l_t) + (1-\gamma^*) s_t^*. \end{aligned} \tag{4.1}$$

Fijémonos en que para este modelo, la componente estacional se computa "hacia el futuro" en lugar de 'hacia el pasado'. Para continuar trabajando con la notación del Capítulo 3, reescribiremos la componente estacional atrasándola m instantes ($s_{t+m}^* \equiv s_t$), pues es mera cuestión de notación. Fijémonos también en que la componente de nivel, l_{t-1}, aparece 2 veces en la ecuación de observación de (4.1) y, en una de ellas, se encuentra elevada a cierto exponente λ. En teoría, el término $\xi_2 l_{t-1}^{\lambda}$ pretende recoger el comportamiento global de la tendencia de la serie (tendencia global) mientras que $l_{t-1} + \xi_1 b_{t-1}$ vendría a representar el comportamiento local (tendencia local) de la misma.

Comencemos por reescribir la ecuación asociada a la componente estacional, que depende de l_t, para que dependa de estados previos. Así,

$$\begin{aligned} s_t &= \gamma^*(Y_t - l_t) + (1-\gamma^*)s_{t-m} = \gamma^*(Y_t - \alpha^*(Y_t - s_{t-m}) - (1-\alpha^*)l_{t-1}) + (1-\gamma^*)s_{t-m}, \\ &= \gamma^*(1-\alpha^*)(Y_t - l_{t-1}) + (1-\gamma^*(1-\alpha^*))s_{t-m} \equiv \gamma(Y_t - l_{t-1}) + (1-\gamma)s_{t-m}, \end{aligned}$$

donde $\gamma = \gamma^*(1-\alpha^*)$. Ahora, sustituyendo la ecuación de observación en las ecuaciones de actualización de las distintas componentes de (4.1),

$$\begin{aligned} l_t &= \alpha^*(l_{t-1} + \xi_1 b_{t-1} + \xi_2 l_{t-1}^\lambda + e_t) + (1-\alpha^*)l_{t-1} = l_{t-1} + \alpha^*\xi_1 b_{t-1} + \alpha^*\xi_2 l_{t-1}^\lambda + \alpha^* e_t, \\ b_t &= \beta^*(\alpha^*\xi_1 b_{t-1} + \alpha^*\xi_2 l_{t-1}^\lambda + \alpha^* e_t) + (1-\beta^*)b_{t-1} = (1-\beta^*(1-\alpha^*\xi_1))b_{t-1} + \beta^*\alpha^*\xi_2 l_{t-1}^\lambda + \beta^*\alpha^* e_t, \\ s_t &= \gamma(\xi_1 b_{t-1} + \xi_2 l_{t-1}^\lambda + s_{t-m} + e_t) + (1-\gamma)s_{t-m} = s_{t-m} + \gamma\xi_1 b_{t-1} + \gamma\xi_2 l_{t-1}^\lambda + \gamma e_t. \end{aligned}$$

En Ng et al. (2020) no se especifica si se normaliza la componente estacional, de modo que asumiremos que no ocurre. De todos modos, podremos sacar conclusiones sobre el modelo con componentes normalizadas.

Denotando por $\mathbf{X}_t = [l_t \quad b_t \quad s_t \quad \cdots \quad s_{t-m+1}]'$, podemos reescribir el modelo LGT como un modelo SSOE general con la siguiente representación

$$\begin{aligned} Y_t &= w(\mathbf{X}_{t-1}) + e_t, \\ \mathbf{X}_t &= \mathbf{M}(\mathbf{X}_{t-1}) + \mathbf{v}e_t, \end{aligned} \tag{4.2}$$

con w, $\mathbf{M}$ y $\mathbf{v}$ tales que, denotando por $\mathbf{X}_{t-1,1}$ a $[1 \quad 0 \quad \cdots \quad 0]\mathbf{X}_{t-1}$,

$$\mathbf{M}(\mathbf{X}_{t-1}) = \begin{bmatrix} 1 & \alpha^*\xi_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & (1-\beta^*(1-\alpha^*\xi_1)) & 0 & 0 & \cdots & 0 & 0 \\ 0 & \gamma\xi_1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \mathbf{X}_{t-1} + \begin{bmatrix} \alpha^*\xi_2 \\ \beta^*\alpha^*\xi_2 \\ \gamma\xi_2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} (\mathbf{X}_{t-1,1})^\lambda$$

$$w(\mathbf{X}_{t-1}) = [1 \quad \xi_1 \quad 0 \quad \cdots \quad 0 \quad 1]\mathbf{X}_{t-1} + \xi_2(X_{t-1,1})^\lambda \equiv \mathbf{w}'\mathbf{X}_{t-1} + \xi_2(\mathbf{X}_{t-1,1})^\lambda,$$

$$\mathbf{v} = [\alpha^* \quad \beta^*\alpha^* \quad \gamma \quad 0 \quad \cdots \quad 0]'.$$

Denotemos por $\mathbf{M}_1$ a la matriz anterior en la expresión de $\mathbf{M}$. Veamos si podemos estudiar la estabilidad de este modelo SSOE, en principio, no lineal. Para ello, despejemos el error en la ecuación de observación de (4.2) e introduzcámoslo en la ecuación de estados:

$$\mathbf{X}_t = \mathbf{M}(\mathbf{X}_{t-1}) + \mathbf{v}(Y_t - w(\mathbf{X}_{t-1})) = \mathbf{M}(\mathbf{X}_{t-1}) + \mathbf{v}Y_t - \mathbf{v}\mathbf{w}'\mathbf{X}_{t-1} - \mathbf{v}\xi_2(\mathbf{X}_{t-1,1})^\lambda.$$

Basta con fijarse en que $\mathbf{v}\xi_2(\mathbf{X}_{t-1,1})^\lambda$ coincide con la parte no lineal de la función $\mathbf{M}$, de modo que

$$\mathbf{X}_t = (\mathbf{M}_1 - \mathbf{v}\mathbf{w}')\mathbf{X}_{t-1} + \mathbf{v}Y_t,$$

y la estabilidad del método LGT estará determinada por los autovalores de la matriz $\mathbf{M}_1 - \mathbf{v}\mathbf{w}'$,

$$\mathbf{M}_1 - \mathbf{v}\mathbf{w}' = \begin{bmatrix} 1-\alpha^* & 0 & 0 & 0 & \cdots & 0 & -\alpha^* \\ -\beta^*\alpha^* & 1-\beta^* & 0 & 0 & \cdots & 0 & -\beta^*\alpha^* \\ -\gamma & 0 & 0 & 0 & \cdots & 0 & 1-\gamma \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & \cdots & 1 & 0 \end{bmatrix}.$$

Uno de los autovalores de esta matriz es el 1, cuto autovector asociado es $[-1 \;\; 0 \;\; 1 \;\; \cdots \;\; 1]$, por lo que el método LGT sin normalización no es estable. Sin embargo, por analogía al caso ETS(A, A, A) estudiado por Hyndman et al. (2008, p.155-160), se tendrá que el método LGT con componente estacional normalizada será estable si el resto de autovalores tienen módulo menor que 1. Se puede demostrar (por inducción en $m \geq 1$), que el polinomio característico asociado a $\mathbf{M}_1 - \mathbf{v}\mathbf{w}'$ tiene la forma

$$\det(\mathbf{M}_1 - \mathbf{v}\mathbf{w}' - x\mathbf{I}) = (-1)^{m+1}(1-x)(1-\beta^*-x)(x^{m+1} + \alpha^* x^m + \cdots + \alpha^* x + \gamma + \alpha^* - 1),$$

de modo que para garantizar la estabilidad habría que imponer que $0 < \beta^* < 2$. Las condiciones de estabilidad asociadas a α^* y γ ya no son inmediatas de deducir, sin embargo, en base a la búsqueda de soluciones numéricas para el polinomio $x^{m+1} + \alpha^* x^m + \cdots + \alpha^* x + \gamma + \alpha^* - 1$, considerando varios valores de m, la región clásica de la forma $0 < \alpha^* < 1$ y $0 < \gamma < 1 - \alpha^*$ ($\iff 0 < \gamma^* < 1$) aparenta estar contenida en la región de estabilidad del método.

Pasemos a hablar sobre el tratamiento propuesto por los autores para parámetros y variables del modelo LGT. El enfoque que ofrecen es pseudo-bayesiano. Por una parte, asumen que las innovaciones son i.i.d con distribución T de Student generalizada[1] con media 0, ϑ grados de libertad y varianza σ^2, e imponen como distribución a priori para σ una semi-T de student generalizada[2] de media 0, con 1 grado de libertad y varianza σ_0, con σ_0 calculado como $\sigma_0 = \text{máx}\{Y_1, \cdots, Y_n\}/30$. Para poder emplear λ con valores reales, imponen que l_t y Y_t sean siempre positivos. No explicitan el uso de distribuciones a priori para α^*, β^*, γ^*, aunque restringen los valores que pueden tomar al intervalo $[0, 1]$ (de hecho $\alpha^* \in [0.0001, 1]$). Lo mismo ocurre para ϑ, cuyo valor se restringe a $[5, 40]$ por defecto. Para ξ_2, ξ_1 y λ, solamente podemos afirmar que los dos últimos toman valores positivos, aparantemente en $[0, 1]$ en base a pruebas con la librería, mientras que el primero toma tanto valores positivos como negativos, quizá entre -1 y 1. Con respecto a los estados iniciales, hasta donde sabemos no se preocupan por definir l_0 ni b_0. Siempre se asigna el valor 0 a b_1, mientras que $l_1 + s_1^* = l_1 + s_{-m+1} = Y_1$. Además, $s_1^* + \cdots + s_m^* = 0$ ($\iff s_{-m+1} + \cdots + s_0 = 0$). Los $m-1$ primeros valores de la componente estacional se inicializan simulando a partir de una distribución normal de media 0 y desviación típica 0.05, truncando los valores que se salen del intervalo $[-1, 1]$ a los extremos de dicho intervalo. A partir de $t = 2$ el modelo evoluciona de acuerdo con (4.1).

Si tuviésemos que aventurar el procedimiento de estimación, diríamos que o bien se eligen valores heurísticamente determinados para los parámetros sin distribución a priori establecida (por ejemplo, partiendo de los valores iniciales ya comentados para la componente estacional), o bien se le asocia

[1] Se dice que una variable aleatoria X tiene distribución T de Student generalizada con media μ, varianza σ^2 y ϑ grados de libertad si $(X - \mu)/\sigma^2$ se distribuye de acuerdo con una T de Student con ϑ grados de libertad.

[2] Se dice que una variable aleatoria Y tiene distribución semi-T de Student generalizada con media μ, varianza σ^2 y ϑ grados de libertad si $Y = |X|$ siendo X una variable con distribución T de Student generalizada con esos parámetros.

una distribución uniforme en los intervalos a los que restringen sus valores. A partir de estos valores iniciales, se prosigue con la simulación de valores distribuidos según la distribución a posteriori mediante el método HMC comentado en la Sección 3.2.4.

4.2. Modelo de tendencia local amortiguada

El segundo de los modelos propuesto por Ng et al. (2020) recibe el nombre de "modelo de tendencia local amortiguada" (DLT) y está basado en el modelo de Holt-Winters aditivo con amortiguamiento, ETS(A, A_d, A). En este caso, nos basaremos en la documentación más actualizada de Orbit al respecto (*Damped Local Trend (DLT)*, s.f.).

$$\begin{aligned} Y_t &= D(t) + l_{t-1} + \nu b_{t-1} + \boldsymbol{\beta}\mathbf{Z}_t + s^*_t + e_t, \\ l_t &= \alpha^*(Y_t - D(t) - \boldsymbol{\beta}\mathbf{Z}_t - s^*_t) + (1-\alpha^*)(l_{t-1} + \nu b_{t-1}), \\ b_t &= \beta^*(l_t - l_{t-1}) + (1-\beta^*)\nu b_{t-1}, \\ s^*_{t+m} &= \gamma^*(Y_t - l_t - \boldsymbol{\beta}\mathbf{Z}_t) + (1-\gamma^*)s^*_t. \end{aligned} \tag{4.3}$$

El modelo incorpora por defecto una componente de regresión lineal sobre covariables, $\boldsymbol{\beta}\mathbf{Z}_t$, así como una función que representa una tendencia global determinista, $D(t)$. Del mismo modo que con el modelo LGT, expresaremos el modelo DLT dentro del marco general de los modelos SSOE, y trataremos de sacar alguna conclusión desde el punto de vista teórico, antes de entrar en la implementación disponible. De nuevo, recurriremos al cambio de notación $s_t \equiv s^*_{t+m}$.

Reescribamos la ecuación de estado asociada a la componente estacional para que dependa de estados en instantes previos:

$$\begin{aligned} s_t &= \gamma^*(Y_t - \alpha^*(Y_t - D(t) - \boldsymbol{\beta}\mathbf{Z}_t - s_{t-m}) - (1-\alpha^*)(l_{t-1} + \nu b_{t-1}) - \boldsymbol{\beta}\mathbf{Z}_t) + (1-\gamma^*)s_{t-m} \\ &= \gamma^*(1-\alpha^*)(Y_t - l_{t-1} - \nu b_{t-1} - \boldsymbol{\beta}\mathbf{Z}_t) + (1-\gamma^*(1-\alpha^*))s_{t-m} + \gamma^*\alpha^* D(t). \end{aligned}$$

A continuación, tratemos de incorporar el término de error, e_t, en las ecuaciones de actualización de estados de (4.3) a partir de la ecuación de observación.

$$\begin{aligned} l_t &= \alpha^*(l_{t-1} + \nu b_{t-1} + e_t) + (1-\alpha^*)(l_{t-1} + \nu b_{t-1}) = l_{t-1} + \nu b_{t-1} + \alpha^* e_t, \\ b_t &= \beta^*(\nu b_{t-1} + \alpha^* e_t) + (1-\beta^*)\nu b_{t-1} = \nu b_{t-1} + \beta^*\alpha^* e_t, \\ s_t &= \gamma^*(1-\alpha^*)(D(t) + s_{t-m} + e_t) + (1-\gamma^*(1-\alpha^*))s_{t-m} + \gamma^*\alpha^* D(t) \\ &= s_{t-m} + \gamma^* D(t) + \gamma^*(1-\alpha^*)e_t. \end{aligned}$$

Finalmente, denotando por $\mathbf{X}_t = [l_t \quad b_t \quad s_t \quad \cdots \quad s_{t-m+1} \quad \boldsymbol{\beta}']'$ podemos expresar el modelo DLT mediante la representación en forma SSOE

$$\begin{aligned} Y_t &= w_t(\mathbf{X}_{t-1}) + e_t, \\ \mathbf{X}_t &= \mathbf{M}_t(\mathbf{X}_{t-1}) + \mathbf{v}e_t, \end{aligned} \tag{4.4}$$

caracterizada por

$$\mathbf{M}_t(\mathbf{X}_{t-1}) = \begin{bmatrix} 1 & \nu & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \nu & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \mathbf{X}_{t-1} \quad + \quad \begin{bmatrix} 0 \\ 0 \\ \gamma^* \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} D(t), \quad \mathbf{v} = \begin{bmatrix} \alpha^* \\ \beta^*\alpha^* \\ \gamma^*(1-\alpha^*) \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

$$w_t(\mathbf{X}_{t-1}) = [1 \quad \nu \quad 0 \quad \cdots \quad 0 \quad 1 \quad \mathbf{Z}_t]\mathbf{X}_{t-1} + D(t) \equiv (\mathbf{w}_t)'\mathbf{X}_{t-1} + D(t).$$

Si denotamos por $\mathbf{M}_1$ a la matriz presente en la expresión de $\mathbf{M}$ como función de $\mathbf{X}_{t-1}$, además de $\tilde{\mathbf{M}}_t = \mathbf{M}_1 - \mathbf{v}(\mathbf{w}_t)'$, $\tilde{\mathbf{M}}_0 = \mathbf{I}$ y $\mathbf{v}_1 = [-\alpha^* \quad -\beta^*\alpha^* \quad \gamma^*\alpha^* \quad 0 \quad \cdots \quad 0]' = [0 \quad 0 \quad \gamma^* \quad 0 \quad \cdots \quad 0]' - \mathbf{v}$, obtenemos la siguiente relación de recursividad entre vectores de estados:

$$\begin{aligned} \mathbf{X}_t &= \mathbf{M}_t(\mathbf{X}_{t-1}) + \mathbf{v}(Y_t - w_t(\mathbf{X}_{t-1})) = \tilde{\mathbf{M}}_t\mathbf{X}_{t-1} + \mathbf{v}_1 D(t) + \mathbf{v}Y_t \\ &= \cdots = \left(\prod_{j=1}^{t} \tilde{\mathbf{M}}_j\right)\mathbf{X}_0 + \sum_{i=0}^{t-1}\left(\prod_{j=0}^{i} \tilde{\mathbf{M}}_j\right)(\mathbf{v}_1 D(t-i) + \mathbf{v}Y_{t-i}). \end{aligned}$$

Así pues, por analogía con lo visto en la Sección 3.2.2, en caso de que el modelo incorpore efectivamente la componente regresiva, no será estable. Si dicha componente se elimina, podemos analizar la estabilidad del método en función de los autovalores de $\tilde{\mathbf{M}}$, la matriz construida de forma idéntica a $\tilde{\mathbf{M}}_t$ cuando $\mathbf{X}_t = [l_t \quad b_t \quad s_t \quad \cdots \quad s_{t-m+1}]'$, que además coincide con la matriz que determina la estabilidad del método ETS(A, A_d, A). En dicho escenario, el método será estable siempre y cuando la componente estacional se normalice y los parámetros α^*, β^*, γ^* y ν verifiquen ciertas restricciones que se pueden consultar en Hyndman et al. (2008, p.157), pero que no se corresponden con las restricciones usuales de la forma $0 < \alpha^*, \beta^*, \gamma^*, \nu < 1$.

Otra cuestión teóricamente relevante es la función $D(t)$. En concreto, las opciones que proponen los autores son cuatro.

1. Constante : $D(t) = \delta_0$,
2. Lineal : $D(t) = \delta_0 + \delta_1 t$,
3. Log-Lineal : $D(t) = \delta_0 + \delta_1 \log(t)$.
4. Logística : $D(t) = \text{Min} + \frac{\text{Max}-\text{Min}}{1+e^{-\delta_0-\delta_1 t}}$.

Incorporar esta componente estacional supone añadir parámetros adicionales como δ_0 y δ_1, que deberá ser estimador juntos a los demás parámetros del modelo. En el caso particular de la función logística, Min y Max serían cantidades introducidas manualmente por el usuario, representando las asíntotas horizontales asociadas a dicha curva función. Aunque a priori pueda parecer que esta componente

complementa el modelo ETS(A, A_d, A), lo cierto es que no parece aportar nada adicional al modelo, pues como vimos en las Secciones 2.2, los modelos de suavización exponencial clásicos son capaces de recoger tendencias lineales y logísticas para series de tiempo, si se escogen adecuadamente sus parámetros. La tendencia log-lineal podría ser más dificil de captar a largo plazo, algo que no suele resultar preocupante, dado que el poder predictivo de los modelos de series de tiempo pierde peso a medida que aumenta el horizonte de predicción. Así pues, es difícilmente justificable pagar el precio de incorporar más parámetros al modelo para lograr algo que, en principio, ya se podía contemplar con modelos más sencillos. Por supuesto, esta idea deberá ser validada en estudio de simulación que llevaremos a cabo más adelante. Por otro lado, la incorporación de algunas de esas tendencias globales dan lugar a problemas de identificabilidad del modelo. Veámoslo con el caso más sencillo, particularizando el modelo al caso sin componente regresiva, sin componente estacional y sin tendencia local ($b_0 = 0 = \beta^*$), con $D(t)$ constante. Se tiene entonces que

$$\begin{aligned} Y_t &= \delta_0 + l_{t-1} + e_t, \\ l_t &= l_{t-1} + \alpha^* e_t. \end{aligned}$$

Para cualesquiera valores de δ_0 y l_0, siempre que la suma de ambos permanezca constante, el modelo caracterizado por estos parámetros es idéntico. En efecto, para $t > 0$ se verifica

$$\begin{aligned} \delta_0 + l_t &= \delta_0 + l_{t-1} + \alpha^* e_t = \delta_0 + l_{t-1} + \alpha^*(Y_t - \delta_0 - l_{t-1}) = (1-\alpha^*)(\delta_0 + l_{t-1}) + \alpha^* Y_t \\ &= \cdots = (1-\alpha^*)^t(\delta_0 + l_0) + \sum_{i=0}^{t-1} \alpha^*(1-\alpha^*)^i Y_{t-i}, \end{aligned}$$

de modo que $\hat{Y}_{t|t-1} = \delta_0 + l_t$ no depende de los valores específicos que tomen δ_0 y l_0, si no varía la suma. Tanto ajustes como predicciones a horizonte h verificarán esta propiedad, lo único que variará será la secuencia de estados $l_0, \cdots, l_t$, que cumplirán $l_t = (1-\alpha^*)^t(\delta_0 + l_0) + \sum_{i=0}^{t-1} \alpha^*(1-\alpha^*)^i Y_{t-i} - l_0$; esto es, la componente que suele representa el nivel de la serie, l_t, pasará a representar una variación respecto a δ_0. A efectos de estimación esto supone un problema, pues cualquier modelo que ofrezca los mismos ajustes, y por tanto los mismos residuos, dará lugar a sumas de residuos al cuadrado idénticas (estimación máximo verosimil gaussiana).

Finalmente, tal y como hemos deducido hasta llegar a la representación (4.4), la componente estacional en un instante dependerá del valor de $D(t)$ en dicho instante, a través del parámetro de suavizado γ. Esto no parece muy razonable, pues estaremos trasladando a la componente estacional estimada por el modelo una parte de la tendencia global de la serie, algo que nada tiene que ver con su comportamiento periódico.

Pasemos a comentar los detalles de la implementación en el paquete Orbit del modelo DLT. Lo más relevante resulta ser que, en base a los resultados que se obtienen aplicando la correspondiente función del paquete para ajustar el modelo DLT, se deduce que la formulación teórica expuesta en (4.3) es incorrecta, así como la que aparece en Ng et al. (2020). Aparentemente, se ha corregido esa dependencia de la componente estacional y la tendencia global, $D(t)$, previamente comentada. Se mantiene el mismo modelo salvo por la siguiente ecuación:

$$s^*_{t+m} = \gamma^*(Y_t - l_t - \boldsymbol{\beta}\mathbf{Z}_t - D(t)) + (1-\gamma^*)s^*_t = s^*_t + \gamma^*(1-\alpha^*)e_t.$$

Afortunadamente, las deducciones teóricas sobre la estabilidad del método se mantienen pues, como ya hemos visto, $D(t)$ no afecta al impacto que tenían las condiciones iniciales sobre el resto de vectores de estados. En cuanto a la formulación como modelo SSOE, es idéntica a la expuesta en (4.4) salvo por la simplificación de $\mathbf{M}_t(\mathbf{X}_{t-1})$ que pasa a ser simplemente $\mathbf{M}_1\mathbf{X}_{t-1}$ (y $\mathbf{v}_1$ pasará a ser $-\mathbf{v}$).

Por otro lado, la tendencia global se implementa con el tiempo normalizado en $\left[0, \frac{n-1}{m}\right]$ de modo que, para el instante $t = 1, \cdots, n$, se computará $D(\frac{t-1}{m})$. Por ejemplo, si lo esperable para la tendencia lineal fuese que $D(1) = \delta_0 + \delta_1$ y $D(n) = \delta_0 + \delta_1 n$, lo que ocurre en realidad es $D(1) = \delta_0$ y

$D(n) = \delta_0 + \delta_1 \frac{n-1}{m}$. La única excepción es el caso log-lineal, en el que se toma $D(\frac{t-1}{m} + 1)$ para no calcular el logaritmo de 0.

De igual modo que con el modelo LGT, no se definen b_0 ni l_0, se toma $b_1 = 0$ y se escogen l_1 y $s_1^*, \cdots, s_m^*$ tales que $Y_1 = l_1 + s_1 + \boldsymbol{\beta}\mathbf{Z}_1 + D(0)$ (o $Y_1 = l_1 + s_1 + \boldsymbol{\beta}\mathbf{Z}_t + D(1)$ en el caso log-lineal), con $\sum_{i=1}^{m} s_i^* = 0$.

La distribución de las innovaciones coincide con la mencionada al hablar del modelo LGT, verificando ϑ y σ_0 las mismas propiedades. Los parámetros de suavizado α^*, β^* y γ^* también se restringen a $[0, 1]$, y los primeros $m - 1$ valores de la componente estacional se inicializan a partir de la misma distribución a priori normal centrada en 0, con desviación típica 0.05 y truncada a partir del intervalo $[-1, 1]$. El parámetro de amortiguamiento ν no se estima, si no que se introduce a nivel de usuario (por defecto tomando el valor 0.8) del mismo modo que Min y Max, en caso de elegir una tendencia de tipo logístico (por defecto tomando los valores 0 y 1 respectivamente). No hemos encontrado ningún tipo de información sobre una posible distribución a priori para δ_0 y δ_1, ni como se inicializan sus valores, por lo que el enfoque bayesiano no parece afectar a todos los parámetros del modelo. Finalmente, a las componentes de $\boldsymbol{\beta}$, β_j con $j = 1, \cdots, \dim(\boldsymbol{\beta})$, se les asocian distribuciones a priori gaussianas con media μ_j y varianza σ_j^2, introducidas por el usuario (por defecto media 0 y varianza 1). Tampoco podemos asegurar como se produce exactamente el ajuste de los modelos, aunque se emplea el método HMC y seguramente se parta de determinados valores iniciales fijos para aquellos parámetros que se vayan a ajustar pero que no se les haya especificado una distribución a priori (α^*, β^*, γ^*, $s_1^*, \cdots s_{m-1}^*$, ν, δ_0, δ_1).

4.3. Análisis de sensibilidad

Tal y como se ha implementado el modelo DLT en el paquete `Orbit-ml`, en algunos parámetros se toman valores por defecto, en lugar de ser estimados. La selección de un tipo u otro de tendencia global determinista (lineal, logística, etc.) también queda a discrección del usuario. Por tanto, una de las primeras cuestiones que podemos plantear es: ¿el modelo se ve muy influenciado por los valores que tomen estos parámetros?

En esta sección vamos a llevar a cabo un análisis de sensibilidad básico, para justificar la decisión que tomaremos de cara al estudio de simulación del Capítulo 5. Esta decisión consiste en dejar que todos los parámetros que no se estimen automáticamente mantengan los valores que los autores han considerado como adecuados.

En la Sección 5.1.2 se establece cómo hemos simulado algunas series de tiempo, de entre las que se encuentran series con DLT como modelo generador. De entre esas series generadas por un modelo DLT, con una posible componente estacional de periodo $m = 24$, se han seleccionado diez de ellas aleatoriamente y, para cada una de ellas, hemos ajustado modelos DLT, tanto en base a la estimación por cadenas de Markov Monte Carlo como por la estimación máxima a posteriori, variando los valores de: el parámetro de amortiguamiento ν, el tipo de tendencia global $D(t)$ y la presencia o no de componente estacional. En concreto, para cada serie hemos ajustado un modelo en cada uno de los escenarios donde,

1. ν toma valores equiespaciados entre 0 y 1, con salto 0.05,
2. $D(t)$ es constante, lineal, log-lineal o logística,
3. se considera una componente estacional o no,

variando en cada ocasión solamente una de las tres componentes del modelo, y dejando las otras dos en sus valores por defecto ($\nu = 0.8$, $D(t)$ lineal y presencia de componente estacional). Tendremos, por tanto, un estudio muy limitado sobre la sensibilidad del modelo frente a variaciones de estas componentes, al tener solamente resultados sobre diez series distintas, y considerando el efecto que produce la variación de cada una de ellas por separado (es decir, sin considerar un posible efecto de

intervención al modificarlas de forma simultánea). No obstante, creemos que es suficiente para nuestros propósitos.

Se ha calculado el SMAPE (definido en la Sección 3.2.3) sobre las predicciones que ofrecen los distintos modelos a horizonte $h = 24$ (reservando la parte final de las series simuladas para tal fin, en lugar de ajustar los modelos con todos los datos) y, posteriormente, se han relativizado los resultados con respecto a los que ofrecen los ajustes por defecto. Por ejemplo, para estudiar la sensibilidad frente al parámetro de amortiguamiento, fijada una de las diez series elegidas, se ha dejado $D(t)$ como lineal y se ha tomado una componente estacional de periodo 24, se han calculado los SMAPE para los ajustes con $\nu = 0, 0.05, \cdots, 1$ y, finalmente, se han dividido todos los SMAPE entre el SMAPE asociado al ajuste con $\nu = 0.8$. El motivo por el que se relativizan los SMAPE, así como el procedimiento de simulación y ajuste de las series se describen con detalle a lo largo de las Secciones 5.1.1, 5.1.2 y 5.1.3.

En las Figuras 4.1 y 4.2 se representan los diagramas de caja construidos con los SMAPE relativizados, tal y como hemos descrito, para los distintos valores de ν, tanto para el ajuste por cadenas de Markov como para el ajuste por estimación máxima a posteriori.

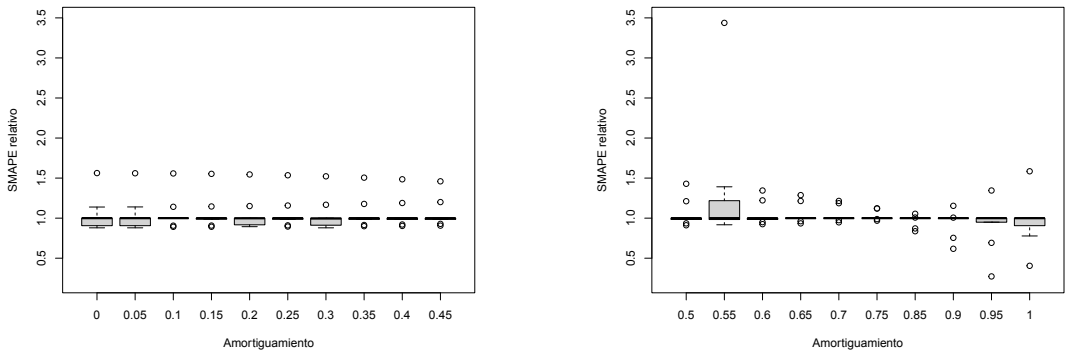

Figura 4.1: Diagramas de caja de los SMAPE relativos, para ν variando entre 0 y 1 con paso 0.05 (izquierda: de 0 a 0.45; derecha: de 0.5 a 0.75 y de 0.85 a 1), con respecto a los ajustes con $\nu = 0.8$, empleando estimación máxima a posteriori.

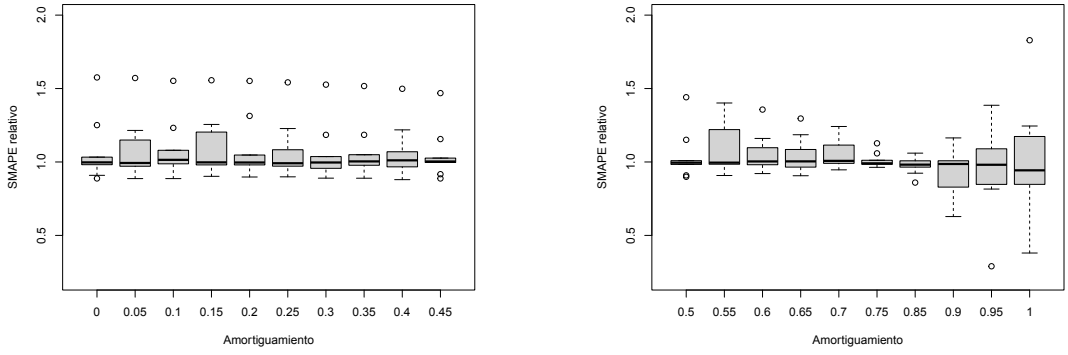

Figura 4.2: Diagramas de caja de los SMAPE relativos, para ν variando entre 0 y 1 con paso 0.05 (izquierda: de 0 a 0.45; derecha: de 0.5 a 0.75 y de 0.85 a 1), con respecto a los ajustes con $\nu = 0.8$, empleando estimación por cadenas de Markov Monte Carlo.

Para los dos tipos distintos de ajuste se deducen las mismas conclusiones, a la vista de las Figuras 4.2 y 4.3. En primer lugar, la mediana de los SMAPE relativos se encuentra siempre muy próxima a 0, de modo que el rendimiento predictivo del modelo DLT es, en términos generales, el mismo para los distintos valores del parámetro de amortiguamiento que para el valor por defecto de 0.8. No obstante, en algunos casos la selección de un valor distinto de 0.8 da lugar a un SMAPE notablemente diferente, como reflejan los bigotes o los "atípicos"para los distintos diagramas de caja.

Este mismo efecto (rendimiento general indistinguible pero algunos casos particulares con variaciones importantes en función del valor que tomen las componentes no estimadas del modelo), se aprecia en las Figuras 4.3 y 4.4, en las que se está midiendo la sensibilidad frente a la tendencia global y la presencia de componente estacional.

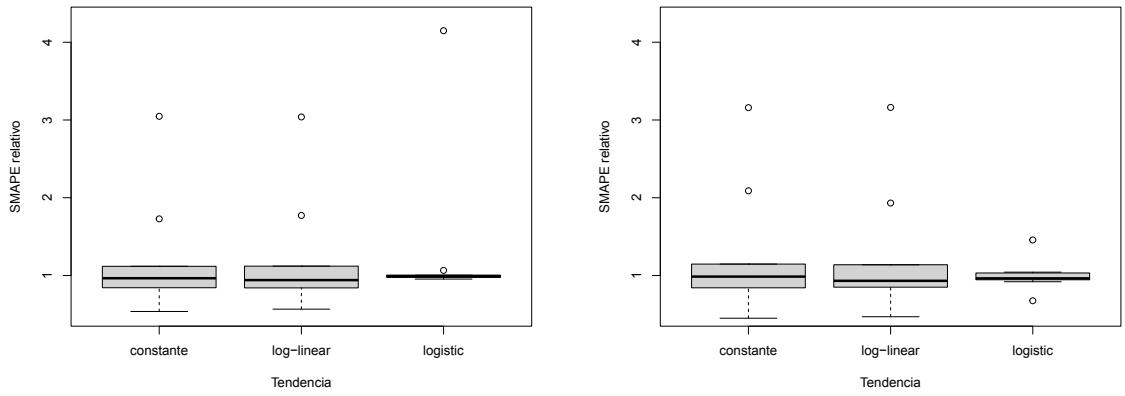

Figura 4.3: Diagramas de caja de los SMAPE relativos, para $D(t)$ constante, log-linear y logística, con respecto a los ajustes con tendencia global lineal, empleando estimación máxima a posteriori (izquierda) y por cadenas de Markov Monte Carlo (derecha).

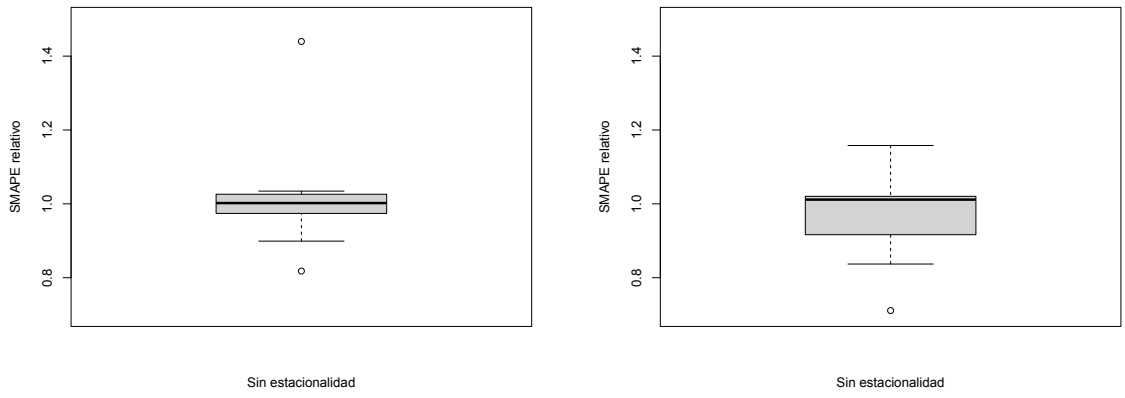

Figura 4.4: Diagramas de caja de los SMAPE relativos, para los ajustes sin componente estacional con respecto a los ajustes con componente estacional de periodo $m = 24$, empleando estimación máxima a posteriori (izquierda) y por cadenas de Markov Monte Carlo (derecha).

Así pues, se pueden sacar dos conclusiones. La primera es que, la variación sistemática (aplicada de forma general a todas las series que se ajustan) de las componentes estudiadas, pasando a tomar valores diferentes de los que los autores han seleccionado como valores por defecto, no produce resultados, de forma agregada, muy diferentes a los que se obtienen sin modificar dichas componentes. La segunda es que, si tuviésemos algún mecanismo para seleccionar, en cada caso, los valores adecuados, podríamos reducir notablemente el error cometido por el ajuste DLT.

Parece razonable plantearse la siguiente pregunta: ¿existe, actualmente, alguna forma de determinar, de forma automática, los valores para ν, el tipo de tendencia más adecuada o la presencia de componente estacional? El paquete `Orbit-ml` ofrece dos posibles mecanismos para esta seleccionar valores para estas componentes (*Backtest*, s.f.).

La primera alternativa consiste en ajustar los modelos mediante estimación máxima a posteriori (independientemente de cual sea el método de estimación que hayamos considerado inicialmente) y seleccionar el valor de ν, el tipo de tendencia global o la presencia de componente estacional en función de cual haya sido el modelo ajustado que minimice el criterio BIC.

La segunda alternativa pasa por recurrir a las ideas de validación cruzada. Se considera un fragmento de la serie de tiempo original al que, o bien se le van añadiendo cada vez más observaciones ("expanding window"), o bien se le añaden nuevas observaciones y se le retiran algunas observaciones iniciales ("rolling window"). Para cada sub-serie, se ajustan los modelos con los distintos valores disponibles para las componentes que se pretenden auto-ajustar, se calculan predicciones a cierto horizonte y se computa algún criterio de error predictivo, como el SMAPE. Finalmente se promedian los valores del criterio para todas las sub-series.

El primer enfoque no termina de ser coherente en el caso de que se pretenda estimar la serie empleando procedimientos basados en cadenas de Markov Monte Carlo, pues se estarían eligiendo los "mejores"valores para la componentes que quedan a discreción del usuario minimizando el criterio BIC para ajustes por estimación máxima a posteriori. Tiene la ventaja, sin embargo, de ser un procedimiento mucho menos costoso computacionalmente que la segunda alternativa. Esta última además, depende a su vez de cuestiones como el tamaño de las sub-series a considerar o cuantas observaciones se añaden o se retiran de la serie en cada caso.

Para ninguna de las dos alternativas se ha presentado algún tipo de resultado teórico que garantice que, por ejemplo, si una serie ha sido generada por un modelo DLT, entonces la selección automática de valores para estas componentes da lugar, si el tamaño de la serie es suficientemente elevado, a:

- La elección del verdadero valor para el tipo componente estacional.
- La correcta identificación de la presencia de componente estacional.
- Una propuesta, $\hat{\nu}$, para el parámetro de amortiguamiento, que se encuentre lo más próxima a ν de entre los posibles valores que puede elegir el método (que habitualmente consistirán en un mallado equiespaciado del intervalo $[0, 1]$).

Para determinar si se trata de mecanismos de selección útiles, hemos empleado las diez series mencionadas anteriormente para testear los resultados predictivos, en términos del SMAPE, de los ajustes basados en ambos procedimientos (BIC y validación cruzada), relativizados con respecto a los resultados de los ajustes en los que se dejan por defecto estas componentes. Concretamente, hemos utilizado el procedimiento de validación cruzada de tipo ("rolling window"), con subseries de tamaño 216 (cada una de las diez series que se pretenden ajustar tienen tamaño 336), moviéndose 24 instantes de tiempo de cada vez (la primera sub-serie abarca las 216 primeras observaciones de la serie original, la segunda va desde la observación número 25 hasta la 240, etc), y fijando un horizonte de predicción de $h = 24$ (mismo horizonte que el de la serie original).

En la Tabla 4.1 están recogidas las medianas de los SMAPE relativizados, para los dos mecanismos de selección, y para los dos procedimientos de estimación posterior. Como se puede apreciar, el procedimiento basado en el criterio BIC, que necesitó de unos quince minutos para completarse sobre el conjunto de diez series de tiempo (en la Sección 1.2 se describe el equipo informático empleado, para

tener una referencia de su capacidad de procesamiento), parece ofrecer, a grandes rasgos, los mismos resultados que dejar por defecto los valores que los autores han considerado. En cambio, podemos afirmar sin ningún tipo de dudas que, por lo menos tal y como lo hemos implementado nosotros, el procedimiento de validación cruzada, que tardó hasta catorce horas en llevarse a cabo, no da lugar a una selección de parámetros que lleven a ajustes con menos error que los que se derivan de los valores por defecto para el parámetro de amortiguamiento, el tipo de tendencia global y la presencia de componente estacional.

	MAP	**MCMC**
BIC	1.030	0.996
CV	1.798	1.875

Tabla 4.1: Medianas de los SMAPE relativos para los ajustes por estimación máxima a posteriori (MAP) y estimación por cadenas de Markov Monte Carlo (MCMC) seleccionando el parámetro de amortiguamiento, el tipo de tendencia global y la presencia (o no) de componente estacional de acuerdo con el criterio BIC o por validación cruzada (CV), con respecto a los resultados de los respectivos ajustes sin ningún procedimiento de selección previo (dejando los valores por defecto).

Capítulo 5

Comparación de modelos con datos reales y simulados

En este capítulo presentaremos los resultados comparativos del rendimiento que han ofrecido los modelos ARMA, ARIMA, ETS, LGT y DLT, por un lado, sobre un conjunto amplio de series de tiempo simuladas, y por otro, en 3 series de tiempo reales, cedidas por la empresa colaboradora para este trabajo.

El estudio de simulación se compone fundamentalmente de dos partes diferenciadas. La primera, que llamaremos estudio benchmark, consiste en comparar la precisión de los distintos modelos, en términos del error predictivo y la cobertura de sus intervalos de predicción, cuando estos son ajustados sobre series de tiempo que realmente se correspondan con trayectorias parciales de los procesos estocásticos subyacentes a los modelos para series de tiempo que hemos considerado. En la segunda, se emplean series de tiempo que en principio no se corresponden con ningún modelo concreto, y que además incluyen una serie de características, habituales en la práctica, que las aleja del marco teórico de los modelos objeto de estudio. Por otro lado, cada una de las 3 series reales presenta particularidades que habrán sido cubiertas por el estudio de simulación, de modo que los resultados de dicho estudio justificarán que se empleen algunos de los modelos tratados, en lugar de todos ellos.

El objetivo de este capítulo es tratar de verificar si, a pesar de los problemas que presentan desde un punto de vista teórico los modelos Orbit (discutidos en las Secciones 4.1 y 4.2), estos pueden ofrecer un rendimiento comparable al de los modelos clásicos o incluso superarlos en alguna área concreta del análisis de series de tiempo.

5.1. Estudio benchmark

5.1.1. Procedimiento de simulación

Comencemos por aclarar cómo hemos llevado a cabo la simulación de las series de tiempo. En total se han generado, con la ayuda del lenguaje de programación R, 2000 series para cada modelo, de las cuales 1000 series serán "series cortas", de tamaño 63 y otras 1000 son "series largas", de tamaño 360. De esas 1000 series cortas o largas, 500 habrán sido construidas con errores gaussianos de media 0 y otras 500 con errores distribuidos según una distribución T de Student generalizada, también con media 0. Algunas series incorporarán dependencia y/o componente estacional, de periodo 7 (en el caso de las series cortas) o 24 (en el caso de las series largas), mientras que otras series no tendrán esta estructura.

Para las series asociadas a modelos ARMA, ARIMA y ETS, nos hemos valido del paquete `smooth` (v3.1.6; Svetunkov, 2022) de R, que consta de funciones específicamente diseñadas para este fin (`sim.ssarima` y `sim.es`). En el caso concreto de las series ARMA (y por extensión ARIMA, pues

estas últimas se han construido a partir de series ARMA), se ha limitado la estructura de dependencia, entre otras cosas para evitar posibles problemas de simulación, de modo que el número máximo de parámetros del modelo ARMA generador (sin contar la desviación típica de las innovaciones o los grados de libertad de la distribución T) fuese 4. Además, p y q serán menores o iguales que 2, mientras que P y Q serán menores o iguales que 1, y se ha impuesto que el modelo ARMA sea causal e invertible. Para las series ARIMA, el número máximo de diferenciaciones regulares, d, será 2, mientras que el número máximo de diferenciaciones estacionales, D, será 1. En el caso de que $d + D > 1$, las series se habrán generado sin constante, para evitar una tendencia global polinómica de grado 2 o superior. Por último, se ha evitado generar modelos ARMA$(0, 0)$ (ruido blanco). En el caso de los modelos de suavización exponencial, se ha impuesto que los parámetros del modelo satisfagan las restricciones que garanticen estabilidad, como modelos SSOE. Se ha restringido los posibles modelos generadores a los modelos puramente aditivos (ETS(A, N, N), ETS(A, A, N), ETS(A, N, A), ETS(A, A, A), ETS(A, Ad, N) y ETS(A, Ad, A)).

Para los modelos LGT y DLT no hay herramientas que permitan simular, de forma automática, series de tiempo. Hemos tenido que construirlas nosotros de acuerdo con la estructura de carácter iterativo presentada en las Secciones 4.1 y 4.2. Así pues, simular una serie de tiempo para esos modelos se reduce a generar los estados iniciales y los parámetros de los mismos. A modo de resumen, para las series LGT, l_0 y b_0 se han sorteado de una distribución uniforme en $(100, 200)$ y $(-10, 10)$, respectivamente, mientras que $s_{-m+1}, \cdots, s_{-1}$ (con m el periodo estacional) se sortean de una distribución normal de media 0 y desviación típica la desviación típica de las innovaciones de la serie (se normaliza la componente estacional, de modo que $s_0 = -s_{-m+1} - \cdots - s_{-1}$). Los parámetros α, β, γ, λ y ξ_1 se han obtenido simulando uniformes en $(0, 1)$, mientras que ξ_2 se sortea de una uniforme en $(-1, 1)$. Para las series DLT se ha procedido análogamente, con la diferencia de que l_0 se sortea de una uniforme en $(-100, 100)$ (notemos que el modelo LGT está diseñado para series positivas, con $l_t > 0\ \forall t$, a diferencia del modelo DLT) y el papel de λ lo ocupa θ. Con respecto a la tendencia global, $D(t)$, se han generado series incorporando aleatoriamente uno u otro tipo, con δ_0 y δ_1 simulados de acuerdo con una distribución uniforme en $(-10, 10)$. En el caso particular de la tendencia global de tipo logístico, L será igual a l_0 más el resultado de una variable uniforme en $(-100, 100)$, mientras que $U = l_0 - (L - l_0)r_u$ donde r_u tendrá distribución uniforme en $(0.5, 3)$. Esta selección de L y U permite que l_0, que viene a ser el nivel inicial de la serie de tiempo, se encuentre entre L y U, que representan las asíntotas horizontales asociadas a una curva logística.

Por último, ya que es común para todas las series, la desviación típica de las innovaciones se ha seleccionado aleatoriamente, de acuerdo con una distribución uniforme, entre 1 y 3. Los grados de libertad de la distribución T, también se han sorteado de acuerdo con una distribución uniforme entre 2 y 10 (notemos que así tenemos garantizada la existencia de varianza finita, pero las colas de la distribución T siguen siendo notablemente más pesadas que las de la distribución normal estándar).

5.1.2. Procedimiento de ajuste y predicción

Una vez hemos generado series de tiempo, que teóricamente se corresponden con modelos ARMA, ARIMA, ETS, LGT o DLT, hemos procedido a ajustarlas con todos y cada uno de esos modelos. En concreto, para las series cortas se guardarán las últimas 7 observaciones, y para las series largas las últimas 24, que se emplearán posteriormente para construir medidas de error de predicción. Las series serán ajustadas, por tanto, con el resto de observaciones previas.

Los ajustes para los modelos ARMA, ARIMA y ETS se obtienen mediante las funciones auto.ssarima y es del paquete `smooth`. En los 3 casos la búsqueda del "mejor modelo"para cada serie se realiza en base al criterio BIC. Para los modelos ARMA y ARIMA, la búsqueda se restringe a modelos con $P, Q \leq 1$ y $p, q \leq 2$. Evidentemente, ajustar modelos ARMA supone fijar $d = D = 0$. También se ha limitado d y D en el caso de los modelos ARIMA, de modo que $d \leq 2$ y $D \leq 1$. Además, en el caso de las series largas, la búsqueda del modelo ARIMA en base al criterio BIC se ha efectuado con el argumento opcional `fast=TRUE`. De acuerdo con la documentación del paquete `smooth`, se reduce el número total de modelos a comparar mediante un algoritmo de ramificación y acotación. El objetivo

de estas limitaciones es un ahorro notable del tiempo computacional, al reducir considerablemente el número máximo de modelos de entre los que seleccionar el más adecuado. Para los modelos ETS, la búsqueda se limita a los modelos puramente aditivos (con el argumento `model = "XXX"`). Por último, también para las series largas, se hará uso de otro argumento opcional, común a `auto.ssarima` y `es` : `initial = "backcasting"`. De este modo, en lugar de considerar el vector de estados iniciales como parámetros adicionales del modelo, que deberán ser estimados conjuntamente con el resto de parámetros, se emplea una alternativa consistente en invertir el orden de las observaciones de la serie de tiempo, ajustar un modelo a la serie invertida y considerar los estados finales de la serie invertida ajustada como estados iniciales de la serie original. El motivo fundamental por el que recurrimos a esta alternativa radica en que, de no hacerlo, el criterio BIC nunca seleccionaría como modelo adecuado aquel que incorpore estacionalidad a la serie, pues esto supone incrementar (a grandes rasgos) el tamaño del vector de estados en m unidades (o más) siendo m el periodo estacional. Notemos que en el caso de las series largas $m = 24$, por lo que la penalización impuesta por el criterio BIC es demasiado elevada como para que compense incluir una componente (o dependencia) estacional. En cambio, al obtenerse valores para el vector de estados iniciales de forma previa al ajuste del modelo, el criterio BIC no considera ese vector como un vector de parámetros, y permite escoger modelos estacionales cuando corresponde. Si bien no hemos encontrado una justificación formal a por qué esta alternativa no estropeará las buenas propiedades teóricas de los modelos SSOE, empíricamente los parámetros ajustados de acuerdo a las dos posibilidades son muy similares, cuando el tamaño muestral es suficientemente grande. Además, recordemos que asumiendo que un modelo SSOE sea estable (ocurre en este caso), la influencia del vector de estados iniciales desaparece con el tiempo, de modo que, de nuevo, si el tamaño de la serie es suficientemente elevado, intuitivamente no debería influir demasiado lo que ocurra con ese vector inicial a la hora de obtener predicciones.

En base al análisis de sensibilidad llevado a cabo en la Sección 4.3, a la hora de ajustar los modelos Orbit con el paquete `Orbit-ml` para el lenguaje de programación Python, se han dejado por defecto aquellos parámetros que quedan a discreción del usuario. Teniendo en cuenta esto, y que los modelos Orbit son únicos, en lugar de una familia de modelos (como ocurre con ARIMA y ETS), en este caso no se produce una búsqueda del "mejor modelo". En cualquier caso, se ajustan los modelos LGT y DLT, empleando las funciones `DLT` y `LGT`, respectivamente, tanto en base a la estimación máxima a posteriori (con el argumento `estimator = "stan-map"`) como en base a cadenas de Markov Monte Carlo (con el argumento `estimator = "stan-mcmc"`). Como detalle a tener en cuenta, para cada serie se han generado 4 cadenas de Markov distintas, y cada una de ellas ha contado con 1000 iteraciones de calentamiento y 2500 de muestreo (con los argumentos `num_warmup = 4000` y `num_sample = 10000`). Esta selección de hiperparámetros se ha comportado aparentemente bien en la práctica, a la hora de evitar divergencias en la estimación por cadenas de Markov (por ejemplo, si el estimador $\hat{R}$ introducido en la Sección 3.2.4 toma un valor superior a 1.1). Así pues, en total tendremos "muestras"de tamaño 10000 de la distribución a posteriori de los parámetros, que debería ser lo suficientemente elevadas para nuestros propósitos.

Una vez ajustado un modelo, se obtienen predicciones a horizonte 7 o 24, e intervalos de predicción del 95 %. Para los modelo ARMA, ARIMA y ETS serán predicciones en media junto con los intervalos de predicción teóricos basados en la distribución gaussiana de los errores (así como en que el modelo ajustado es realmente el modelo generador de los datos). En el caso de los modelos Orbit ajustados de acuerdo con la estimación máxima a posteriori, las predicciones se obtienen siguiendo las ecuaciones de observación y estados asociada a la estructura de los métodos, que coincide con la predicción en media para el modelo DLT, pero no ocurre lo mismo con el modelo LGT (debido a la presencia de términos no lineales en el vector de estados en la ecuación de observación). Para obtener intervalos de predicción, la única opción disponible en este caso se basa en el bootstrap de los residuos (*Methods of Estimations and Predictions*, s.f.). Los autores no especifican qué tipo de bootstrap emplean , pero teniendo en cuenta que se obtienen estimaciones de los parámetros que caracterizan el proceso de innovaciones, probablemente se trate de un bootstrap paramétrico; esto es, se obtienen simulaciones de la distribución T de student generalizada que tiene como parámetros las estimaciones de los mismos obtenidos a partir de la serie con la que se trabaja. En el caso del ajuste por cadenas de Markov Monte

Carlo, se obtendrán tantas posibles trayectorias futuras para la serie como valores se hayan simulado de la distribución a posteriori de los parámetros, obteniendo simulaciones del proceso de error, de nuevo, en base al remuestreo. A partir de esta colección de trayectorias, se computa en cada instante, la media y los cuantiles del 2.5 % y 97.5 % como predicción puntual y extremos del intervalo de predicción del 95 %.

5.1.3. Medición del rendimiento de los modelos

En primer lugar, hemos guardado el tiempo computacional asociado a todo el procedimiento de selección, ajuste y predicción. En segundo lugar, a partir de las predicciones puntuales, teniendo en cuenta que conocemos la "verdadera evolución"de las series (las últimas observaciones simuladas para las series, que nos reservamos previamente), computaremos el SMAPE, el MSE y una modificación propuesta por la empresa colaboradora, que denominaremos RMAPE ("robust MAPE"), y que no es más que el MAPE original, pero empleando una media truncada, en lugar de una media con todos los datos. Así, si denotamos por $e_1, \cdots, e_h$ los errores de predicción a horizonte h y por $\mathrm{e}_1^* \leq \cdots \leq \mathrm{e}_h^*$ los mismos errores ordenados, con $y_1^*, \cdots, y_h^*$ los valores de la serie de tiempo asociados,

$$\mathrm{RMAPE} = 100 \frac{1}{h-2N} \sum_{i=N+1}^{h-N} \frac{|\mathrm{e}_t^*|}{|y_t^*|},$$

de modo que se calcula el MAPE con los $h-2N$ errores centrales (eliminando los N más grandes y más pequeños). Desde ese punto de vista, se trata de una media truncada (simétricamente), eliminando el $200N/h\,\%$ de los datos (para que tenga sentido, hay que imponer $2N < h$). En nuestro caso, hemos empleado $N = 1$ para las series cortas (con $h = 7$) y $N = 2$ para las series largas (con $h = 24$). Por último, se guardará la variable indicadora de si el verdadero valor de la serie al final del horizonte de predicción estuvo dentro del intervalo de predicción del 95 %, y la longitud de dicho intervalo.

Tenemos por tanto, para cada serie simulada y para cada ajuste, varias medidas del error de predicción puntual, medidas relativas a los intervalos de predicción, para analizar la cobertura de los mismos y una medida del tiempo empleado en todo el tratamiento de la serie. ¿Qué haremos con estos datos? En primer lugar diferenciaremos, como es evidente, entre las series cortas y las series largas, pues el paradigma es completamente diferente. También diferenciaremos los resultados en función de cual ha sido el modelo generador de la serie de tiempo, y si las innovaciones son o no gaussianas. Una vez separadas los datos en base a estos criterios, promediaremos los tiempos computacionales y las variables indicadoras para obtener un tiempo promedio de ajuste y predicción, así como una estimación global de la cobertura de los modelos, respectivamente. Con respecto a la longitud de los intervalos de predicción y las medidas de error, no podemos simplemente promediar resultados, pues los valores numéricos que se obtengan para cada serie dependerán, en gran medida, de la escala de las series con las que se trabaja. Por ejemplo, el MSE asociado a una serie que oscila en torno a las 1000 unidades tenderá a ser muy superior al MSE de una serie que oscila en torno a las 10 unidades. Con los criterios de error relativo (SMAPE y RMAPE) ocurre algo parecido, tendiendo a tomar valores mucho más elevados cuando la serie tome valores muy cercanos a 0. Si promediásemos resultados, podríamos ocultar el hecho de que un modelo lo haga particularmente mal cuando la serie tiene una escala muy pequeña o muy grande, frente a como lo hagan los demás para esas series. Veamos esto con un ejemplo.

Ejemplo 5.1 (Argumento en contra del promedio de medidas de error)**.** Supongamos que disponemos de 10 series de tiempo, para las que hemos ajustado dos modelos distintos. Medimos el error de predicción a través del SMAPE (en tanto por uno), y obtenemos los siguientes resultados:

1. Modelo 1 : 0.06, 0.001, 0.026, 0.025, 0.028, 0.11, 0.24, 0.45, 0.31, 0.7.
2. Modelo 2 : 0.138, 0.003, 0.062, 0.06, 0.082, 0.257, 0.564, 0.2, 0.11, 0.3.

El modelo 1 ha obtenido mejores resultados en 7 de las 10 series, cometiendo entre 2 y 3 veces menos error que el modelo 2. En cambio, en las últimas 3 series el modelo 1 ha cometido entre 2 y 3 veces

más error que el modelo 2. Ahora bien, el modelo 1 tiene un error promedio del 19.5 % mientras que el modelo 2 del 17.76 %. ¿Es razonable creer que el modelo 1 comete, en general, menos error que el modelo 2? En nuestra opinión, no.

Para tratar de solventar esta cuestión, hemos optado por relativizar los resultados con respecto a los que se han obtenido, para cada serie, con el modelo ajustado asociado a la familia de modelos que incluye el modelo generador de la serie (en el caso de las series generadas por modelos Orbit, se tomará como punto de referencia los resultados de los modelos ajustados por cadenas de Markov Monte Carlo). Por ejemplo, si la serie ha sido generada por un modelo ARMA, una vez ha sido ajustada por los modelos ARMA, ARIMA, ETS, LGT y DLT, se obtendrán los SMAPE para cada uno de esos 5 ajustes, y se calculará el cociente entre esas cantidades y el SMAPE del modelo ARMA. Se obtiene de esta manera una medida del error relativo cometido por todos los métodos frente al método que, en teoría, debería ser óptimo. Una vez tenemos estos cocientes, se calculará la mediana para los resultados de cada modelo. El motivo de emplear la mediana es que se trata de una medida de localización que presenta la siguiente propiedad : si X es una variable aleatoria absolutamente continua y positiva, entonces $\text{Med}(1/X) = 1/\text{Med}(X)$. La demostración de dicha propiedad es inmediata, y es una característica interesante en este contexto, al calcular cocientes de variables que, en principio, son positivas (en ningún caso se ha obtenido un error de predicción 0), puesto que esto nos permite afirmar que si el comportamiento relativo del modelo A frente al modelo B en términos de un criterio de error es, en mediana, inferior a 1 (y por tanto que la probabilidad de que uno cometa menos error que el otro sea superior a 0.5), entonces el comportamiento relativo del modelo B frente al modelo A en términos de ese criterio de error será, en mediana, superior a 1. Esto no se sostiene con las medias, pues puede ocurrir que $\mathbb{E}[X] > 1$ y $\mathbb{E}[1/X] > 1$ simultáneamente.

Ejemplo 5.2 (Argumento a favor de la mediana de medidas de error relativizadas)**.** Considerando los mismos datos que en el Ejemplo 5.1, los resultados relativizados con respecto al modelo 1, por ejemplo, serían:

1. Modelo 1 : 1, 1, 1, 1, 1, 1, 1, 1, 1, 1.

2. Modelo 2 : 2.3, 3, 2.385, 2.4, 2.928, 2.336,2.35, 0.444, 0.355, 0.429.

Calculando ahora medianas, evidentemente para el modelo 1 obtendremos mediana con valor 1, y para el modelo 2, la mediana toma el valor 2.343 (promedio entre 2.336 y 2.35). De acuerdo con este criterio, la mediana es bastante mayor que 1. Por tanto, la probabilidad de que el modelo 2 obtenga mayores errores que el modelo 1 es superior al 50 % (efectivamente, en este caso es del 70 %). Además, es necesario tomar 2.343 veces el error cometido por el modelo 1 para garantizar que el error del modelo 2 será igual o inferior a esa cantidad, con una probabilidad del 50 %.

Este enfoque tiene limitaciones, puesto que habrá casos en los que aparentemente un modelo supere a otro en rendimiento relativo con respecto a un tercero, pero que al comparar los dos primeros entre si el resultado no sea consistente. Más adelante aportaremos una justificación a por qué tomaremos como umbrales para decir cuando un modelo comete un menor error que otro (a pesar de que los resultados estén relativizados con respecto a algún tercer modelo) en 5 centésimas para los criterios SMAPE y RMAPE y 1 décima para el MSE. Del mismo modo, diremos que las longitudes de los intervalos de predicción serán significativamente superiores para un modelo que para otro si hay una diferencia de 5 centésimas o más.

En las siguientes secciones, iremos presentando y comentando los resultados obtenidos y resumidos, de acuerdo con estos criterios, mediante tablas que los agrupan en función del modelo generador de las series, su tamaño, y el tipo de error. Evidentemente, las interpretaciones más fiables se obtendrán a partir de las series con mayor tamaño muestral, puesto que tanto la selección basada en el criterio BIC, como los estimadores de los parámetros presentan sus buenas propiedades en el escenario asintótico. Por otro lado, es de esperar que el modelo que menos error cometa en cada caso sea el modelo asociado a la familia que contiene el verdadero modelo generador de los datos. Por último, al menos para los

modelos ARMA, ARIMA y ETS, sería razonable que la cobertura de los intervalos sea menor cuando las innovaciones no son realmente gaussianas, puesto que los intervalos se construyen de acuerdo con dicha hipótesis.

5.1.4. Series cortas con errores gaussianos

Comencemos por analizar las tablas que resumen los resultados para las series cortas con errores gaussianos. La Tabla 5.1 contiene la información asociada a las series cortas generadas por un modelo ARMA.

	SMAPE	**RMAPE**	**MSE**	**Longitud**	**Cobertura**	**Tiempo**
ARMA	1.000	1.000	1.000	1.000	0.936	0.723
ARIMA	1.000	1.000	1.000	1.000	0.940	3.798
ETS	1.033	1.056	1.080	1.192	0.950	0.119
LGT	1.122	1.171	1.321	1.223	0.918	3.969
LGT MAP	1.090	1.132	1.252	0.982	0.872	0.652
DLT	1.150	1.186	1.419	1.416	0.928	3.078
DLT MAP	1.125	1.169	1.353	1.027	0.880	0.674

Tabla 5.1: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARMA generador de las series cortas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

Como podemos apreciar en esta primera tabla, ningún modelo consigue mejorar significativamente el error cometido por el modelo ARMA, en términos de MSE, SMAPE y RMAPE. El rendimiento de ARIMA sería indistinguible del de ARMA (algo que no sorprende, pues uno es un caso particular del otro, y la comparación se reduce a si el criterio BIC decidió que era adecuado diferenciar la serie o no) y no queda claro si ETS comete un mayor error que ARMA, pues solamente el RMAPE relativo supera el umbral de 1.05. Los modelos Orbit, ajustados tanto por MAP (de ahora en adelante DLT-MAP y LGT-MAP) como por MCMC (de ahora en adelante simplemente DLT y LGT) cometen un mayor error que ARMA, error que también es significativamente mayor al que comete ETS, aunque no hay diferencias significativas entre si. Si nos fijamos en la cobertura de los intervalos de predicción, todos los métodos superan el 90 % salvo por los modelos DLT-MAP y LGT-MAP, aunque solamente ETS y ARIMA consiguen una mayor cobertura que ARMA, el primero de los cuales lo paga con una longitud casi un 20 % superior. El tiempo computacional necesario para ETS es el menor, seguido por los modelos DLT-MAP, LGT-MAP y por ARMA. Finalmente, tanto ARIMA como DLT y LGT necesitan un tiempo mucho mayor para ofrecer resultados.

En la Tabla 5.2 tenemos los resultados de las series cortas generadas por un modelo ARIMA. En este caso, el modelo que peor rendimiento presenta en base a todas las métricas, salvo el tiempo computacional, es el modelo ARMA, lo cual no es de extrañar (y de hecho se repetirá en todos los

	SMAPE	RMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.283	1.240	1.656	0.866	0.680	1.904
ARIMA	1.000	1.000	1.000	1.000	0.862	3.751
ETS	1.020	1.024	1.046	1.098	0.894	0.172
LGT	0.996	1.015	1.010	1.010	0.888	4.631
LGT MAP	1.000	1.017	1.015	0.856	0.820	0.649
DLT	0.998	1.002	0.995	1.039	0.870	3.300
DLT MAP	0.994	1.006	0.999	0.845	0.800	0.672

Tabla 5.2: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA generador de las series cortas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

casos en los que las series a ajustar no sean generadas por un modelo ARMA), puesto que se trata de un modelo fundamentado en la hipótesis de estacionariedad de las series, que no se respeta en este caso. No obstante, en términos de SMAPE, RMAPE y MSE, el rendimiento de los distintos métodos es esencialmente el mismo. Las longitudes de los intervalos de predicción son similares para ARIMA, DLT y LGT, logrando estos últimos una cobertura ligeramente superior. La cobertura de ETS es la mayor, aunque también lo es la longitud de los intervalos de predicción. De nuevo, los modelos que menos tiempo emplean son ETS y los modelos DLT-MAP y LGT-MAP, seguidos de ARMA, y luego por ARIMA y los modelos ajustados por MCMC.

Vayamos con la Tabla 5.3 y los resultados para las series generadas por ETS. En términos de error, ARIMA, ETS, LGT y LGT-MAP, resultan indistinguibles, y ARMA lo hace peor que los primeros cinco. En el caso de DLT y DLT-MAP, de acuerdo con los criterio RMAPE y MSE, también ofrecen un peor rendimiento que ETS y ARIMA, aunque no se pueda ser tan categórico cuando se comparan los modelos DLT y DLT-MAP con LGT y LGT-MAP (sobre todo el ajuste DLT frente al ajuste LGT-MAP). El que peor lo vuelve a hacer, con diferencia, es ARMA. Si nos fijamos en la cobertura de los intervalos de predicción, solamente ETS consigue superar el 91 % siendo el resto de coberturas notablemente inferiores. En algún caso, como para el modelo DLT, se obtiene una menor cobertura con intervalos de mayor longitud. Los tiempos, de nuevo, respetan el orden anteriormente mencionado.

Podemos recopilar aquellos patrones que hemos ido observando, y que se van a repetir en las dos tablas que veremos a continuación, para evitar estar repitiendo las mismas ideas una y otra vez. En primer lugar, ETS es el método que menos tarda en el procedimiento de ajustar modelos, seleccionar uno de ellos y obtener predicciones, seguido de los modelos DLT-MAP y LGT-MAP, así como por el modelo ARMA. Los modelos ARIMA, LGT y DLT tardan considerablemente más en completar este procedimiento. Por otro lado, el modelo ARMA debe restringirse a series estacionarias, pues en otros casos comete un error significativamente superior al del resto de modelos. Por su parte, el ajuste por cadenas de Markov y en base a la estimación máxima a posteriori ofrece resultados consistentemente similares, en términos de errores de predicción, aunque los intervalos de predicción para los modelos ajustados por MAP tienden a quedarse cortos. Además, los ajustes LGT, LGT-MAP, DLT y DLT-MAP son también muy similares entre si.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.118	1.122	1.244	0.886	0.756	0.780
ARIMA	1.000	1.000	1.000	0.989	0.852	3.698
ETS	1.000	1.000	1.000	1.000	0.912	0.172
LGT	1.005	1.014	1.027	0.921	0.830	4.752
LGT MAP	1.023	1.040	1.056	0.787	0.776	0.632
DLT	1.046	1.070	1.106	1.093	0.852	3.263
DLT MAP	1.047	1.063	1.154	0.825	0.790	0.647

Tabla 5.3: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ETS generador de las series cortas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

Continuemos con la Tabla 5.4, en la que se recogen los resultados para las series generadas de acuerdo con el modelo LGT. Al margen de ARMA, el modelo peor parado en términos de error es ARIMA en este caso, siendo significativamente peor que LGT y LGT-MAP e incluso que DLT, si nos fijamos en el MSE. No se puede decir lo mismo entre ARIMA y ETS, sin embargo.

Por su parte, en términos de error, ETS y las dos versiones de LGT y DLT presentan un rendimiento similar. En cuanto a la cobertura de los intervalos, es muy similar entre los ajustes LGT y DLT con entorno a un 93 %, que a su vez son algo superiores a la de ETS con un 90 %. La cobertura de ARIMA es la menor, a pesar de que la longitud de los intervalos sea comparable a la de los previamente mencionados, a excepción de DLT, que tiene los intervalos más amplios.

Finalmente, para este grupo de series de tiempo, tenemos en la Tabla 5.5 los resultados para las series generadas según el modelo DLT. En este caso, ETS y ARIMA resultan, en términos de error, indistinguibles entre si y presentan peores resultados que los modelos Orbit. La cobertura que logra DLT, es del 94 % superior a la del resto de modelos, que se quedan en un 91 % para ETS y LGT, y no llega al 87 % para ARIMA. También es cierto que la longitud de los intervalos de predicción es menor en todos los casos que la que presenta DLT.

5.1.5. Series cortas con errores T de Student generalizados

En lugar de analizar pormenorizadamente cada una de las tablas con los datos de las series cortas con errores T (Tabla 5.6, Tabla 5.7, Tabla 5.8, Tabla 5.9 y Tabla 5.10), lo que haremos será remarcar las diferencias que se encuentran con respecto a lo ya vista en las tablas de datos con errores gaussianos. No habrá muchas discrepancias, teniendo en cuenta que la distribución T es también simétrica respecto al 0, de modo que los parámetros estimados para los modelos deberían ser sean similares a los que se estimarían bajo normalidad, con excepción de la desviación de las innovaciones, que pasará a incorporar la parte de la variabilidad extra asociada a la distribución T, respecto de la normal.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.177	1.151	1.319	1.009	0.864	0.755
ARIMA	1.057	1.057	1.143	1.009	0.870	3.624
ETS	1.024	1.030	1.049	1.032	0.902	0.138
LGT	1.000	1.000	1.000	1.000	0.928	4.518
LGT MAP	1.003	1.002	1.008	0.846	0.858	0.614
DLT	1.017	1.016	1.031	1.111	0.930	3.213
DLT MAP	1.012	1.014	1.024	0.854	0.852	0.630

Tabla 5.4: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo LGT generador de las series cortas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.296	1.319	1.728	0.847	0.798	0.814
ARIMA	1.068	1.092	1.114	0.851	0.866	3.740
ETS	1.052	1.060	1.094	0.901	0.908	0.174
LGT	0.991	0.985	0.972	0.898	0.912	6.256
LGT MAP	1.003	1.008	1.015	0.755	0.858	0.626
DLT	1.000	1.000	1.000	1.000	0.942	3.182
DLT MAP	1.000	0.996	0.990	0.782	0.890	0.642

Tabla 5.5: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo DLT generador de las series cortas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.000	1.000	1.000	1.000	0.944	0.755
ARIMA	1.000	1.000	1.000	1.000	0.942	3.833
ETS	1.037	1.067	1.094	1.134	0.962	0.117
LGT	1.124	1.167	1.294	1.124	0.950	4.006
LGT MAP	1.088	1.114	1.208	0.927	0.910	0.652
DLT	1.150	1.207	1.366	1.331	0.952	3.069
DLT MAP	1.130	1.193	1.310	1.005	0.892	0.668

Tabla 5.6: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARMA generador de las series cortas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.274	1.289	1.624	0.852	0.686	1.897
ARIMA	1.000	1.000	1.000	1.000	0.882	3.748
ETS	1.027	1.047	1.073	1.057	0.894	0.171
LGT	1.016	1.046	0.991	0.967	0.900	4.613
LGT MAP	1.013	1.029	1.017	0.837	0.862	0.643
DLT	1.023	1.061	1.022	1.009	0.882	3.431
DLT MAP	1.028	1.059	1.039	0.824	0.836	0.655

Tabla 5.7: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA generador de las series cortas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.143	1.124	1.317	0.910	0.744	0.781
ARIMA	1.000	1.000	1.000	1.000	0.876	3.705
ETS	1.000	1.000	1.000	1.000	0.892	0.171
LGT	1.017	1.038	1.063	0.917	0.852	4.649
LGT MAP	1.017	1.037	1.047	0.782	0.806	0.617
DLT	1.016	1.058	1.071	1.052	0.856	3.285
DLT MAP	1.033	1.055	1.120	0.814	0.800	0.638

Tabla 5.8: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ETS generador de las series cortas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.139	1.195	1.363	1.017	0.864	0.742
ARIMA	1.087	1.113	1.186	1.041	0.892	3.566
ETS	1.026	1.028	1.076	1.047	0.920	0.137
LGT	1.000	1.000	1.000	1.000	0.922	4.292
LGT MAP	1.014	1.011	1.023	0.855	0.868	0.617
DLT	1.019	1.017	1.034	1.105	0.934	3.226
DLT MAP	1.021	1.020	1.041	0.860	0.864	0.629

Tabla 5.9: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo LGT generador de las series cortas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.186	1.215	1.414	0.866	0.786	0.796
ARIMA	1.029	1.037	1.088	0.870	0.858	3.706
ETS	1.004	0.990	0.998	0.939	0.928	0.160
LGT	0.987	0.985	0.980	0.900	0.916	5.666
LGT MAP	0.991	0.992	0.989	0.759	0.852	0.621
DLT	1.000	1.000	1.000	1.000	0.948	3.142
DLT MAP	0.994	0.990	1.003	0.774	0.866	0.641

Tabla 5.10: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo DLT generador de las series cortas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

En el caso de las series generadas por un modelo ARIMA, la longitud de los intervalos de predicción para las series ajustadas por ETS pasa a ser comparable con las longitudes de los modelos ARIMA y Orbit ajustados por MCMC, y simultáneamente la cobertura pasa a ser esencialmente la misma. Considerando las series generadas por ETS, ahora no hay ningún método que consiga superar el 90 % de cobertura, aunque ETS sigue siendo el más certero en ese aspecto con un 89 %. Si nos fijamos en los resultados de las series generadas por el modelo LGT, el error cometido por el ajuste ARIMA pasa a ser significativamente peor que el error cometido por el análogo ETS Por último, a diferencia de lo que ocurría con las series generadas por el modelo DLT para errores gaussianos, el rendimiento de ARIMA y ETS ha resultado ser el mismo que el de los modelos Orbit.

Cabe destacar que, al contrario de lo que cabría esperar, la cobertura no ha resultado ser siempre menor en las series con errores T frente a lo que observamos en las series con errores gaussianos, para los modelos ARMA, ARIMA y ETS. En algunos casos aumenta ligeramente, como ocurre con las series generadas por un modelo ARMA y ajustadas también por algún modelo ARMA, pues se pasa de una cobertura de 0.936 cuando los errores son gaussianos a una de 0.944 cuando se trata de errores T. No obstante, en ningún caso se han producido variaciones de la cobertura superiores a 2.5 centésimas, y no olvidemos que existe un error de Monte Carlo no despreciable, teniendo en cuenta que cada tabla lleva asociadas solamente 500 series. Aún así, podemos tratar de dar una explicación heurística a por qué podría ocurrir un fenómeno como este, si descartamos la hipótesis del error de aproximación de la cobertura. Puesto que la variabilidad del error aumenta al emplear errores T generalizados frente a errores gaussianos, y dicha variabilidad será recogida en los modelos ARMA, ARIMA y ETS en la desviación típica del error, cabe pensar que en general, la longitud de los intervalos de predicción será mayor en el caso de los errores T que en el de los errores gaussianos. Si el horizonte de predicción no es muy elevado frente al tamaño de la serie ajustada (7 frente a 56 en el caso de las series cortas), puede darse el caso de que el proceso de innovaciones no tenga oportunidad de generar alguna observación "extrema", a pesar de que si haya aparecido alguna para la parte de la serie que se emplee a la hora de obtener el modelo ajustado.

5.1.6. Series largas con errores gaussianos

Continuaremos disctutiendo los resultados obtenidos para las series largas, centrándonos en esta Sección en aquellas que fueron construidas con errores gaussianos.

De acuerdo con la Tabla 5.11, los ajustes ARMA y ARIMA son totalmente indistinguibles sobre series generadas por un modelo ARMA, del mismo modo que para las series cortas análogas. En términos de error, ETS y los modelos de Orbit rinden por igual y lo hacen peor que ARMA y ARIMA. En cuanto a las longitudes de los intervalos de predicción, los más amplios son los de ETS, seguidos por los de los modelos DLT, LGT, DLT-MAP, y LGT-MAP. Finalmente, los más cortos son los de ARMA y ARIMA, aunque no podamos asegurar que sean significativamente más cortos que los de los ajustes MAP. Notemos que la cobertura de los intervalos de predicción para todos los métodos excepto para DLT-MAP y tanto LGT como LGT-MAP es similar, superando el 94 %, de modo que no se ve compensado el incremento de las longitudes de los intervalos frente a los de ARMA, en ese sentido. Finalmente, y esto es algo que se repetirá en el resto de tablas, volvemos a tener la misma ordenación en términos del tiempo computacional necesario para seleccionar y ajustar un modelo a los datos, así como para predecir. El método más rápido es ETS, seguido de LGT-MAP y DLT-MAP, ARMA, ARIMA, DLT y LGT. No obstante, hay una diferencia mucho mayor entre el tiempo necesario para obtener el ajuste ARIMA y los ajustes DLT y LGT que el que existía en el caso de las series cortas (ventaja que proviene de emplear un algoritmo de ramificación y acotación para limitar la búsqueda de posibles modelos, como ya comentamos previamente). Además, LGT tarda considerablemente más (aproximadamente el doble de tiempo) que DLT.

	sMAPE	**rMAPE**	**MSE**	**Longitud**	**Cobertura**	**Tiempo**
ARMA	1.000	1.000	1.000	1.000	0.942	4.923
ARIMA	1.000	1.000	1.000	1.000	0.944	8.384
ETS	1.081	1.116	1.273	1.248	0.966	0.262
LGT	1.102	1.136	1.331	1.123	0.928	37.663
LGT MAP	1.101	1.148	1.328	1.048	0.906	3.404
DLT	1.105	1.146	1.341	1.184	0.942	18.656
DLT MAP	1.097	1.152	1.313	1.065	0.916	3.739

Tabla 5.11: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARMA generador de las series largas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

En la Tabla 5.12 están los resultados para las series largas generadas por un modelo ARIMA. De acuerdo con los 3 criterios de error, los ajustes ARIMA, LGT y LGT-MAP se comportan igual, siendo los que obtienen mejores resultados. Por su parte, si nos fjiamos en las columnas de RMAPE y MSE, los ajustes ETS y DLT lo hacen peor ARIMA, aunque no se podría asegurar que realmente rindan peor que LGT. El ajuste ARMA, del mismo modo que cuando revisamos las series cortas no estacionarias, es el que comete más error, con diferencia (algo que se ocurrirá de ahora en adelante, por lo que no

lo mencionaremos explícitamente). Las longitudes de los intervalos de predicción son similares para ARMA, ARIMA y ETS, que a su vez son superiores a las de los modelos Orbit, que son comparables entre si. La cobertura de los intervalos de predicción, que resulta razonablemente buena para el ajuste ARIMA (93 %), es algo menor en el caso de ETS (87 %) y francamente desastrosa en el resto de casos, no superando el 73 % para ningún otro modelo y siendo particularmente mala en el caso de LGT, que no supera el 63 % en ninguna de sus dos versiones.

	sMAPE	**rMAPE**	**MSE**	**Longitud**	**Cobertura**	**Tiempo**
ARMA	1.504	1.470	2.312	1.000	0.606	4.573
ARIMA	1.000	1.000	1.000	1.000	0.928	13.213
ETS	1.049	1.055	1.118	1.026	0.876	0.422
LGT	1.010	1.035	1.037	0.872	0.622	42.913
LGT MAP	1.012	1.033	1.027	0.819	0.616	3.441
DLT	1.029	1.055	1.075	0.893	0.730	27.297
DLT MAP	1.035	1.056	1.108	0.827	0.724	3.767

Tabla 5.12: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA generador de las series largas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

Centrémonos ahora en las series largas generadas de acuerdo con el modelo ETS, correspondientes con la Tabla 5.13. Los ajustes ARIMA, ETS y LGT cometen el mismo error, según cualquiera de los 3 criterios que empleamos, mientras que los modelos DLT y DLT-MAP rinden peor que ARIMA y ETS. No podemos, sin embargo, establecer una diferencia significativa de rendimiento entre LGT y DLT. La longitud de los intervalos de predicción es casi idéntica para ARMA, ARIMA, ETS y DLT, y algo inferior en el caso de LGT, LGT-MAP y DLT-MAP. La mejor cobertura para los intervalos de predicción se obtiene con el ajuste ETS, del 92 %, muy similar a la que ofrece ARIMA, del 91 %, y a su vez superior a la de los ajustes LGT (85 %), DLT (80 %) y ARMA (75 %). En este ejemplo particular destaca que los ajustes MAP y por cadenas de Markov Monte Carlo de los modelos Orbit ofrecen la misma cobertura de intervalos de predicción (a pesar de la mayor longitud de intervalos de los modelos ajustados por cadenas de Markov), y que la cobertura asociada a LGT sea superior a la de DLT, cuando este segundo modelo tenía intervalos de mayor longitud que el primero.

Los resultados asociados a las series generadas por modelos LGT se pueden consultar en la Tabla 5.14. De acuerdo con los distintos criterios de error, el rendimiento de los modelos ETS, LGT y DLT es similar, y en el caso de los dos últimos, superior al rendimiento de ARIMA y ARMA. Las longitudes de los intervalos de predicción son similares para los modelos de Orbit, pero inferiores a las ARIMA y

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.158	1.143	1.376	0.994	0.750	4.858
ARIMA	1.000	1.000	1.000	1.000	0.910	13.889
ETS	1.000	1.000	1.000	1.000	0.922	0.629
LGT	1.026	1.033	1.042	0.896	0.852	42.057
LGT MAP	1.033	1.047	1.063	0.837	0.844	3.411
DLT	1.063	1.093	1.139	0.985	0.802	24.650
DLT MAP	1.063	1.092	1.159	0.900	0.790	3.725

Tabla 5.13: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ETS generador de las series largas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.198	1.222	1.455	1.151	0.850	3.807
ARIMA	1.077	1.087	1.156	1.143	0.882	9.758
ETS	1.023	1.036	1.057	1.111	0.892	0.653
LGT	1.000	1.000	1.000	1.000	0.880	36.253
LGT MAP	0.997	0.996	0.992	0.939	0.868	3.319
DLT	1.004	1.004	1.009	1.049	0.776	20.511
DLT MAP	1.002	1.000	1.003	0.943	0.736	3.659

Tabla 5.14: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo LGT generador de las series largas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

ETS. Se observan coberturas parecidas para los ajustes ETS, ARIMA y LGT, aunque la peor cobertura se obtiene para DLT quien obtiene un resultado aun peor que la cobertura que ofrece ARMA.

Finalmente la Tabla 5.15 contiene los resultados de los ajustes sobre las series generadas por un modelo DLT. En términos de error, todos los modelos se comportan por igual. Las longitudes de los intervalos de predicción son similares para ARMA, ARIMA, ETS y DLT, y son menores en el caso de DLT-MAP, LGT-MAP y LGT. Curiosamente, la cobertura más acertada es la de ETS (aproximadamente un 95 %) aunque no podamos realmente establecer diferencias con la de DLT (casi un 94 %) o ARIMA (casi un 93 %). La cobertura del ajuste LGT ya es ligeramente inferior, con un 90 %.

	sMAPE	**rMAPE**	**MSE**	**Longitud**	**Cobertura**	**Tiempo**
ARMA	1.028	1.028	1.049	0.985	0.798	2.912
ARIMA	0.993	0.990	0.977	0.988	0.926	6.746
ETS	0.985	0.979	0.964	1.017	0.952	0.384
LGT	1.001	1.002	1.003	0.934	0.900	44.608
LGT MAP	1.004	1.007	1.014	0.860	0.868	3.405
DLT	1.000	1.000	1.000	1.000	0.936	24.389
DLT MAP	1.001	1.001	1.004	0.887	0.904	3.719

Tabla 5.15: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo DLT generador de las series largas con innovaciones gaussianas empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

5.1.7. Series largas con errores T de Student generalizados

De manera análoga con lo expuesto en la Sección 5.1.5, no repasaremos una por una las tablas con los resultados de las series largas construidas con errores T, si no que, al ser muy similares a los resultados de las tablas para las series con errores gaussianos, nos centraremos en constatar las diferencias que haya en relación a los criterios de error, entre tablas para series con el mismo modelo generador. También mencionaremos los cambios producidos en términos de la cobertura de los intervalos de predicción. Todos los datos de esta sección se pueden consultar en las Tablas 5.16, 5.17, 5.18, 5.19 y 5.20.

Las series largas generadas por modelos ARMA, ARIMA y LGT con errores T han dado lugar a resultados sobre el rendimiento de los distintos modelos y la longitud de los intervalos de predicción esencialmente idénticos a los resultados de las series del mismo tipo con errores gaussianos. Ahora bien, en el caso de las series generadas por un modelo ARMA, la cobertura que proporciona el modelo DLT pasa a estar a medio camino entre la de el ajuste análogo para LGT y el ajuste ARMA, aunque todos los modelos ofrecen una cobertura superior al 93 %. Por su parte, para las series generadas por un modelo LGT, la cobertura que ofrece este método (87.4 %) ha resultado ligeramente inferior a la

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.000	1.000	1.000	1.000	0.958	4.415
ARIMA	1.000	1.000	1.000	1.000	0.960	7.566
ETS	1.111	1.197	1.377	1.279	0.968	0.262
LGT	1.129	1.211	1.393	1.119	0.938	37.736
LGT MAP	1.130	1.210	1.393	1.031	0.930	3.393
DLT	1.135	1.225	1.438	1.173	0.948	19.184
DLT MAP	1.128	1.215	1.386	1.057	0.934	3.749

Tabla 5.16: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARMA generador de las series largas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.519	1.562	2.460	1.003	0.672	4.395
ARIMA	1.000	1.000	1.000	1.000	0.914	12.337
ETS	1.066	1.063	1.127	1.043	0.880	0.421
LGT	1.013	1.037	1.036	0.890	0.646	40.652
LGT MAP	1.015	1.028	1.045	0.849	0.638	3.441
DLT	1.038	1.068	1.122	0.885	0.792	23.222
DLT MAP	1.032	1.053	1.107	0.814	0.770	3.767

Tabla 5.17: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA generador de las series largas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.264	1.244	1.502	0.994	0.788	3.804
ARIMA	1.000	1.000	1.000	1.000	0.912	10.630
ETS	1.000	1.000	1.000	1.000	0.918	0.686
LGT	1.011	1.019	1.027	0.852	0.868	38.403
LGT MAP	1.009	1.024	1.042	0.811	0.854	3.382
DLT	1.034	1.056	1.081	0.939	0.804	23.058
DLT MAP	1.039	1.058	1.069	0.866	0.796	3.726

Tabla 5.18: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ETS generador de las series largas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.208	1.247	1.455	1.244	0.850	4.298
ARIMA	1.071	1.083	1.171	1.206	0.906	11.001
ETS	1.028	1.036	1.054	1.164	0.890	0.670
LGT	1.000	1.000	1.000	1.000	0.874	32.525
LGT MAP	0.999	0.999	0.997	0.960	0.866	3.385
DLT	1.002	1.001	1.003	1.042	0.758	18.891
DLT MAP	0.999	0.999	0.998	0.963	0.744	3.722

Tabla 5.19: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo LGT generador de las series largas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARMA	1.035	1.041	1.061	1.025	0.794	2.900
ARIMA	1.001	1.001	1.006	1.047	0.916	6.854
ETS	0.987	0.989	0.983	1.076	0.938	0.385
LGT	1.000	1.000	1.002	0.937	0.876	41.653
LGT MAP	1.005	1.004	1.008	0.879	0.848	3.428
DLT	1.000	1.000	1.000	1.000	0.902	23.151
DLT MAP	1.002	1.000	1.005	0.897	0.880	3.718

Tabla 5.20: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo DLT generador de las series largas con innovaciones T de Student empleadas. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

que ofrece ARIMA (90.6 %).

En el caso de las series generadas por un modelo ETS con errores T, hay una ligera diferencia a la hora de comparar el rendimiento de DLT con ARIMA y ETS, pues en este caso solamente el criterio RMAPE da lugar a diferencias significativas entre el primer modelo y los dos últimos. Además, la longitud de los intervalos de predicción pasa a ser menor en el caso de DLT frente a ARIMA y ETS, aunque sigue siendo mayor que la longitud de los intervalos de predicción de LGT.

Para las series largas con errores T generadas por un modelo DLT, la longitud de los intervalos de predicción de ETS resultó ser ligeramente superior a la de los intervalos de DLT, aunque siga siendo comparable a la longitud de los intervalos de predicción que ha ofrecido ARIMA.

Si nos fijamos en el comportamiento general de la cobertura de los intervalos de predicción destacan dos cuestiones. Por un lado, la cobertura de estos intervalos en el caso de las series generadas por modelos ARMA aumenta ligeramente para todos los ajustes al pasar de errores gaussianos a errores T. Este comportamiento es difícil de justificar cuando trabajamos con series con un tamaño relativamente elevado (se han ajustado los modelos con series de tamaño 336, y el horizonte de predicción es 24). No es sencillo explicar por qué ha ocurrido solamente para estos datos, pues el argumento empleado para las series cortas en la Sección 5.1.5 no parece que se sostenga al aumentar tanto el tamaño de las series. Afortunadamente, este efecto no se mantiene en el resto de las tablas sino que, al menos en el caso de la cobertura de los intervalos de predicción para los modelos que pertenecen a la familia que contiene al modelo generador de las series, se ve reducida cuando se pasa a los errores con colas más pesadas. Así pues, la cobertura que proporcionan los intervalos del modelo ARIMA para series generadas por algún modelo ARIMA pasa de un 92.8 % a un 91.4 %, y en el caso de las series generadas por algún modelo ETS, la cobertura de los intervalos del ajuste ETS para de un 92.2 % a un 91.8 %. De todos modos, no ocurre esto último de manera generalizada, puesto que el ajuste ETS para series generadas por un modelo ARIMA pasa de un 87.6 % a un 88 %. En el caso particular de los modelos Orbit, al asumir por defecto que las innovaciones siguen una distribución T generalizada, siendo el caso gaussiano una particularización cuando los grados de libertad de la distribución T toman valores muy elevados, no tendría por qué ocurrir un aumento o una disminución consistente al tratar con uno u otro tipo de distribución de las innnovaciones, y eso es exactamente lo que pasa. Por ejemplo, la cobertura de LGT

sobre las series generadas por un modelo LGT es de un 88 % cuando las innovaciones son gaussianas y de un 87.4 % cuando se distribuyen según una distribución T generalizada, mientras que la cobertura de DLT sobre series generadas de acuerdo con dicho modelo pasa de un 93.6 % a un 90.2 % al cambiar la distribución de las innovaciones.

5.1.8. Conclusiones del estudio benchmark

Vamos a tratar de identificar las características de los distintos modelos que se hayan mantenido consistentemente a lo largo de las distintas comparativas que hemos efectuado a lo largo de la Sección 5.1.

En primer lugar, en caso de estar convencidos de que se trabaja con series estacionarias, nada supera el rendimiento de ARMA, específicamente diseñado para tal fin. No obstante, si las series no presentan esta propiedad, podemos esperar que los resultados que ofrezca un ajuste ARMA no sean los adecuados.

Si nos centramos en los modelos de Orbit, resulta reseñable que, en términos de error de predicción, no hemos encontrado (en ningún caso) diferencias relevantes entre los ajustes por cadenas de Markov Monte Carlo, y los ajustes basados en estimación máxima a posteriori. Por tanto, si el objetivo del análisis de series de tiempo fuese exclusivamente obtener predicciones puntuales, teniendo en cuenta el tiempo computacional mucho más elevado requerido por el primero de los métodos de ajuste, habría que decantarse por el ajuste MAP. Lo que sí es cierto es que, en general, los intervalos de predicción asociados al asjute por estimación máxima a posteriori tienden a ser demasiado estrechos, dando lugar a coberturas limitadas. En este sentido, sí es razonable recurrir a la estimación basada en cadenas de Markov. Por otro lado, tampoco hemos encontrado ningún caso en el que se puedan establecer diferencias significativas entre el rendimiento de LGT y DLT, ajustados de la misma forma, aunque en alguna ocasión si que ha existido cierta ambiguedad, al ser LGT simultáneamente comparable a DLT y otro tercer método, que a su vez obtenía resultados significativamente mejores que DLT. Esto es un problema insalvable al basarse en comparaciones 2 a 2 (y ocurre lo mismo en otros contextos, como por ejemplo a la hora de comparar medias de grupos de acuerdo con el modelo ANOVA).

Ha destacado también el papel de ETS, modelo que solamente en un caso (serie cortas generadas por un modelo DLT con errores gaussianos) ha obtenido peores resultados en términos de error de predicción que los modelos de Orbit, y de hecho, es el único caso en que no ha tenido un rendimiento comparable al del "mejor modelo"para las series con las que se estuviese trabajando. Este hecho, ligado a que resulta ser el modelo computacionalmente más eficiente, dificulta mucho poder justificar el empleo de los modelos Orbit, frente a ETS (al menos a la hora de obtener predicciones puntuales).

El rendimiento general de ARIMA ha sido similar al de los modelos de Orbit (y a su vez inferior al de ETS). Sin embargo, recordemos que no hemos permitido una total libertad para escoger el mejor modelo ARIMA en función del criterio BIC, si no que hemos limitado considerablemente la máxima estructura que podría tener el método (con el objetivo de reducir el tiempo computacional necesario para obtener los resultados que hemos analizado), limitando a su vez el rendimiento que podría llegar a ofrecer en las series generadas por ETS, LGT y DLT (no así en las series generadas por ARMA y ARIMA, cuyo modelo generador se encontraba dentro del conjunto de modelos que se podrían llegar a haber ajustado).

Que la distribución de las innovaciones sea gaussiana o una distribución T generalizada no parece tener un efecto relevante sobre la cobertura de los intervalos de predicción ni sobre los errores de predicción. Es más, en el caso de que las series de tiempo tengan un tamaño reducido, el efecto de considerar innovaciones con colas más pesadas puede ser el opuesto al esperado, incrementándose la cobertura de los métodos que asumen distribución gaussiana.

Por último, cabe destacar que en ningún caso se ha efectuado un análisis de residuos para verificar que el modelo ajustado automáticamente en función del criterio BIC fuese capaz de recoger adecuadamente la estructura de dependencia de los datos, y en otro caso buscar otro modelo alternativo que si fuese capaz de esto último. Esta práctica podría haber sido beneficiosa en el caso de ARIMA y ETS (no para los modelos de Orbit, en los que no hay proceso de selección) a la hora de reducir

los errores predictivos, teniendo en cuenta que el criterio BIC es un criterio razonable de selección de modelos desde el punto de vista asintótico, pero no necesariamente para tamaños muestrales pequeños o moderados.

Antes de dar por cerrado el estudio benchmark, fijémonos en la Tabla 5.21. Se trata de resultados asociados a las series largas generadas por un modelo ARIMA con errores T generalizados. En concreto, es una tabla resumen de los errores de predicción medidos según el SMAPE, de modo que cada columna revela los resultados que se hubiesen obtenido si, en lugar de tomar como referencia el modelo ARIMA, se tomase como referencia el modelo que da nombre a la columna de la tabla. Nos servirán para justificar los umbrales que hemos empleado a lo largo de este capítulo para considerar como significativas las diferencias entre distintos valores de MSE, SMAPE, RMAPE o longitudes de los intervalos de predicción.

	ARMA	**ARIMA**	**ETS**	**LGT**	**LGT MAP**	**DLT**	**DLT MAP**
ARMA	1.000	1.519	1.341	1.580	1.540	1.361	1.358
ARIMA	0.658	1.000	0.938	0.987	0.985	0.963	0.969
ETS	0.746	1.066	1.000	1.057	1.049	1.022	1.030
LGT	0.633	1.013	0.946	1.000	0.999	0.997	0.993
LGT MAP	0.649	1.015	0.953	1.001	1.000	0.996	0.999
DLT	0.735	1.038	0.979	1.003	1.004	1.000	1.000
DLT MAP	0.736	1.032	0.971	1.007	1.001	1.000	1.000

Tabla 5.21: Tabla comparativa con el rendimiento de los distintos métodos, medido de acuerdo con el SMAPE, para series largas generadas por un modelo ARIMA con innovaciones T de Student. Cada columna contiene la mediana de los SMAPE relativizados con respecto al modelo que le da nombre.

Si nos fiásemos de la primera columna, tomando como referencia el modelo ARMA, el error de predicción cometido por el modelo ARIMA sería superior (en 2.5 centésimas) al error de predicción cometido por el ajuste LGT. En cambio, para el resto de columnas, esta relación se invierte. Si, en cambio, se respeta el umbral de 5 centésimas de diferencia, esos modelos serían indistinguibles en términos de rendimiento, lo cual sería coherente con el hecho de que la interpretación de los resultados en este caso dependa de qué modelo se toma como referencia. Experimentalmente, esa cantidad de 5 centésimas para los criterios SMAPE y RMAPE, 1 décima para el MSE y 5 centésimas para la longitud de los intervalos de confianza son valores que consiguen que, en caso de que para el modelo que se considere como referencia se consiga una diferencia superior a esas cantidades entre los resultados de dos modelos distintos, entonces se respete (como mínimo) la relación de orden entre los resultados si se hubiese tomado como referencia cualquier otro modelo.

5.2. Series simuladas con estacionalidad multiplicativa, innovaciones multiplicativas o intervenciones de salto

En esta sección nos centraremos en la segunda parte del estudio de simulación comparativo. La idea es medir el rendimiento de los distintos modelos sobre series de tiempo que, por una parte no hayan

sido generadas por ninguno de los modelos que estamos poniendo a prueba, y que además incluirán alguna de las siguientes características: innovaciones de tipo multiplicativo, componente estacional multiplicativa o una intervención de salto, ya sea de tipo transitorio o permanente.

5.2.1. Procedimiento de simulación

Todas las series simuladas para esta parte se basan en un mismo procedimiento inicial: seleccionar aleatoriamente series ARIMA, ETS y DLT de entre las generadas para el estudio benchmark, reescalarlas (para que tengan aproximadamente la misma escala) y sumarlas. El proceso de reescalado consiste en dividir la serie por el dato que tuviese el máximo valor absoluto, y multiplicarla posteriormente por 100 (la elección de este último número es completamente arbitraria).

Una vez obtenidas nuevas series por el procedimiento antes mencionado, se transformarán para incorporar aquellos fenómenos que interesen, en cada caso. Por ejemplo, para crear series con innovaciones de tipo multiplicativo, si denotamos por $Y_1, \cdots, Y_n$ a la serie de partida, hemos llevado a cabo lo siguiente:

1. $Y_t \leftarrow Y_t - \text{mín}\{Y_1, \cdots, Y_n\} + 1, \ \forall 1 \leq t \leq n.$
2. $Y_t \leftarrow Y_t(1 + e_t), \ \forall 1 \leq t \leq n$, donde $e_1 \cdots e_n$ es una muestra aleatoria simple con distribución normal de media 0 y desviación típica 0.3.
3. $Y_t \leftarrow Y_t - \text{mín}\{Y_1, \cdots, Y_n\} + 10, \ \forall 1 \leq t \leq n.$

De este modo, con el primer paso se obtiene una serie estrictamente positiva; con el segundo paso se logra una serie de tiempo para la que la variabilidad aparente del error aumenta con el nivel de la misma (la desviación típica se ha elegido como aquel valor para el que las representaciones de las series se correspondiesen con lo que estamos buscando, no hay una interpretación estadística subyacente); y con el tercer paso trasladamos de nuevo la serie hasta obtener otra serie positiva y no demasiado cercana al valor 0 (de otro modo podrían surgir problemas al ajustar algún método). En la Figura 5.1 podemos ver una de las series construidas de acuerdo con este esquema. Se puede apreciar como, efectivamente, capta esa idea de variabilidad de las innovaciones de la serie incrementándose con el nivel de la misma.

Para generar series con una componente estacional de tipo multiplicativo el procedimiento que hemos seguido es algo más laborioso. La idea que perseguimos en este contexto es generar series que presenten un patrón repetitivo evidente, que dicho patrón sea la característica predominante de la serie, y que además tenga una amplitud (la componente estacional) que se incremente con el nivel de la serie. Lo resumimos en los siguientes pasos:

1. Generar la componente estacional a través de una suma finita de términos que imiten un desarrollo en serie de Fourier. Esto es, se sortean valores a_1 de una distribución uniforme en $(0.3, 0.4)$ y a_2, a_3, b_1, b_2, b_3 de sendas uniformes, independientes, en $(-0.3, 0.3)$. Posteriormente se multiplican a_1 por 1 o -1 con probabilidad 0.5 (cada valor) y el resto por 0 o 1 con la misma probabilidad. Se construye la componente estacional como

   $$s_t = a_1 \cos\left(\frac{2t\pi}{m}\right) + a_2 \cos\left(\frac{4t\pi}{m}\right) + a_3 \cos\left(\frac{6t\pi}{m}\right) + b_1 \sin\left(\frac{2t\pi}{m}\right) + b_2 \sin\left(\frac{4t\pi}{m}\right) + b_3 \sin\left(\frac{6t\pi}{m}\right),$$

   con m el periodo estacional deseado.

2. Suavizar la serie de tiempo mediante un procedimiento basado en regresión no paramétrica local (lowess con ancho de banda 0.2) para quedarnos únicamente con el "comportamiento general"de la serie, eliminando el ruido del proceso de innovaciones y una posible componente estacional previa. Denotemos por $Y_1^*, \cdots, Y_n^*$ a la serie suavizada.

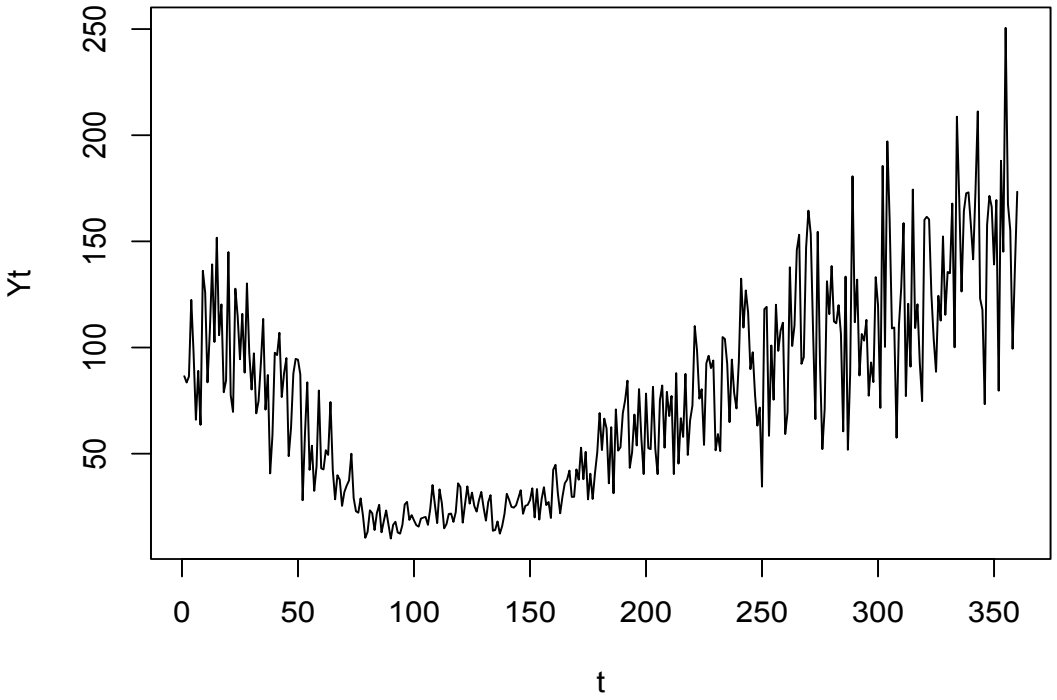

Figura 5.1: Ejemplo de serie simulada con innovaciones de tipo multiplicativo

3. Tomar $Y_t = Y_t^*(1 + s_t) + e_t,\ \forall 1 \leq t \leq n$, con e_t siguiendo una distribución normal de media 0 y con varianza igual al cuantil empírico del 10 % asociado a $Y_t^*(1 + s_t)$ con $t = 1, \cdots, n$.

4. $Y_t \leftarrow Y_t - \text{mín}\{Y_1, \cdots, Y_n\} + 10,\ \forall 1 \leq t \leq n$.

Fijémonos en que el paso 1 se asegura de que al menos haya un término de la suma que tenga una escala no despreciable (a_1). Los extremos de las distribuciones uniforme, de nuevo, han sido elegidos de modo que las series resultantes tuviesen "buen aspecto". No hay una justificación formal al empleo de un método de regresión local, ni a la elección del parámetro de suavizado asociado más allá de que efectivamente se consigue construir series con las características buscadas. Del mismo modo, la elección de la desviación típica de las innovaciones añadidas en el paso 3 permite que, cuando las observaciones de la serie se encuentren en zonas con un nivel elevado, las innovaciones sean despreciables frente a la componente estacional, mientras que cuando se encuentren en zonas de bajo nivel, la variabilidad de las innovaciones se deje entrever. En la Figura 5.2 aparece representada una serie construida de esta manera. Resulta ser un ejemplo representativo de una serie de tiempo con componente estacional multiplicativa.

Por último, explicaremos como hemos obtenido series con una intervención de salto. Para ello debemos explicar qué entendemos por una "intervención". Por analogía con el enfoque clásico del análisis para series de tiempo asociado a los modelos Box-Jenkins, entenderemos una intervención como una perturbación de la serie correspondiente a un shock externo, cuyo origen en principio es conocido. En concreto, asumiremos que dicha perturbación tiene como consecuencia variar los valores de la serie, a partir del instante en que tiene lugar el efecto de la intervención, durante un periodo finito (salto transitorio, se puede observar el fin del efecto de la intervención en la propia serie) o indefinido (salto permanente, no se llega a observar el final del efecto de la intervención). Esa variación será

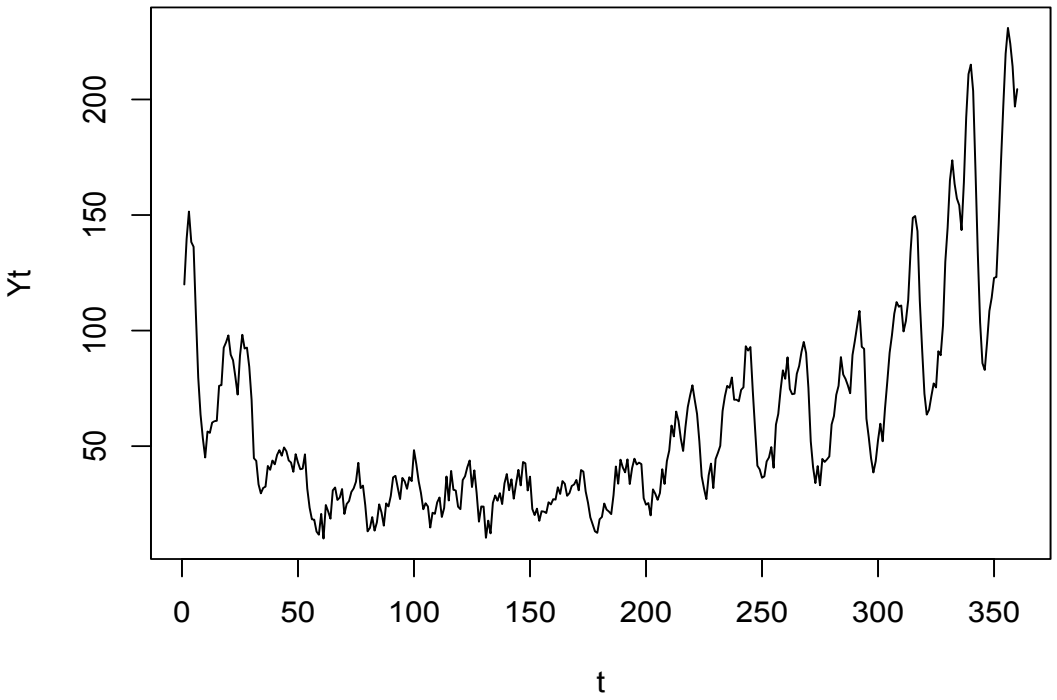

Figura 5.2: Ejemplo de serie simulada con componente estacional de tipo multiplicativo

la misma para todos los valores de la serie que se vean afectados. Formalmente, a partir de la serie original, $Y_1, \cdots, Y_n$, se obtiene la serie perturbada $Y_1^*, \cdots, Y_n^*$ verificando

$$Y_t^* = Y_t + \beta \mathbf{1}(r_1 \leq t \leq r_2),\ \forall 1 \leq t \leq n,$$

con $1 \leq r_1 < r_2 \leq n$ y $\beta \in \mathbb{R}$ el coeficiente que regula el impacto de la perturbación, en caso de que el efecto de la misma sea transitorio, y

$$Y_t^* = Y_t + \beta \mathbf{1}(r \leq t),\ \forall 1 \leq t \leq n,$$

para cierto $1 \leq r \leq n$, en caso de que el efecto sea permanente.

Esto es, a grandes rasgo, lo que hemos hecho a partir de las series construidas previamente, sin más que escoger el coeficiente β asociado a la intervención y los puntos en los que se produce el inicio y el final del efecto sobre la serie. La selección de β ha consistido en tomar 3 veces la desviación típica de la serie diferenciada regularmente hasta que pasase un test de estacionalidad (Said y Dickey, 1984), que se aproximaría a 3 veces la desviación típica del error asociado a un modelo ARIMA ajustado sobre dicha serie, con el signo escogido aleatoriamente. De este modo se logran "saltos"no despreciables. Los extremos se escogen aleatoriamente, dejando las $m-1$ primeras y últimas (en caso de intervención transitoria) observaciones sin perturbar para que exista un punto de referencia a la hora de que los modelos traten de captar esas intervenciones. En las Figuras 5.3 y 5.4 se puede ver un ejemplo representativo de ambos tipos de intervención.

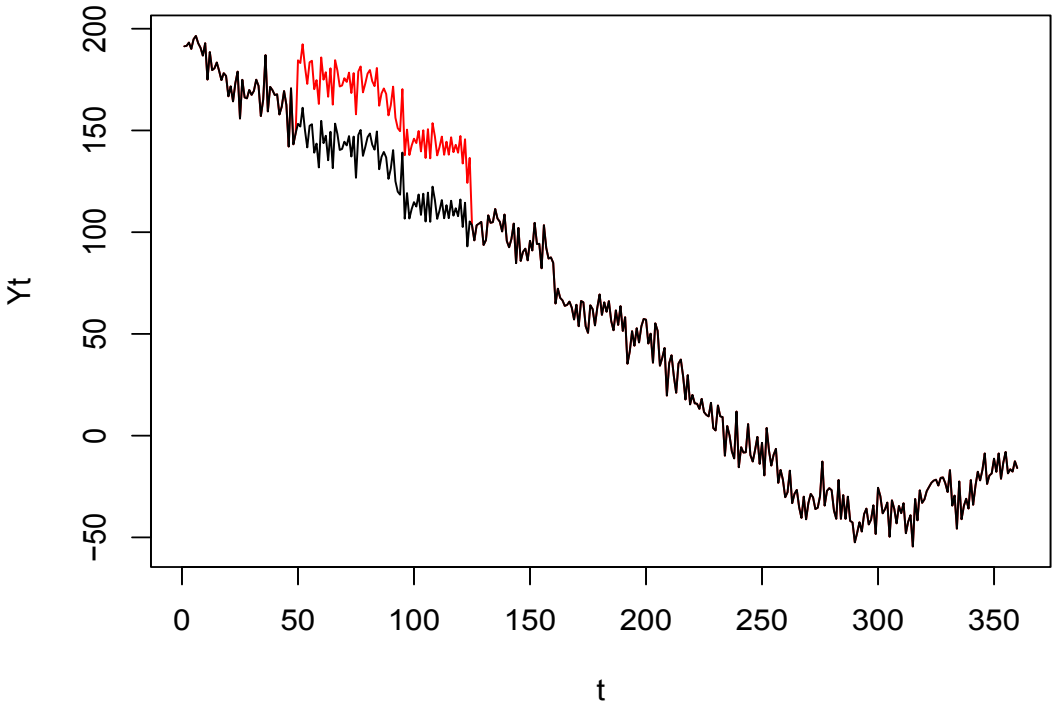

Figura 5.3: Ejemplo de serie simulada con una intervención de salto transitoria. La serie original (en negro) y la perturbada (en rojo) están superpuestas

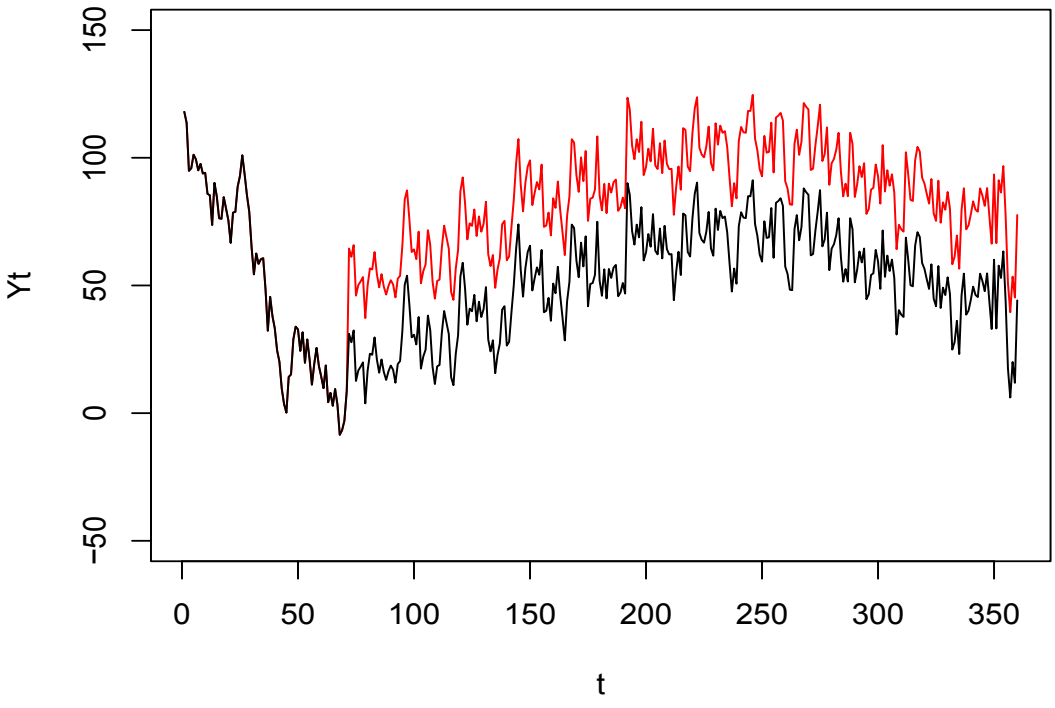

Figura 5.4: Ejemplo de serie simulada con una intervención de salto permanente. La serie original (en negro) y la perturbada (en rojo) están superpuestas

Para cada uno de estos 4 escenarios (errores multiplicativos, componente estacional multiplicativa, intervenciones de salto transitorio y permanente) se han construido 500 series cortas (tamaño 63 reservando las últimas 7 observaciones para medir el error de predicción y periodo $m = 7$) y otras 500 largas (tamaño 360 reservando las últimas 24 observaciones para medir el error de predicción y periodo $m = 24$).

5.2.2. Procedimiento de ajuste y predicción

En base a los resultados del estudio benchmark, teniendo en cuenta que las series simuladas no serán, en general, estacionarias, no ajustaremos modelos ARMA. Por contra, para las series con innovaciones multiplicativas o componente estacional multiplicativa incorporaremos al estudio las versiones multiplicativas del modelo ETS, mencionadas en la Sección 3.2. Creemos que puede resultar interesante determinar si merece la pena aumentar el total de modelos de entre los que ETS puede escoger cual ajustar a una serie de tiempo, incorporando los modelos multiplicativos (de la forma ETS(M, X, X), ETS(X, M, X), ETS(X, X, M) con X representando cualquier valor del conjunto $\{N, A, Ad, M\}$ según corresponda).

Las ideas fundamentales detrás del ajuste y predicción ligados a los distintos modelos son las que ya hemos tratado en la Sección 5.1.2. Las series con innovaciones multiplicativas serán transformadas mediante un logaritmo (por este motivo nos hemos asegurado de que tomen valore positivos) antes de ser ajustadas por los modelos ARIMA, ETS aditivo, LGT y DLT, tal y como se procedería en la práctica con series de este tipo (han sido generadas con esta idea en mente, y el logaritmo efectivamente logra estabilizar la varianza/escala de la componente estacional). Se obtendrán las predicciones e intervalos de predicción sobre la serie transformada, y se devolverán a la escala original a través de la función exponencial. En el caso del modelo ETS no aditivo, la serie no será transformada, pues en teoría los modelo de este tipo están diseñados para tratar con series que presenten estas características.

En cuanto a las series con intervención, ya sea transitoria o permanente, requieren de modelos que permitan incorporar una componente regresiva, de modo que el modelo LGT queda descartado. Se ajustarán, por tanto, los modelos ARIMA, ETS y DLT. Ahora bien, la idea que seguiremos a la hora de tratar las intervenciones es la siguiente: se conoce el momento en el que tiene lugar el shock externo causante de la perturbación de la serie, pero no se conoce con exactitud (a priori) en qué momento este shock externo tiene un efecto sobre la serie de tiempo. De acuerdo con esta premisa, para cada serie de tiempo con intervención se ajustarán modelos que admitan un posible salto con extremo inicial y final (en caso de intervención transitoria) que tenga lugar en 3 posibles instantes distintos: el instante en el que realmente ha tenido lugar (que conocemos por construcción de las series simuladas) y los instantes inmediatamente anterior y posterior. Además, se ajustará un modelo que no considere ningún efecto sobre la serie. Posteriormente se escogerá el mejor modelo en términos del criterio BIC, de entre las 3 o 9 posibles combinaciones de modelos con intervención (en función de si es permanente o transitoria, respectivamente) y el modelo adicional que no incorpora intervención de ningún tipo. De este modo, que el modelo finalmente escogido sea aquel que no incorpora intervención quiere decir que, bajo el paradigma de la familia de modelos con la que se trabaje, se puede considerar el "salto"de la serie como una evolución natural de una serie de tiempo que haya sido generada por uno de los modelos de la familia.

Notemos que una selección mediante criterio BIC de modelos con intervención frente a modelos sin intervención no es el mecanismo estándar en este contexto, al trabajar con modelos Box-Jenkins. Lo habitual es ajustar un modelo con intervención y efectuar un contraste de significación del coeficiente de regresión asociado a la perturbación. Sin embargo, ese enfoque no es viable para nuestro estudio por dos motivos. En primer lugar, los contrastes de significación usuales carecen de sentido desde el punto de vista bayesiano, en el que se enmarca el modelo DLT. En segundo lugar, las funciones que permiten ajustar los modelos ARIMA y ETS del paquete `smooth` tienen problemas a la hora de calcular la matriz de información de Fisher cuando se incorpora una componente de regresión, de modo que tampoco resulta posible efectuar contrastes de significación para estos modelos (no conocemos la existencia de ningún otro paquete que permita ajustar ARIMA y ETS como modelos SSOE y que además incorpore

la posibilidad de introducir la parte de regresión).

Por último, tendremos que limitar el ajuste de DLT a su versión estimada en base a la estimación máxima a posteriori. Esto se debe a que, en el momento en el que redactamos este trabajo, no está implementada en la librería la posibilidad de calcular el criterio BIC (ni ningún otro criterio de información) para los modelos ajustados por cadenas de Markov Monte Carlo. Por tanto, no hay herramientas de selección de modelos que no se basen en algún procedimiento de validación cruzada, algo que descartamos en base a los resultados del análisis de sensibilidad para los parámetros del modelo DLT.

5.2.3. Medición del rendimiento de los modelos

Del mismo modo que para el estudio benchmark, se calcularán el SMAPE, MSE y RMAPE (definido como en la Sección 5.1.3), así como la longitud del intervalo de predicción y la indicadora de pertenencia de la verdadera observación de la serie de tiempo al intervalo de predicción en el instante final, y el tiempo que se tarda en todo el proceso de selección de modelos, ajuste y predicción, para todas las series de tiempo. Estas medidas serán resumidas mediante las medianas de los criterios de error de predicción y longitudes de los intervalos de predicción relativizados (procedimiento explicado en la Sección 5.1.3) y el promedio de las variables indicadoras y los tiempos computacionales. Puesto que en este caso no hay un modelo de referencia evidente (al no haber sido generadas las series a partir de un modelo de los que se testea en concreto), se tomará a ARIMA como modelo base, con respecto al cual se calculan las medidas relativizadas.

Adicionalmente, para las series de tiempo con intervención transitoria y permanente se calcularán la frecuencia relativa con la que el modelo finalmente seleccionado considera que hay alguna perturbación, la frecuencia relativa con la que, en caso de considerar que hay una perturbación, se ha acertado en los instantes de inicio y final de la misma, así como el promedio del porcentaje del salto que el modelo ha sido capaz de capturar (suponiendo que el modelo asume la existencia de una perturbación) medido como el cociente entre el coeficiente de regresión β estimado y el verdadero salto introducido en la serie.

5.2.4. Series con innovaciones de tipo multiplicativo

Vamos a analizar los resultados obtenidos para las series cortas y largas simuladas con innovaciones multiplicativas, correspondientes con la Tabla 5.22 y la Tabla 5.23.

Como podemos apreciar, en el caso de las series cortas, a excepción del modelo ETS multiplicativo ningún otro modelo obtiene errores de predicción significativamente menores a los errores que comete algún otro modelo. En el caso particular de ETS multiplicativo, solamente el criterio SMAPE sugiere que hay una reducción del error, y dicha reducción se podría asegurar solamente respecto a ARIMA. Las longitudes de los intervalos de predicción son similares para ARIMA, ETS aditivo y LGT, inferiores para ETS multiplicativo, DLT-MAP y LGT-MAP, y superiores para DLT. Las coberturas de los intervalos de predicción son muy parecidas para ARIMA, ETS aditivo, LGT y DLT, superando en todos los casos el 95 %; y son inferiores en el caso de ETS multiplicativo, LGT-MAP y DLT-MAP. En particular, la cobertura del modelo ETS multiplicativo no supera el 90 %. En cuanto al tiempo computacional, el modelo que menos necesita sigue siendo ETS, tanto en su versión puramente aditiva como la versión que admite modelos multiplicativos, seguido de LGT-MAP y DLT-MAP. Los modelos que más tiempo tardan en el proceso de selección, ajuste y predicción son ARIMA, DLT y LGT.

Si nos fijamos ahora en los resultados para las series largas, llegamos a las mismas conclusiones en términos de error de predicción, con la salvedad de que ahora ya no existen dudas acerca de si el modelo ETS multiplicativo obtiene un mejor rendimiento que ARIMA, cosa que no ocurre. Las longitudes de los intervalos ahora son similares para todos los modelos, excepto para LGT-MAP, con intervalos ligeramente más pequeños que los que se obtienen con ARIMA, ETS aditivo o DLT.

	sMAPE	rMAPE	MSE	longitud	Cobertura	Tiempo
ARIMA	1.000	1.000	1.000	1.000	0.968	2.431
ETS	0.999	0.994	1.000	0.980	0.974	0.195
ETS M	0.945	0.966	0.906	0.722	0.896	0.415
LGT	0.971	1.033	0.975	1.021	0.956	3.289
LGT MAP	0.988	1.055	1.000	0.766	0.910	0.628
DLT	0.965	1.005	0.960	1.230	0.968	3.020
DLT MAP	0.973	1.029	0.999	0.813	0.908	0.647

Tabla 5.22: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA para las series cortas simuladas con innovaciones de tipo multiplicativo. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARIMA	1.000	1.000	1.000	1.000	0.966	1.586
ETS	0.997	0.994	0.994	1.007	0.978	0.338
ETS M	0.970	0.977	0.938	0.964	0.964	0.565
LGT	0.992	1.008	0.981	0.961	0.944	19.530
LGT MAP	0.988	1.006	0.964	0.935	0.928	3.593
DLT	0.988	1.006	0.977	1.078	0.956	19.665
DLT MAP	0.990	1.011	0.980	0.962	0.932	4.039

Tabla 5.23: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA para las series largas simuladas con innovaciones de tipo multiplicativo. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

En términos de cobertura, las diferencias se reducen considerablemente, siendo solamente los intervalos asociados a LGT-MAP y DLT-MAP inferiores al 94 %. Los tiempos computacionales siguen el mismo orden para los ajustes de las series largas que para los ajustes de las series cortas, con la excepción de que ARIMA ha resultado ser más rápido que tanto DLT-MAP como LGT-MAP.

5.2.5. Series con componente estacional multiplicativa

Nos fijaremos ahora en las tablas asociadas a las series con componente estacional de tipo multiplicativo, presentes en las Tablas 5.24 y 5.25.

Comenzando por los resultados para las series cortas, vemos como, de hecho, todos los modelos ofrecen un error significativamente menor que ARIMA. Además, también se pueden encontrar diferencias significativas entre el rendimiento de la versión aditiva de ETS, y ETS multiplicativo, LGT-MAP, DLT y DLT-MAP. Por su parte, DLT comete menores errores de predicción que el resto de modelos, a excepción de LGT-MAP. La longitud de los intervalos de predicción es similar para DLT, ETS y ARIMA, ligeramente inferior para LGT, y notablemente inferior para LGT-MAP y DLT-MAP y, sobre todo, para ETS multiplicativo. Las coberturas son parecidas para ARIMA, ETS, LGT y DLT y, a su vez, superan las coberturas de LGT-MAP y DLT-MAP, así como de ETS multiplicativo, que no logra superar ni el 85 %.

Si nos centramos en los resultados de las series largas, tenemos que ARIMA y ETS multiplicativo han obtenido los mayores errores de predicción, superando a ETS aditivo y a los modelos Orbit, por ese orden. Las longitudes de los intervalos de predicción son, para ARIMA, las mayores de todos los modelos, con diferencia. Además, los intervalos de ETS aditivo superan en longitud a los de DLT, LGT, DLT-MAP, LGT-MAP y, por último, ETS multiplicativo. En cuanto a coberturas tenemos dos grupos diferenciados. Por una parte ARIMA, ETS y DLT, con coberturas próximas al 95 %; por otra, el resto de modelos, con coberturas en torno al 92 %.

	sMAPE	**rMAPE**	**MSE**	**longitud**	**Cobertura**	**Tiempo**
ARIMA	1.000	1.000	1.000	1.000	0.958	3.467
ETS	0.937	0.935	0.839	0.964	0.966	0.168
ETS M	0.875	0.868	0.727	0.401	0.832	0.733
LGT	0.888	0.898	0.797	0.889	0.950	3.900
LGT MAP	0.851	0.866	0.761	0.735	0.924	0.643
DLT	0.820	0.846	0.673	0.978	0.980	3.736
DLT MAP	0.819	0.831	0.686	0.715	0.932	0.660

Tabla 5.24: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA para las series cortas simuladas con componente estacional de tipo multiplicativo. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

No hemos hecho mención a los tiempos computacionales, pues se comportan como en la sección

	sMAPE	rMAPE	MSE	Longitud	Cobertura	Tiempo
ARIMA	1.000	1.000	1.000	1.000	0.974	2.546
ETS	0.904	0.885	0.790	0.807	0.966	0.516
ETS M	0.997	0.986	1.004	0.565	0.920	2.519
LGT	0.835	0.848	0.664	0.681	0.922	27.218
LGT MAP	0.795	0.799	0.609	0.623	0.926	3.629
DLT	0.805	0.793	0.638	0.715	0.942	28.299
DLT MAP	0.808	0.789	0.653	0.652	0.918	3.887

Tabla 5.25: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza con respecto a los resultados del modelo ARIMA para las series largas simuladas con componente estacional de tipo multiplicativo. También se incluye la cobertura media y tiempo de ejecución (en segundos) promedio.

anterior, si comparamos series cortas o largas con innovaciones multiplicativas y sus análogas con componente estacional multiplicativa.

5.2.6. Series con intervención de salto

Finalmente, echaremos un vistazo a las series con intervenciones de salto, tanto transitorio como permanente. Las Tablas 5.26 y 5.27 contienen los resultados asociados a las intervenciones transitorias mientras que las Tablas 5.28 y 5.29 se refieren a los resultados para las series simuladas con intervenciones de salto permanente.

	sMAPE	rMAPE	MSE	Long.	Cobert.	Tiempo	Det.	Ef. Inc.	Ext.1	Ext.2
ARIMA	1.000	1.000	1.000	1.000	0.810	14.453	0.996	0.978	0.968	0.966
ETS	1.024	1.011	1.009	1.049	0.860	2.241	0.972	0.991	0.969	0.977
DLT MAP	0.963	0.958	0.907	0.960	0.840	0.229	0.996	0.968	0.988	0.978

Tabla 5.26: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza (Long.) con respecto a los resultados del modelo ARIMA para las series cortas simuladas con intervención de salto transitorio. También se incluye la cobertura media (Cobert.) y tiempo de ejecución (en segundos) promedio, así como la frecuencia relativa de identificación del efecto de intervención (Det.), el promedio de la magnitud del efecto de intervención incorporado con respecto al tamaño real del salto (Ef. Inc.) y las frecuencias relativas de correcta indetificación de los instantes de inicio y final del efecto de intervención (Ext.1 y Ext.2).

Podemos comentar simultáneamente las tablas de las intervenciones de tipo transitorio para series cortas y largas, puesto que los resultados son casi idénticos. Tanto en términos de error de predicción y longitud de intervalos de predicción (exceptuando DLT-MAP, que obtiene intervalos algo más pequeños que los otros dos modelos), como en función de la frecuencia relativa de identificación del efecto de la intervención, del tanto por uno del salto incorporado en caso de detectar intervención, y de la frecuencia relativa de correcta identificación del principio y final del efecto de intervención, los 3 modelos tienen aproximadamente el mismo rendimiento. Quizá podríamos argumentar que ETS ha logrado una mayor cobertura que ARIMA o DLT, pero honestamente ninguno de los 3 métodos logra una cobertura superior al 90 %, de modo que no habría que fiarse de los intervalos de predicción de ninguno de ellos. Lo que destaca, sin embargo, es que DLT-MAP requiere de menos tiempo computacional para llegar a ajustar el modelo que considera como óptimo, y obtener predicciones con él, que los otros 2 modelos.

	sMAPE	**rMAPE**	**MSE**	**Long.**	**Cobert.**	**Tiempo**	**Det.**	**Ef. Inc.**	**Ext.1**	**Ext.2**
ARIMA	1.000	1.000	1.000	1.000	0.822	101.217	0.982	0.989	0.965	0.941
ETS	0.983	0.986	0.968	1.001	0.890	6.360	0.968	0.992	0.977	0.959
DLT MAP	0.953	0.947	0.913	0.884	0.802	4.054	0.996	0.981	0.986	0.972

Tabla 5.27: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza (Long.) con respecto a los resultados del modelo ARIMA para las series largas simuladas con intervención de salto transitorio. También se incluye la cobertura media (Cobert.) y tiempo de ejecución (en segundos) promedio, así como la frecuencia relativa de identificación del efecto de intervención (Det.), el promedio de la magnitud del efecto de intervención incorporado con respecto al tamaño real del salto (Ef. Inc.) y las frecuencias relativas de correcta indetificación de los instantes de inicio y final del efecto de intervención (Ext.1 y Ext.2).

	sMAPE	**rMAPE**	**MSE**	**Long.**	**Cobert.**	**Tiempo**	**Det.**	**Ef. Inc.**	**Ext.1**
ARIMA	1.000	1.000	1.000	1.000	0.828	5.884	0.942	1.005	0.960
ETS	0.997	0.995	0.985	1.065	0.882	1.036	0.836	1.005	0.976
DLT MAP	0.930	0.940	0.871	0.916	0.834	0.158	0.962	0.953	0.983

Tabla 5.28: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza (Long.) con respecto a los resultados del modelo ARIMA para las series cortas simuladas con intervención de salto permanente. También se incluye la cobertura media (Cobert.) y tiempo de ejecución (en segundos) promedio, así como la frecuencia relativa de identificación del efecto de intervención (Det.), el promedio de la magnitud del efecto de intervención incorporado con respecto al tamaño real del salto (Ef. Inc.) y las frecuencias relativas de correcta indetificación del instante de inicio del efecto de intervención (Ext.1).

	sMAPE	rMAPE	MSE	Long.	Cobert.	Tiempo	Det.	Ef. Inc.	Ext.1
ARIMA	1.000	1.000	1.000	1.000	0.826	35.691	0.848	1.002	0.953
ETS	0.991	0.996	0.973	1.000	0.890	1.818	0.788	1.028	0.980
DLT MAP	0.963	0.973	0.921	0.884	0.802	3.481	0.936	0.975	0.985

Tabla 5.29: Tabla comparativa con el rendimiento relativo mediano de los distintos métodos en términos de SMAPE, RMAPE, MSE y longitud de los intervalos de confianza (Long.) con respecto a los resultados del modelo ARIMA para las series largas simuladas con intervención de salto permanente. También se incluye la cobertura media (Cobert.) y tiempo de ejecución (en segundos) promedio, así como la frecuencia relativa de identificación del efecto de intervención (Det.), el promedio de la magnitud del efecto de intervención incorporado con respecto al tamaño real del salto (Ef. Inc.) y las frecuencias relativas de correcta indetificación del instante de inicio del efecto de intervención (Ext.1).

En cuanto a las intervenciones de tipo permanente, en términos de errores de predicción solamente se encuentra una mejora por parte de DLT-MAP frente a los otros dos métodos en el caso de las series cortas. Las longitudes de los intervalos de predicción son casi idénticas para ARIMA y ETS y menores para DLT-MAP. La cobertura es insuficiente en todos los casos, aunque algo superior para ETS frente a los otros dos modelos. Destaca un mayor porcentaje de identificación del efecto de intervención para ARIMA y DLT-MAP que para ETS, y en el caso de las series largas, DLT-MAP también detecta con mayor acierto las intervenciones que ARIMA. Ahora bien, en caso de que los modelos detecten el efecto de intervención, los resultados de la proporción del salto que se incorpora, así como la fiabilidad con la que se detecta el inicio de la intervención es la misma en los tres casos. En relación al tiempo computacional, el modelo que más tarda es ARIMA, aunque ahora ETS requiere menos tiempo en el caso de las series largas, pero más en el caso de las series cortas, que el ajuste MAP de DLT.

5.2.7. Conclusiones de la segunda parte del estudio de simulación

Tal y como podíamos esperar, una vez realizado el estudio benchmark, podemos constatar de nuevo que, a grandes rasgos, el ajuste por cadenas de Markov de los modelos Orbit ofrece los mismos resultados, en términos de errores de predicción, que los ajustes basados en estimación máxima a posteriori. No obstante, de nuevo, los intervalos de predicción más fiables se obtienen mediante el ajuste MCMC. Asimismo, no hay aparentemente ningún motivo para emplear el método LGT, pudiendo quedarnos con DLT. En cualquier caso, estos modelos han obtenido unos resultados sorprendentemente buenos para series con componente estacional multiplicativa. Además, teniendo en cuenta que el tiempo necesario para el ajuste DLT-MAP en el caso de series con intervención es comparable al de ETS, y que en el caso de que la intervención sea de tipo permanente detecta, con mayor probabilidad, la presencia del efecto de la intervención, puede tener sentido decantarse por DLT en este contexto.

El modelo ETS multiplicativo, que no estaba presente en el estudio anterior, parece ofrecer cierta ventaja frente a su análogo aditivo, solamente cuando el tamaño de las series es pequeño. Además esta ventaja radica exclusivamente en términos del error de predicción, puesto que necesita de un mayor tiempo computacional (derivado de ampliar el conjunto de posibles modelos a ajustar para incorporar los modelos multiplicativos) y los intervalos de predicción tienden a tener una menor longitud que los de los modelos ETS aditivos, con una pérdida de cobertura. Cuando el tamaño de las series es elevado, desaparece esa mejora del error de predicción (y de hecho se obtienen peores resultados en el caso de componente estacional multiplicativa).

ARIMA ha sido el modelo que ha obtenido unos peores resultados, al hablar de error de predicción,

en el caso de series con componente estacional multiplicativa. No obstante, recordemos lo muy limitado que se encuentra este modelo en nuestro estudio, puesto que solamente se admite una diferenciación estacional y P y Q como máximo toman el valor 1. Con toda seguridad, una mayor flexibilidad a la hora de seleccionar el mejor modelo ARIMA daría lugar a un incremento en su rendimiento.

Por último, si nos encontramos ante series con un innovaciones de tipo multiplicativo, para las que una transformación como el logaritmo logra estabilizar la variabilidad de la serie, entonces, de acuerdo con los resultados de este estudio, no hay ningún modelo preferible para reducir el error de predicción y obtener intervalos de predicción con una cobertura adecuada. Puesto que el tiempo que tarda el modelo ETS aditivo es, en promedio, el menor de todos los modelos puestos a prueba, decantarse por este modelo parece la opción más razonable.

5.3. Series de tiempo con datos reales

Las tres series de tiempo con las que vamos a trabajar en esta sección se basan en datos anonimizados. La información disponible sobre el contexto relativo a la recogida de los datos es muy limitada. No obstante, conocemos la temática principal de las series: demanda de energía, consumo de agua y reserva de plazas para viajar en avión.

La finalidad del estudio comparativo sobre estas series no es un análisis exhaustivo de las mismas de acuerdo con cada una de las metodologías estudiadas a lo largo de este trabajo, sino que pretendemos comparar el rendimiento de los procedimientos de ajuste automático, para ser coherentes con el estudio de simulación previo. No obstante, habrá particularidades por las que nos tendremos que preocupar para poder obtener ajustes razonables. Por ejemplo, todas las series tienen datos faltantes y alguna de ellas deberá incorporar un posible efecto de intervención.

No hemos llegado a discutir el tratamiento de datos faltantes en una serie de tiempo y, hasta donde sabemos, no existe una metodología estándar al respecto. El criterio general que vamos a seguir, consistirá en imputar los datos faltantes a través de algún valor razonable y, simultáneamente, añadir como variable regresora a la función indicadora de aquel instante en el que ha tenido lugar la imputación. Posteriormente, dejaremos que un procedimiento de selección basado en el criterio BIC determine si dicha componente regresiva merece ser incluida. Notemos que, el hecho de necesitar un mecanismo de selección e incorporación de regresión a los modelos de serie de tiempo, condiciona que el único modelo Orbit aplicable sea el modelo DLT, ajustado mediante estimación máxima a posteriori.

Nos centraremos en la obtención de predicciones puntuales, por dos motivos. El primero, es que la empresa colaboradora ha sugerido que su principal interés radica en la predicción en media, y no en los intervalos de predicción. En segundo lugar, ya hemos constatado en el estudio de simulación anterior que los ajustes MAP no dan lugar a intervalos de predicción fiables, en términos de cobertura. Así pues, la comparación del rendimiento de los distintos métodos se basará exclusivamente las medidas MSE, SMAPE y RMAPE de error predictivo.

Con respecto a qué modelos vamos a emplear, evidentemente el modelo DLT-MAP es un candidato indiscutible. Por otro lado, ajustaremos modelos ARIMA y ETS aditivos, una vez transformadas las series para estabilizar la varianza, de ser necesario.

5.3.1. Demanda de energía

La primera de las tres series de datos reales que vamos a tratar es una serie relacionada con demanda energética, medida en kilovatios hora (kWh). En la Figura 5.5 aparece representado su gráfico secuencial.

Se trata de una serie de tiempo de longitud 1487, cuyas mediciones han sido obtenidas de forma horaria (24 medidas cada día). El objetivo en este caso sería tratar de obtener predicciones a 48 horas, de modo que reservaremos los últimos 48 valores de la serie para calcular medidas de error predictivo para las predicciones que obtendremos con el resto de los datos.

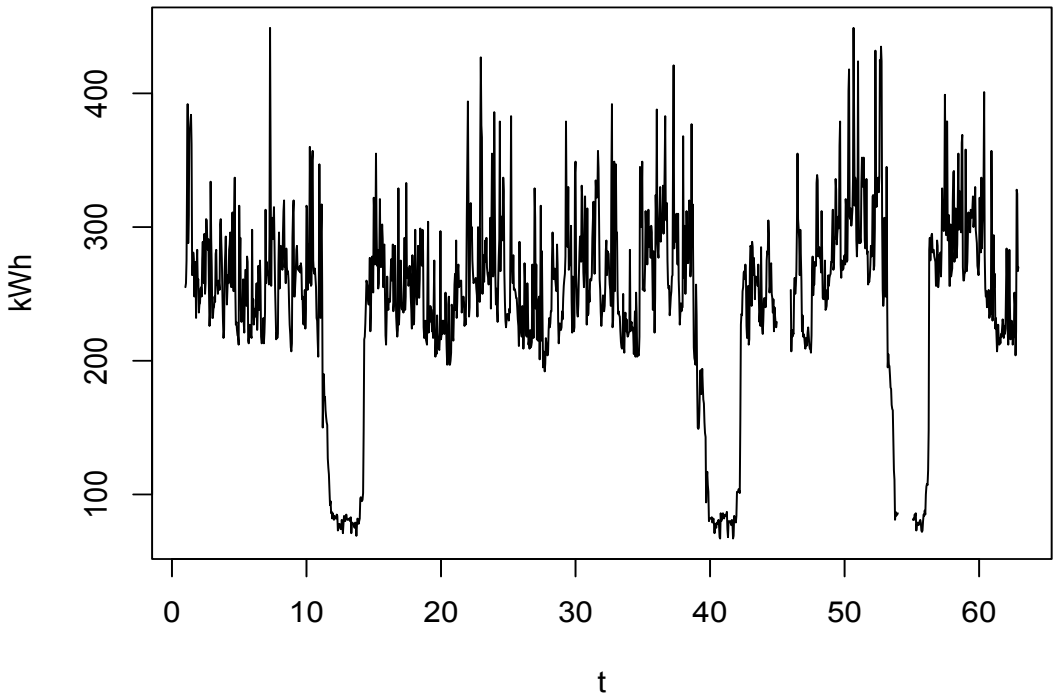

Figura 5.5: Gráfico secuencial de la serie de demanda de energía. El tiempo está expresado en días y la demanda en kilovatios hora.

A la vista del gráfico secuencial, hay un fenómeno que se repite de forma semi-regular consistente en una caída del nivel de demanda hasta unos 80 kWh. Al margen de estas caídas, la demanda parece oscilar en un rango entre 200 y 400 kWh. El motivo por el que nos referimos como semi-regulares a las caídas de demanda se debe a que ni tienen una duración consistente, ni ocurren dejando un mismo lapso de tiempo entre ellas. Según la información de la que disponemos, se asocian con cierres parciales de la industria productora de energía, por lo que la demanda máxima que pueden satisfacer se ve severamente limitadas. Si no tratásemos de corregir de algún modo estas caídas, podría ocurrir que algún modelo tenga dificultades a la hora de ofrecer un buen ajuste. Por tanto, y teniendo en cuenta que los cierres de una planta de producción se pueden considerar como una intervención externa de duración conocida, trataremos de modelar este fenómeno como una intervención. En particular, como una intervención de salto, al ser el único tipo de estos fenómenos que hemos podemos incorporar al modelo DLT-MAP.

Esta serie posee un total de 48 datos faltantes. Por ejemplo, se puede apreciar como algunos de ellos se corresponden con valores de demanda de energía para la tercera caída, entre los días 50 y 60. Además, la variabilidad de la serie es claramente mayor cuanto mayor es el nivel de la misma, destacando la reducción de esta variabilidad durante los periodos de cierre parcial de la industria. Para estabilizar esta varianza (algo necesario para emplear modelos de tipo aditivo que asumen que las innovaciones del modelo son homocedásticas) emplearemos la siguiente transformación: restaremos 40 kWh a todas las observaciones de la serie y aplicaremos la función logaritmo. El motivo de trasladar la serie es que, de otro modo, el efecto del logaritmo es prácticamente imperceptible, al variar la serie original entre 100 y 400 kWh. La Figura 5.6 contiene el gráfico secuencial de la transformada, que aparentemente no presenta diferencias de variabilidad asociadas al nivel.

En la Figura 5.7 aparece representada, más detalladamente, la primera de las tres caídas visibles

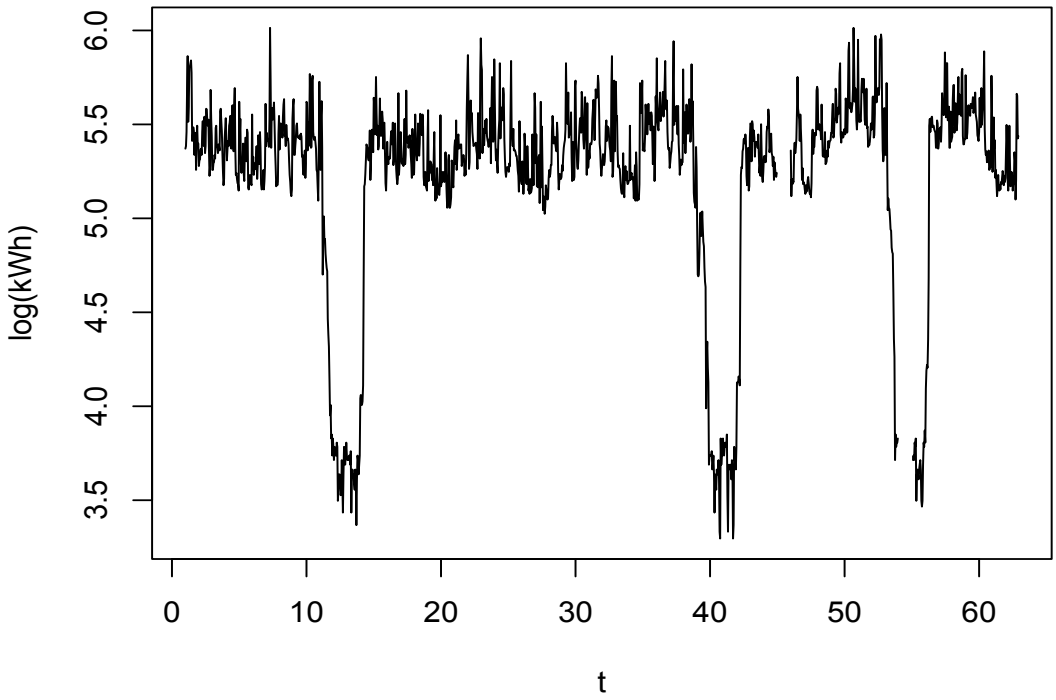

Figura 5.6: Gráfico secuencial de la serie de demanda de energía transformada. El tiempo está expresado en días y la demanda en log-kilovatios hora.

en la Figura 5.5. A la vista de ese gráfico, creemos conveniente considerar una estructura de 3 saltos consecutivos, delimitados por las líneas verticales con colores rojo, azul y verde. En primer lugar, se produce una caída en torno al inicio del día 11 respecto al nivel usual de la serie, seguida de otra caída próxima al mediodía de ese día 11 y poco después, la demanda volvería a caer otra vez más. Hacia el final del día 13 la demanda comienza a subir, terminando así el efecto de la intervención de salto de menor duración (delimitada por líneas verticales verdes). Durante el día 14, tendrían lugar otros dos saltos casi consecutivos, terminandose el efecto de las otras dos intervenciones de salto (delimitadas por las líneas verticales azules y rojas), hasta recuperar el nivel original.

Podría ser cuestionable la necesidad de incluir tres saltos, o quizá sería conveniente añadir más, si uno estuviese especialmente interesado en captar el efecto de la intervención. Con respecto a lo primero, el procedimiento de selección en base al criterio BIC se encargará de dilucidar si alguno de los saltos es redundante al considerar los otros dos. Para la segundo, basta con decir que la inclusión de esta intervención se debe meramente a una cuestión utilitaria, con el objetivo de reducir posibles problemas a la hora de ajustar modelos (como una incorrecta identificación del orden de diferenciación de un modelo ARIMA), de modo que nos bastará con corregir, de un modo más o menos certero, el efecto de estas caídas. Para las otras dos intervenciones, se considerará una estructura de tres saltos similar, y se asumirá que la magnitud de los saltos es la misma en las tres caídas.

Falta por aclarar como imputaremos los datos faltantes. Puesto que no se trata de datos faltantes aislados, sino de dos días de datos faltantes optaremos por tomar la media entre los valores en el día anterior y en el día posterior. No parece razonable emplear un valor medio general teniendo en cuenta la existencia de las caídas de demanda y tampoco queremos favorecer ningún modelo en particular imputando los valores faltantes por predicciones de un modelo ajustado sobre los valores previos de la serie, de modo que nuestra propuesta parece un punto medio aceptable, teniendo en cuenta que se trata de mediciones horarias y va a existir una estructura estacional de periodo 24.

Una vez transformada la serie, imputados los datos faltantes y establecidas las variables regresoras

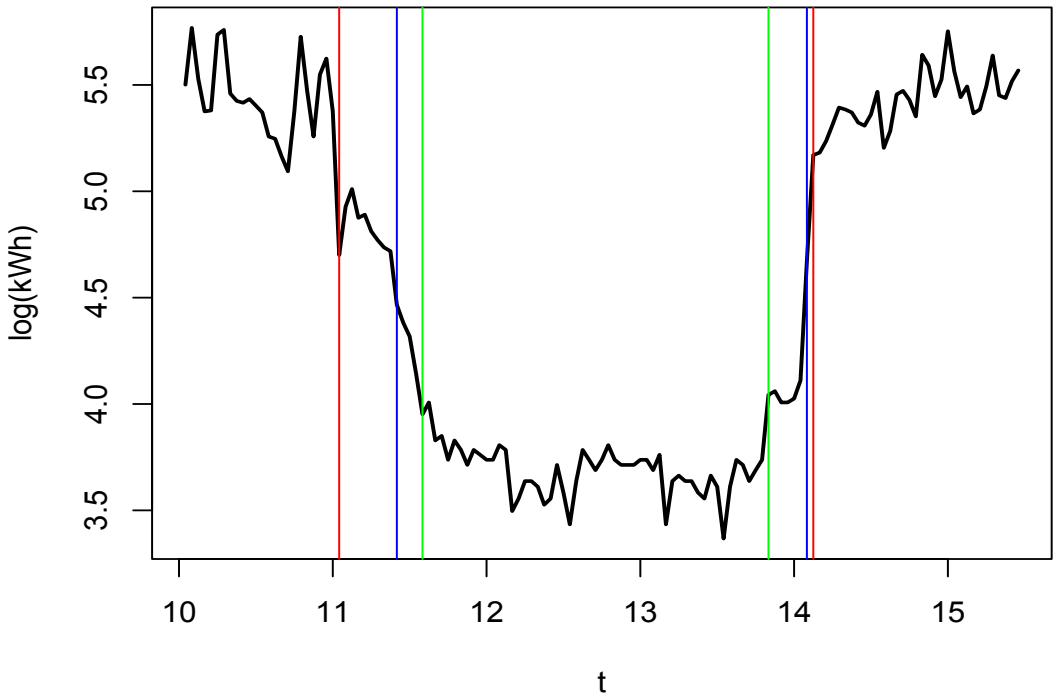

Figura 5.7: Estructura en tres saltos consecutivos de la primera caída de demanda. Los límites de los efectos de la intervención en salto están marcados por las líneas verticales de un mismo color: rojo, azul y verde.

asociadas tanto a las intervenciones de salto como a las imputaciones, hemos procedido al ajuste automático de los modelos ARIMA, ETS y DLT-MAP. En el caso de los dos primeros modelos, se ha empleado el criterio BIC para la selección automática, haciendo uso de las funciones `es` y `auto.ssarima` de la librería `smooth` de R, con los argumentos `initial = "backcasting"` y `bounds ="admisible"` (de igual modo que con las series largas del estudio de simulación). Para todos los modelos, se ha llevado a cabo otro procedimiento de selección, también basado en el criterio BIC, con el objetivo de determinar si era conveniente eliminar alguna variable regresora. En el caso del modelo DLT-MAP este procedimiento se ha llevado a cabo manualmente, partiendo de un modelo con todas las componentes de regresión y eliminando en cada paso aquella que diese lugar a una mayor reducción del valor del BIC (en caso de que eliminar alguna suponga una reducción de ese valor). Para los modelos ARIMA y ETS, esta selección se realiza de manera simultánea con el proceso de selección del propio modelo dentro de la familia correspondiente, mediante el argumento `xregDo = "select"` de las funciones `es` y `auto.ssarima`.

El modelo ARIMA finalmente elegido es un ARIMA$(4,0,2)\times(1,1,4)_{24}$ sin constante, mientras que el modelo ETS seleccionado ha sido un ETS(A,N,A). Para ARIMA y DLT-MAP se han mantenido algunas (no todas) de las variables regresoras. En particular se ha conservado en el modelo la estructura de intervención en tres saltos consecutivos. Para el modelo ETS, en cambio, el proceso de selección ha terminado por eliminar toda componente regresiva, incluyendo las intervenciones de salto. En la Figura 5.8 aparece representada la parte final de la serie de demanda de energía original junto a las predicciones en media para los tres modelos ajustados. Como se puede apreciar en ese gráfico, el modelo que ofrece unas predicciones más alejadas del verdadero comportamiento de la serie de demanda es el modelo ETS (predicciones en color verde) mientras que las predicciones para los modelos ARIMA y DLT-MAP son similares (colores rojo y azul, respectivamente).

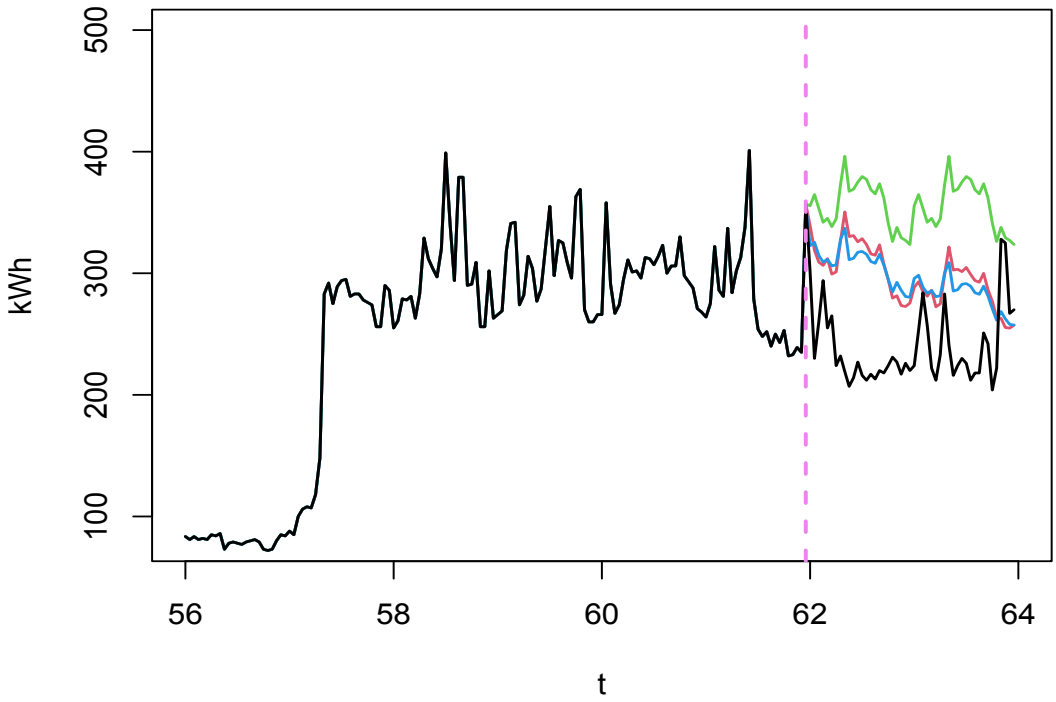

Figura 5.8: Serie real de demanda de energía (negro) junto a las predicciones a horizonte 48, a partir de la línea discontinua color morado, para los modelos ARIMA (rojo), ETS (verde) y DLT-MAP (azul) ajustados.

En la Tabla 5.30 aparecen recogidos los valores para el SMAPE, RMAPE y MSE asociados a las predicciones de los tres modelos ajustados. El modelo que comete un menor error, de acuerdo con los tres criterios es DLT-MAP, seguido de cerca por ARIMA. El modelo ETS comete bastante más error que los otros dos, casi el doble en términos de SMAPE o RMAPE y en torno al triple en términos de MSE.

	SMAPE	**RMAPE**	**MSE**
ARIMA	12.449	28.997	5361.083
ETS	19.971	53.142	15413.74
DLT-MAP	11.963	27.950	4742.824

Tabla 5.30: SMAPE, RMAPE y MSE para las predicciones obtenidas con los ajustes ARIMA, ETS y DLT-MAP para la serie de demanda de energía.

5.3.2. Consumo de agua

La segunda serie de datos reales que emplearemos para comparar el rendimiento de los modelos ARIMA, ETS y DLT-MAP se relaciona con el consumo de agua medido por un contador, en metros cúbicos por segundo (m^3/s), cada 15 minutos. Se trata, por tanto, de una serie de tiempo cuarto-horaria, y en la Figura aparece representado su gráfico secuencial.

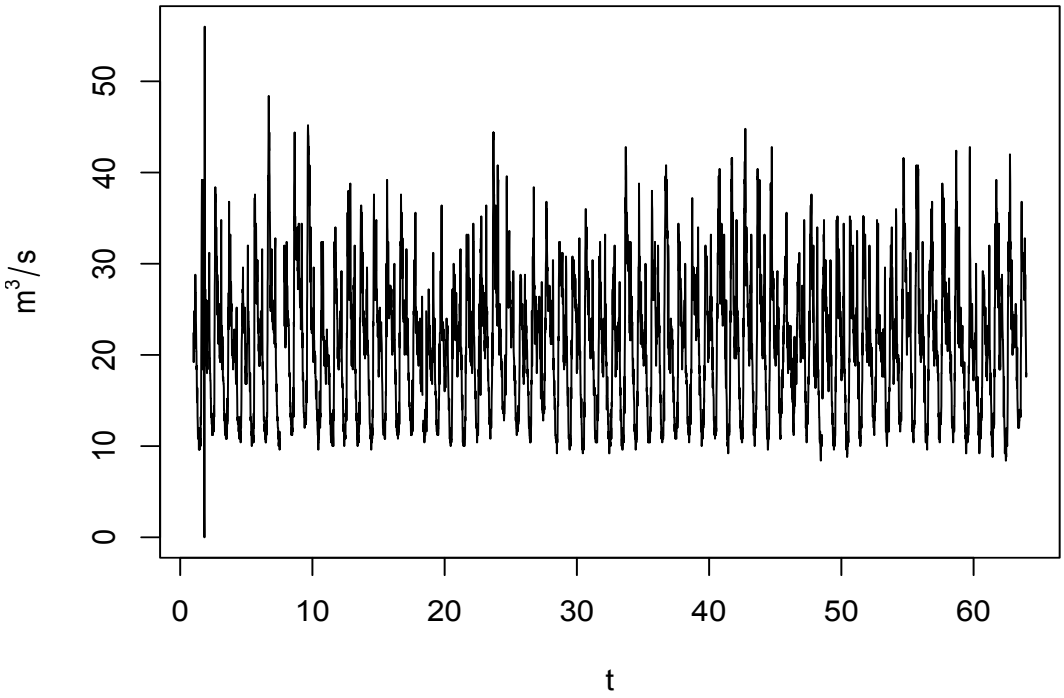

Figura 5.9: Gráfico secuencial de la serie de consumo de agua. El tiempo está expresado en días y el consumo en métros cúbicos por segundo.

Esta serie de consumo de agua tiene longitud 6048, correspondiente a sesenta y tres días de mediciones cuarto-horarias. El interés en este caso radica en ofrecer predicciones a una semana vista o, equivalentemente, el horizonte de predicción es $m = 672$ cuartos de hora. Guardaremos por tanto las últimas 672 observaciones y procederemos con el análisis asociado a las otras 5376. Cabe destacar que trabajar con series tan largas para predecir a un horizonte tan elevado no suele ser recomendable, puesto que las series de datos reales no suelen mantener una estructura constante en el tiempo, si no que hay ligeras variaciones que, si bien permiten que al ajustar modelos sobre series de tamaño moderado se obtengan ajustes razonablemente buenos, pueden estropear por completo el rendimiento de los modelos en otro caso. No obstante, esta serie de tiempo es bastante regular, por lo que no debería ser demasiado problemático una predicción a tanto tiempo vista.

En el gráfico secuencial de la Figura 5.9 podemos comprobar como, usualmente, la serie oscila entre 10 y 40 m^3/s. No obstante, al inicio hay una observación que supera los 50 metros cúbicos por segundo y otras cuatro que, o bien toman el valor 0, o bien están muy próximas a 0. De acuerdo con la información disponible, se trata de errores puntuales de medición, de modo que serán tratados como valores atípicos. Por otro lado, hay hasta 40 observaciones faltantes que deberemos imputar y, en este caso, optaremos por un promedio entre los valores correspondientes a la misma hora del día anterior y del día siguiente (no hay dos datos faltantes a distancia 96 entre sí), respetando una componente

estacional que se deja entrever en el gráfico secuencial. Finalmente, no resulta demasiado evidente que la variabilidad de la serie cambie con el nivel, por lo que no emplearemos una transformación logarítmica en este caso, en principio.

Del mismo modo que con la serie de demanda energética, hemos empleado procedimientos de ajuste automático y selección de variables regresoras (asociadas a los datos atípicos y faltantes) basados en la minimización del criterio BIC. Las funciones y argumentos empleados para ello son completamente análogos a los usados en la Sección 5.3.1.

El mejor ajuste ARIMA resulta ser un ARIMA$(3,0,3)\times(3,1,0)_{96}$ con constante y el mejor ajuste ETS es un ETS(A,N,A), en ambos casos manteniendo algunas de las variables regresoras, algo que también ocurre para el ajuste DLT-MAP. En la Figura 5.10 se representan cuatro gráficos consistentes en la parte final de la serie real de consumo de agua y las predicciones a horizonte 672 a partir del día 56 para los ajustes ARIMA, ETS y DLT-MAP previamente mencionados, además de un gráfico adicional análogo (en la esquina inferior derecha), en el que las predicciones han sido construidas transformando previamente la serie desplazada por una unidad (para que no haya observaciones con valor 0) mediante un logaritmo, y deshaciendo la tranformación una vez obtenidas las predicciones sobre la nueva serie.

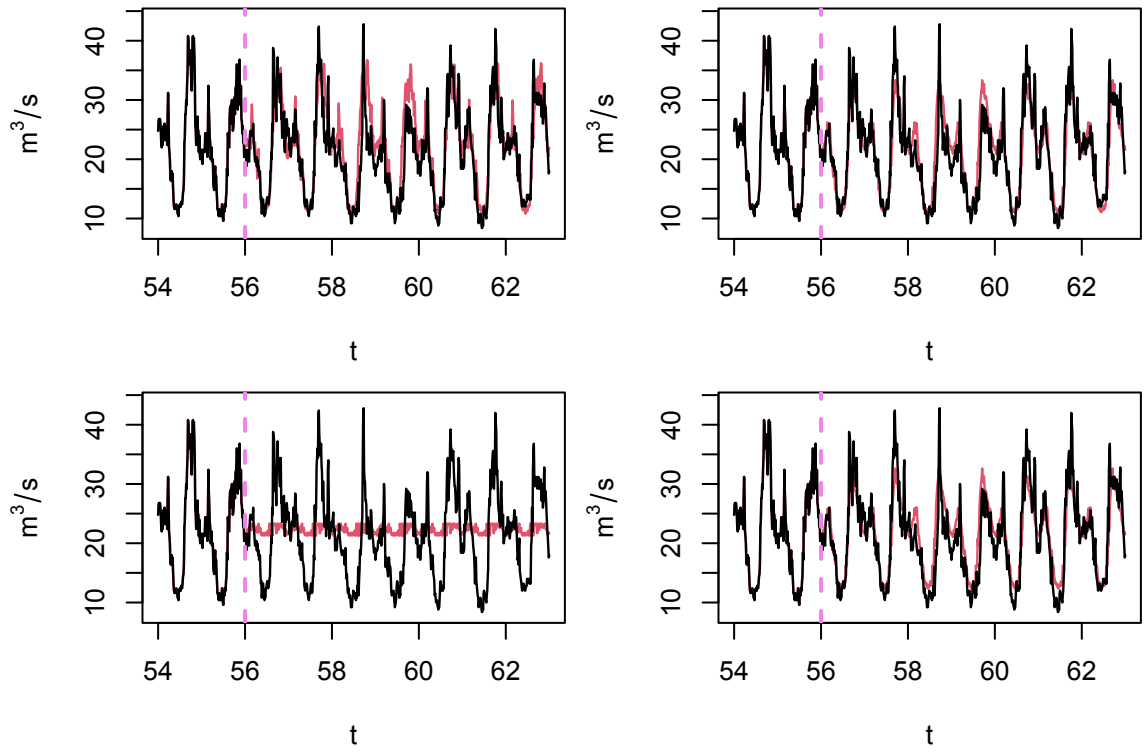

Figura 5.10: Serie real de consumo de agua (negro) junto a las predicciones a horizonte 672 (rojo), a partir de la línea discontinua color morado, para el ajuste ARIMA (arriba izquierda), ETS (arriba derecha), DLT-MAP (abajo izquierda) y DLT-MAP con transformación logarítmica previa (abajo derecha).

El motivo por el que se incluye esa cuarta opción radica en el gráfico de la esquina inferior izquierda, correspondiente al ajuste DLT-MAP sobre la serie original, en el que se ve como claramente el modelo no está obteniendo predicciones razonables, al no ser capaz de capturar la magnitud de la componente estacional. Si tuviésemos que aventurar por qué ocurre esto, podríamos decir que probablemente se deba

a que los valores iniciales que se proporcionan a los algoritmos de optimización para esta componente estacional se sortean de una distribución normal de media 0 y desviación típica 0.05, tal y como se menciona en la Sección 4.2, independientemente de la escala real de la serie. A la hora de obtener los valores de los parámetros y estados iniciales que maximicen la distribución a posteriori, es posible que se produzca una convergencia hacia un máximo local para el que los primeros valores de la componente estacional no disten mucho de los que se emplean en la inicialización, con una escala mucho menor de la que deberían tener. En cualquier caso, esto pone de manifiesto un problema con el procedimiento de estimación asociado a este modelo, con el que se debe tener cuidado.

Resulta difícil comparar las predicciones de unos y otros ajustes en base a los gráficos de la Figura 5.10, pero parece evidente que las predicciones del modelo DLT-MAP ajustado sobre la serie transformada por el logaritmo son similares a las predicciones del modelo ETS, y ligeramente distintas a las del modelo ARIMA. Este último tiene una amplitud asociada a la componente estacional mayor que los otros dos casos.

Se recogen en la Tabla 5.31 los valores para el SMAPE, RMAPE y MSE asociados a las predicciones de los cuatro modelos ajustados (incluyendo el desastroso ajuste DLT-MAP sobre la serie original). El mejor modelo, en base a los criterios de error predictivo, es ETS. ARIMA y el ajuste DLT-MAP sobre la serie transformada obtienen resultados similares en términos de SMAPE y RMAPE, aunque ARIMA tiene un MSE ligeramente mayor DLT-MAP. El ajuste DLT-MAP sobre la serie original, como era de esperar, es comparativamente mucho peor que los otros tres.

	SMAPE	**RMAPE**	**MSE**
ARIMA	6.248	13.174	15.194
ETS	5.624	11.383	11.045
DLT-MAP	14.046	33.815	49.664
DLT-MAP-log	6.309	13.153	12.072

Tabla 5.31: SMAPE, RMAPE y MSE para las predicciones obtenidas con los ajustes ARIMA, ETS y DLT-MAP para la serie de consumo de agua, además de las predicciones obtenidas con el ajuste DLT-MAP sobre la serie de consumo transformada.

5.3.3. Reserva de vuelos

La última de las tres series de tiempo de datos reales que discutiremos en este trabajo es una serie que recoge el número total de reservas de billetes de avión para una determinada aerolínea. En la Figura 5.11 se representa el gráfico secuencial para esta serie de tiempo.

La serie está compuesta por 729 observaciones que se corresponden con las reservas de vuelos registradas a lo largo de 729 días. Se pretende poder ofrecer predicciones con un horizonte máximo de predicción de 31 días, de modo que trabajaremos con los datos de reservas de vuelos de los primeros 698 días, dejando los últimos 31 para construir medidas de error de predicción.

Resulta evidente, a la vista del gráfico secuencial, que hay ciertos periodos en los que la variabilidad es menor que en otros, correspondiendo con aquellos días en los que las reservas de vuelos fueron menores. Hemos empleado la transformación logaritmo para tratar de solventar este problema. En la Figura 5.12 se representa el gráfico secuencial de la serie transformada, revelándonos que la heterocedasticidad

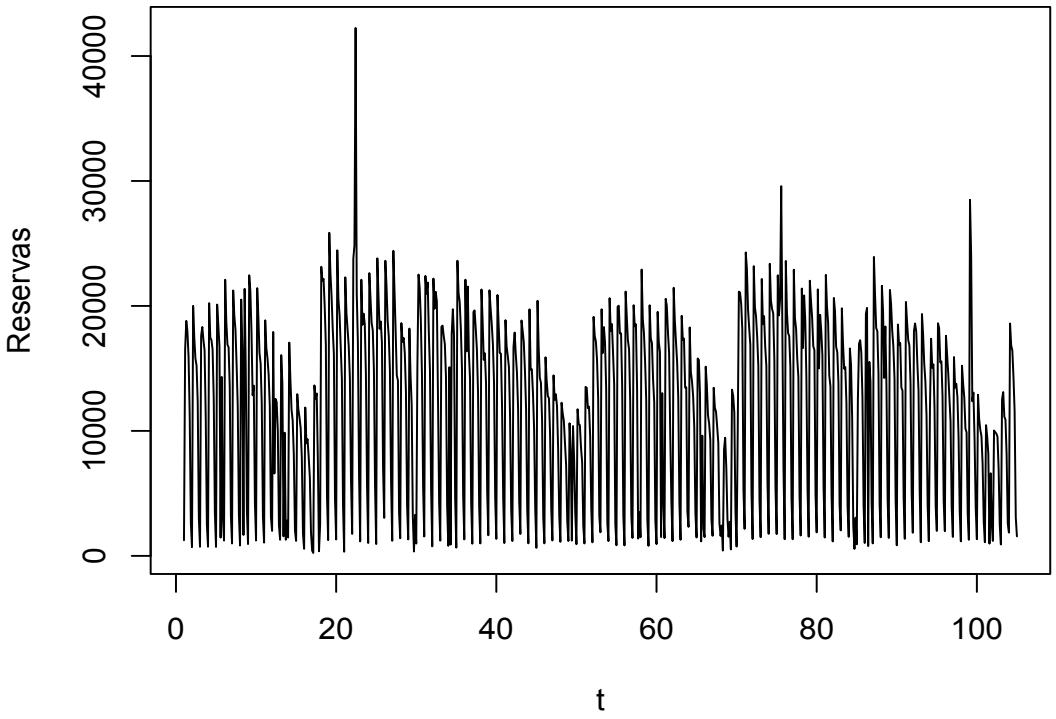

Figura 5.11: Gráfico secuencial de la serie de reserva de vuelos. El tiempo está expresado en semanas.

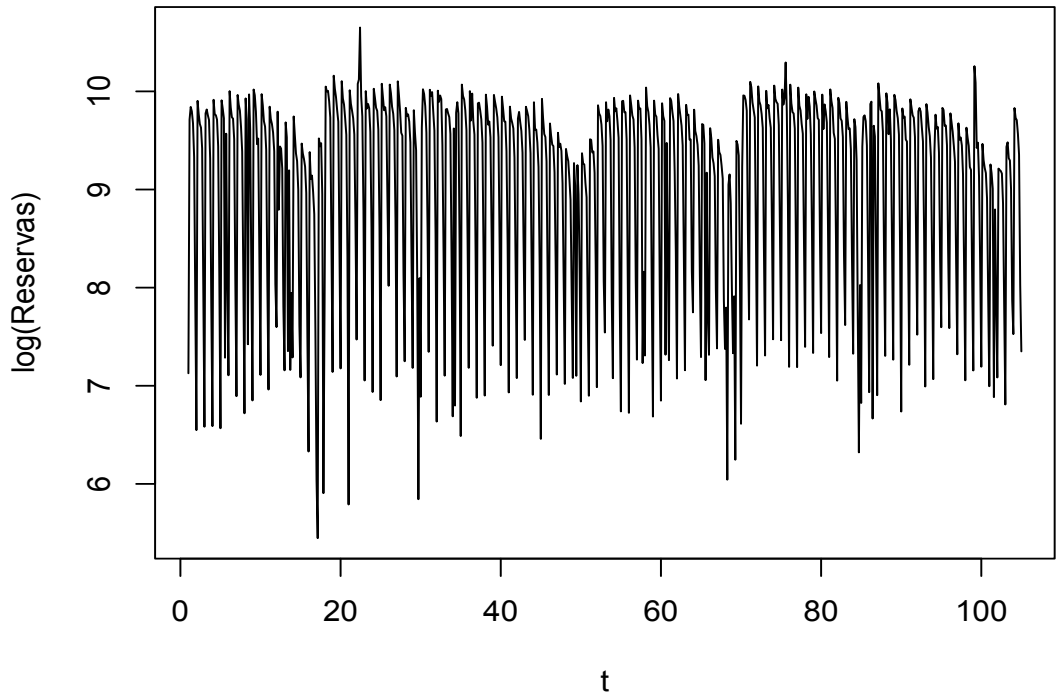

Figura 5.12: Gráfico secuencial de la serie del logaritmo del número de reservas de vuelos. El tiempo está expresado en semanas.

de la serie se reduce considerablemente. Por otro lado, tanto en la serie original como en la serie transformada, se vislumbra una componente estacional de periodo semanal.

Fijémonos en que durante la vigesimoprimera semana se ha reservado un número total de vuelos anormalmente elevado, superándose las 40000 reservas en un mismo día, cuando el total de reservas suele oscilar entorno a las 10000. Este valor que podría considerarse como atípico resulta mucho menos destacable al trabajar con logaritmos puesto que, de hecho, sobresalen en mayor medida aquellos días que han tenido un número menor de reservas (en unidades logarítmicas), por la parte inferior del gráfico secuencial. Por tanto, y teniendo en cuenta el elevado número de días con un número "destacable"de reservas, no parece que se trate de un problema de datos atípicos, sino más bien de que la distribución asociada a las innovaciones para esta serie de tiempo tiene colas pesadas. No obstante, la única alternativa disponible para ajustar los modelos ARIMA y ETS pasa por asumir una distribución gaussiana para las innovaciones. Para el modelo DLT, esto no debería suponer un problema, al asumir que la distribución de las innovaciones es una T de Student generalizada.

En este caso, no hay un problema de datos faltantes. De todos modos, para ser coherentes con la comparación llevada a cabo en las Secciones 5.3.1 y 5.3.2, nos limitaremos a ajustar los modelos ARIMA, ETS y DLT sobre la serie de logaritmos de las reservas de vuelos. Recordemos que de las conclusiones del estudio de simulación se deducía que, a grandes rasgos, los resultados para los modelos DLT y LGT eran similares entre sí. También son parecidos los resultados para los ajustes por cadenas de Markov y mediante la estimación máxima a posteriori. Como el foco está puesto sobre el error de predicción puntual, quedarnos con el ajuste MAP para el modelo DLT no supone una gran pérdida de información. El ajuste y la selección de modelos dentro de las familias ARIMA y ETS se llevará a cabo de manera análoga a lo expuesto en las Secciones 5.3.1 y 5.3.2, basándonos en el criterio BIC.

Los modelos ARIMA y ETS finalmente ajustados han sido un $\text{ARIMA}(2,0,1)\times(0,1,2)_7$ sin constante y un $\text{ETS}(A,N,A)$ respectivamente. De acuerdo con el gráfico de la Figura 5.13, en el que se representa la parte final de la serie de reservas de vuelo original junto a las predicciones para los tres modelos ajustados, las predicciones para los modelos ARIMA y DLT-MAP (en colores rojo y azul, respectivamente) son similares. En ambos casos las predicciones se quedan por encima de los verdaderos valores de la serie de tiempo original, exceptuando el último ciclo semanal, en el que prácticamente se superponen las predicciones con la realidad. Las predicciones ligadas al ajuste ETS, en cambio, presentan una componente estacional de tamaño más moderado, siendo más próximas a la realidad.

Del mismo modo que en las Secciones 5.3.1 y 5.3.2, hemos recopilado en la Tabla 5.32 los errores de predicción medidos según los criterios SMAPE, RMAPE y MSE. Como era de esperar, en base a lo observado en la Figura 5.13, los criterios de error toman valores parecidos para los ajustes ARIMA y DLT-MAP, que a su vez son mayores que los valores de los criterios de error para el mejor modelo en este caso; esto es, el modelo ETS.

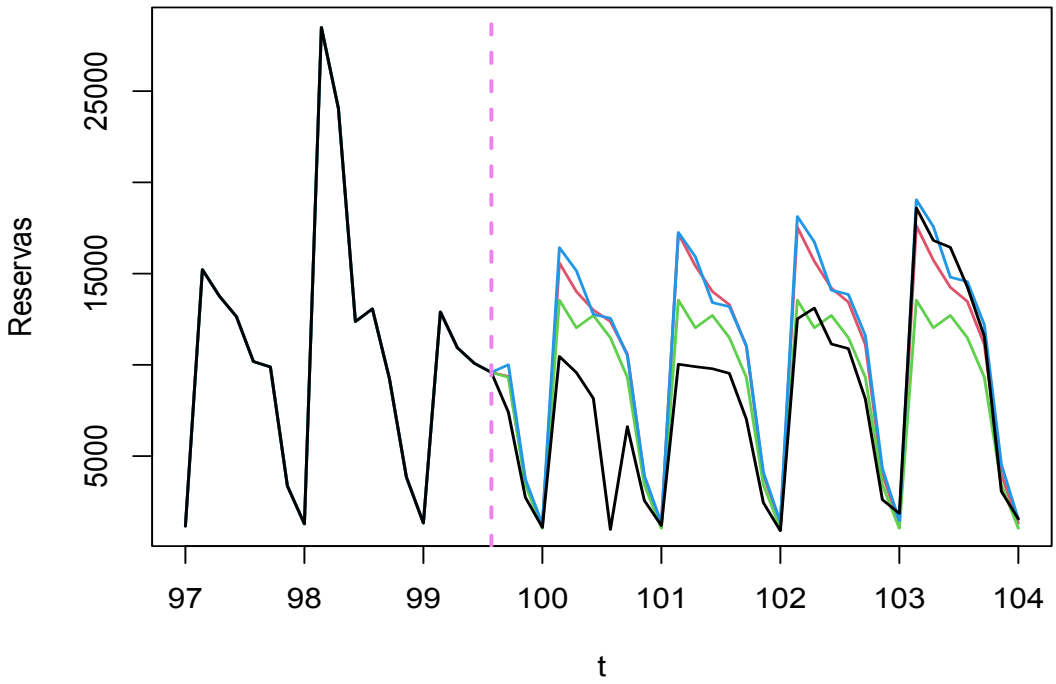

Figura 5.13: Serie real de reservas de vuelos (negro) junto a las predicciones a horizonte 31, a partir de la línea discontinua color morado, para los modelos ARIMA (rojo), ETS (verde) y DLT-MAP (azul) ajustados.

	SMAPE	**RMAPE**	**MSE**
ARIMA	16.479	34.984	13581861
ETS	14.044	24.683	8957510
DLT-MAP	17.091	38.0875	14983398

Tabla 5.32: SMAPE, RMAPE y MSE para las predicciones obtenidas con los ajustes ARIMA, ETS y DLT-MAP para la serie de reservas de vuelos.

Capítulo 6

Conclusiones y posibles extensiones

A la vista de los resultados de los Capítulos 4 y 5, es difícil de justificar el uso de los modelos Orbit frente a los modelos Box-Jenkins o de suavización exponencial.

Por un lado, se echa en falta una mayor transparencia a la hora de establecer cómo están realmente implementados los modelos DLT y LGT en el paquete `Orbit-ml`, puesto que el enfoque de estimación no parece ser ni puramente frecuentista ni completamente bayesiano, y sería pertinente saber qué coeficientes están siendo tratados como parámetros y cuales serían variables aleatorias con sus respectivas distribuciones a priori, algo que ya hemos comentado en las Secciones 4.1 y 4.2. Además, el modelo más completo, en el sentido de permitir incorporar una componente regresiva (necesaria para el tratamiento de efectos de intervención, entre otras cuestiones), es el que peores propiedades presenta, como su falta de estabilidad o la existencia de redundancias en la formulación del modelo que, de ser ignoradas, llevarían a problemas a la hora de estimar valores para los parámetros que lo caracterizan. Cabe destacar, sin embargo, que en la práctica los modelos consiguen obtener ajustes, lo que sugiere que los autores son conscientes del problema de identificabilidad, y habrán impuesto restricciones al conjunto de valores que podrán tomar los parámetros, en consecuencia. De todos modos, el pequeño análisis de sensibilidad de la Sección 4.3 sugiere que, en ausencia de un mecanismo adecuado de determinación de un valor para el parámetro de amortiguamiento, ν, el tipo de tendencia global, $D(t)$, o la presencia de componente estacional, se trata de un modelo sobreparametrizado y con una estructura superflua, puesto que, en general, se obtiene el mismo rendimiento para cualesquiera valores que se les dé.

De acuerdo con el estudio de simulación llevado a cabo en el Capítulo 5, generalmente los modelos Orbit han obtenido un rendimiento similar o inferior a un ajuste basado en suavización exponencial. Ha habido, sin embargo, dos excepciones. La primera de ellas ha sido el caso de las series simuladas con componente estacional de tipo multiplicativo, para las que los modelos Orbit han logrado tanto predicciones puntuales como intervalos de predicción, comparativamente buenos. La segunda, consiste en las series con una intervención de salto de tipo permanente. En este escenario, el rendimiento predictivo de los modelos ARIMA, ETS y DLT-MAP fue similar, pero el último de ellos obtuvo unas tasas de correcta identificación del efecto de intervención, así como de su instante inicial muy buenas, necesitando para ello un tiempo computacional bastante inferior al que necesitaría un ajuste ARIMA, por ejemplo. No obstante, esto solamente será cierto al restringirnos a un ajuste basado en la estimación máxima a posteriori, que da lugar a intervalos de predicción con una cobertura inferior a la esperada, mientras que la estimación por cadenas de Markov Monte Carlo requiere de tiempos computacionales mucho más elevados (comparables a un ajuste ARIMA sin ningún tipo de restricción a sus órdenes).

De las tres series de tiempo reales empleadas para comparar los modelos ARIMA, ETS y DLT-MAP, no se deduce un claro ganador en términos del error de predicción. No obstante, los resultados han sido parecidos entre los ajustes ARIMA y DLT-MAP, en los tres casos.

Una vez resumidos las conclusiones que se pueden extraer de este trabajo, cabe destacar las limitaciones del mismo, así como posibles continuaciones con el objetivo de profundizar en la comparativa. Para empezar, todos los resultados relativos al Capítulo 5 han sido obtenidos con una computadora

personal, lo que ha motivado la necesidad de imponer severas restricciones a la hora de ajustar, sobre todo, los modelos ARIMA. Sería pertinente repetir el estudio permitiendo una mayor libertad en los órdenes que pueden llegar a tomar estos modelos. Por otro lado, para ser coherentes con la metodología de los modelos de espacio de estados, hemos recurrido al paquete `smooth` de R, pero la implementación de los modelos ARIMA y ETS en dicho paquete puede no haber sido la más eficiente, ni la más precisa en la obtención de estimaciones. Por tanto, quizá cabría extender el estudio para incluir los ajustes ARIMA o ETS que proporcionan otros paquetes de R o Python.

Hemos obviado completamente los procedimientos de verificación de las hipótesis subyacentes a los modelos empleados. En ningún caso hemos obtenido gráficos de bondad de ajuste o contrastes formales de hipótesis de media cero e incorrelación para los residuos de los modelos. Desde un punto de vista formal, sería relevante, por ejemplo, estudiar si los modelos revisados a lo largo de este trabajo son capaces, con los ajustes obtenidos por procedimientos de selección automáticos basados en criterios de información, de recoger adecuadamente la estructura de dependencia de los datos, consistentemente. De lo contrario, no sería recomendable fiarse de este tipo de ajustes.

Tampoco sabemos, a ciencia cierta, si el criterio BIC es una herramienta adecuada para contrastar la significación de coeficientes en este tipo de modelos, a pesar de haber tenido que recurrir a esta idea al no existir otra alternativa implementada en los paquetes utilizados. Un estudio adicional en este sentido podría resultar de utilidad.

Bibliografía

[1] *Backtest*. Orbit 1.1.4 dev documentation. (s.f.). Recuperado el 26 de enero de 2023 de https://orbit-ml.readthedocs.io/en/latest/tutorials/backtest.html

[2] Basawa, I. V. y Rao, P. (1980). *Statistical Inference for Stochastic Processes*. Academic Press.

[3] Box, G. E. P., Jenkins, G. M., Reinsel, G. C. y Ljung, G. M. (2015). *Time Series Analysis: Forecasting y Control (5th ed.)*. Wiley.

[4] Brockwell, P. J. y Davis, R. A. (2016). *Introduction to Time Series and Forecasting (3rd ed.)*. Springer.

[5] Brooks, S., Gelman, A., Jones, G., y Meng, X. L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

[6] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., y Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1-32.

[7] *Damped Local Trend (DLT)*. Orbit 1.1.4 dev documentation. (s.f.). Recuperado el 26 de enero de 2023 de https://orbit-ml.readthedocs.io/en/latest/tutorials/dlt.html

[8] Forbes, C. S., Snyder, R. D. y Shami, R. G. (2000) *Bayesian Exponential Smoothing*. Department of Econometrics and Business Statistics Working Paper, Monash University, Australia.

[9] Guerrero, V.M. (1993) Time-series analysis supported by power transformations. *Journal of Forecasting*, 12(1), 37-48.

[10] Hyndman, R., Koehler, A. B., Ord, J. K. y Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach*. Springer.

[11] Kass R. E. y Raftery A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773-795.

[12] *Methods of Estimations and Predictions*. Orbit 1.1.4 dev documentation. (s.f.). Recuperado el 26 de enero de 2023 de https://orbit-ml.readthedocs.io/en/latest/tutorials/model_estimations_predictions.html

[13] Ng, E., Wang, Z., Chen, H., Yang, S., y Smyl, S. (2020). *Orbit: Probabilistic Forecast with Exponential Smoothing*. arXiv: Computation.

[14] Nocedal, Jorge, y Stephen J. Wright. (2006.) *Numerical Optimization (2da ed.)*. Springer.

[15] R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[16] Said, S.E. y Dickey, D. A. (1984) Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71, 599-607.

[17] Svetunkov, I. y Boylan, J. E. (2019). State-space ARIMA for supply-chain forecasting. *International Journal of Production Research*, 58, 1-10.

[18] Svetunkov, I. (2022). *smooth: Forecasting Using State Space Models*. R package version 3.1.6. https://CRAN.R-project.org/package=smooth

[19] Van Rossum, G., y Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.