Traballo Fin de Grao

Comparativa entre Metodoloxías Clásicas e de Aprendizaxe Automática na Análise de Series Temporais

Antón Figueroa Martínez

Xuño, 2025

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Comparativa entre metodoloxías clásicas e aprendizaxe automática na análise de series temporais
Breve descrición do contido
A metodoloxía clásica de análise de series temporais, amplamente empregada en diversas áreas, como economía, meteoroloxía ou medicina, baséase en modelos estatísticos como ARIMA (Modelo Autorregresivo Integrado de Media Móbil) ou modelos de suavizado exponencial. Estes modelos contan cunha base teórica sólida e poden ser interpretados facilmente. Pola súa banda, os métodos de aprendizaxe automática, como redes neuronais, máquinas de vectores de soporte (SVM) ou árbores de decisión, comezaron a gañar popularidade para a predición de series temporais nos últimos anos. Estes métodos ofrecen vantaxes potenciais en termos de capacidade de capturar patróns complexos e non lineares nos datos, adaptabilidade a diferentes tipos de series temporais e capacidade de xerar predicións precisas en contextos dinámicos e cambiantes. O obxectivo deste TFG é facer unha revisión da metodoloxía clásica de series temporais e levar a cabo unha comparativa práctica cos métodos de aprendizaxe automática, analizando as vantaxes e desvantaxes de ambos enfoques.

Índice

Índice de figuras

Resumo

Ao considerar un conxunto de datos, podemos atoparnos con observacións independentes ou con observacións que presenten algún tipo de dependencia espacial ou temporal, como é o caso das series temporais. Ao ter en conta esta dependencia, xorde naturalmente a teoría estatística da análise de series temporais, na que nos adentramos nas seguintes páxinas. O obxectivo deste traballo é a descrición e comparación dos diferentes modelos e metodoloxías de análise de series temporais. Partindo desta base, comparáronse en canto a rendemento, sinxeleza, interpretabilidade e eficiencia computacional, chegando á conclusión de que os modelos máis axeitados varían en cada caso.

Abstract

When taking into consideration a set of data, one can find independent observations or observations that present some kind of spacial or temporal dependence, as we see in the case of time series. By taking this dependence into account, the statistical theory of time series analysis naturally appears, as we will be discussing it over the next pages. The objective of this piece of work is the description and comparison of the different models and methodologies about time series analysis. From this comparison made on the base of the following terms: accuracy, simpleness, interpretability and computational efficiency, I have reached the conclusion that the most appropriate models vary depending on each case.

Introdución

Unha serie temporal é unha sucesión de datos observados en instantes de tempo consecutivos. Esta definición tan simple, fai da análise de series temporais un campo moi amplo dentro da estatística. Entender os mecanismos e as relacións intrínsecas destes conxuntos de datos é de especial relevancia en campos como a economía, a meteoroloxía, a medicina ou as telecomunicacións.

Neste traballo faremos un percorrido polas distintas etapas da análise de series temporais, analizando as súas características, propondo modelos que describan comportamento e estudando os modelos máis relevantes para a predición de novos valores das mesmas. Deste xeito, comezaremos abordando cuestións da teoría estatística clásica, para logo revisar metodoloxías máis recentes da área da aprendizaxe automática.

Primeiramente, no Capítulo 1, facemos unha introdución teórica ás series temporais definindo os procesos estocásticos, a estacionariedade, a autocorrelación... , xunto con nocións básicas sobre a descomposición, as características ou as medidas de erro máis comúns no ámbito das series temporais.

No Capítulo 2 repasamos varios dos modelos clásicos de regresión para series temporais, principalmente o alisado exponencial e os modelos ARMA, ARIMA e SARIMA.

No Capítulo 3 comentamos algúns métodos de aprendizaxe automática, como os procesos gaussianos, que empregan ideas da estatística bayesiana, ou as máis actuais redes neurais recorrentes coas súas sucesivas evolucións: os modelos LSTM, ou mesmo unha introdución da que foi a súa evolución, os transformers.

Por último, no Capítulo 4 levamos a cabo unha comparación dos métodos vistos, analizando as diferenzas de rendemento e a empregabilidade dos mesmos. Incluímos tamén dous apéndices que complementan o contido visto no traballo. No Apéndice A atópanse figuras e representacións gráficas de conceptos e métodos vistos, mentres que no Apéndice B discutimos con certo detalle o algoritmo de descomposición STL.

Capítulo

1 Series de Tempo. Definición e Análise

Neste primeiro capítulo tratamos os temas fundamentais da formulación matemática das series temporais, comezando pola súa formulación en forma de procesos estocásticos, continuando co estudo dos procesos estacionarios, a autocorrelación e a predición lineal de futuros valores dunha serie temporal dada en función dos anteriores.

Ademais, complementamos o anterior con dúas seccións sobre ferramentas útiles para a análise previa á modelización de series temporais como son a descomposición de series temporais e a obtención de características das mesmas, xunto con unha sección sobre as métricas dispoñibles para o estudo dos erros de predición puntual e distribucional de series temporais.

1.1. Series Temporais e Procesos Estocásticos

Comezamos definindo formalmente o que de aquí en diante entenderemos por serie temporal. Esta parte introdutoria está baseada no Capítulo 1 de Brockwell e Davis (1991).

Definición 1 (Serie temporal)**.** Unha serie temporal é un conxunto de datos rexistrados en distintos instantes de tempo. Distinguimos dous tipos:

- As *series temporais continuas*, que denotaremos por $x(t)$, con $t \in [0, T]$, rexístranse ao longo dun intervalo temporal.
- As *series temporais discretas*, que denotaremos por x_t, con $t \in \{1, \dots, T\}$, rexístranse en instantes de tempo discretos, consecutivos e equiespaciados[1].

Neste traballo centrarémonos no estudo das series temporais discretas. Dámonos conta de que, a nivel teórico, a Definición 1 non nos aporta ningunha pista sobre a orixe (a distri-

[1] No caso de que os datos non estean distribuídos de forma homoxénea, estariamos no campo dos *datos transversais* (en inglés: «*cross-sectional data*») que son aqueles que se obteñen sen ter en conta o instante temporal. Non serán obxecto de estudo.

bución) dos datos que compoñen a serie temporal, sendo esta unicamente un conxunto de observacións. Isto motiva a introdución dos procesos estocásticos como un xeito de caracterizar non só os datos concretos dunha serie temporal, senón o procedemento que os xerou, tendo en conta a distribución de posibles series temporais que se poderían obter a partir das distintas realizacións do mesmo.

Definición 2 (Proceso estocástico)**.** Un proceso estocástico é unha familia de variables aleatorias $\{X_t : t \in \mathbb{Z}\}$[II], definidas sobre un espazo de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$.

Definición 3 (Realización dun proceso estocástico)**.** Para un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$, as súas realizacións veñen dadas por $\{X_\cdot(\omega) : \omega \in \Omega\}$ (funcións definidas en $\mathbb{Z}$).

Deste xeito, as series (de datos) temporais son "*camiños mostrais*" (realizacións) xerados en función do proceso estocástico subxacente[III]. Na literatura é habitual referirse mediante a denominación "*series temporais*", tanto aos datos, que usualmente denotaremos en minúscula $\{x_t : 1 \leq t \leq T\}$, con $T \in \mathbb{Z}$, como ao proceso estocástico, denotado con maiúsculas $\{X_t : 1 \leq t \leq T\}$, cuxa realización deu lugar a tales datos. Podemos ver na Figura A.1 unha representación gráfica dun proceso estocástico, xunto con varias realizacións do mesmo.

Para cada subconxunto finito de variables aleatorias pertencentes a un proceso estocástico podemos definir a súa función de distribución.

Definición 4 (Función de distribución dun proceso estocástico)**.** Dados un proceso estocástico con $T \subset \mathbb{R}$ e un vector de índices pertencente a $\mathcal{T} = \{\boldsymbol{t} = (t_1, \dots, t_n)' \in \mathbb{Z}^n : t_1 < \dots < t_n,\ n \in \mathbb{Z}^+\}$ (de dimensión finita). As funcións de distribución de $\{X_t : t \in \mathcal{T}\}$ son as funcións $\{F_{\boldsymbol{t}}(\cdot), \boldsymbol{t} \in \mathcal{T}\}$ definidas como

$$F_{\boldsymbol{t}}(\boldsymbol{x}) = \mathbb{P}\left(X_{t_1} \leq x_1, \dots, X_{t_n} \leq x_n\right), \quad \text{onde } \boldsymbol{x} = (x_1, \dots, x_n)' \in \mathbb{R}^n.$$

Reciprocamente, o teorema (de extensión) de Kolmogorov establece baixo qué condicións de consistencia unha familia de distribucións de dimensión finita, $\{F_{\boldsymbol{t}}(\cdot), \boldsymbol{t} \in \mathcal{T}\}$ nos termos da Definición 4, determina a un proceso estocástico.

Teorema 1 (Teorema de Kolmogorov)**.** Un conxunto de funcións de distribución de probabilidade $\{F_{\boldsymbol{t}}(\cdot) : \boldsymbol{t} \in \mathcal{T}\}$ dado correspóndese co conxunto de funcións de distribución dun proceso estocástico se, e só se, para cada $n \in \{1, 2, \dots\}$, $\boldsymbol{t} = (t_1, \dots, t_n)' \in \mathcal{T}$ e $i \in \{1, \dots, n\}$,

$$\lim_{x_i \to \infty} F_{\boldsymbol{t}}(\boldsymbol{x}) = F_{\boldsymbol{t}^{(i)}}(\boldsymbol{x}^{(i)}),$$

[II] Ao restrinxirnos a series temporais e discretas, o conxunto de índices T será un conxunto de instantes temporais da forma: $\{1, 2, 3, \dots\}$ ou $\{0, \pm 1, \pm 2, \dots\}$.

[III] Citando de Williams e Rasmussen (2006):p. 13: "*falando mal e pronto, un proceso estocástico é unha xeneralización dunha distribución de probabilidade (que describe a unha variable aleatoria de dimensión finita) a funcións*".

onde $\boldsymbol{t}^{(i)}$ e $\boldsymbol{x}^{(i)}$ obtéñense ao eliminarlle a compoñente i-ésima aos vectores $\boldsymbol{t}$ e $\boldsymbol{x}$. Analogamente, empregando a equivalencia entre as funcións de distribución e as funcións características, se $\phi_{\boldsymbol{t}}(\cdot)$ é a función característica de $F_{\boldsymbol{t}(\cdot)}$, i.e.,

$$\phi_{\boldsymbol{t}}(\boldsymbol{u}) = \int_{\mathbb{R}^n} e^{i\boldsymbol{u}'\boldsymbol{x}} F_{\boldsymbol{t}}(dx_1, \dots, dx_n), \quad \boldsymbol{u} = (u_1, \dots, u_n)' \in \mathbb{R}^n,$$

entón as funcións familia $\{\phi_{\boldsymbol{t}} : t \in \mathcal{T}\}$ son funcións características dun proceso estocástico se, e só se,

$$\lim_{u_i \to 0} \phi_{\boldsymbol{t}}(\boldsymbol{u}) = \phi_{\boldsymbol{t}(i)}(\boldsymbol{u}(i)),$$

onde igualmente $\boldsymbol{u}(i)$ é o vector resultante de eliminarlle a compoñente i-ésima ao vector $\boldsymbol{u}$.

Notación Específica para as Series Temporais

A fin de facilitar a notación, é moi habitual na literatura relativa ás series temporais o uso do *backward shift operator*, B, que definimos a continuación.

Definición 5 (Operador de retardo)**.** O operador de retardo B aplicado a unha variable aleatoria pertencente a un proceso estocástico $\{X_t\}$ obtén a variable aleatoria inmediatamente anterior dentro do proceso: $\mathrm{B}\, X_t = X_{t-1}$. Ademais, aplicando repetidas veces o operador, denotamos

$$\mathrm{B}^i\, X_t = X_{t-i}, \quad i \in \mathbb{Z}.$$

1.2. Estacionariedade e Autocorrelación

Posto que os procesos estocásticos abarcan un conxunto potencialmente infinito de variables aleatorias, non ten sentido empregar a matriz de covarianza para estudar a súa interdependencia lineal, xa que unicamente aplica a un conxunto finito de variables aleatorias. Seguimos estando interesados na relación lineal das variables aleatorias dúas a dúas, mais por cuestións de dimensión, falaremos de función de (auto)covarianza en lugar de matriz de covarianza.

Definición 6 (Función de autocovarianza)**.** Para un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$ tal que $\mathrm{Var}\,(X_t) < \infty$, $\forall t \in \mathbb{Z}$, defínese a función de autocovarianza $\gamma_X(\cdot, \cdot)$ de $\{X_t\}$ como

$$\gamma_X(r, s) = \mathrm{Cov}\,(X_r, X_s) = \mathbb{E}\left[(X_r - \mathbb{E}\,(X_r))(X_s - \mathbb{E}\,(X_r))\right], \quad r, s \in \mathbb{Z}.$$

Estacionariedade. Regularidade dos Procesos Estocásticos

De xeito similar ao que acontece noutras ramas das matemáticas, dado que a definición de proceso estocástico é considerablemente ampla, o máis habitual será que traballemos con procesos estocásticos que cumpran unhas certas condicións de regularidade.

Definición 7 (Estacionariedade (débil))**.** Unha serie temporal $\{X_t : t \in \mathbb{Z}\}$ dise estacionaria[IV] (debilmente, ou de xeito condicional) se cumpre as seguintes tres condicións:

1. $\mathbb{E}\left(X_t^2\right) < \infty, \ \forall t \in \mathbb{Z}$, 2. $\mathbb{E}\left(X_t\right) = m, \ \forall t \in \mathbb{Z}$, 3. $\gamma_X(r,s) = \gamma_X(r+t, s+t), \ \forall r,s,t \in \mathbb{Z}$,

Definición 8 (Estacionariedade (forte))**.** Unha serie temporal $\{X_t : t \in \mathbb{Z}\}$ dise estacionaria (fortemente, ou de xeito estrito) se cumpre que para calquera $k, h \in \mathbb{Z}$ e $t_1, \dots, t_k \in \mathbb{Z}$ entón os vectores $(X_{t_1}, \dots, X_{t_k})'$ e $(X_{t_1+h}, \dots, X_{t_k+h})'$ teñen a mesma distribución conxunta.

Intuitivamente, unha serie temporal será estacionaria cando as súas propiedades estatísticas sexan constantes respecto do tempo, tendo en consecuencia unha aparencia semellante en todo momento, como se pode apreciar na Figura A.2. A diferenza entón entre a estacionariedade débil e a forte radica en ata que punto esta condición se debe garantir. A estacionariedade débil asegura a independencia temporal dos momentos de ata orde 2 e require da súa existencia, mentres que a estacionariedade forte asegura que, se existen, os momentos de calquera orde son independentes do tempo.

É habitual referirse con estacionariedade á definición débil da mesma, xa que garante que para todo vector finito de variables aleatorias do proceso estocástico a súa matriz de covarianza é idéntica.

Autocorrelación. Correlación en procesos Estacionarios

En presenza de estacionariedade tense que $\gamma_X(r,s) = \gamma_X(r-s, 0), \ \forall r,s \in \mathbb{Z}$, polo que, para procesos estacionarios, redefínese a función de autocovarianza como

$$\gamma_X(h) := \gamma_X(h,0) = \mathrm{Cov}\left(X_{t+h}, X_t\right), \ \forall t,h, \in \mathbb{Z}.$$

Proposición 1 (Propiedades da función de autocovarianza)**.** Se $\gamma(\cdot)$ é a función de autocovarianza dun proceso estacionario $\{X_t : t \in \mathbb{Z}\}$, entón

$$\gamma(0) \geq 0, \qquad \gamma(0) \geq |\gamma(h)|, \ \forall h \in \mathbb{Z}, \quad \text{e} \quad \gamma(h) = \gamma(-h), \ \forall h \in \mathbb{Z}.$$

Estamos xa en disposición de caracterizar ás funcións de autocovarianza dos procesos estacionarios, a través do Teorema 2, para o que primeiro introducimos a seguinte definición.

Definición 9 (Función semidefinida positiva)**.** Unha función $\mu : \mathbb{Z} \to \mathbb{R}$ dise que é *semidefinida positiva* se, e só se,

$$\sum_{i,j=1}^{n} a_i \mu(t_i - t_j) a_j \geq 0,$$

[IV] É importante ter clara a diferenza entre estacionariedade (do inglés: «*stationarity*») e estacionalidade (do inglés: «*seasonality*»), xa que en galego as súas raíces son moi semellantes a diferenza do que acontece en inglés.

para todos os enteiros positivos n e todos os vectores $\boldsymbol{a} = (a_1, \dots, a_n)' \in \mathbb{R}^n$ e $\boldsymbol{t} = (t_1, \dots, t_n)' \in \mathbb{Z}^n$.

Teorema 2 (Caracterización das funcións de autocovarianza)**.** Unha función $\gamma : \mathbb{Z} \to \mathbb{R}$ é unha función de autocovarianza dunha serie temporal estacionaria se, e só se, é par e semidefinida positiva.

Demostración. "$\Longrightarrow$" Dada unha serie temporal $\{X_t\}$ entón, para os vectores $\boldsymbol{a} = (a_1, \dots, a_n)' \in \mathbb{R}^n$ e $\boldsymbol{t} = (t_1, \dots, t_n)' \in \mathbb{Z}^n$, definimos $\boldsymbol{Z_t} = (X_{t_1} - \mathbb{E}(X_{t_1}), \dots, X_{t_n} - \mathbb{E}(X_{t_n}))$. Tense que

$$\begin{aligned} 0 \leq \operatorname{Var}\left(\boldsymbol{a}'\boldsymbol{t_t}\right) = \mathbb{E}\left(\boldsymbol{a}'\boldsymbol{Z_t}\right) = \mathbb{E}\left[\left(\boldsymbol{a}'\boldsymbol{Z_t}\right)^2\right] - \cancel{\mathbb{E}\left(\boldsymbol{a}'\boldsymbol{Z_t}\right)^2} = \boldsymbol{a}'\,\mathbb{E}\left(\boldsymbol{Z_t}\boldsymbol{Z_t'}\right)\boldsymbol{a} = \boldsymbol{a}'\Gamma_n\boldsymbol{a} \\ = \sum_{i,j=1}^{n} a_i\gamma(t_i - t_j)a_j, \end{aligned}$$

onde Γ_n é a matriz de covarianza de $(X_{t_1}, \dots, X_{t_n})'$.

"$\Longleftarrow$" Sexa $\mu : \mathbb{Z} \to \mathbb{R}$ unha función par e semidefinida positiva. Veremos agora que existe un proceso estacionario tal que teña a $\mu(\cdot)$ como función de autocovarianza empregando o teorema de Kolmogorov. Para cada $n \in \mathbb{Z}^+$ e cada $\boldsymbol{t} = (t_1, \dots, t_n)' \in \mathcal{T}$ (empregando a notación da Definición 4), sexa $F_{\boldsymbol{t}}$ a función de distribución en $\mathbb{R}^n$ da función característica

$$\phi_{\boldsymbol{t}}(\boldsymbol{u}) = \exp(-\boldsymbol{u}'K\boldsymbol{u}/2),$$

onde $\boldsymbol{u} = (u_1, \dots, u_n)' \in \mathbb{R}^n$ e $K = [\mu(t_i - t_j)]_{i,j=1}^n$. Ao ser μ semidefinida positiva, tamén o será K, sendo entón $\phi_{\boldsymbol{t}}$ a función característica dunha distribución normal multivariante con media cero e matriz de covarianza K, logo tense que

$$\phi_{\boldsymbol{t}(i)}(\boldsymbol{u}(i)) = \lim_{\boldsymbol{u}_i \to 0} \phi_{\boldsymbol{t}}(\boldsymbol{u}), \quad \forall \boldsymbol{t} \in \mathcal{T},$$

polo que, aplicando o teorema de Kolmogorov, tense a existencia dunha serie temporal $\{X_t\}$ con funcións de distribución $F_{\boldsymbol{t}}$ e funcións características $\phi_{\boldsymbol{t}}$, con $t \in \mathcal{T}$. Tendo comprobado que, ao ser K a matriz de covarianza de $\{X_t\}$, entón $\operatorname{Cov}(X_i, X_j) = \mu(i - j)$, sendo logo $\mu(\cdot)$ a función de autocovarianza de $\{X_t\}$ como queríamos ver. □

O habitual será empregar no lugar da función de autocovarianza a función de autocorrelación, xa que esta non depende da escala dos datos da serie temporal estudada.

Definición 10 (Función de autocorrelación (ACF))**.** Para un proceso estocástico estacionario $\{X_t : t \in \mathbb{Z}\}$, defínese a súa función de autocorrelación como

$$\rho_X(h) := \gamma_X(h)/\gamma_X(0) = \operatorname{Corr}(X_{t+h}, X_t), \quad t, h \in T.$$

En ocasións pode resultar útil considerar a correlación entre X_t e X_{t+h} excluíndo a aportación das variables aleatorias $X_{t+1}, \dots, X_{t+h-1}$. Para isto emprégase a función de autocorrelación parcial, como podemos ver na Figura A.3.

Definición 11 (Función de autocorrelación parcial (PACF))**.** Para un proceso estocástico estacionario $\{X_t : t \in \mathbb{Z}\}$, defínese a súa función de autocorrelación parcial como

$$\begin{aligned}\alpha(1) &= \operatorname{Corr}(X_2, X_1) = \rho(1),\\ \alpha(h) &= \operatorname{Corr}\left(X_{h+1} - \mathrm{P}_{\operatorname{span}\{1,X_2,\dots,X_h\}} X_{h+1}, X_1 - \mathrm{P}_{\operatorname{span}\{1,X_2,\dots,X_h\}} X_1\right), \quad h \geq 2,\end{aligned}$$

onde $P_{\operatorname{span}\{1,X_2,\dots,X_h\}}X_1$ denota a proxección[v] de X_1 sobre o subespazo $\operatorname{span}\{1, X_2, \dots, X_h\}$ de $L^2(\Omega, \mathcal{A}, \mathbb{P})$, formado polas combinación lineais de $\{1, X_2, \dots, X_h\}$, i.e.,

$$\mathrm{P}_{\operatorname{span}\{X_1,\dots,X_h\}} X_{h+1} = \sum_{i=1}^{h} \phi_{hi} X_{h+1-i}. \tag{1.1}$$

Aínda que precisa, a definición anterior da función de autocorrelación parcial non é moi práctica á hora de calcular o valor da mesma. Como alternativa, partindo das ecuacións

$$\left\langle X_{h+1} - \mathrm{P}_{\operatorname{span}\{X_1,\dots,X_h\}} X_{h+1}, X_i \right\rangle = 0, \quad i = h, \dots, 1,$$

desenvolvendo e empregando a definición da función de autocorrelación, séguese a obtención do seguinte sistema de ecuacións en forma matricial

$$\begin{pmatrix} \rho(0) & \rho(1) & \dots & \rho(h-1) \\ \rho(1) & \rho(2) & \dots & \rho(h-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(h-1) & \rho(h-2) & \dots & \rho(0) \end{pmatrix} \begin{pmatrix} \phi_{h1} \\ \phi_{h2} \\ \vdots \\ \phi_{hh} \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(h) \end{pmatrix}, \quad h \geq 1, \tag{1.2}$$

de onde podemos despexar ϕ_{hh} en función de $\rho(i)$, con $i = 0, \dots, h$ e de ϕ_{hj}, con $j = 0, \dots, h-1$. Como probamos máis adiante na Proposición 3, temos que $\alpha(h) = \phi_{hh}$, con $h \geq 1$, polo que podemos obter recursivamente a función de autocorrelación parcial en función da función de autocorrelación.

[v] Neste punto estamos empregando sen mencionalo o *teorema de proxección ortogonal* ou *teorema de proxección de Hilbert*, que establece que para un espazo de Hilbert, no noso caso $\mathcal{H} = L^2(\Omega, \mathcal{A}, \mathbb{P})$ con $\langle X, Y \rangle = \mathbb{E}(XY)$, para calquera subconxunto pechado, convexo e non baleiro $K \subseteq \mathcal{H}$ e para un $\boldsymbol{x} \in \mathcal{H}$, entón existe un único $\mathrm{P}_K \boldsymbol{x} \in K$ tal que é o elemento de K máis próximo a $\boldsymbol{x}$, cumprindo $\|\boldsymbol{x} - \mathrm{P}_K \boldsymbol{x}\| = d(\boldsymbol{x}, K)$. Ademais, tense que

$$\boldsymbol{x} - \mathrm{P}_K \boldsymbol{x} \in K^{\perp} \iff \langle \boldsymbol{x} - \mathrm{P}_K \boldsymbol{x}, \boldsymbol{k} \rangle = 0, \ \forall \boldsymbol{k} \in K. \tag{1.3}$$

Estimación da ACF para unha serie temporal dada

Para unha serie temporal de datos $\{x_t : 1 \leq t \leq T\}$ o estimador mostral, e ao que nos restrinxiremos no ámbito deste traballo, para a función de autocovarianza é

$$\hat{\gamma}(h) = T^{-1} \sum_{t=1}^{T-h} (x_t - \overline{x}_T)(x_{t+h} - \overline{x}_T), \;\; 0 \leq h \leq T-1.$$

A partir del obteremos como $\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$ o estimador mostral da función de autocorrelación. Mentres que substituíndo en (1.2) os valores de $\hat{\rho}(\cdot)$, podemos despexar o estimador da función de autocorrelación parcial como $\hat{\alpha}(h) = \hat{\phi}_{hh}$.

1.3. Predición para Procesos Estacionarios

Ao longo do traballo trataremos repetidamente a cuestión de, coñecidos $\{X_1, \dots, X_T\}$, predicir $\{X_t : t \geq T+1\}$ para un proceso estocástico $\{X_t\}$. Se denotamos por $\mathcal{M}$ ao subespazo pechado de $L^2(\Omega, \mathcal{A}, \mathbb{P})$ que contén ás funcións de $\{X_1, \dots, X_T\}$, entón o mellor preditor de X_{T+h} será o elemento de $\mathcal{M}$ cuxa distancia cadrática respecto del sexa menor[VI], i.e., segundo v, $\mathrm{P}_{\mathcal{M}} X_{T+h}$, ou equivalentemente, $\mathbb{E}(X_{T+h} \mid X_1, \dots, X_T)$, segundo a definición da esperanza condicionada.

A Esperanza Condicionada como Proxección

A fin de ter clara a notación, ímonos deter levemente na esperanza condicionada. Comezamos definindo a esperanza dunha variable aleatoria condicionada respecto dun subespazo pechado $\mathcal{N}$ de $L^2(\Omega, \mathcal{A}, \mathbb{P})$.

Definición 12 (Operador da esperanza condicionada)**.** Se $\mathcal{N}$ é un subespazo pechado de $L^2(\Omega, \mathcal{A}, \mathbb{P})$ tal que contén ás funcións constantes[VII], entón, para un $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, a esperanza condicionada de X respecto de $\mathcal{M}$ coincide coa súa proxección

$$\mathbb{E}_{\mathcal{N}} X = \mathrm{P}_{\mathcal{N}} X.$$

Séguese entón que $\mathbb{E}_{\mathcal{N}}$ é un operador de proxección e ademais, por (1.3), tense que $\mathbb{E}_{\mathcal{N}} X$ é un único elemento de $\mathcal{N}$ tal que

$$\mathbb{E}(W \, \mathbb{E}_{\mathcal{N}} X) = \mathbb{E}(WX), \;\; \forall W \in \mathcal{N}.$$

[VI] Este é o criterio de optimalidade máis estendido, mais non é o único. Neste traballo será este o criterio estándar que empregaremos. Segundo se indica en Brockwell e Davis (1991):p. 166, este criterio, para procesos con momentos de segunda orde finitos da lugar a unha teoría: "*simple, elegante e útil na práctica*".

[VII] Se o subespazo $\mathcal{N} = \mathcal{M}_0$ está formado unicamente polas funcións constantes, entón $\mathbb{E}_{\mathcal{M}_0} X = \mathbb{E}(X)$.

Polo tanto, a esperanza condicionada respecto dun número finito de variables aleatorias será un caso particular da Definición 12, e virá dada de xeito análogo, condicionando agora respecto do subespazo xerado polas mesmas, i.e., $\mathcal{M}(Z_1, \dots, Z_n)$.

Definición 13 (Esperanza condicionada)**.** Se $Z_1, \dots, Z_n$ son variables aleatorias en $(\Omega, \mathcal{A}, \mathbb{P})$ e $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ entón a *esperanza condicionada* de X dadas $Z_1, \dots, Z_n$ defínese como

$$\mathbb{E}(X \mid Z_1, \dots, Z_n) = \mathbb{E}_{\mathcal{M}(Z_1,\dots,Z_n)} X \quad (= \mathrm{P}_{\mathcal{M}(Z_1,\dots,Z_n)} X),$$

onde $\mathcal{M}(Z_1, \dots, Z_n)$ é un subespazo pechado de $L^2(\Omega, \mathcal{A}, \mathbb{P})$ ao que pertencen todas as funcións que se poden poñer como $\phi(Z_1, \dots, Z_n)$, con $\phi : \mathbb{R}^n \to \mathbb{R}$ unha función medible.

Cálculo de Preditores (Lineais) para Procesos Estacionarios

Volvendo ao problema da predición, a dificultade reside en que o cálculo de $\mathrm{P}_{\mathcal{M}(X_1,\dots,X_T)} X_{T+h}$ non é en absoluto sinxelo, polo que, rebaixando as expectativas pedagóxicas, podemos restrinxir $\mathcal{M}$ considerando menos funcións preditoras pero simplificando os cálculos. É por iso, polo que traballaremos sobre $\operatorname{span}\{1, X_1, \dots, X_T\} \subsetneq \mathcal{M}(X_1, \dots, X_T)$, tratando no que resta de sección os preditores lineais dos procesos estacionarios.

Coa notación $\mathcal{H}_T = \operatorname{span}\{X_1, \dots, X_T\}$, centrarémonos da predición para X_{T+1}, xa que o caso X_{T+h} é análogo. Baseándonos no anterior temos que

$$\hat{X}_{T+1} = \begin{cases} 0 & \text{se } T = 0, \\ \mathrm{P}_{\mathcal{H}_T} X_{T+1} & \text{se } T \geq 1. \end{cases} \implies \hat{X}_{T+1} = \phi_{T1} X_T + \cdots + \phi_{TT} X_1, \ \ T \geq 1. \tag{1.4}$$

onde os coeficientes $\phi_{T1}, \dots, \phi_{TT}$ veñen determinados polas "*ecuacións de predición*" (1.3),

$$\left\langle \sum_{i=1}^{T} \phi_{Ti} X_{T+1-i}, X_{T+1-j} \right\rangle = \langle X_{T+1}, X_{T+1-j} \rangle \implies \sum_{i=1}^{T} \phi_{Ti} \gamma(i-j) = \gamma(j), \ \ j = 1, \dots, T,$$
$$\implies \Gamma_T \boldsymbol{\phi}_T = \boldsymbol{\gamma}_T, \tag{1.5}$$

onde $\Gamma_T = [\gamma(i-j)]_{i,j=1,\dots,T}$, $\boldsymbol{\gamma}_T = (\gamma(1), \dots, \gamma(T))'$ e $\boldsymbol{\phi}_T = (\phi_{T1}, \dots, \phi_{TT})'$. Deste xeito, as ecuacións (1.4) e (1.5) coñécense como as *ecuacións de predición de paso un* para procesos estacionarios.

Aínda que (1.5) poida ter varias solucións, todas elas darán lugar ao mesmo estimador mediante (1.4). No seguinte resultado, interesámonos polo caso no que Γ_n é non singular, xa que entón a solución pode obterse facilmente como: $\boldsymbol{\phi}_T = \Gamma_T^{-1} \boldsymbol{\gamma}_T$.

Proposición 2 (Unicidade do preditor lineal para procesos estacionarios)**.** Sexa $\{X_t : t \in \mathbb{Z}\}$ un proceso estacionario con media cero e con función de autocovarianza $\gamma(\cdot)$ tal que $\gamma(0) > 0$ e onde $\gamma(h) \xrightarrow{h\to\infty} 0$. Entón a matriz de covarianza $\Gamma_n = [\gamma(i-j)]_{i,j=1,\dots,n}$ de $(X_1, \dots, X_n)'$ é non singular $\forall n \in \mathbb{Z}^+$.

Demostración. Procederemos por redución ao absurdo, supoñendo que Γ_n é singular para algún n e chegando a unha contradición coas condicións de que $\gamma(0) > 0$ e que $\gamma(h) \xrightarrow{h\to\infty} 0$.

Por hipótese, $\{X_1, \dots, X_n\}$ non son independentes, logo $\exists r \in \mathbb{Z}$ tal que $1 \leq r < n$ para o que existen constantes $a_1, \dots, a_r$ de xeito que

$$X_{r+1} = \sum_{i=1}^{r} a_i X_i, \text{ con } \Gamma_r \text{ non singular} \xrightarrow{\{X_t\} \text{ estacionario}} X_{r+h} = \sum_{i=1}^{r} a_i X_{i+h-1}, \ \forall h \geq 1.$$

Logo, para todos os $n \geq r+1$ existirán constantes $a_1^{(n)}, \dots, a_r^{(n)}$ tal que

$$X_n = \sum_{i=1}^{r} a_i^{(n)} X_i = (a_1^{(n)}, \dots, a_r^{(n)})' \cdot (X_1, \dots, X_r) = \boldsymbol{a}^{(n)\prime} \cdot \boldsymbol{X}_r.$$

Buscaremos agora a contradición vendo por unha banda que

$$\gamma(0) = \boldsymbol{a}^{(n)\prime} \Gamma_r \boldsymbol{a}^{(n)} \implies \gamma(0) \geq \lambda_{\min} \boldsymbol{a}^{(n)\prime} \boldsymbol{a}^{(n)} \implies |a_i^n| \text{ limitados } \forall i,$$

sendo $\lambda_{\min}$ o menor valor propio do Γ_r que, ao ser unha matriz de covarianza, é simétrica e semidefinida positiva. Alternativamente, vemos que

$$\gamma(0) = \operatorname{Cov}\left(X_n, \sum_{i=1}^{r} a_i^{(n)} X_i\right) \implies \gamma(0) \leq \sum_{i=1}^{r} \underbrace{\left|a_i^{(n)}\right|}_{\text{limitado}} \underbrace{|\gamma(n-i)|}_{\xrightarrow{n\to\infty} 0} \xrightarrow[❶]{n\to\infty} 0,$$

onde estamos empregando en ❶ a hipótese $\gamma(h) \xrightarrow{h\to\infty} 0$, chegando deste xeito a unha contradición con que $\gamma(0) > 0$. □

Cabe logo preguntarnos, dado que estamos predicindo o valor de X_{T+1}, cal é o erro da predición que estamos cometendo. Expresámolo en termos do MSE, que virá dado por

$$\begin{aligned}
\varepsilon_{T+1} &= \mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})^2\right] = \operatorname{Var}\left(X_{T+1} - \hat{X}_{T+1}\right) \\
&= \operatorname{Var}\left(X_{T+1}\right) + \operatorname{Var}\left(\hat{X}_{T+1}\right) - 2\operatorname{Cov}\left(X_{T+1}, \hat{X}_{T+1}\right) \\
&= \gamma(0) + \operatorname{Var}\left(\boldsymbol{\phi}_T' \boldsymbol{X}_T\right) - 2\operatorname{Cov}\left(X_{T+1}', \boldsymbol{\phi}_T \boldsymbol{X}_T\right) \\
&= \gamma(0) + \boldsymbol{\phi}_T' \Gamma_T \boldsymbol{\phi}_T - 2\boldsymbol{\phi}_T' \boldsymbol{\gamma}_T \\
&= \gamma(0) + \boldsymbol{\gamma}_T' \cancel{\Gamma_T^{-1}} \cancel{\Gamma_T} \Gamma_T^{-1} \boldsymbol{\gamma}_T - 2\boldsymbol{\gamma}_T' \Gamma_T^{-1} \boldsymbol{\gamma}_T \\
&= \gamma(0) - \boldsymbol{\gamma}_T' \Gamma_T^{-1} \boldsymbol{\gamma}_T.
\end{aligned}$$

Tendo demostrado a Proposición 2, estamos xa en disposición de probar a equivalencia entre as dúas definicións da función de autocorrelación parcial ($\alpha(h) = \phi_{hh}$).

Proposición 3 (Equivalencia entre as dúas definicións da función de autocorrelación parcial)**.** Sexa $\{X_t : t \in \mathbb{Z}\}$ un proceso estacionario con media cero e con función de autocovarianza $\gamma(\cdot)$ tal que $\gamma(0) > 0$ e onde $\gamma(h) \xrightarrow{h\to\infty} 0$. Para a definición dos coeficientes ϕ_h. dada en (1.1), tense que

$$\phi_{hh} = \operatorname{Corr}\left(X_{h+1} - \mathrm{P}_{\operatorname{span}\{X_2,\dots,X_h\}} X_{h+1}, X_1 - \mathrm{P}_{\operatorname{span}\{X_2,\dots,X_h\}} X_1\right).$$

Demostración. Denotamos $\mathcal{H}_1 = \operatorname{span}\{X_2, \dots, X_h\}$ e $\mathcal{H}_2 = X_1 - \mathrm{P}_{\mathcal{H}_1} X_1$, logo

$$\begin{aligned}\operatorname{Corr}\left(X_{h+1} - \mathrm{P}_{\mathcal{H}_1} X_{h+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\right) &\overset{❶}{=} \frac{\langle X_{h+1} - \mathrm{P}_{\mathcal{H}_1} X_{n+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\rangle}{\|X_{h+1} - \mathrm{P}_{\mathcal{H}_1} X_{h+1}\| \cdot \|X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\|} \\ &\overset{❷}{=} \frac{\langle X_{h+1} - \mathrm{P}_{\mathcal{H}_1} X_{n+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\rangle}{\|X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\|^2} \\ &\overset{❸}{=} \frac{\langle X_{h+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\rangle}{\|X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\|^2},\end{aligned}$$

onde ❶ tense pola definición usual da covarianza, en ❷ estamos empregando que o proceso $\{X_t\}$ é estacionario, e en ❸ empregamos que $\mathrm{P}_{\mathcal{H}_1} X_{h+1} \perp (X_1 - \mathrm{P}_{\mathcal{H}_1} X_1)$. Alternativamente temos que

$$\sum_{i=1}^{h} \phi_{hi} X_{h+1-i} = \mathrm{P}_{\mathcal{H}_1 \cup \mathcal{H}_2} X_{h+1} \overset{❹}{=} \mathrm{P}_{\mathcal{H}_1} X_{h+1} + \mathrm{P}_{\mathcal{H}_2} X_{h+1} = \mathrm{P}_{\mathcal{H}_1} X_{h+1} + k \cdot (X_1 - \mathrm{P}_{\mathcal{H}_1} X_1), \quad k \in \mathbb{R} \tag{1.6}$$

onde ❹ tense dado que $\mathcal{H}_1 \cup \mathcal{H}_2 = \operatorname{span}\{X_1, \dots, X_h\}$ e $\mathcal{H}_1 \perp \mathcal{H}_2$. Vemos que a expresión co coeficiente k será

$$k = \frac{\langle X_{h+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\rangle}{\|X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\|^2}. \tag{1.7}$$

Empregando de novo a estacionariedade de $\{X_t\}$ temos que os vectores $(X_1, \dots, X_h)'$, $(X_h, \dots, X_1)'$ e $(X_2, \dots, X_{h+1})'$ teñen a mesma matriz de covarianza, logo

$$\mathrm{P}_{\mathcal{H}_1} X_1 = \phi_{h-1,1} X_2 + \phi_{h-1,2} X_3 + \cdots + \phi_{h-1,h-1} X_h, \tag{1.8}$$

$$\mathrm{P}_{\mathcal{H}_1} X_{h+1} = \phi_{h-1,1} X_h + \phi_{h-1,2} X_{h-1} + \cdots + \phi_{h-1,h-1} X_2. \tag{1.9}$$

Finalmente, substituíndo (1.8) e (1.9) en (1.6) temos

$$\mathrm{P}_{\mathcal{H}_1 \cup \mathcal{H}_2} X_{h+1} = kX_1 + \sum_{i=1}^{h-1} \left[\phi_{h-1,i} - k\phi_{h-1,h-i}\right] X_{h+1-i}. \tag{1.10}$$

Como por Proposición 2 a representación de $\hat{X}_{h+1}$ ten que ser única, comparando (1.4) e (1.10) temos que

$$\phi_{hh} = k = \operatorname{Corr}\left(X_{h+1} - \mathrm{P}_{\mathcal{H}_1} X_{h+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\right).$$

□

Cálculo de preditores lineais para paso h

Se queremos predicir X_{T+h} con $h \geq 1$ a partir de $X_1, \dots, X_T$ o desenvolvemento é incriblemente semellante ao anterior, tendo igualmente que

$$\hat{X}_{T+h} = \mathrm{P}_{\mathcal{H}_T} X_{T+h} = \phi_{T1}^{(h)} X_T + \cdots + \phi_{TT}^{(h)} X_1, \quad \text{con } T, h \geq 1,$$

é tal que $\boldsymbol{\phi}_T^{(h)} = (\phi_{T1}^{(h)}, \dots, \phi_{TT}^{(h)})'$ é a solución do sistema en forma matricial

$$\Gamma_T \boldsymbol{\phi}_T^{(h)} = \boldsymbol{\gamma}_T^{(h)}, \quad \text{onde } \boldsymbol{\gamma}_T^{(h)} = (\gamma(h), \dots, \gamma(n+h-1))'.$$

Métodos Recursivos para o Cálculo de Preditores Lineais

Como xa vimos, no cálculo de $\hat{X}_{T+1}$ debemos resolver un sistema de T ecuacións con T incógnitas, o cal para valores altos de T (o cal é o habitual na práctica) é moi custoso computacionalmente. Para atallar este hándicap introduciremos a continuación algoritmos recursivos para o cálculo dos coeficientes $\boldsymbol{\phi}_T$ sen necesidade de ter que resolver o sistema lineal. Ademais, polo mero feito de seren recursivos, ao engadir novas observacións podemos reaproveitar os cálculos xa feitos.

Algoritmo Recursivo de Durbin-Levinson

Este algoritmo baséase na descomposición de $\mathcal{H} = \operatorname{span}\{X_1, \dots, X_T\}$ en $\mathcal{H}_1 = \operatorname{span}\{X_2, \dots, X_T\}$ e $\mathcal{H}_2 = X_1 - \mathrm{P}_{\mathcal{H}_1} X_1$. Procédese calculando primeiro o valor do coeficiente ϕ_{TT}, que se corresponde coa autocorrelación parcial, asociado a X_1, para logo obter os valores dos demais coeficientes $\boldsymbol{\phi}_T = (\phi_{T1}, \dots, \phi_{T,T-1})'$ a partir del.

Proposición 4 (Algoritmo de Durbin-Levinson)**.** Sexa $\{X_t : t \in \mathbb{Z}\}$ un proceso estacionario con media cero e con función de autocovarianza $\gamma(\cdot)$ tal que $\gamma(0) > 0$ e onde $\gamma(h) \xrightarrow{h\to\infty} 0$. Entón os coeficientes de $\boldsymbol{\phi}_T$ e o correspondente MSE ε_T veñen dados por

$$\phi_{11} = \gamma(1)/\gamma(0), \qquad \phi_{TT} = \left[\gamma(T) - \sum_{i=1}^{T-1} \phi_{T-1,i}\gamma(T-i)\right] \varepsilon_{T-1}^{-1},$$

$$\begin{pmatrix} \phi_{T1} \\ \vdots \\ \phi_{T,T-1} \end{pmatrix} = \begin{pmatrix} \phi_{T-1,1} \\ \vdots \\ \phi_{T-1,T-1} \end{pmatrix} - \phi_{TT} \begin{pmatrix} \phi_{T-1,T-1} \\ \vdots \\ \phi_{T-1,1} \end{pmatrix}, \qquad \varepsilon_0 = \gamma(0), \qquad \varepsilon_T = \varepsilon_{T-1}[1 - \phi_{TT}^2].$$

Demostración. Procédese empregando as mesmas ideas que xa vimos na demostración da

Proposición 3. Partindo da factorización (1.7) e substituíndo a expresión (1.8) temos

$$\begin{aligned}\phi_{TT} &= \frac{\langle X_{T+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\rangle\, X_1}{\|X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\|^2} \overset{❶}{=} \frac{\left\langle X_{T+1}, X_1 - \sum_{i=1}^{T-1} \phi_{T-1,i} X_{i+1}\right\rangle}{\varepsilon_{T-1}} \\ &= \left[\gamma(T) - \sum_{i=1}^{T-1} \phi_{T-1,i}\gamma(T-i)\right] \varepsilon_T^{-1},\end{aligned}$$

onde en ❶ estamos empregando a definición do erro $\varepsilon_T = \mathbb{E}(X_{T+1} - \hat{X}_{T+1})^2$, xa que $\mathrm{P}_{\mathcal{H}_1} X_1$ é a predición de X_1 en función das $T-1$ variables aleatorias de $\mathcal{H}_1$. Por outra banda, aplicando a Proposición 2 que nos garante que as factorizacións (1.4) e (1.10) son iguais temos que

$$\phi_{Tj} = \phi_{T-1,j} - \phi_{TT}\phi_{T-1,T-j}, \quad \text{para } j = 1, \dots, T-1.$$

Onde, por último, a expresión do erro de predición vén dada por

$$\begin{aligned}\varepsilon_T &= \left\|X_{T+1} - \hat{X}_{T+1}\right\|^2 \\ &\overset{❷}{=} \left\|X_{T+1} - \mathrm{P}_{\mathcal{H}_1} X_{T+1} - \mathrm{P}_{\mathcal{H}_2} X_{T+1}\right\|^2 \\ &= \left\|X_{T+1} - \mathrm{P}_{\mathcal{H}_1} X_{T+1}\right\|^2 + \left\|\mathrm{P}_{\mathcal{H}_2} X_{T+1}\right\|^2 - 2\left\langle X_{T+1} - \mathrm{P}_{\mathcal{H}_1} X_{T+1}, \mathrm{P}_{\mathcal{H}_2} X_{T+1}\right\rangle \\ &\overset{❸}{=} \left\|X_{T+1} - \mathrm{P}_{\mathcal{H}_1} X_{T+1}\right\|^2 + \|\phi_{TT}(X_1 - \mathrm{P}_{\mathcal{H}_1} X_1)\|^2 - 2\left\langle X_{T+1} - \mathrm{P}_{\mathcal{H}_1} X_{T+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\right\rangle \\ &\overset{❹}{=} \varepsilon_{T-1} + \phi_{TT}^2\varepsilon_{T-1} - 2\phi_T\left\langle X_{T+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\right\rangle \\ &\overset{❺}{=} \varepsilon_{T-1}(1 - \phi_{TT}^2)\end{aligned}$$

onde en ❷ empregamos que $\mathcal{H}_1 \cup \mathcal{H}_2 = \mathrm{span}\{X_1, \dots, X_h\}$ con $\mathcal{H}_1 \perp \mathcal{H}_2$, o punto ❸ séguese da demostración da Proposición 3, en ❹ estamos aplicando a definición de ε_{T-1} e a relación $\mathrm{P}_{\mathcal{H}_1} X_{h+1} \perp (X_1 - \mathrm{P}_{\mathcal{H}_1} X_1)$ e, por último, en ❺ empregamos (1.7) para ver que $\left\langle X_{T+1}, X_1 - \mathrm{P}_{\mathcal{H}_1} X_1\right\rangle = \phi_{TT}\varepsilon_{T-1}$. □

Neste caso, a ecuación do algoritmo de Durbin-Levinson para o erro mostra claramente como ao engadir unha observación o cálculo do erro de predición diminúe nun factor de $1 - \phi_{TT}^2$, sendo ϕ_{TT}, como xa vimos, a función de autocorrelación parcial.

Algoritmo Recursivo de *Innovacións*

O termo *innovación* circunscríbese unicamente ao contexto das series temporais e refírese á compoñente de variabilidade nova e impredicible a partir dos valores anteriores da serie que se "engade" en cada instante temporal.

Definición 14 (Innovación)**.** Para un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$ non necesariamente estacionario, defínense as súas *innovacións* como proceso estocástico $\{X_t - \hat{X}_t : t \in \mathbb{Z}\}$.

Proposición 5. Sexa $X_t : t \in \mathbb{Z}$ un proceso estocástico e $\mathcal{H}_T = \text{span}\{X_1, \dots, X_T\}$, entón

$$\mathcal{H}_T = \text{span}\{X_1 - \hat{X}_1, \dots, X_T - \hat{X}_T\}, \text{ con } T \geq 1.$$

Demostración. Obtense directamente dado que as innovacións son ortogonais, pois

$$(X_i - \hat{X}_i) \subset \text{span}\{X_1 - \hat{X}_1, \dots, X_{j-1} - \hat{X}_{j-1}\} \perp (X_j - \hat{X}_j), \quad i < j.$$

□

Polo que as innovacións xeran o mesmo espazo de Hilbert que as variables aleatorias que compoñen un proceso estocástico dado. Polo tanto, xa que $\hat{X}_{T+1} \in \mathcal{H}_T$, podemos considerar alternativamente á factorización (1.4) a seguinte[VIII]

$$\hat{X}_{T+1} = \sum_{i=1}^{T} \theta_{Ti}(\underbrace{X_{T+1-i} - \hat{X}_{T+1-i}}_{\text{innovacións}}), \text{ con } T \geq 1. \tag{1.11}$$

Consecuentemente, o obxectivo do algoritmo das innovacións será o de obter de xeito recursivo o valor dos coeficientes $\boldsymbol{\theta}_T = (\theta_{T1}, \dots, \theta_{TT})'$. Este ademais ten a particularidade de que non precisa que o proceso estocástico sexa estacionario, polo tanto, en vez de traballar coa función de autocorrelación empregaremos directamente a función de autocovarianza do proceso $\{X_t\}$ que denotaremos, a fin de facer esta distinción máis evidente, por

$$\kappa(r, s) = \gamma_X(r, s) = \langle X_r, X_s \rangle = \mathbb{E}(X_r X_s).$$

Proposición 6 (Algoritmo das innovacións)**.** Sexa $\{X_t : t \in T\}$ un proceso estocástico de media cero e covarianza $\kappa(\cdot, \cdot)$, onde a matriz $[\kappa(i, j)]_{i,j=1}^{T}$ é non singular para todo $T \in \mathbb{Z}^+$. Logo, os coeficientes das predición dadas por (1.11), i.e., $\boldsymbol{\theta}_T$ e o correspondente erro de predición (MSE) ε_T, obtéñense de xeito recursivo como

$$\theta_{T,T-j} = \varepsilon_j^{-1}\left(\kappa(T+1, j+1) - \sum_{i=0}^{j-1} \theta_{j,j-i}\theta_{T,T-i}\varepsilon_i\right), \quad j = 0, 1, \dots, n-1.$$

$$\varepsilon_0 = \kappa(1, 1), \qquad \varepsilon_T = \kappa(T+1, T+1) - \sum_{i=0}^{n-1} \theta_{T,T-i}^2 \varepsilon_i.$$

Demostración. Pola Proposición 5 sabemos que $\{X_1 - \hat{X}_1, \dots, X_T - \hat{X}_T\}$ é ortogonal. Partimos agora da ecuación (1.11) buscando despexar os coeficientes $\boldsymbol{\theta}_T$ da factorización en

[VIII] É igualmente frecuente, e mesmo máis sinxela, a factorización dada por $\hat{X}_{T+1} = \sum_{i=0}^{T-1} \theta_{T,T-i}(X_{i+1} - \hat{X}_{i+1})$.

innovacións.

$$(1.11) \xrightarrow{\times(X_{j+1}-\hat{X}_{h+1}),\ \text{para } 0\leq j<T} \left\langle \hat{X}_{T+1}, X_{j+1}-\hat{X}_{j+1} \right\rangle = \theta_{T,T-j}\varepsilon_j$$

$$\xrightarrow{(X_{T+1}-\hat{X}_{T+1})\perp(X_{j+1}-\hat{X}_{j+1})} \theta_{T+1,T-j} = \varepsilon_j^{-1}\left\langle X_T, X_{j+1}-\hat{X}_{j+1}\right\rangle. \tag{1.12}$$

Agora, aplicándolle (1.11) a $\hat{X}_{j+1}$ chegamos á expresión buscada.

$$\begin{aligned}
\theta_{T+1,T-j} &= \varepsilon_j^{-1}\left\langle X_T, X_{j+1}-\sum_{i=1}^{j}\theta_{ji}(X_{j+1-i}-\hat{X}_{j+1-i})\right\rangle \\
&= \varepsilon_j^{-1}\left(\kappa(T+1,j+1)-\sum_{i=1}^{j}\theta_{ji}\left\langle X_{T+1}, X_{j+1-i}-\hat{X}_{j+1-i}\right\rangle\right) \\
&\overset{❶}{=} \varepsilon_j^{-1}\left(\kappa(T+1,j+1)-\sum_{i=0}^{j-1}\theta_{j,j-i}\underbrace{\left\langle X_{T+1}, X_{i+1}-\hat{X}_{i+1}\right\rangle}_{\star}\right) \\
&\overset{❷}{=} \varepsilon_j^{-1}\left(\kappa(T+1,j+1)-\sum_{i=0}^{j-1}\theta_{j,j-i}\theta_{T,T-i}\varepsilon_i\right),
\end{aligned}$$

onde en ❶ aplicamos un cambio de índice no sumatorio, tomando $i = j - i$, para logo en ❷ poder aplicar (1.12) en $\star$ simplificando a expresión. Finalmente os erros de predición obtéñense directamente como

$$\varepsilon+_T = \left\|X_{T+1}-\hat{X}_{T+1}\right\|^2 \overset{❸}{=} \|X_{T+1}\|^2 - \left\|\hat{X}_{T+1}\right\|^2 \overset{❹}{=} \kappa(T+1,T+1) - \sum_{i=0}^{T-1}\theta_{T,T-i}^2\varepsilon_i,$$

onde en ❸ estamos empregando o teorema de Pitágoras e en ❹ calculamos a función de autocovarianza para a factorización en innovacións (1.11). □

1.4. Descomposición e Compoñentes das Series Temporais

Como paso previo á modelización dunha serie temporal $\{X_t : t \in \mathbb{Z}\}$ é habitual, e recomendable, facer unha análise preliminar das compoñentes que conforman a variabilidade da mesma, dividíndose na tendencia (T_t), a compoñente estacional (S_t, pode haber varias correspondéndose con distintos períodos de tempo) e a compoñente cíclica (cuxa variabilidade adoita integrarse na compoñente da tendencia). Ademais, logo de extraerlle á serie temporal as compoñentes T_t e S_t, restará a compoñente residual, que denotaremos por R_t.

Vemos facilmente que hai distintos xeitos de combinar estas compoñentes para dar lugar á serie orixinal. En particular, a descomposición aditiva $X_t = T_t + S_t + R_t$ e a descomposición multiplicativa: $X_t = T_t \cdot S_t \cdot R_t$.

A descomposición aditiva é máis fácil de interpretar, xa que, por exemplo, as compoñentes comparten unidades coa propia serie temporal, mentres que a multiplicativa recolle mellor

a variabilidade cando a magnitude das oscilacións da compoñente estacional varía na mesma escala que a tendencia, i.e., a maior valor da tendencia maior amplitude das oscilacións da compoñente estacional como, por exemplo, podemos ver nos valores orixinais da Figura A.4. Fixámonos en que podemos pasar da descomposición multiplicativa a unha aditiva empregando unha transformación logarítmica

$$X_t = T_t \cdot S_t \cdot R_t \longrightarrow \log X_t = \log T_t + \log S_t + \log R_t.$$

En presencia de heterocedasticidade na serie temporal orixinal é desexable buscar unha transformación que a reduza, co fin de eliminar fontes de variabilidade coñecidas e simplificar o máis posible a serie a modelar. Para este fin, destaca a familia das transformacións Box-Cox, xa que parametrizan unha familia de transformacións potenciais, entre as que se inclúe a logarítmica, que se adaptan a unha ampla variedade de casos prácticos. Podemos ver un exemplo da súa aplicación na Figura A.4.

Definición 15 (Transformacións Box-Cox)**.** Propostas en Box e Cox (1964). É unha familia de transformacións que engloba tanto á transformación logarítmica como a distintas transformacións potenciais. Defínese en función dun parámetro $\lambda \in \mathbb{R}$, e está dada por

$$w_t = \begin{cases} \log x_t & \text{se } \lambda = 0, \\ \dfrac{\text{signo}(x_t)\,|x_t|^{\lambda} - 1}{\lambda} & \text{noutro caso.} \end{cases}$$

Existen múltiples metodoloxías para realizar a descomposición da variabilidade nas tres compoñentes. Por cuestións de espazo e interese, restrinxímonos unicamente a dúas delas, a descomposición clásica, que é a máis sinxela de todas sendo a cambio pouco eficaz, e a descomposición STL, que é máis potente e versátil e é de uso moi habitual na literatura cando se precisa unha metodoloxía de descomposición xeral. Se o noso problema é máis específico, só considerando, por exemplo, datos mensuais ou trimestrais, existen métodos máis potentes como o SEATS (do inglés: «*Seasonal Extraction in ARIMA Time Series*»), desenvolto no Banco de España, ou o método X-11.

Descomposición Clásica

Estimación da tendencia a través de medias móbiles

Para estimar a tendencia empréganse medias móbiles. O valor de T_t virá dado por unha media ponderada dos valores da serie temporal adxacentes, é dicir

$$\hat{T}_t = \frac{1}{n} \sum_{i=-k}^{k} x_{t+i}, \quad \text{onde } n = 2k + 1.$$

Neste caso, empregamos unha media móbil de orde n (denótase por n-MA) centrada e simétrica. En función da lonxitude do ciclo estacional pode ser preferible empregar unha media móbil de orde par, para o que se utilizan medias móbiles ponderadas ou, equivalentemente, medias móbiles de medias móbiles. Por exemplo, denotamos por 2×4-MA a aplicar unha 4-MA seguido de unha 2-MA, obtendo unha media móbil ponderada dada por

$$\begin{aligned}\hat{T}_t &= \frac{1}{2}\left[\frac{1}{4}(x_{t-1}+x_t+x_{t+1}+x_{t+2})+\frac{1}{4}(x_{t-2}+x_{t-1}+x_t+x_{t+1})\right]\\ &= \frac{1}{8}x_{t-2}+\frac{1}{4}x_{t-1}+\frac{1}{4}x_t+\frac{1}{4}x_{t+1}+\frac{1}{8}x_{t+2}.\end{aligned}$$

Isto permite que se o patrón estacional ten lonxitude 4, a media móbil pondere con $\frac{1}{4}$ cada "etapa" do patrón, axudando que a tendencia sexa independente do patrón estacional.

Polo tanto, para estimar a tendencia, se o patrón estacional é par e de orde m empregarase unha media móbil de orde $m+1$ (2×m-MA); mentres que se o patrón é impar empregarase unha media móbil simple de orde m (m-MA).

Estimación da compoñente estacional

Para estimar a compoñente estacional, S_t, calculamos para a serie $x_t - \hat{T}_t$ a media dos valores de cada estación, para posteriormente trasladalos de xeito que a suma de todas as estacións sexa cero, onde, se denotamos $q = t \pmod m$, isto é

$$\hat{S}_t = \lambda + \frac{m}{T}\sum_{k=1}^{T/m}\left(x_{q+m(k-1)} - \hat{T}_{q+m(k-1)}\right), \text{ onde o parámetro } \lambda \text{ permite que } \sum_{t=1}^{m}\hat{S}_t = 0.$$

Estimación da compoñente residual

Unha vez extraídas as estimacións da tendencia e da compoñente estacional, obtemos a estimación da compoñente residual $\hat{R}_t = x_t - \hat{T}_t - \hat{S}_t$, completando a descomposición da serie temporal.

Neste caso, describimos a versión aditiva ($X_t = T_t + S_t + R_t$) da descomposición clásica. A versión multiplicativa ($X_t = T_t \cdot S_t \cdot R_t$) obtense de xeito análogo, garantindo na estimación da compoñente estacional que $\sum_{t=1}^{m}\hat{S}_t = m$.

Limitacións da descomposición clásica

Ao traballarmos coa descomposición clásica enfrontamos principalmente dúas limitacións:

- Ao empregar medias móbiles para a estimación da tendencia, non teremos descomposición para os primeiros e últimos $\lfloor m/2 \rfloor$ valores.

- A compoñente estacional é constante para cada período, o cal só é certo en casos particulares.

Ademais, tamén presenta fraquezas á hora de captar variacións abruptas na tendencia ou períodos estacionais atípicos, polo que é un método pouco robusto.

Descomposición STL

O modelo STL, do inglés: «*Seasonal and Trend decomposition using LOESS*», foi proposto en Cleveland et al. (1990)[IX]. A diferenza doutros modelos de descomposición modernos permite descompoñer series temporais con calquera frecuencia estacional maior que un (non só datos trimestrais ou mensuais, como os métodos SEATS e X-11), así como descompoñer series nas que haxa datos faltantes. Ademais, STL permite realizar descomposicións robustas das distintas compoñentes, evitando a excesiva influencia de datos aberrantes.

Podemos ver un exemplo de aplicación da descomposición STL na Figura A.5 respecto da descomposición clásica (sobre os mesmos datos) na Figura A.6. Fixámonos en que, aínda que a descomposición clásica deixa remanentes máis pequenos, non é capaz de separar correctamente a variabilidade, copiando demasiado á serie na tendencia e omitindo as distintas frecuencias estacionais que presenta.

Co obxectivo de manter o ritmo do traballo, omitimos a definición do algoritmo STL, que se pode consultar, xunto cunha revisión da regresión LOESS, no Apéndice B.

Descomposición *vs.* Predición

Aínda que pode resultar evidente, é importante sinalar que a descomposición das series temporais en compoñentes é unha ferramenta útil de análise das mesmas, mais en ningún caso é unha ferramenta aplicable na predición. Obviando a inclusión dos pesos de robustez, a función que dá lugar a cada compoñente a partir da serie temporal orixinal é un filtro lineal que illa a variabilidade correspondente, pero sen realizar en ningún momento modelización algunha da mesma.

Deste xeito, podemos ver na Figura A.7 como a suma das compoñentes (tendencia máis compoñentes estacionais) que vimos na Figura A.5 está lonxe de aproximar á serie orixinal. Isto débese a que o obxectivo dunha descomposición non é diminuír os erros, neste caso residuos, de predición, senón que trata de extraerlles toda a variabilidade posible que se corresponda con algunha das compoñentes.

[IX] En Cleveland et al. (1990):p. 1 defínese ao STL como: "*un proceso de filtrado para a descomposición de series temporais estacionais*".

1.5. Características das Series Temporais

Referímonos nesta sección ao estudo das características dunha serie temporal. Estas son indicadores numéricos que resumen algunha calidade ou aspecto da serie temporal. Algunhas características xerais son a media, a mediana ou os cuantís dos datos, mais no contexto particular das series temporais existen indicadores máis específicos.

Algúns exemplos de características de uso común que se seguen de conceptos xa vistos neste traballo son: o primeiro coeficiente da autocorrelación ou a suma ao cadrado dos dez primeiros coeficientes da autocorrelación, en ambos casos tanto para a serie orixinal, como para a serie diferenciada e mesmo dúas veces diferenciada. A continuación, trataremos varias características habituais na análise preliminar de series temporais.

Características das Compoñentes

O feito de descompoñer unha serie temporal nas súas compoñentes, $X_t = T_t + S_t + R_t$, permítenos obter información sobre a mesma en forma de características. Podemos estimar, por exemplo, cal é a forza ou mesmo a significatividade de cada unha delas, que se expresa en termos da "forza dunha compoñente".

Definición 16 (Forza da tendencia)**.** Para unha serie temporal $X_t = T_t + S_t + R_t$ defínese a forza da súa tendencia como

$$F_T = \text{máx}\left(0, 1 - \frac{\text{Var}\left(R_t\right)}{\text{Var}\left(T_t + R_t\right)}\right)$$

Definición 17 (Forza da compoñente estacional)**.** Para unha serie temporal $X_t = T_t + S_t + R_t$ defínese a forza da súa compoñente estacional como

$$F_S = \text{máx}\left(0, 1 - \frac{\text{Var}\left(S_t\right)}{\text{Var}\left(S_t + R_t\right)}\right)$$

En ambos casos a interpretación é a mesma: se as devanditas compoñentes son pouco significativas, entón as varianzas do numerador e denominador serán semellantes e a forza correspondente próxima a 0; e noutro caso, a forza será próxima a 1, indicando que a variabilidade da compoñente correspondente é moito maior que a variabilidade non explicada, o que dá unha idea da estrutura da serie temporal obxecto de estudo.

Ademais, existen multitude de outras características baseadas nas compoñentes. Mencionamos a continuación só unha delas que consideramos particularmente orixinal.

Definición 18 (Espiñosidade ou prevalencia de picos)**.** Para unha serie temporal $X_t = T_t + S_t + R_t$ defínese $V_t = \text{Var}\left(R_1, \dots, R_{t-1}, R_{t+1}, \dots, R_T\right)$, i.e., a varianza dos residuos deixando

fóra unha observación. A "*espiñosidade*", tradución literal do inglés: «*spikiness*», defínese como

$$S = \mathrm{Var}\,(V_t)\,, \quad \text{con } 1 \leq t \leq T.$$

Deste xeito, unha espiñosidade alta indica que os valores da serie $\{V_t\}$ son dispersos, tendo varianza alta, implicando que a varianza dos residuos é moi sensible ao feito de eliminar unha observación, o que revela a presenza de observacións atípicas, picos, na serie temporal.

Expoñente de Hurst

Foi introducido orixinalmente por Hurst (1951) como unha medida da "*persistencia*" dunha serie temporal.

Definición 19 (Persistencia e anti-persistencia)**.** Unha serie temporal dise *persistente* se tende a manter a tendencia inmediatamente anterior. O efecto contrario é a *anti-persistencia*, na que unha tendencia ascendente adoita anteceder a unha descendente e viceversa[x].

Deste xeito o expoñente de Hurst, $H \in (0, 1)$, codifica esta calidade tomando valores en $(0, 05)$ cando a serie é anti-persistente e valores en $(05, 1)$ cando é persistente. Se $H = 05$, entón enténdese que a serie temporal é aleatoria.

Ademais da información sobre a persistencia da serie, o expoñente de Hurst tamén trasladda información sobre a preditibilidade da serie, posto que, para poder modelar satisfactoriamente unha serie temporal, precisamos que o seu comportamento non sexa totalmente aleatorio, o cal virá indicado cando o expoñente estea próximo a 05.

A expresión do coeficiente de Hurst pode deducirse empregando varios métodos. O máis empregado, segundo se indica en Qian e Rasheed (2004), é a *análise do rango reescalado* ou análise R/S (do inglés: «*reescaled range analysis*»), que se describe na seguinte definición.

Definición 20 (Análise R/S)**.** Dada unha serie temporal $X_1, \ldots, X_n$, obtéñense as seguintes series produto da orixinal ata chegar á serie do rango reescalado $(R/S)_t$, con $t = 1, \ldots, n$.

1. Serie axustada á media: $Y_t = X_t - m$, onde $m = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. Serie coa desviación acumulada: $Z_t = \sum_{i=1}^{t} Y_i$.
3. Serie do rango: $R_t = \mathrm{máx}(Z_1, \ldots, Z_t) - \mathrm{mín}(Z_1, \ldots, Z_n)$.
4. Serie das desviacións estándar: $S_t = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(X_i - u)^2}$, onde $u = \frac{1}{t}\sum_{i=1}^{t} X_i$.

[x] En inglés é habitual referirse ás series temporais anti-persistentes como «*mean-reverting series*», xa que adoitan ter tendencia a oscilar ao redor da media.

5. Serie do rango reescalado: $(R/S)_t = R_t/S_t$.

Pódese ver que a serie R/S medra segundo a expresión

$$(R/S)_t = c \cdot t^H,$$

onde $c \in \mathbb{R}$ e H é o *expoñente de Hurst*, que podemos despexar de xeito sinxelo como

$$H = \frac{\log((R/S)_t \cdot c^{-1})}{\log(t)},$$

é dicir, correspóndese coa pendente da recta de regresión de $\log((R/S)_t)$ sobre $\log(t)$.

1.6. Medidas de Erro

Medidas de Erro Puntuais

Ao facer as predicións, $\{\hat{x}_{T+1|T}, \dots, \hat{x}_{T+h|T}\}$, dos novos valores descoñecidos, $\{x_{T+1}, \dots, x_{T+h}\}$, a partir dos valores dunha serie temporal $\{x_1, \dots, x_T\}$, cométense uns certos erros $e_{T+h} = x_{T+h} - \hat{x}_{T+h|T}$, respecto do valor observado. Representan a "parte impredicible" da observación. Para medir a precisión dun modelo pódese resumir a información dos erros de distintos xeitos.

Medidas dependentes da escala

Se as medidas dependen directamente dos propios erros, estarán na escala dos datos orixinais, o que nos restrinxe á hora de comparar a precisión de modelos sobre distintos conxuntos de datos. As medidas máis habituais deste tipo son as seguintes:

Definición 21 (Erro absoluto medio)**.** O MAE (do inglés: «*mean absolute error*») é unha medida de resumo do erro facilmente interpretable dada por

$$\text{MAE} = h^{-1} \sum_{i=1}^{h} |e_{T+i}| \,.$$

Definición 22 (Erro cadrático medio)**.** O MSE (do inglés: «*mean squared error*») é unha medida de resumo do erro moi habitual en case todos os ámbitos da estatística. Vén dado por

$$\text{MSE} = h^{-1} \sum_{i=1}^{h} e_{T+i}^2.$$

Aínda que noutros contextos adoita empregarse directamente o MSE, no contexto das series temporais é máis habitual o uso do RMSE («*root mean squared error*») dado por $\text{RMSE} = \sqrt{\text{MSE}}$.

Malia a maior interpretabilidade do MAE respecto do RMSE, este último é máis habitual polo feito de que, ao minimizar o MAE, convérxese á mediana, mentres que, ao minimizar o RMSE, converxese á media. Unha explicación intuitiva deste feito dáse en Hanley et al. (2001).

Medidas independentes da escala

Procurar medidas de resumo dos erros independentes da escala non é algo trivial. De entre as múltiples propostas a máis inmediata é a seguinte:

Definición 23 (Erro porcentual absoluto medio)**.** O MAPE (do inglés: «*mean absolute percentage error*») defínese como

$$\text{MAPE} = h^{-1} \sum_{i=1}^{h} 100 \left| \frac{e_{T+i}}{x_{T+i}} \right|.$$

Mais esta resulta unha medida pouco satisfactoria. En xeral, as medidas que se basean en porcentaxes presentan problemas de estabilidade cando o rango dos datos se achega a cero. Ademais, é unha medida condicionada á significatividade do cero da escala que se estea a empregar[xi].

Unha alternativa que elude estes inconvenientes é a dos *erros escalados*, propostos en Hyndman e Koehler (2006). Calcúlase mediante o cociente entre o erro rexistrado polo modelo e o erro que se cometería mediante unha predición naíf de un paso[xii]. Deste xeito, se o valor é menor que 1 o modelo comportarase de xeito máis preciso que o modelo naíf, e viceversa. Ademais, o proceso de escalado non varía en función da medida (dependente da escala) á que se lle aplique.

Definición 24 (Erro escalado absoluto medio)**.** O MASE (do inglés: «*mean absolute scaled error*») defínese a partir dos coeficientes

$$q_i = \frac{e_i}{h^{-1} \sum_{j=T+1}^{T+h} |x_j - x_{j-m}|},$$

onde m é o período da predición naíf estacional, como

$$\text{MASE} = h^{-1} \sum_{i=T+1}^{T+h} |q_i|.$$

[xi] Por exemplo, carece de sentido empregar o MAPE como medida de precisión dun modelo cuxos datos se correspondan con temperaturas expresadas en graos Fahrenheit ou Celsius, posto que en ambos casos, a elección da temperatura que se corresponde cos 0° é arbitraria, é dicir, non é significativo onde se atopa o 0. Exemplo tomado da sección 5.8 de Hyndman e Athanasopoulos (2021).

[xii] Que tratamos na Sección 2.1.

Definición 25 (Raíz do erro escalado cadrático medio)**.** O RMSSE (do inglés: «*root mean squared scaled error*») está definida a partir dos coeficientes

$$q_i^2 = \frac{e_i^2}{h^{-1}\sum_{j=T+1}^{T+h}(x_j - x_{j-m})^2},$$

como

$$\text{RMSSE} = \sqrt{h^{-1}\sum_{i=T+1}^{T+h} q_i^2}.$$

Medidas de Erro Distribucionais

No caso de que as predicións sexan en distribución e non unicamente de xeito puntual, necesitamos especificar medidas de erro que valoren a coincidencia da distribución predita respecto do valor real. Tomamos $f_{p,h}$ como o cuantil de probabilidade p da distribución de predición, é dicir, o valor tal que esperamos que x_{T+h} sexa menor que $f_{p,h}$ con probabilidade p.

Definición 26 (Puntuación cuantil)**.** A puntuación cuantil (en inglés: «*Quantile Score*», QS), tamén coñecida como a "función de perda do pinball" (do inglés: «*pinball loss function*») pola súa gráfica que recorda a unha pelota rebotando, defínese como

$$Q_{p,h} = \begin{cases} 2(1-p)(f_{p,h} - x_{T+h}), & \text{se } x_T + h < f_{p,h}, \\ 2p(x_{T+h} - f_{p,h}), & \text{se } x_T + h \geq f_{p,h}. \end{cases}$$

É unha sorte de MAE (de feito, se $p = 0,5$ coinciden) no que o erro absoluto está ponderado, penalizando en maior medida os erros menos probables.

Para o caso no que a predición unicamente nos traslade os intervalos de predición, dados por $[l_{\alpha,h}, u_{\alpha,h}]$, onde a probabilidade de que $\hat{x}_{T+h}$ estea no intervalo anterior é $1-\alpha$, definimos a seguinte medida

Definición 27 (Puntuación de Winkler)**.** Coa notación anterior, defínese como

$$W_{\alpha,t} = \begin{cases} (u_{\alpha,h} - l_{\alpha,h}) + \frac{2}{\alpha}(l_{\alpha,h} - x_{T+h}), & \text{se } x_{T+h} < l_{\alpha,h}, \\ (u_{\alpha,h} - l_{\alpha,h}), & \text{se } l_{\alpha,h} \leq x_{T+h} \leq u_{\alpha,h}, \\ (u_{\alpha,h} - l_{\alpha,h}) + \frac{2}{\alpha}(x_{T+h} - u_{\alpha,h}), & \text{se } x_{T+h} > u_{\alpha,h}. \end{cases}$$

Se ademais $l_{\alpha,h} = f_{\alpha/2,h}$ e $u_{\alpha,h} = f_{1-\alpha/2,h}$, vén dada por

$$W_{\alpha,t} = \frac{(Q_{\alpha/2,h} + Q_{1-\alpha/2,h})}{\alpha}.$$

Polo de agora, só definimos medidas dada unha certa probabilidade p ou $1-\alpha$ de referencia. Introducimos entón unha medida que teña en conta a precisión respecto de calquera cuantil.

Definición 28 (*«Continuous Ranked Probability Score»*)**.** A CRPS defínese a partir da integral das respectivas puntuacións cuantís para todas as probabilidades p. Sendo $F_h(\cdot)$ a función de distribución predita para o tempo $T+h$, defínese como

$$\begin{aligned}\mathrm{CRPS}(F_h, x_{T+h}) &= \int (F_h(z) - \mathbf{1}_{\{z \geq x_{T+h}\}})^2 dz \\ &= 2\int_0^1 Q_{p,h} dp.\end{aligned}$$

Desta forma tamén podemos comparar a mellora no rendemento relativo de varios métodos, dados por $\{M_1, M_2, \dots, M_i, \dots\}$, respecto doutro peor M, a través do cociente

$$\frac{\mathrm{CRPS}_M - \mathrm{CRPS}_{M_i}}{\mathrm{CRPS}_M}.$$

As proporcións desta forma, como medidas comparativas entre métodos, coñécense como *«skill scores»* e son aplicables tanto a predicións puntuais como en distribución.

Capítulo

2 Modelización. Metodoloxías Clásicas

Neste segundo capítulo abordaremos a modelización de series temporais dende o punto de vista da teoría clásica. En particular, trataremos dúas familias de modelos: os modelos de alisado exponencial e os modelos ARIMA. Os primeiros describen unha metodoloxía de predición baseada na estimación de compoñentes da serie temporal. Modelos coma o de Holt-Winters pertencen a esta familia e son de uso moi común na literatura. Por outra banda, os modelos ARIMA baséanse na teoría sobre os procesos estocásticos autorregresivos de media móbil (ARMA), polo que a predición pasa a ter un papel subordinado, xa que o que se busca é o modelo orixinal que "xerou" os datos, para logo facer as predicións baseándose nel, aínda que na práctica non teña por que existir necesariamente.

A diferenza do que acontece nun contexto máis amplo de regresión, no eido das series temporais tratamos de explicar unha serie temporal en función de si mesma, a partir dos seus valores en instantes temporais anteriores. Deste xeito, acometemos unha *auto-regresión*, non contando co esquema de variable resposta *vs.* variable explicativa habitual nun contexto de regresión usual.

2.1. Tipos de Modelos e Exemplos Triviais

A continuación, introducimos algunhas metodoloxías triviais de predición que se mencionan en Brockwell e Davis (1991), cuxa única utilidade práctica é a de servir de comparación respecto dos modelos máis complexos para avaliar os seu rendemento. Polo tanto, dada unha serie temporal de datos $\{x_t : 1 \leq t \leq T\}$, defínense como

Axuste á media Calquera valor futuro da serie predise como a media dos valores observados

$$\hat{x}_{T+h} = \frac{x_1 + \cdots + x_T}{T}.$$

Axuste naíf Os valores futuros da serie predinse como o último dos valores observados

$$\hat{x}_{T+h} = x_T.$$

Axuste naíf estacional No caso de ter unha serie estacional empregamos como preditor o último valor observado da mesma estación que se está predicindo. Polo tanto, sendo m o período estacional[I]

$$\hat{x}_{T+h} = x_{T+h-m(\lfloor \frac{h-1}{m} \rfloor +1)}.$$

Axuste da derivada Neste caso, a partir do último valor observado, supoñemos unha evolución lineal da serie en función da súa taxa de variación media

$$\hat{x}_{T+h} = x_T + \frac{h}{T-1}\sum_{t=2}^{T} x_t - x_{t-1} = x_T + h\frac{x_T - x_1}{T-1}.$$

Podemos ver un exemplo de predición coas 4 metodoloxías anteriores na Figura A.8.

2.2. Modelos de Alisado Exponencial (ETS)

Os modelos de alisado ou suavizado exponencial[II] representan á serie temporal como unha suma ponderada dos seus valores anteriores, onde os pesos da mesma diminúen exponencialmente co tempo. Comezaremos primeiro introducindo o alisado exponencial como unha metodoloxía de predición, familiarizándonos coas compoñentes empregadas e a notación da mesma, para logo expoñer os modelos estatísticos subxacentes ás devanditas metodoloxías.

Como xa vimos na Sección 1.4, os modelos de descomposición non están pensados para a predición de futuros valores da serie temporal. O alisado exponencial (ETS), baséase en ideas da descomposición de series temporais[III], mais céntranse exclusivamente na predición, polo que non ten por obxectivo separar axeitadamente as distintas fontes de variabilidade da serie.

Metodoloxías de Predición

Consideramos entón unha serie temporal de datos $\{x_t : 1 \leq t \leq T\}$, co obxectivo de predicir o valor de x_T por $\hat{x}_T$. O *alisado exponencial simple* é o máis elemental dos métodos que compoñen a familia do alisado exponencial. Vén dado por

$$\hat{x}_{T+1} = \alpha x_T + \alpha(1-\alpha)x_{T-1} + \alpha(1-\alpha)^2 x_{T-2} + \dots,$$

[I] Por exemplo, se o período é mensual teríamos que $m = 12$, mentres que se o período é cuadrimestral entón $m = 3$.

[II] Do inglés, «*exponential smoothing*».

[III] Precisamente as siglas ETS veñen das tres compoñentes en inglés: «*Error, Trend, Seasonal*».

onde $\alpha \in [0,1]$ é o *parámetro de suavizado*. Será moi útil a nivel de notación comprobar como a expresión anterior é equivalente á que se obtén de considerar recursivamente para cada $t \in [1,T]$ a estimación (de paso un[IV]) dada por $\hat{x}_{t+1} = \alpha x_t + (1-\alpha)\hat{x}_t$,

$$\begin{aligned}\hat{x}_{T+1} &= \alpha x_T + (1-\alpha)\hat{x}_T \\ &= \alpha x_T + \alpha(1-\alpha)x_{T-1} + (1-\alpha)^2\hat{x}_{T-1} \\ &= \alpha x_T + \alpha(1-\alpha)x_{T-1} + \alpha(1-\alpha)^2 x_{T-2} + (1-\alpha)^3\hat{x}_{T-3} \dots .\end{aligned}$$

Polo que estamos estimando cada novo valor por unha media ponderada da estimación anterior e o último dato observado. O habitual será empregar a notación recursiva equivalente e denotar os modelos ETS en *forma compoñente*, calculando primeiro as compoñentes da serie para despois obter a estimación en función delas. Logo, para o alisado exponencial simple teríamos unha representación en forma compoñente dada polas ecuacións

$$\begin{aligned}\text{Ecuación de predición} \qquad & \hat{x}_{t+1} = \ell_t, \\ \text{Ecuación de nivel} \qquad & \ell_t = \alpha x_t + (1-\alpha)\ell_{t-1}.\end{aligned}$$

Nos métodos de alisado exponencial empréganse tres compoñentes: o nivel, a pendente e a compoñente estacional, sendo a primeira delas intrínseca a todos os modelos de alisado exponencial. Cada unha delas busca representar as variacións en tempo t de cada unha das compoñentes da serie para poder telas en conta na predición seguinte.

As compoñentes estacional e da pendente poden participar ou non no modelo. De facelo, existen distintas implementacións posibles das mesmas, sendo a expresión do nivel dependente das elixidas para as demais compoñentes. A expresión da pendente dependerá do parámetro de suavizado da pendente $0 \leq \beta \leq 1$. Conta con dúas posibles implementacións:

- Pendente aditiva (A): calcúlase a pendente en función da variación do nivel o último intervalo temporal. Isto permite anticiparse a cambios no nivel da serie.

  $$\begin{aligned}\text{Ecuación de predición} \qquad & \hat{x}_{t+h} = \ell_t + hb_t \\ \text{Ecuación de nivel} \qquad & \ell_t = \alpha x_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ \text{Ecuación de pendente} \qquad & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1},\end{aligned}$$

  este modelo en concreto coñécese como o *método (de pendente lineal) de Holt*.

- Pendente aditiva atenuada (A_d): neste caso engádese un parámetro de atenuado, $0 < \phi < 1$[V], que multiplica a b_t de xeito que, no caso de realizar predicións a moi longo

[IV] Aquí estamos empregando unha notación algo abreviada. Ao escribirmos $\hat{x}_{t+h}$, estamos denotando a predición de x_{t+h} con paso h, é dicir, a partir dos valores coñecidos $\{x_1, \dots, x_t\}$, sendo, por exemplo, $\hat{x}_{t+1}$ unha predición de x_{t+1} con paso 1.

[V] O máis habitual é que $h \in [0{,}8, 0{,}98]$.

prazo ($h \gg 1$), a pendente non provoque que $\hat{x}_{T+h} \xrightarrow{h\to\infty} \pm\infty$.

$$\begin{aligned}
\text{Ecuación de predición} \quad & \hat{x}_{t+h} = \ell_t + (\phi + \phi^2 + \cdots + \phi^h) b_t \\
\text{Ecuación de nivel} \quad & \ell_t = \alpha x_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
\text{Ecuación de pendente} \quad & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1},
\end{aligned}$$

co cal conséguese que $\hat{x}_{T+h} \xrightarrow{h\to\infty} \ell_T + \frac{\phi}{1+\phi} b_T \in \mathbb{R}$, sendo máis razoable empiricamente.

Por outra banda, a expresión da compoñente estacional dependerá tamén do seu parámetro de suavizado, $0 \leq \gamma^* \leq 1$, e do período estacional $m \in \mathbb{N}$. Calcúlase como

$$s_t = \gamma^*(y_t - \ell_t) + (1-\gamma^*)s_{t-m} \xrightarrow{\gamma=\gamma^*(1-\alpha)} s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},$$

onde, para obter unha expresión explícita, substituímos a compoñente do nivel e realizamos o cambio de parámetro anterior, tal que $0 \leq \gamma \leq 1-\alpha$. As implementacións posibles son:

- Compoñente estacional aditiva (A): exprésase a compoñente estacional en termos absolutos, sendo máis axeitado cando as variacións estacionais son homoxéneas ao longo da serie.

$$\begin{aligned}
\text{Ecuación de predición} \quad & \hat{x}_{t+h} = \ell_t + h b_t + s_{t+h-m(\lfloor \frac{h-1}{m} \rfloor + 1)} \\
\text{Ecuación de nivel} \quad & \ell_t = \alpha(x_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Ecuación de pendente} \quad & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta) b_{t-1} \\
\text{Ecuación estacional} \quad & s_t = \gamma(x_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}.
\end{aligned}$$

- Compoñente estacional multiplicativa (M): exprésase a compoñente estacional en termos relativos, sendo máis axeitado cando as variacións estacionais son proporcionais ao nivel da serie. Calcúlase de xeito análogo ao caso anterior, obtendo a implementación seguinte

$$\begin{aligned}
\text{Ecuación de predición} \quad & \hat{x}_{t+h} = (\ell_t + h b_t) s_{t+h-m(\lfloor \frac{h-1}{m} \rfloor + 1)} \\
\text{Ecuación de nivel} \quad & \ell_t = \alpha \frac{x_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Ecuación de pendente} \quad & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta) b_{t-1} \\
\text{Ecuación estacional} \quad & s_t = \gamma \frac{x_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}.
\end{aligned}$$

Estes dous modelos coñécense como o *método aditivo/multiplicativo de Holt-Winters*. Ademais, en ambos métodos, podemos supor que a compoñente da pendente estea atenuada, tendo o *método multiplicativo atenuado de Holt-Winters* un uso moi estendido na predición

de series estacionais por ser preciso, robusto e sinxelo de aplicar. A súa formulación é a seguinte

$$\begin{aligned}
\text{Ecuación de predición} \quad & \hat{x}_{t+h} = \left[\ell_t + (\phi + \phi^2 + \cdots + \phi^h) b_t\right] s_{t+h-m(\lfloor \frac{h-1}{m} \rfloor + 1)} \\
\text{Ecuación de nivel} \quad & \ell_t = \alpha \frac{x_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
\text{Ecuación de pendente} \quad & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \\
\text{Ecuación estacional} \quad & s_t = \gamma \frac{x_t}{(\ell_{t-1} + \phi b_{t-1})} + (1-\gamma) s_{t-m}.
\end{aligned}$$

Modelos de Espazo de Estados

Para cada unha das metodoloxías de predición vistas ata agora nesta sección podemos definir varios modelos estatísticos subxacentes. Estes involucran variables aleatorias en vez de observacións e polo tanto, non só achegan predicións puntuais senón tamén en distribución. Neste contexto, as compoñentes pasan a denotarse estados, polo que os modelos que obteremos serán modelos de espazo de estados.

Con todo, para cada un dos modelos consideramos tamén os erros cometidos no tempo $t+1$, ε_{t+1}, como unha variable aleatoria máis, podendo ser estes

$$\underbrace{\varepsilon_{t+1} = x_{t+1} - \hat{x}_{t+1}}_{\text{Aditivos}} \qquad \text{ou} \qquad \underbrace{\varepsilon_{t+1} = \frac{x_{t+1} - \hat{x}_{t+1}}{\hat{x}_{t+1}}}_{\text{Multiplicativos}}.$$

Deste xeito, a familia de modelos que podemos formar conterá a $3 \cdot 3 \cdot 2 = 18$ modelos segundo a tipoloxía dos erros, da pendente e da compoñente estacional. Por exemplo, o modelo de espazo de estados para alisamento exponencial de Holt (con pendente aditiva) con erros multiplicativos virá determinado polas ecuacións

$$\begin{aligned}
X_t &= (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t) \\
\ell_t &= (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t) \\
b_t &= b_{t-1} + \alpha\beta(\ell_{t-1} + b_{t-1})\varepsilon_t.
\end{aligned}$$

Comentarios finais sobre os modelos ETS

Non trataremos aquí en profundidade dous temas importantes, como son as estimación dos parámetros do modelo e a propia selección do modelo. Respecto da estimación dos parámetros, precisamos estimar tanto os parámetros de suavizado α, β e ϕ (se procede) como os estados iniciais ℓ_0, b_0, s_0, s_{-1}, ... , S_{-m+1}. En canto aos procedementos de estimación, é habitual empregar tanto máxima verosimilitude como mínimos cadrados, só sendo ambos enfoques equivalentes no caso no que os erros sexan aditivos. En canto á selección do

modelo, o máis común será empregar criterios de información como o Criterio de Información de Akaike (AIC), a súa versión corrixida para valores pequenos de T ou o Criterio de Información de Bayes (BIC).

Para rematar a sección, vemos na Figura A.9 o resultado da predición con varias metodoloxías da familia dos modelos de alisado exponencial.

2.3. Modelos Autorregresivos de Media Móbil (ARMA)

A importancia dos modelos ARMA, do inglés: «*AutoRegressive Moving Average*», para procesos estacionarios, propostos en Box, Jenkins et al. (1970), radica na súa simplicidade, capacidade preditiva e asentada teoría. O obxectivo será o de aproximar unha serie temporal (estacionaria) mediante un proceso ARMA, servíndonos deste último para realizar predicións ou análises da serie.

A falta dos resultados concretos, é razoable pensar que mediante un proceso ARMA imos poder aproximarnos a calquera outro procesos estacionarios dado. En particular, temos o seguinte resultado, que demostraremos máis adiante.

Proposición 7. Sexa $\gamma(\cdot)$ unha función de autocovarianza tal que $\lim_{h\to\infty}\gamma(h)=0$. Entón para calquera $k\in\mathbb{Z}^+$ é posible atopar un proceso ARMA, $\{X_t\}$, con función de autocovarianza $\gamma_X(\cdot)$, tal que

$$\gamma_X(h)=\gamma(h),\quad \text{onde } h=0,1,\dots,k.$$

Procesos ARMA Estacionarios

Comezamos estudando os procesos ARMA que conforman unha familia de procesos estocásticos estacionarios. Como se desprende do seu nome, os procesos ARMA están constituídos á súa vez por outros dous: os procesos AR e os procesos MA.

É importante ter en consideración que na definición destes procesos, dados por $\{X_t : t\in T\}$, intervirá outro, dado por $\{Z_t : t\in T\}$, que representará á parte aleatoria do modelo. Dependendo do contexto, os Z_t desígnanse como erros, máis correctamente como innovacións ou mesmo en inglés como *«shocks»*. Aínda que non é a única posibilidade, ao longo deste traballo, e segundo se indica en Brockwell e Davis (1991):p. 77–78, o proceso $\{Z_t\}$ seguirá unha distribución de ruído branco, $\mathrm{WN}(0,\sigma^2)$, que se define como segue.

Definición 29 (Distribución de ruído branco)**.** Un proceso estocástico $\{Z_t\}$ segue unha distribución de ruído branco con media cero e varianza σ^2, i.e., $\mathrm{WN}(0,\sigma^2)$, se, e só se,

$$\mathbb{E}\,(Z_t)=0,\ \ \forall t,\qquad \text{e}\qquad \gamma_Z(h)=\begin{cases}\sigma^2 & \text{se } h=0,\\ 0 & \text{se } h\neq 0.\end{cases}$$

Comezamos introducindo os procesos de media móbil nos que X_t se expresa a partir dunha combinación lineal das compoñentes aleatorias $\{Z_t, \dots, Z_{t-q}\}$.

Definición 30 (Procesos $\mathrm{MA}(q)$)**.** Un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$ dise que é un proceso de media móbil de orde q, denotado por $\mathrm{MA}(q)$, se é estacionario e cumpre

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} = \theta(\mathrm{B}) Z_t, \tag{2.1}$$

onde $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ e con $\{Z_t\} \sim WN(0, \sigma^2)$.

Podémonos facer unha idea do grao de xeneralidade que acadan os procesos $\mathrm{MA}(q)$, en termos da variedade de funcións de autocovarianza que admiten unha representación como un procesos $\mathrm{MA}(q)$, a partir do resultado seguinte.

Proposición 8. Sexa $\{X_t\}$ un proceso estacionario con media cero e cuxa función de autocovarianza cumpre que $\gamma(h) = 0$, para $|h| > q$, e $\gamma(q) \neq 0$. Entón $\{X_t\}$ admite unha representación como un proceso $\mathrm{MA}(q)$.

Alternativamente, temos os procesos autorregresivos nos que a variable aleatoria X_t se explica a partir dunha combinación das p variables anteriores máis un erro Z_t que, neste caso, é a única compoñente non determinista do proceso.

Definición 31 (Procesos $\mathrm{AR}(p)$)**.** Un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$ dise que é un proceso autorregresivo de orde p, denotado por $\mathrm{AR}(p)$, se é estacionario e cumpre

$$\phi(\mathrm{B}) X_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \tag{2.2}$$

onde $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ e con $\{Z_t\} \sim WN(0, \sigma^2)$.

Como xa imos vendo, o uso do operador de retardo B será recorrente na formulación destes procesos, xa que todos eles involucran combinacións lineais de instantes anteriores dos mesmos. Á súa vez, a análise dos polinomios $\theta(\cdot)$ e $\phi(\cdot)$ será de especial utilidade na análise dos procesos ARMA.

Deste xeito, como conxugación natural dos procesos autorregresivos e os procesos de media móbil xorden os procesos ARMA.

Definición 32 (Procesos $\mathrm{ARMA}(p, q)$)**.** Un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$ dise que é un proceso $\mathrm{ARMA}(p, q)$ se é estacionario e cumpre

$$\begin{aligned} X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} &= Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \\ \phi(\mathrm{B}) X_t &= \theta(\mathrm{B}) Z_t, \quad \forall t \in \mathbb{Z}. \end{aligned} \tag{2.3}$$

onde $\{Z_t\} \sim WN(0, \sigma^2)$. Dise que ten media μ se $\{X_t - \mu\}$ é un proceso $\mathrm{ARMA}(p, q)$.

Na Figura A.10 podemos ver exemplos destes tres procesos estocásticos. Obsérvase a simple vista a aparencia estacionaria dos procesos.

Existencia e Unicidade de Solución Estacionaria para Procesos ARMA

Para os procesos expostos con compoñente autorregresiva, a súa definición é implícita como o proceso estocástico que cumpre a ecuación en diferenzas (2.2) ou (2.3). Polo tanto, é conveniente procurar un resultado que nos garanta, baixo certas condicións de regularidade, a existencia e unicidade de solución estacionaria para as mencionadas ecuacións en diferenzas. No caso dos procesos MA, o proceso $\{X_t\}$ vén determinado en (2.1) de forma explícita en función das innovacións $\{Z_t\}$, polo que temos garantida a súa existencia e unicidade.

No caso dos procesos AR a cousa non está tan clara. Valéndonos da recorrencia sobre X_t dada en (2.2), podemos escribir un proceso $\mathrm{AR}(p)$ como un proceso $\mathrm{MA}(\infty)$. Por exemplo, para un proceso $\mathrm{AR}(1)$ temos

$$\begin{aligned} X_t &= Z_t + \phi_1 X_{t-1} = Z_t + \phi_1 Z_{t-1} + \phi_1^2 X_{t-2} = Z_t + \phi_1 Z_{t-1} + \phi_1^2 Z_{t-2} + \phi_1^3 Z_{t-3} + \dots \\ &= \sum_{i=0}^{\infty} \phi_1^i Z_{t-i}. \end{aligned} \tag{2.4}$$

Polo tanto, se somos capaces de comprobar que o proceso $\mathrm{MA}(\infty)$ resultante é converxente, teríamos garantida a existencia e unicidade de solución estacionaria para o proceso $\mathrm{AR}(p)$ asociado. Para o caso do proceso $\mathrm{AR}(1)$ anterior, podemos comprobar polo criterio de Cauchy para a converxencia en media cadrática que se $|\phi_1| < 1$ o proceso $\mathrm{MA}(\infty)$ converxe. No caso[VI] $|\phi_1| > 1$ non temos converxencia coa representación dada en (2.4), mais si para a representación

$$\begin{aligned} X_t &= Z_t + \phi_1 X_{t-1} = -\phi_1^{-1} Z_{t+1} + \phi_1^{-1} X_{t+1} = -\phi_1^{-1} Z_{t+1} - \phi_1^{-2} Z_{t+2} - \phi_1^{-3} Z_{t+3} - \dots \\ &= -\sum_{i=0}^{\infty} \phi_1^{-i} Z_{t+i}, \end{aligned}$$

empregando os mesmos argumentos que no caso anterior. Esta solución estacionaria resulta antinatural polo feito contraintuitivo de que X_t estea correlacionado coas innovacións futuras $\{Z_s : s > t\}$. Deste xeito, introducimos naturalmente os procesos ARMA *causais* como os procesos nos que podemos asegurar que o proceso $\mathrm{MA}(\infty)$ asociado é converxente sen necesidade de involucrar ás innovacións futuras.

Definición 33 (Procesos ARMA causais)**.** Un proceso $\mathrm{ARMA}(p,q)$ dado por $\phi(\mathrm{B})X_t = \theta(\mathrm{B})Z_t$ dise *causal* se existe unha sucesión de variables aleatorias $\{\psi_i\}$ tal que $\sum_{i=0}^{\infty} |\psi_i| < \infty$ e

$$X_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i} = \psi_i B^i Z_t = \psi(B) Z_t, \quad t \in \mathbb{Z}. \tag{2.5}$$

[VI] Para o caso $|\phi_1| = 1$ non existe solución estacionaria do proceso $\mathrm{AR}(1)$ orixinal, polo que non se considera.

É importante ter en conta que a causalidade (igual que a invertibilidade, que veremos máis adiante) non é unha propiedade exclusiva de $\{X_t\}$, dependendo tamén do proceso $\{Z_t\}$. Como paso previo á caracterización dos procesos causais detémonos no seguinte resultado, que garante a convexencia e a estacionariedade do proceso resultante.

Proposición 9. Se $\{X_t\}$ é un proceso estacionario e $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$, entón $\forall i \in \mathbb{Z}$ o proceso $\{Y_t\} = \psi(B)X_t$ é convexente e estacionario con función de autocovarianza dada por

$$\gamma_Y(h) = \sum_{i,j=-\infty}^{\infty} \psi_i \psi_j \gamma_X(h - i + j).$$

Demostración. A demostración da convexencia, tanto en media cadrática como en probabilidade, pode consultarse en Brockwell e Davis (1991):Prop. 3.1.1.

Para probar a estacionariedade de $\{Y_t\}$ comprobamos que

$$\mathbb{E}(Y_t) = \sum_{i=-\infty}^{\infty} \psi_i \mathbb{E}(X_{t-i}) = \mu \sum_{i=-\infty}^{\infty} \psi_i < \infty, \text{ onde } \mu = \mathbb{E}(X_t),$$

é independente do tempo. Ademais

$$\begin{aligned}
\operatorname{Cov}(Y_{t+h}, Y_t) &= \mathbb{E}(Y_{t+h}Y_t) - \mathbb{E}(Y_{t+h})\mathbb{E}(Y_t) \\
&= \mathbb{E}\left[\left(\sum_{i=-\infty}^{\infty} \psi_i X_{t+h-i}\right)\left(\sum_{j=-\infty}^{\infty} \psi_j X_{t-j}\right)\right] + \sum_{i,j=-\infty}^{\infty} \psi_i \psi_j \mu^2 \\
&= \sum_{i,j=-\infty}^{\infty} \psi_i \psi_j \left(\mathbb{E}(X_{t+h-i}X_{t-j}) + \mu^2\right) = \sum_{i,j=-\infty}^{\infty} \psi_i \psi_j \gamma(h - i + j).
\end{aligned}$$

□

Teorema 3 (Caracterización dos procesos ARMA causais)**.** Se $\{X_t\}$ é un ARMA(p, q) e $\phi(\cdot)$ e $\theta(\cdot)$ non teñen ceros en común, entón $\{X_t\}$ é causal se, e só se, $\phi(z) \neq 0$, con $z \in \mathbb{R}$ tal que $|z| \leq 1$. Os coeficientes $\{\psi_i\}$ de (2.5) veñen determinados pola relación

$$\psi(z) = \sum_{i=0}^{\infty} \psi_i z^i = \theta(z)/\phi(z), \quad |z| \geq 1. \tag{2.6}$$

Demostración. "$\Longleftarrow$" Por hipótese podemos asegurar que existe unha expansión en serie de potencias de $1/\phi(z)$ como

$$\frac{1}{\phi(z)} = \sum_{i=0}^{\infty} \xi_i z^i = \xi(z), \quad |z| < 1 + \varepsilon.$$

Polo tanto, temos que $\xi(z)\phi(z) = 1$, para $|z| \leq 1$. Multiplicando por $\xi(B)$ a ambos lados de (2.3) obtemos a representación (2.6) que estabamos buscando, sendo o proceso causal

$$X_t = \xi(B)\theta(B)Z_t = \sum_{i=0}^{\infty} \phi_i Z_{t-i}.$$

"$\Longrightarrow$" Por hipótese, podemos escribir (2.3) como

$$\theta(B)Z_t = \phi(B)X_t \implies \theta(B)Z_t = \phi(B)\psi(B)Z_t.$$

Logo, se denotamos por $\eta(z) = \phi(z)\psi(z) = \sum_{i=0}^{\infty} \eta_i z^i$, para $|z| \leq 1$, podemos reescribir o anterior como

$$\sum_{i=0}^{q} \theta_i Z_{t-i} = \sum_{i=0}^{\infty} \eta_i Z_{t-i}.$$

Ao seguir $\{Z_t\}$ unha distribución de ruído branco podemos comparar directamente os coeficientes (equivalente a multiplicar por Z_{t-k}), tendo que $n_k = \theta_k$, para $k = 0, \dots, q$, e $\eta_k = 0$, para $k > q$, logo

$$\theta(z) = \eta(z) = \phi(z)\psi(z), \quad |z| \leq 1.$$

Polo tanto, se $\phi(z_0) = 0$ para algún $|z_0| \leq 1$, teríamos que $\theta(z_0) = 0$, chegando a unha contradición ao teren $\phi(\cdot)$ e $\theta(\boldsymbol{\theta})$ ceros en común. Entón chegamos a que $\phi(z) \neq 0$, para $|z| \leq 1$. □

De xeito análogo á condición de causalidade dos procesos ARMA, temos a condición de *invertibilidade*. Deste xeito, un proceso ARMA dirase invertible se se pode poñer como un proceso AR(∞).

Definición 34 (Procesos ARMA invertibles)**.** Un proceso ARMA(p, q) dado por $\phi(\mathrm{B})X_t = \theta(\mathrm{B})Z_t$ dise *invertible* se existe unha sucesión de variables aleatorias $\{\pi_i\}$ tal que $\sum_{i=0}^{\infty} |\pi_i| < \infty$ e

$$Z_t = \sum_{i=0}^{\infty} \pi_i X_{t-i}, \quad t \in \mathbb{Z}.$$

Teorema 4 (Caracterización dos procesos ARMA invertibles)**.** Se $\{X_t\}$ é un ARMA(p, q) e $\phi(\cdot)$ e $\theta(\cdot)$ non teñen ceros en común, entón $\{X_t\}$ é invertible se, e só se, $\theta(z) \neq 0$, con $z \in \mathbb{R}$ tal que $|z| \leq 1$. Os coeficientes $\{\pi_i\}$ de $X_t = \sum_{i=0}^{\infty} \pi_i Z_{t-i}$ veñen determinados pola relación

$$\psi(z) = \sum_{i=0}^{\infty} \pi_i z^i = \theta(z)/\phi(z), \quad |z| \geq 1.$$

Demostración. Análoga á demostración do Teorema 3. □

Deste xeito, unicamente coa hipótese de que $\phi(z)\theta(z) \neq 0$, para $|z| \leq 1$, temos garantida a existencia e unicidade de solución estacionaria e a converxencia das expresións

$$X_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i}, \qquad \text{e} \qquad Z_t = \sum_{i=0}^{\infty} \pi_i X_{t-i},$$

onde $\sum_{i=0}^{\infty} \psi_i z^i = \theta(z)/\phi(z)$ e $\sum_{i=0}^{\infty} \pi_i z^i = \phi(z)/\theta(z)$, para $|z| \leq 1$.

Predición en Procesos ARMA

A predición de futuros valores para un proceso ARMA realízase apoiándose fundamentalmente no xa visto na Sección 1.3. En particular, faremos uso do algoritmo das innovacións, visto na Proposición 6. Este aplícase sobre unha transformación do proceso, para simplificar significativamente as contas. O desenvolvemento concreto pode atoparse en Brockwell e Davis (1991):p. 175.

Como produto da aplicación do algoritmo das innovacións e do desenvolvemento posterior, sendo $m = \max(p, q)$, a predición de paso un para o proceso ARMA $\{X_t\}$ vén dada por

$$\begin{cases} \hat{X}_{T+1} = \displaystyle\sum_{i=1}^{T} \theta_{Ti}(X_{T+1-i} - \hat{X}_{T+1-i}), & 1 \leq T \leq m, \\ \hat{X}_{T+1} = \phi_1 X_T + \cdots + \phi_p X_{T+1-p} + \displaystyle\sum_{i=1}^{q} \theta_{Ti}(X_{T+1-j} - \hat{X}_{T+1-i}), & T \geq m, \end{cases}$$

ademais, o erro de predición é o seguinte

$$\mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})^2\right] = \varepsilon_T = \sigma^2 \cdot \tilde{\varepsilon}_T, \tag{2.7}$$

onde os parámetros θ_{Ti} e $\tilde{\varepsilon}_T$ obtéñense coa aplicación do algoritmo.

Predición en procesos ARMA para paso h

Cun razoamento análogo ao visto na Sección 1.3, podemos calcular as predicións para o instante $T + h$ do proceso estocástico $\{X_t : t \leq T\}$. Para iso, obtemos co procedemento para paso un (visto xusto antes) os preditores $\{\hat{X}_t : t \leq T + 1\}$, coas que calculamos a susodita predición de paso h,

$$\hat{X}_{T+h} = \begin{cases} \displaystyle\sum_{i=h}^{T+h-1} \theta_{T+h-1}(X_{T+h-i} \hat{X}_{T+h-i}), & 1 \leq h \leq m - T, \\ \displaystyle\sum_{i=1}^{p} \phi_i \hat{X}_{T+h-i} + \sum_{i=h}^{q} \theta_{T+h-1,i}(X_{T+h-i} \hat{X}_{T+h-i}), & h > m - T. \end{cases} \tag{2.8}$$

Ademais, suposto que $\{X_t\}$ sexa invertible, para valores grandes de T, cando $T \to \infty$, podemos aproximar o erro cadrático medio de predición de paso h para procesos ARMA, $\sigma_T^2(h)$,

por

$$\sigma_T^2(h) \simeq \sigma^2 \sum_{i=0}^{h-1} \psi_i^2,$$

onde $\psi(z) = \theta(z)/\phi(z)$, para $|z| \leq 1$.

Estimación para Procesos ARMA

Dado que a selección da orde, i.e., de p e q, a abordaremos no apartado Sección 2.4 no contexto máis xeral dos modelos ARIMA, centrarémonos na estimación dos parámetros necesarios para poder modelar unha serie temporal polo proceso ARMA que mellor a aproxime. Estes parámetros son a media, os coeficientes $\{\phi_i : i = 1, \dots, p\}$ e $\{\theta_i : i = 1, \dots, q\}$ e a varianza do ruído branco σ^2. Como o estimador da media é simplemente a media mostral, centrarémonos nos restantes parámetros.

Existen multitude de enfoques e metodoloxías para abordar a estimación: método do momentos, método de máxima verosimilitude, método de mínimos cadrados... Cada un cunha eficiencia, unha distribución e un grao de complexidade distinto, o que leva a que cada un deles acabe especializándose nun uso concreto. No que resta da sección faremos un repaso polos máis relevantes.

Método dos momentos. Ecuacións de Yule-Walker

Este método só resulta de utilidade para un proceso autorregresivo, i.e., con $q = 0$, xa que, para o caso dun proceso ARMA xeral, a estimación resultante non é eficiente (ten maior varianza que a obtida con outros estimadores). Alternativamente a como xa fixemos en (1.2), podemos obter as ecuacións de Yule-Walker partindo de (2.2) e multiplicando a ambos lados por X_{t-j}, con $j = 0, \dots, p$, para logo tomar esperanzas, obtendo

$$\begin{aligned}
\mathbb{E}\left(X_t X_{t-j}\right) &= \sum_{i=1}^{p} \phi_i \, \mathbb{E}\left(X_{t-i}, X_{t-j}\right) + \mathbb{E}\left(Z_t X_{t-j}\right), \quad \text{para } j = 0, \dots, p. \\
\gamma(j) &= \sum_{i=1}^{p} \phi_i \gamma(i-j) + \mathbb{E}\left(Z_t \left(\sum_{k=0}^{\infty} \psi_{j+k} Z_{t-j-k}\right)\right), \quad \text{para } j = 0, \dots, p. \\
\gamma(j) &= \sum_{i=1}^{p} \phi_i \gamma(i-j) + \cancel{\psi_t} \sigma^2 \delta_{t,t-j}, \quad \text{para } j = 0, \dots, p,
\end{aligned} \tag{2.9}$$

onde na segunda liña empregamos a hipótese de causalidade do modelo. Entón, para $j = 1, \dots, p$ temos o sistema

$$\begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix} = \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(p-1) \\ \gamma(1) & \gamma(2) & \dots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \dots & \gamma(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} \implies \Gamma_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p, \tag{2.10}$$

mentres que para o caso $j = 0$, obtemos a ecuación sobre a varianza do ruído branco seguinte

$$\sigma^2 = \gamma(0) - \sum_{i=1}^{p} \phi_i \gamma(i) = \gamma(0) - \boldsymbol{\phi}' \cdot \boldsymbol{\gamma}_p.$$

Deste xeito, substituíndo as autocovarianzas $\gamma(\cdot)$ polas súas estimacións mostrais $\hat{\gamma}(\cdot)$, chegamos á expresión das ecuacións de Yule-Walker para obter a estimación do vector de parámetros, dada por $\hat{\boldsymbol{\phi}}$, dun proceso $\mathrm{AR}(p)$. Dámonos de conta de que estamos formulando unha sorte de método dos momentos, posto que xorde de igualar os momentos de orde dous, as autocovarianzas neste caso.

Empregando agora a Proposición 2, que vimos no Capítulo 1 para a predición de procesos estacionarios, temos que se $\hat{\gamma}(0) > 0$ entón Γ_p é non singular, podendo despexar

$$\hat{\boldsymbol{\phi}} = \Gamma_p^{-1} \hat{\boldsymbol{\gamma}}_p \qquad \text{e} \qquad \hat{\sigma}^2 = \hat{\gamma}(0) - \boldsymbol{\phi}' \hat{\boldsymbol{\gamma}}_p,$$

chegando á expresión do proceso axustado[VII]

$$X_t - \hat{\phi}_1 X_{t-1} - \cdots - \hat{\phi}_p X_{t-p} = Z_t, \;\; \{Z_t\} \sim WN(0, \sigma^2).$$

Podemos comprobar ademais que a función de autocovarianza do proceso axustado terá que cumprir as ecuacións resultantes de substituír en (2.9) os parámetros estimados $\hat{\boldsymbol{\phi}}$, polo que, ao coincidir coas ecuacións de Yule-Walker, temos que coincidirá coa función de autocovarianza estimada $\hat{\gamma}(\cdot)$.

Método dos momentos. Estimación recursiva para procesos AR e MA

Na práctica teremos unha serie temporal de datos $\{x_1, \dots, x_T\}$ dun proceso que suporemos de media cero e estacionario. Logo, se $\hat{\gamma}(0) > 0$, poderemos axustar un proceso $\mathrm{AR}(m)$ para calquera $m < T$, estando este dado por

$$X_t - \hat{\phi}_{m1} X_{t-1} - \cdots - \hat{\phi}_{mm} X_{t-m} = Z_t, \;\; \text{con } \{Z_t\} \sim \mathrm{WN}(0, \hat{\varepsilon}_m). \tag{2.11}$$

[VII] Aínda que non se traballou, pode comprobarse que o polinomio $\hat{\phi}(z) = 1 - \hat{\phi}_1 z - \cdots - \phi_p z^p \neq 0$ para todo $|z| \leq 1$, polo que o proceso axustado manterá igualmente a condición de causalidade.

Polo tanto, o que faremos será axustar todos estes posibles modelos, determinando os seus parámetros $\hat{\boldsymbol{\phi}}'_m = (\hat{\phi}_{m1}, \dots, \hat{\phi}_{mm})'$ e os sucesivos erros de predición[VIII] $\hat{\varepsilon}_m$, para todos os posibles valores de m. Para salvar o elevado custo computacional que tería resolver as ecuacións de Yule-Walker para cada valor de m, podemos empregar o algoritmo de Durbin-Levinson sobre as autocovarianzas mostrais, visto na Proposición 4, para calcular as estimacións de forma recursiva, obtendo ademais, como xa vimos, a estimación da función de autocorrelación parcial, dada por $\hat{\alpha}(m) = \hat{\phi}_{mm}$.

Por último, a partir dos parámetros estimados, dos erros de predición e da análise da autocorrelación podemos determinar a idoneidade do axuste do dun modelo AR xunto coa determinación da súa orde, p.

Analogamente, se quixésemos axustar un modelo proceso MA(m) aos datos $\{x_1, \dots, x_T\}$, este estaría dado por

$$X_t = Z_t + \hat{\theta}_{m1} Z_{t-1} + \cdots + \hat{\theta}_{mm} Z_{t-m}, \;\; \{Z_t\} \sim \mathrm{WN}(0, \hat{\varepsilon}_m),$$

para todos os posibles ordes m. Podendo calcular recursivamente a estimación dos parámetros aplicando o algoritmo das innovacións novamente sobre as autocovarianzas mostrais, visto na Proposición 6. Igual que acontecía no caso autorregresivo co algoritmo de Durbin-Levinson, neste caso as estimacións obtidas xunto cunha análise da autocorrelación permite determinar a identificación do modelo como un proceso de media móbil, así como a determinación da orde, q.

Método dos momentos. Estimación preliminar para procesos ARMA

Dise "preliminar" porque a eficiencia das estimacións obtidas polo método dos momentos para procesos ARMA non é comparable á obtida a través dos outros métodos, polo que as estimacións resultantes só serán útiles como solucións aproximadas que sirvan de iterantes iniciais en esquemas numéricos.

Para un proceso ARMA(p, q) causal de media cero dado por (2.3), temos pola caracterización da causalidade (2.6) que $\psi(z)\phi(z) = \theta(z)$. Igualando os coeficientes da expresión chegamos ás ecuacións seguintes

$$\psi_j - \sum_{i=1}^{j} \phi_i \psi_{j-i} = \theta_j, \;\; \text{con } 0 \leq j < \max(p, q+1),$$
$$\psi_j - \sum_{i=1}^{p} \phi_i \psi_{j-i} = 0, \;\; \text{con } j > \max(p, q+1),$$

[VIII] Para xustificar que $\hat{\varepsilon}_m$ é o erro de predición deberíamos introducir algúns resultados que non tratamos por cuestión de tempo, extensión do traballo e dificultade dos mesmos. En todo caso, o resultado tense ao probar que $\hat{\boldsymbol{\phi}}'_m \boldsymbol{X}_m = (\hat{\phi}_{m1}, \dots, \hat{\phi}_{mm}) \cdot (X_{m,\dots,X_1})'$ é o mellor preditor lineal de X_{m+1}.

que podemos reescribir, empregando a notación de que $\theta_j = 0$ para $j > q$ e que $\phi_j = 0$ para $j > p$, como

$$\psi_0 = 1, \qquad \psi_j = \theta_j + \sum_{i=1}^{\text{mín}(j,p)} \phi_i \psi_{j-i}, \ \text{ con } j = 1, 2, \dots$$

Ao seren os coeficientes $\psi_1, \dots, \psi_{p+q}$ os dun proceso MA(∞), podemos estimalos empregando o algoritmo das innovacións, descrito no apartado anterior, obtendo as estimacións $\hat{\theta}_{m1}, \dots, \hat{\theta}_{m,p+q}$. Deste xeito chegamos ás ecuacións

$$\hat{\theta}_{mj} = \theta_j + \sum_{i=1}^{\text{mín}(j,p)} \phi_i \hat{\theta}_{m,j-i}, \ \text{ con } j = 1, 2, \dots, p+q, \tag{2.12}$$

da que buscamos obter os estimadores $\hat{\boldsymbol{\phi}}$ e $\hat{\boldsymbol{\theta}}$ dos parámetros do modelo. Para as ecuacións con $j = q+1, \dots, q+p$, de (2.12) obtemos o sistema

$$\begin{pmatrix} \hat{\theta}_{m,q+1} \\ \hat{\theta}_{m,q+2} \\ \vdots \\ \hat{\theta}_{m,q+p} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_{m,q} & \hat{\theta}_{m,q-1} & \dots & \hat{\theta}_{m,q+1-p} \\ \hat{\theta}_{m,q+1} & \hat{\theta}_{m,q} & \dots & \hat{\theta}_{m,q+2-p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}_{m,q+p-1} & \hat{\theta}_{m,q+p} & \dots & \hat{\theta}_{m,q} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix},$$

co que podemos despexar $\hat{\boldsymbol{\phi}}$, obtendo posteriormente $\hat{\boldsymbol{\theta}}$ como

$$\hat{\theta}_j = \hat{\theta}_{mj} - \sum_{i=1}^{\text{mín}(j,p)} \hat{\phi}_i \hat{\theta}_{m,j-i}, \ \ j = 1, 2, \dots, q.$$

Por último, a estimación varianza do ruído branco $\hat{\sigma}^2$ coincide co erro de predición obtido ao aplicar o algoritmo de innovacións, $\hat{\varepsilon}_m$.

Método de máxima verosimilitude e de mínimos cadrados

A estimación por máxima verosimilitude é o método máis eficiente para a estimación de procesos ARMA, é dicir, para unha mesma serie temporal estacionaria, dada polo vector $\boldsymbol{X}_T = (X_1, \dots, X_T)'$, a varianza do estimador resultante será estritamente menor á dos demais estimadores. Para calcular a verosimilitude de $\boldsymbol{X}_T$ supoñemos que este segue unha distribución normal multivariante[ix], polo que virá dada por

$$\mathcal{L}(\Gamma_T) = (2\pi)^{-T/2} (\det \Gamma_T)^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{X}_T' \Gamma_T^{-1} \boldsymbol{X}_T \right).$$

Deste xeito, os estimadores de $\boldsymbol{\phi}$, $\boldsymbol{\theta}$ e σ^2 serán aqueles que maximicen $\mathcal{L}(\Gamma_T)$. O habitual, para evitarnos o cálculo de Γ_T^{-1}, é expresar a verosimilitude en termos das innovacións dadas

[ix] Sendo entón $\{X_t : 1 \leq t \leq T\}$ un proceso gaussiano. A suposición de normalidade é menos restritiva do que semella, posto que incluso se $\{X_t\}$ non é gaussiano, a distribución asintótica dos estimadores $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$ é a mesma que no caso no que si o é, Bartlett (2007):p. 18.

polas predicións de paso un, $X_i - \hat{X}_i$, e dos seus erros de predición, $\mathbb{E}[(X_i - \hat{X}_i)^2] = \varepsilon_i$. Para isto definimos a matriz C, triangular inferior, seguinte que relaciona as observacións en función dos coeficientes do algoritmo de innovacións, é dicir, seguindo as ecuacións[x] (1.11),

$$\underbrace{\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix}}_{\boldsymbol{X}_T} = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \theta_{11} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{T1} & \theta_{T2} & \dots & 1 \end{pmatrix}}_{C} \underbrace{\begin{pmatrix} X_1 - \hat{X}_1 \\ X_2 - \hat{X}_2 \\ \vdots \\ X_T - \hat{X}_T \end{pmatrix}}_{\boldsymbol{U}_T}. \tag{2.13}$$

Se definimos ademais, $D = \operatorname{diag}(\varepsilon_1, \dots, \varepsilon_T)$, podemos comprobar como calculando as covarianzas a ambos lados de (2.13) chegamos a que $\Gamma_n = CDC'$, logo $\Gamma_n^{-1} = C^{-1\prime} D^{-1} C^{-1}$. Polo tanto,

$$\boldsymbol{X}_T' \gamma_T^{-1} \boldsymbol{X}_T = \boldsymbol{U}_T' C' \Gamma_T^{-1} C \boldsymbol{U}_T = \boldsymbol{U}_T' C' C^{-1\prime} D^{-1} C^{-1} C \boldsymbol{U}_T = \boldsymbol{U}_T' D^{-1} \boldsymbol{U}_T = \sum_{i=1}^{T} \frac{(X_i - \hat{X}_i)^2}{\varepsilon_i},$$

sendo ademais o determinante

$$\det(\Gamma_T) = \cancel{\det(C)^2} \det(D) = \varepsilon_1 \cdot \dots \cdot \varepsilon_T.$$

Polo que a verosimilitude resultante é da forma

$$\begin{aligned} \mathcal{L}(\Gamma_T) &= (2\pi)^{-T/2} (\varepsilon_1 \cdot \dots \cdot \varepsilon_T)^{-1/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{T} \frac{(X_i - \hat{X}_i)^2}{\varepsilon_i} \right) \\ &= (2\pi\sigma^2)^{-T/2} (\tilde{\varepsilon}_1 \cdot \dots \cdot \tilde{\varepsilon}_T)^{-1/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{T} \frac{(X_i - \hat{X}_i)^2}{\tilde{\varepsilon}_i} \right) = \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2), \end{aligned}$$

onde na segunda igualdade estamos empregando que $\varepsilon_i = \sigma^2 \cdot \tilde{\varepsilon}_i$, visto en (2.7). De agora en diante empregaremos a seguinte notación

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{i=1}^{T} \frac{(X_i - \hat{X}_i)^2}{\tilde{\varepsilon}_i}.$$

Como adoita ser habitual neste tipo de procedementos, continuamos tomando logaritmos

$$\log(\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)) = -\frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \log(\tilde{\varepsilon}_i) - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma^2}. \tag{2.14}$$

Derivando agora respecto de σ^2, que é independente de θ_{ij} e $\tilde{\varepsilon}_i$, obtemos a expresión para $\hat{\sigma}^2$ en función de $\hat{\boldsymbol{\phi}}$ e $\hat{\boldsymbol{\theta}}$,

$$\frac{T}{2\sigma^2} = \frac{S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{2\hat{\sigma}^4} \implies \sigma^2 = \frac{S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{T}.$$

[x] Lembramos neste punto que estamos considerando como notación $\hat{X}_1 = 0$.

Polo que deducimos que en (2.14) o terceiro termo é constante, sendo entón $\hat{\phi}$ e $\hat{\boldsymbol{\theta}}$ as solucións do problema de minimización

$$\text{minimizar} \qquad \log\left(\frac{S(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})}{T}\right)+\frac{1}{T}\sum_{i=1}^{T}\log(\tilde{\varepsilon}_i).$$

Para resolver o susodito problema empréganse métodos numéricos. Malia todo, a optimización é complexa e require que os valores iniciais dos estimadores sexan razoablemente bos. Esta é precisamente a función principal dos métodos dos momentos que vimos antes. Ademais, a modo de simplificación do problema, pódese empregar que, segundo figura en Bartlett (2007):p. 25, cando $T\to\infty$ tense que

$$\varepsilon_t\to\sigma^2 \implies \tilde{\varepsilon}_t\to 1 \implies \frac{1}{T}\sum_{i=1}^{T}\log(\tilde{\varepsilon}_i)\to 0. \tag{2.15}$$

Alternativamente, pódese considerar a estimación por "mínimos cadrados" resultante de minimizar a suma ponderada de cadrados $S(\boldsymbol{\phi},\boldsymbol{\theta})$, respecto de $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$. Igualmente, podemos observar que, segundo T vaia medrando, os estimadores de máxima verosimilitude e de mínimos cadrados van converxendo ao volverse equivalentes ambas minimizacións.

2.4. Modelos Autorregresivos Integrados de Media Móbil (ARIMA)

Nesta sección, exporemos a extensión dos modelos ARMA para series temporais que non teñan que ser necesariamente estacionarias. Para iso teremos que realizar un paso previo, diferenciando os procesos estocásticos ata conseguir que acaden a condición de estacionariedade.

Diferenciación e Integración de Series Temporais

A diferenciación dunha serie temporal $\{X_t\}$, que recibe tal nome pola similitude coa diferenciación de funcións reais de variable real, consiste na obtención dunha nova serie dada por

$$\nabla X_t = X_t - X_{t-1} = (1-\mathrm{B})X_t,$$

onde B é o operador de retardo e ∇ é o *operador de diferenciación*. Deste xeito, podemos derivar múltiples veces empregando a notación

$$\nabla^d X_t = \nabla(\nabla^{d-1}(X_t)) = (1-\mathrm{B})^d X_t,\ \ i\geq 1.$$

Diferenciando un certo número k de veces podemos eliminar tendencias polinómicas de grao k dos datos, o que permite obter procesos estocásticos de media constante.

No caso de que esteamos traballando con series estacionais, a diferenciación anteriormente descrita non será de utilidade. Polo que introducimos a diferenciación con retardo d e o seu operador diferencial asociado, ∇_d, tal que

$$\nabla_d X_t = X_t - X_{t-d} = (1 - \mathrm{B}^d)X_t.$$

Deste xeito, se tomamos $d = m$ como o período estacional da serie, estamos diferenciando entre estacións, eliminando en consecuencia a variabilidade estacional asociada. Combinando os operadores $\nabla^{\cdot}$ e $\nabla_{\cdot}$ en conxunto, poderemos eliminar a variabilidade asociada á compoñente estacional e á tendencia dunha serie temporal dada.

No caso particular dos procesos ARIMA non consideramos series con compoñente estacional, xa que as trataremos independentemente ao final desta sección. Tendo isto en conta, introducimos a seguinte definición:

Definición 35 (Serie integrada de orde d)**.** Unha serie temporal $\{X_t\}$ dise integrada con orde d, $I(d)$ se a serie resultante de diferenciala d veces, $\nabla^d X_t = (1 - \mathrm{B})^d X_t$, é estacionaria.

É habitual referirse as series $I(1)$ como series cunha *raíz unitaria*, xa que, por exemplo, no caso dun proceso AR ou ARMA o feito de ser integrada de orde 1 implica que $\phi(z) = 0$, con $|z| = 1$, tendo $\phi(\cdot)$ unha raíz.

Definición dos Procesos ARIMA

En vista do exposto no apartado anterior, xa estamos en disposición de introducir os procesos ARIMA como procesos ARMA que son aplicables a series integradas de calquera orde, previa diferenciación das mesmas, é dicir:

Definición 36 (Procesos $\mathrm{ARIMA}(p, d, q)$)**.** Un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$ dise que é un proceso $\mathrm{ARIMA}(p, d, q)$ se o proceso diferenciado d veces, $(1-\mathrm{B})^d X_t$ é un proceso $\mathrm{ARMA}(p, q)$ causal. Isto equivale a dicir que $\{X_t\}$ satisfai a ecuación en diferenzas dada por

$$\phi^*(\mathrm{B})X_t = \phi(\mathrm{B})(1 - \mathrm{B})^d X_t = \theta(\mathrm{B})Z_t, \quad \forall t \in \mathbb{Z}. \tag{2.16}$$

onde $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ e $\phi(z) \neq 0$, para $|z| \geq 1$. Dise que ten media μ se $\{X_t - \mu\}$ é un proceso $\mathrm{ARMA}(p, q)$.

Confección dun Modelo ARIMA

Temos que ter en conta que, para unha serie temporal $\{X_t\}$ non necesariamente estacionaria, as casuísticas que se poden dar á hora de estacionarizar a serie son moi diversas, polo que requiren dunha análise previa e sistemática que vai máis alá da diferenciación da mesma.

O primeiro paso, como en calquera análise de series temporais, é representar graficamente a serie buscando signos que denoten falta de estacionariedade: tendencia, estacionalidade, heterocedasticidade, valores atípicos, saltos bruscos... En base a isto podemos determinar a aplicación de transformacións para homoxeneizar a varianza, como as vistas na Figura A.4.

Tests de estacionariedade

En presenza de tendencia, o que implica que a función de autocorrelación diminúe lentamente, procédese diferenciando a serie. Para determinar a orde de diferenciación pódense empregar tests sobre a estacionariedade da serie, coñecidos como tests de raíz unitaria. Os máis comúns son o *test de Kwiatkowski-Phillips-Schmidt-Shin* (KPSS) cuxa hipótese nula é que a serie sexa estacionaria (podendo ter unha tendencia determinista), o *test de Phillips-Perron* ou o *test de Dickey-Fuller* cuxa hipótese nula é que a serie teña raíces unitarias, non sendo estacionaria. Deste xeito, podemos ir realizando diferenciacións na serie ata que poidamos comprobar a súa estacionariedade mediante os tests anteriores.

Identificación e selección da orde do modelo

Unha vez estacionarizada a serie podemos proceder como vimos na sección anterior, Sección 2.3. Deste xeito, estimaremos os parámetros ϕ, $\boldsymbol{\theta}$ e σ^2 por máxima verosimilitude para varias ordes (p, q) empregando criterios de información como o *Criterio de Información de Akaike* (AIC)

$$AIC = -2\ln(\mathcal{L}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)) + 2(p + q + 1),$$

o *AIC Corrixido* (AICC), indicado cando T é pequeno en comparación co número de parámetros $(p + q + 1)$,

$$AICC = -2\ln(\mathcal{L}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)) + 2(p + q + 1)\underbrace{\frac{T}{T - p - q - 2}}_{\text{termo corrector}},$$

ou o *Criterio de Información de Bayes* (BIC)

$$BIC = -2\ln(\mathcal{L}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)) + (p + q + 1)(\ln(T) - 2).$$

Estes criterios penalizan aos modelos con valores altos de p e q, favorecendo aos modelos máis simples.

Diagnose e Validación

Unha vez axustado un modelo ARMA(p, q) sobre a serie temporal estacionarizada $\{X_t\}$, só resta comprobar que os residuos, que para os estimadores de máxima verosimilitude veñen

dados por

$$\hat{W}_t = \frac{X_t - \hat{X}_t(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{\sqrt{\hat{\varepsilon}_t(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}}, \quad t = 1, \dots, T,$$

son consistentes co modelo escollido, é dicir, seguen unha distribución $\mathrm{WN}(0, \hat{\sigma}^2)$ de forma aproximada (xa que son os erros e non os residuos os que teñen unha distribución de ruído branco), sendo, polo tanto, incorrelados. Deste xeito, a función de autocorrelación mostral da serie dos valores axustados $\{\hat{X}_t\}$ non debera ser significativamente distinta de cero, polo que, novamente, existen tests para contrastar dita hipótese. Se cadra, o máis coñecido é o *test de Ljung-Box* proposto en Ljung e Box (1978), que xurdiu como mellora do *test de Box-Pierce* proposto oito anos antes en Box e Pierce (1970), e contrasta a hipótese nula de que a serie estea incorrelada, empregando o estatístico

$$Q = T(T+2) \sum_{i=1}^{h} \frac{\hat{\rho}_h^2}{T-i},$$

onde h é o número de retardos considerados para o test, que segue baixo a hipótese nula unha distribución $\chi^2(h-p-q)$.

Malia que neste traballo só consideramos as innovacións dunha distribución de ruído branco, $\mathrm{WN}(0, \sigma^2)$, tamén é habitual considerar erros *independentes e identicamente distribuídos*, i.i.d. Un dos motivos é que dita condición actúa como hipótese necesaria en varios resultados sobre a converxencia asintótica dos estimadores dos parámetros dos modelos ARMA.

Entón, se en (2.16) supoñemos $\{Z_t\}$ i.i.d. con distribución normal $\mathrm{N}(0, 1/T)$, os residuos $\{\hat{W}_t\}$ seguirán aproximadamente[xi] a mesma distribución, polo que, non séndoo tecnicamente, estarán próximos a ser i.i.d. Logo, para comprobar que $\{\hat{W}_t\}$ *son* i.i.d. existen unha variedade de tests que se coñecen como *tests de aleatoriedade*. Describimos a continuación algúns deles.

Test dos puntos de inflexión Este baséase na frecuencia de aparición de *puntos de inflexión*[xii], máximos ou mínimos locais tal que $\hat{W}_{t-1} \leqslant \hat{W}_t$ e $\hat{W}_{t-1} \geqslant \hat{W}_t$, para $2 \leq t \leq T-1$. Sexa P o número de puntos de inflexión de $\{\hat{W}_1, \dots, \hat{W}_T\}$. Entón, se $\{\hat{W}_t\}$ fosen i.i.d., a probabilidade ter un punto de inflexión no instante t sería de $2/3$, polo que

$$\mu_P = \mathbb{E}(P) = \frac{2}{3}(T-2), \qquad \text{e} \qquad \sigma_P^2 = \mathrm{Var}(P) = \frac{16T-29}{90}.$$

[xi] Segundo se expón en Brockwell e Davis (1991):p. 308 a distribución dos residuos $\{\hat{W}_t\}$ neste caso é asintoticamente normal con media cero cunha varianza heterocedástica que para valores pequenos de t é menor que $1/T$, achegándose a este valor cando t aumenta.

[xii] Do inglés, «*turning points*».

Considérase entón un contraste onde a hipótese nula é a aleatoriedade da serie, i.e., que $\{\hat{W}_t\}$ sexa i.i.d., baseándose en que, baixo a hipótese nula, P segue asintoticamente unha normal $\mathrm{N}(\mu_P, \sigma_P^2)$.

Test do cambio de signo Este emprega a frecuencia coa que a serie derivada cambia de signo, é dicir, a frecuencia coa que $\hat{W}_i > \hat{W}_{i-1}$, para $2 \leq t \leq T$. Sexa S o número de cambios de signo da serie derivada. Se a serie dos residuos é aleatoria, a probabilidade de que haxa un cambio de signo no tempo t sera de $1/2$, polo que

$$\mu_S = \mathbb{E}(S) = \frac{1}{2}(T-1), \qquad \text{e} \qquad \sigma_S^2 = \mathrm{Var}(S) = \frac{T-1}{12}.$$

En consecuencia, formúlase un contraste onde a hipótese nula é a aleatoriedade de $\{\hat{W}_t\}$, empregando que baixo a hipótese nula S segue asintoticamente unha normal $\mathrm{N}(\mu_S, \sigma_S^2)$.

Test do rango Este defínese a partir da frecuencia de aparición de pares de índices (i, j) tal que $\hat{W}_j > \hat{W}_i$, con $j > i$ para $i = 1, \ldots, T-1$. Sexa R o número de pares de índices que cumpre a condición anterior. Existen $\binom{T}{2} = T(T-1)/2$ pares de índices tal que $j > i$, e baixo a aleatoriedade da serie, a probabilidade de que $\hat{W}_j > \hat{W}_i$ é de $1/2$, polo que

$$\mu_R = \mathbb{E}(R) = \frac{1}{4}T(T-1), \qquad \text{e} \qquad \sigma_R^2 = \mathrm{Var}(R) = \frac{T(T-1)(2T+5)}{8}.$$

Polo tanto, seguindo a secuencia dos casos anteriores, establécese un contraste onde a hipótese nula é que a serie $\{\hat{W}_t\}$ sexa aleatoria, servíndonos de que baixo a hipótese nula, R segue asintoticamente unha normal $\mathrm{N}(\mu_R, \sigma_R^2)$.

Modelos SARIMA

Ata agora non tratamos sobre a aplicación dos modelos ARMA a series que, non só non son estacionarias, senón que son estacionais. Veremos a continuación de forma introdutoria como ser capaces de caracterizar tamén a variabilidade estacional a partir das técnicas xa vistas para o modelado de series temporais con procesos ARMA e procesos ARIMA.

Para poder modelar esta estacionalidade teremos que empregar simultaneamente dous modelos ARMA. Denotaremos por s o período estacional (número de estacións) da serie. Primeiramente, formulamos un modelo "intra-estacional" dado por un modelo $\mathrm{ARMA}(P, Q)$ para cada subserie estacional:

$$\Phi(\mathrm{B}^s)X_t = \Theta(\mathrm{B}^s)U_t,$$

con $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$, $\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$ e $\{U_{j+12t} \sim \mathrm{WN}(0, \sigma_U^2)\}$ para cada $j = 1, \ldots, s$. Desta forma, cada subserie estacional estaría xerada polo mesmo proceso

ARMA. Porén, dada a distribución de ruído branco das innovacións, este modelo non permite que exista correlación entre as distintas estacións, o cal na práctica resulta dificilmente crible. Para modelar esta relación consideramos un novo modelo "inter-estacional" entre as innovacións das distintas estacións. Este novo modelo será un ARMA(p,q), de xeito que o proceso $\{U_t\}$ virá dado por

$$\phi(\mathrm{B}^s)U_t = \theta(\mathrm{B}^s)Z_t,$$

con $\phi(z) = 1-\phi_1 z-\cdots-\phi_p z^p$, $\theta(z) = 1+\theta_1 z+\cdots+\theta_q z^q$ e $\{Z_t\} \sim \mathrm{WN}(0,\sigma^2)$. Considerando este modelo, permítese que haxa correlación entre valores consecutivos de U_t, non só captando o patrón estacional da serie, senón tamén entre as innovacións pertencentes á mesma subserie estacional (como: $U_1, U_{1+s}, U_{1+2s}, \ldots$), dando pé a que, deste xeito, o patrón estacional varíe entre cada un dos seus ciclos.

Poñendo en común o razoamento anterior xunto coa teoría xa vista de diferenciación para a estacionarización de series temporais, estamos en disposición de introducir a definición dos proceso SARIMA.

Definición 37 (Procesos SARIMA(p,d,q))**.** Un proceso estocástico $\{X_t : t \in \mathbb{Z}\}$ dise que é un proceso SARIMA$(p,d,q)\times(P,D,Q)_s$ con período s se o proceso diferenciado

$$Y_t := (1-\mathrm{B})^d(1-\mathrm{B}^s)^D X_t = \nabla^d \nabla_s^D X_t,$$

é un proceso ARMA(p,q) causal tal que

$$\phi(\mathrm{B})\Phi(\mathrm{B}^s)Y_t = \theta(\mathrm{B})\Theta(\mathrm{B}^s)Z_t, \quad \forall t \in \mathbb{Z}.$$

onde $\{Z_t\} \sim \mathrm{WN}(0,\sigma^2)$ e $\phi(z) \neq 0$ e $\Phi(z) \neq 0$, para $|z| \geq 1$. Os polinomios $\phi(z)$, $\Phi(z)$, $\theta(z)$ e $\Theta(z)$ coinciden cos definidos nos parágrafos anteriores.

Predición con Modelos ARIMA

Se $d = 0$, a teoría de predición concorda exactamente coa vista na Sección 2.3. Para valores de $d \geq 1$ debemos ter en conta que a ecuación (2.16) non determina a media $\mathbb{E}(X_t)$ nin o momento de orde dous $\mathbb{E}(X_{t+h}X_t)$ debido a que está diferenciada. Para resolver a indeterminación anterior é necesario engadir unha nova hipótese, de xeito que, sendo $\{Y_t\}$ o proceso estocástico "derivado"

$$Y_t = (1-\mathrm{B})^d X_t, \quad t = 1, 2, \ldots, \tag{2.17}$$

onde ademais $\{Y_t\}$ é un proceso ARMA(p,q) causal, o vector[xiii] $(X_{1-d}, \ldots, X_0)'$ está incorrelado con Y_t, para todo $t > 0$. Deste xeito, despexamos X_t na ecuación (2.17) empregando o

[xiii] Realízase, de ser necesario, un reetiquetado da serie temporal $\{X_t\}$ para garantir que os valores observados da mesma son: $\{X_{1-d}, X_{2-d}, \ldots, X_T\}$, sendo en consecuencia os valores observados da serie diferenciada: $\{Y_1, \ldots, Y_T\}$.

binomio de Newton como

$$X_t = Y_t - \sum_{i=1}^{d} \binom{d}{i} (-1)^i X_{t-i}, \quad t = 1, 2, \dots . \tag{2.18}$$

A partir dela, buscamos atopar o mellor preditor lineal de X_{T+h}, $\hat{X}_{T+h}$, polo que, definindo os espazos de Hilbert: $S_T = \text{span}\{X_{1-d}, \dots, X_T\}$, $P_T = \text{span}\{X_{1-d}, \dots, X_0\}$ e $R_T = \text{span}\{Y_1, \dots, Y_T\}$, temos que virá dado por $\hat{X}_{T+h} := \text{P}_{S_T} X_{T+h}$. Vemos agora que pola hipótese de incorrelación entre Y_t e $(X_{1-d}, \dots, X_0)$ temos que $P_T \perp R_T$, polo que, tomando proxeccións sobre (2.18) obtemos

$$\begin{aligned}
\hat{X}_{T+h} := \text{P}_{S_T} X_{T+h} &= \text{P}_{S_T} Y_{T+h} - \sum_{i=1}^{d} \binom{d}{i} (-1)^i \text{P}_{S_T} X_{T+h-i} \\
&\overset{❶}{=} \cancel{\text{P}_{P_T} Y_{T+h}} + \text{P}_{R_T} Y_{T+h} - \sum_{i=1}^{d} \binom{d}{i} (-1)^i \text{P}_{S_T} X_{T+h-i} \\
&\overset{❷}{=} \text{P}_{R_T} Y_{T+h} - \sum_{i=1}^{d} \binom{d}{i} (-1)^i \text{P}_{S_T} X_{T+h-i}, \quad t = 1, 2, \dots ,
\end{aligned}$$

onde en ❶ empregamos que $P_T \perp R_T$ e en ❷ empregamos a hipótese en incorrelación. Deste xeito, as predicións $\hat{Y}_{T+h} = \text{P}_{R_T} Y_{T+h}$ poden obterse directamente a partir de (2.8), mentres que os preditores $\hat{X}_{T+1}$, $\hat{X}_{T+2}$, ... , obténense recursivamente a partir de (3.1). Deste xeito, as ecuacións de predición de paso h para procesos ARIMA veñen dadas por

$$\begin{aligned}
\hat{X}_{T+h} &= \sum_{i=1}^{p} \phi_i \hat{Y}_{T+h-i} + \sum_{i=h}^{q} \theta_{T+h-1,i} (X_{T+j-i} - \hat{X}_{T+j-i}) \\
&= \sum_{i=1}^{p+d} \phi_i^* \hat{X}_{T+h-i} + \sum_{i=h}^{q} \theta_{T+h-1,i} (X_{T+j-i} - \hat{X}_{T+j-i}),
\end{aligned}$$

para $h \geq 1$ e $T > \text{máx}(p, q)$, onde $\phi^*(z) = (1 - z)^d \phi(z)$. Neste caso, de forma análoga ao que pasaba coa predición para procesos ARMA, se $\{X_t\}$ é invertible, para valores grandes de T, cando $T \to \infty$, podemos aproximar o erro cadrático medio de predición con paso h para procesos ARIMA, $\sigma_T^2(h)$, por

$$\sigma_T^2 \simeq \sigma^2 \sum_{i=0}^{h-1} \psi_i^2,$$

onde $\psi(z) = \sum_{i=0}^{\infty} \psi_i z^i = (\phi^*(z))^{-1} \theta(z)$, para $|z| < 1$.

Podemos algúns exemplos de predición de series temporais con modelos ARIMA na Figura A.11 e na Figura A.12.

Capítulo

3 Modelización. Metodoloxías de Aprendizaxe Automática

Neste capítulo trataremos tres metodoloxías relativamente recentes de predición e regresión para series temporais. Comezando polos Procesos Gaussianos, que constitúen unha metodoloxía bayesiana de regresión tipo kernel, pasando polas redes neurais recorrentes e as súas variantes, ata chegar aos transformers, que a día de hoxe se atopan á vangarda na predición de series temporais.

3.1. Procesos Gaussianos (GPs)

A diferenza do visto ata agora, os Procesos Gaussianos (GP) conforman unha metodoloxía de regresión non paramétrica e bayesiana. É importante diferenciar os procesos gaussianos como procesos estocásticos cunha función de distribución gaussiana,

Definición 38 (Proceso gaussiano)**.** Dise do proceso estocástico cuxas funcións de distribución son normais multivariantes[1].

dos Procesos Gaussianos (de agora en diante GP) como método para asignar probabilidades sobre un espazo de funcións, que serán obxecto de estudo desta sección. Os GPs son de uso xeral, non estando pensados especificamente para o modelado de series temporais, aínda que si é este un dos seus principais campos de aplicación. Unha das súas principais vantaxes é que ofrecen predicións distribucionais e non só puntuais.

Estatística Bayesiana *vs.* Frecuentista

Na estatística distínguense dous enfoques ou paradigmas que son equivalentes: o frecuentista e o bayesiano. O enfoque teórico dos GPs emprega un paradigma bayesiano. A diferenza entre ambos ten un punto filosófico, e radica na interpretación que se fai do concepto de probabilidade. Na práctica ambos enfoques producen os mesmos resultados, aínda que en función do problema a tratar un dos dous adoita resultar máis axeitado.

[1] Nos procesos gaussianos temos que a estacionariedade débil implica a estacionariedade forte.

Por unha banda, na estatística frecuentista defínese a probabilidade, obxectivamente, como a frecuencia de ocorrencia dun suceso a longo prazo, é dicir, tras múltiples realizacións dun experimento concreto. Neste sentido, os parámetros que definimos sobre os datos, como a media ou a varianza, son valores descoñecidos e fixos, que podemos estimar.

Por outra banda, a estatística bayesiana define a probabilidade, dun xeito un tanto máis subxectivo, como o grao de confianza ou de certeza dun evento, sendo este susceptible a cambiar coa presenza de nova información dispoñible sobre o mesmo. Ademais, neste caso, os parámetros trátanse como variables aleatorias, tendo as súas propias distribucións de probabilidade e representando estas a incerteza que se ten sobre os mesmos. Esta definición máis aberta permite un maior grao de flexibilidade metodolóxica a cambio dunha maior complexidade tanto teórica como computacional.

É habitual na estatística bayesiana partir dunha distribución de probabilidade "previa", *prior*, que se lle supón aos datos e que se transforma ao ter en conta nova información na distribución de probabilidade "posterior", *posterior*.

Definición e Aplicacións a Series Temporais

Como vimos antes, os GPs poden aplicarse a conxuntos de datos calquera, $\{(\boldsymbol{x}_i, y_i)\}$, mais tendo en conta os obxectivos deste traballo, a formulación empregada estará adaptada ao contexto das series temporais. Unha formulación máis xeral dos GPs pódese atopar en Williams e Rasmussen (2006) ou en Murphy (2023). Deste xeito, consideramos unha serie temporal $\{X_t : 1 \leq t \leq T\} \equiv \{(t, X_t) : 1 \leq t \leq T\}$, que supoñemos coñecida, cumprindo que

$$X_t = f(t) + \varepsilon_t, \text{ tal que } \varepsilon_t \sim \mathrm{N}(0, \sigma^2),$$

onde $f : \mathbb{R} \to \mathbb{R}$ é unha función descoñecida que modela os datos da serie temporal e ε_t e o erro. Se non tiveramos en conta os erros ε_t no modelo, estariamos obrigando á función f a interpolar os datos. O noso obxectivo será o de obter a distribución de f para os puntos $\{T + h : 1 \leq h \leq H\}$, dándonos isto a predición para a distribución dos novos valores da serie, $\hat{X}_{T+h} = f(T + h)$.

Para obter a distribución de $f(T + h)$ estudaremos o "grao de semellanza" de $T + h$ respecto de $\{1, \dots, T\}$, para poder valernos así dos valores coñecidos $\{X_t : 1 \leq t \leq T\}$, mediante o uso da función kernel, $K(\cdot, \cdot)$, que será a que determine a tipoloxía das funcións f que consideraremos. A elección do kernel, así como dos seus parámetros, será determinante na predición obtida e tratarémola máis adiante na Sección 3.1. Coa notación

$$\mathcal{T} = (1, \dots, T), \quad \mathcal{T}^* = (T + 1, \dots, T + H), \quad \boldsymbol{f} = (f(1), \dots, f(T)), \quad \boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_T),$$
$$\boldsymbol{X} = (X_1, \dots, X_T) = (f(1) + \varepsilon_1, \dots, f(T) + \varepsilon_T), \quad \boldsymbol{X}_* = (f(T + 1), \dots, f(T + H)),$$

e definindo as matrices de covarianza

$$\mathbf{K}_{\bullet\bullet} = [K(i,j)]_{i,j\in\mathcal{T}}\,,\ \ \mathbf{K}_{\bullet *} = [K(i,j)]_{i\in\mathcal{T}}^{j\in\mathcal{T}^*}\,,\ \ \mathbf{K}_{*\bullet} = [K(i,j)]_{j\in\mathcal{T}}^{i\in\mathcal{T}^*}\,,\ \ \mathbf{K}_{**} = [K(i,j)]^{i,j\in\mathcal{T}^*}\,.$$

Estamos en disposición de introducir a hipótese fundamental na que se basean os GPs, que é a de supoñerlle unha distribución normal multivariante ao vector formado por $\boldsymbol{X}$ e $\boldsymbol{X}^*$, é dicir, estamos supondo que o vector $(X_1,\dots,X_T,\hat{X}_{T+1},\dots,\hat{X}_{T+H})$ formado polos valores observados e polas predicións segue unha distribución normal multivariante, sendo este entón un proceso gaussiano. En particular, sendo $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X})$ e $\boldsymbol{\mu}_* = \mathbb{E}(\boldsymbol{X}^*)$, temos

$$\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}_*\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}_*\end{pmatrix}, \begin{pmatrix}\mathbf{K}_{\bullet\bullet} + \sigma^2\boldsymbol{I}_T & \mathbf{K}_{\bullet *}\\ \mathbf{K}'_{*\bullet} & \mathbf{K}_{**}\end{pmatrix}\right). \tag{3.1}$$

Isto permítenos obter a distribución de $\boldsymbol{X}^*$ condicionando a distribución $(T+H)$-normal anterior. Precisamente, por tratarse dunha distribución normal, podemos obter a distribución condicionada de forma explícita como

$$\boldsymbol{X}_* \mid \boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{\mu}_*\big|_X, \boldsymbol{\Sigma}_*\big|_X\right),$$

onde, sendo $\tilde{\boldsymbol{K}}_{\bullet\bullet} = \mathbf{K}_{\bullet\bullet} + \sigma^2\boldsymbol{I}_T$,

$$\begin{aligned}\boldsymbol{\mu}_*\big|_X &= \boldsymbol{\mu}_* + \boldsymbol{K}'_{\bullet *}\tilde{\boldsymbol{K}}_{\bullet\bullet}^{-1}(X-\boldsymbol{\mu}),\\ \boldsymbol{\Sigma}_*\big|_X &= \boldsymbol{K}_{**} - \boldsymbol{K}'_{\bullet *}\tilde{\boldsymbol{K}}_{\bullet\bullet}^{-1}\boldsymbol{K}_{\bullet *}.\end{aligned}$$

Como podemos comprobar, o cálculo da distribución condicionada de $\boldsymbol{X}^*$ require da inversión dunha matriz $T\times T$, o cal non anticipa un rendemento computacional moi bo para o método. Unha das contrapartidas dos GPs é que ten $\mathcal{O}(T^3)$ en tempo, podendo mellorarse aplicando esquemas aproximados ata $\mathcal{O}(TH^2)$.

Enxeñería dos Kernels

Pasamos a falar agora dos kernels que, como vimos antes, condensan a capacidade preditiva dos GPs, xa que, en función do kernel empregado, as funcións que obteremos como realizacións de (3.1) terán unha forma ou outra, incidindo por ende na precisión das predicións obtidas.

O habitual cando consideramos índices $\mathcal{T}$ reais, como é o noso caso, é empregar kernels estacionarios, i.e., tales que $K(i,j) = K(r)$, onde $r = |i-j|$. Deste xeito, o "grao de semellanza" a priori entre dúas observacións só dependerá da distancia dos seus índices. Pasamos a enunciar algúns dos kernels máis comúns. O kernel dado pola *Función Radial Básica* (do inglés: «*Radial Basis Function*», RBF) adoita ser o máis común, polo menos no que aos exemplos da bibliografía se refire. Defínese como

$$K(r,\ell) = \sigma^2\exp\left(-\frac{r^2}{2\ell^2}\right),$$

onde ℓ é o parámetro de escala do kernel. É habitual referirse ao kernel RBF con multitude de outros nomes como "kernel exponencial cadrático" ou "kernel gaussiano". Para modelar series temporais cunha compoñente estacional empréganse kernels periódicos como o seguinte

$$K(r, \ell, p) = \sigma^2 \exp\left(-\frac{2}{\ell^2}\operatorname{sen}^2\left(\pi\frac{r}{p}\right)\right),$$

con ℓ como parámetro de escala e p como a periodicidade das oscilacións. Aínda que existe multitude de kernels, na práctica o máis habitual é, a partir de kernels coñecidos, combinalos (sumándoos ou multiplicándoos) para obter novos kernels que se axusten mellor aos datos. Á súa vez, a partir dun kernel $\overline{K}(i,j)$ dado, podemos obter outro mediante as seguintes operacións

- $K(i,j) = c\overline{K}(i,j), \quad \forall c \in \mathbb{R}^+$.
- $K(i,j) = f(i)\overline{K}(i,j)f(j)$, para toda función f.
- $K(i,j) = q(\overline{K}(i,j))$, sendo q un polinomio con coeficiente principal non negativo.
- $K(i,j) = \exp(\overline{K}(i,j))$.

Malia todo, esta é a parte que supón un maior reto á hora de axustar GPs, posto que non hai unha boa forma de automatizar a elección do kernel, dependendo esta da subxectividade de quen especifica o modelo.

Estimación dos Parámetros dun Proceso Gaussiano

Unha vez especificada unha función kernel, só resta estimar os parámetros do mesmo para que se axusten o mellor posible aos datos. De xeito semellante a como procediamos cos métodos clásicos, trataremos de maximizar a verosimilitude. Mais neste caso, como o método é bayesiano, ao maximizar a verosimilitude, realmente estamos maximizando a verosimilitude *marxinal*, é dicir, a probabilidade dos datos condicionada respecto dos parámetros dados polo vector $\boldsymbol{\theta}$ (ata aquí nada novo), pero tendo tamén en conta a probabilidade de cada unha das realizacións (funcións f) segundo o *prior*. Polo tanto, ao calcular a verosimilitude, non estamos asumindo ningunha función f como verdadeira, senón integrando respecto de todas elas ponderadas polas súas respectivas probabilidades. Matematicamente defínese como

$$\mathbb{P}\left(\boldsymbol{X} \mid \mathcal{T}, \boldsymbol{\theta}\right) = \int \mathbb{P}\left(\boldsymbol{X} \mid \boldsymbol{f}, \sigma^2\right) \mathbb{P}\left(\boldsymbol{f} \mid \mathcal{T}, \boldsymbol{\theta}\right) d\boldsymbol{f}.$$

Ao estar traballando con procesos gaussianos e seguir, polo tanto, unha distribución normal, podemos calcular a integral anterior explicitamente. Tomando logaritmos tense

$$\begin{aligned}\log \mathbb{P}\left(\boldsymbol{X} \mid \mathcal{T}, \boldsymbol{\theta}\right) &= \mathrm{N}(\boldsymbol{X} \mid \mathbf{0}, \tilde{\boldsymbol{K}}_{\bullet\bullet}) \\ &= -\frac{1}{2}\underbrace{(\boldsymbol{X}-\boldsymbol{\mu})'\tilde{\boldsymbol{K}}_{\bullet\bullet}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}_{❶} - \frac{1}{2}\underbrace{\log\left|\tilde{\boldsymbol{K}}_{\bullet\bullet}\right|}_{❷} - \frac{T}{2}\log(2\pi).\end{aligned}$$

Vemos que en ❶ temos a distancia de Mahalanobis entre $\boldsymbol{X}$ e $\boldsymbol{\mu}$ ao cadrado, penalizando este termo que as funcións f xeradas polos parámetros $\boldsymbol{\theta}$ non se axusten aos datos. Mentres que en ❷ temos o determinante da matriz de covarianza dada polo kernel empregado, que á súa vez depende dos parámetros $\boldsymbol{\theta}$. Polo tanto, o termo ❷ penaliza que o determinante de $\tilde{\boldsymbol{K}}_{\bullet\bullet}$ sexa alto, o cal se relaciona cunha maior complexidade do modelo.

Exemplos de Modelado con Procesos Gaussianos

Na Figura A.13 hai varios exemplos de GPs aplicados á predición de series temporais. Neles, a xeneralidade da formulación dos GPs permítelles ser moi flexibles á hora de tratar con datos faltantes.

Como complemento, é moi recomendable o recurso web Görtler, Kehlbeck e Deussen (2019), onde se fai un repaso introdutorio dos fundamentos matemáticos sobre os que asentan os GPs acompañados de múltiples representacións interactivas dos mesmos que clarifican, por exemplo, a relación do determinante da matriz de covarianza coa complexidade do modelo ou a combinación de distintos tipos de kernels.

3.2. Redes Neurais Recorrentes (RNNs)

Xa dende os anos 50s e 60s comezouse a traballar no desenvolvemento de modelos baseados en redes neurais como aproximadores universais de funcións. Por cuestións de espazo e tempo non imos facer unha introdución "paso a paso" das redes neurais usuais (en inglés: *«feedforward neural networks»*). Suponse entón certo coñecemento básico sobre o funcionamento dunha rede neural. Para

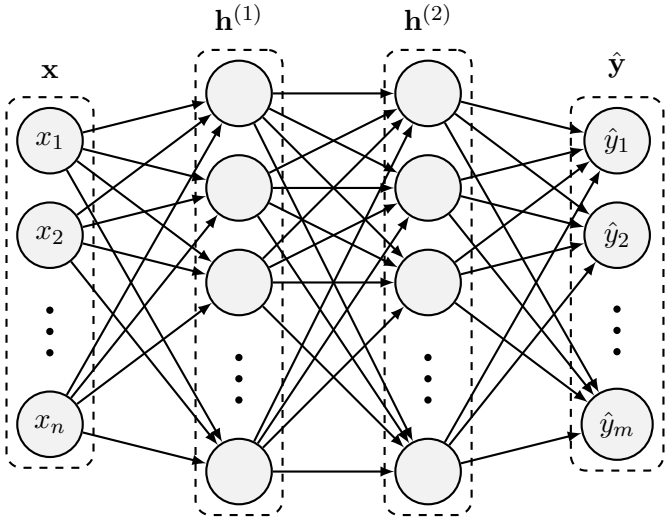

Figura 3.1: Esquema dunha rede neural con dúas capas ocultas. Fonte: Love (2024).

refrescarnos a memoria, podemos ver na Figura 3.1 unha representación dunha rede neural con dúas capas ocultas.

Volvendo ás series temporais, estes avances levaron a que a finais dos anos 80 se propuxeran as redes neurais recorrentes (RNN). Estas especialízanse no tratamento de secuencias de datos. Protagonizaron a primeira gran irrupción das redes neurais no campo, ata a chegada dos transformers en 2017 como veremos na Sección 3.4.

Cabe mencionar que non hai unha única definición posible para o que é unha rede neural recorrente. Existen multitude de arquitecturas posibles considerando distintas conexións entre os nodos da rede. Deste xeito, seguindo o criterio empregado na Sección 3.1, empregaremos unha notación adaptada ao contexto deste traballo, podendo consultarse unha formulación máis xeral das mesmas en Bengio, Goodfellow, Courville et al. (2017):pp. 373-420.

Consideraremos entón unha serie temporal de datos dada polo vector de observacións $\boldsymbol{x} = (x_1, \dots, x_T)$, e buscaremos predicir o valor de x_{T+1} por $\hat{x} = \hat{x}_{T+1}$. Podemos ver na Figura 3.2 un esquema dunha rede neural recorrente básica asociada a este problema.

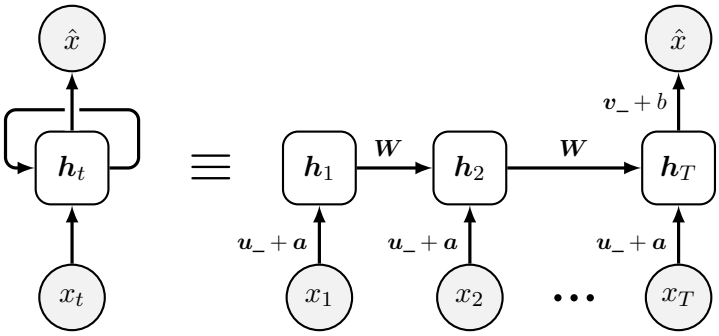

Figura 3.2: Esquema dunha rede neural recorrente. Adaptación de Love (2024).

Como podemos observar, as principais diferenzas das RNNs respecto do que viamos na Figura 3.1 son o esquema recorrente e a compartición de parámetros. Neste caso, unicamente temos un vector oculto por cada elemento da serie $\{\boldsymbol{h}_1, \dots, \boldsymbol{h}_T\}$ tal que $\boldsymbol{h}_t \in \mathbb{R}^\tau$, actualizándose cada un deles en función do seu valor anterior e dun novo valor da serie temporal. Desta forma, os vectores $\boldsymbol{h}_t$ actúan como unha memoria dos valores anteriores da serie. Ademais, os parámetros empregados para calcular os vectores $\boldsymbol{h}_t$ son comúns a toda a rede, facendo que o cálculo de $\boldsymbol{h}_t$ sexa un proceso recorrente. Describimos a continuación as ecuacións que definen á rede da Figura 3.2, definindo polo camiño os parámetros da mesma:

$$\begin{aligned} \boldsymbol{h}_t &= \tanh\left(\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{u}x_t + \boldsymbol{a}\right), && \text{onde } \boldsymbol{W} \in \mathcal{M}_{T\times\tau} \text{ e } \boldsymbol{u}, \boldsymbol{a} \in \mathbb{R}^\tau, \\ \hat{x} &= \operatorname{softmax}\left(\boldsymbol{v}\cdot\boldsymbol{h}_T + b\right), && \text{onde } \boldsymbol{v} \in \mathbb{R}^\tau \text{ e } b \in \mathbb{R}. \end{aligned}$$

Deste xeito, o cálculo recursivo de $\boldsymbol{h}_t$ faise aplicándolle a $\tanh$ a unha función afín en $\boldsymbol{h}_{t-1}$ e x_t. Hai que ter en conta que nos modelos de redes neurais é case obrigado traballar con datos normalizados, para evitar que a rede deba contrarrestar a magnitude de certas entradas.

É por iso polo que na saída da mesma se emprega a función softmax, que como no noso caso estamos traballando con escalares, equivale a unha función loxística. Podemos recordar a gráfica de ambas funcións de activación na Figura 3.3.

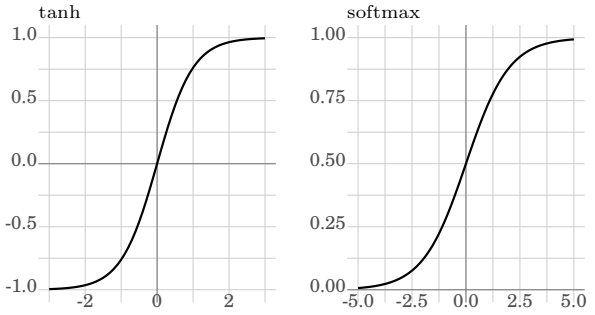

Figura 3.3: Gráfica das funcións tanh e softmax. Elaboración propia con ggplot2 en R.

Para levar a cabo a estimación dos parámetros do modelo, emprégase o algoritmo de retropropagación no tempo (BPTT) que podemos ver representado na Figura 3.4. Mediante o BPTT calculamos dende o final da rede "cara atrás" as derivadas da función de perda L, que pode estar determinada polo MSE de predición ou pola menos a log-verosimilitude, respecto dos parámetros do modelo. Todo isto co obxectivo de aplicar posteriormente algún algoritmo de descenso do gradiente que axuste os parámetros do modelo para minimizar o valor de L.

En xeral, o uso principal das redes neurais recorrentes é o *«sequence-to-sequence»*, isto é, tomar unha secuencia ou serie de datos de entrada, e obter outra secuencia de datos de saída, $\boldsymbol{y} = (y_1, \ldots, y_R)$. Neste sentido, un dos primeiros usos que recibiron este tipo de redes foi o da tradución de textos, nos que entra unha sucesión de palabras nun idioma e sae outra sucesión traducida a outro idioma. Cando ambas secuencias, de entrada e saída, teñen a mesma lonxitude ($T = R$), a estrutura da rede pode deducirse a partir da vista na Figura 3.2 engadindo un nodo de saída sobre cada un dos vectores ocultos. Mais, no caso habitual de que as secuencias de entrada e saída sexan de distinta lonxitude, é común ver estruturas de tipo *codificador-decodificador*, onde primeiramente temos unha RNN que codifica a información da secuencia de entrada, para logo pasarlla a outra RNN que a partir dela obtén a secuencia de saída. Podemos ver un exemplo de dita arquitectura na Figura A.14.

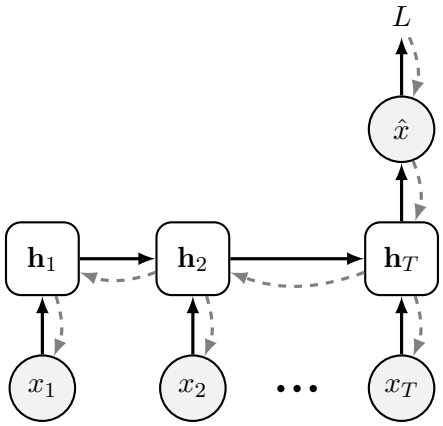

Figura 3.4: Esquema do proceso de retropropagación nunha rede neural recorrente. Adaptación de Love (2024).

3.3. LSTM (Long Short-Term Memory)

Un dos principais problemas que enfrontan as redes neurais recorrentes é o esvaecemento/explosión do gradiente debido á aplicación da mesma recorrencia en múltiples pasos. Isto redunda nunha falta de memoria a longo prazo que dificulta que a rede poida ter en conta patróns dos datos que non sexan razoablemente recentes.

Existen múltiples propostas para atallar este problema. Unha das máis directas é a de engadir conexións "a través do tempo", relacionando puntos da rede separados por varios pasos de tempo, recuperando o gradiente anterior e mitigando o esvaecemento que se puidera ter provocado entre ambos. Seguindo con esa filosofía xorden as *«Long Short-Term Memory»* (LSTM) RNNs, que conforman a arquitectura baseada en RNNs que mellor desenvolvemento acadou atallando este problema, segundo se indica en Bengio, Goodfellow, Courville et al. (2017):p. 412.

Xeneralizando o enfoque visto na Figura 3.2, os LSTMs están organizados en células nas que se actualiza o valor do vector oculto $\boldsymbol{h}_t$, e que se concatenan para obter a estrutura recursiva da rede. Na Figura 3.5 podemos ver unha esquema que describe un funcionamento dunha destas "células".

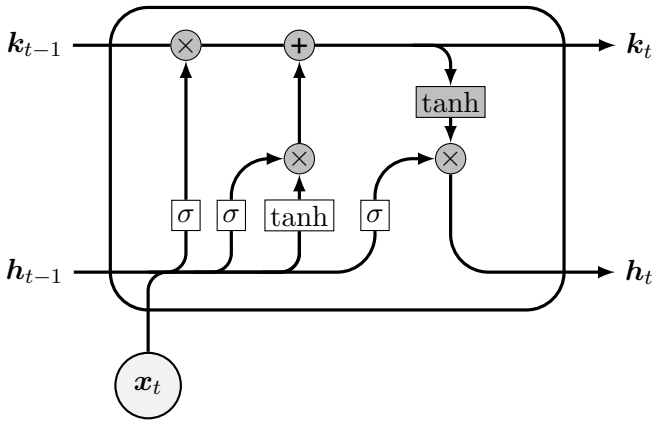

Figura 3.5: Esquema dunha célula dun LSTM. Adaptado de Love (2024).

A idea detrás dos LSTMs é a de empregar un novo vector oculto $\{\boldsymbol{k}_1, \dots, \boldsymbol{k}_T\}$, con $\boldsymbol{k} \in \mathbb{R}^\kappa$, que almacene a "memoria a longo prazo", caracterizando $\boldsymbol{h}_t$ unicamente a "memoria a curto prazo". Isto matematicamente redunda en que, a través das *portas* (cada unha das tres operacións entre $\boldsymbol{h}_t$ e $\boldsymbol{k}_t$ que vemos en vertical na Figura 3.5) da célula, o vector $\boldsymbol{k}_t$ determina dinamicamente en cada iteración da recorrencia en que medida se lle engade ao vector oculto $\boldsymbol{h}_t$ información do pasado. Isto foi un gran avance respecto das distintas tipoloxías de RNNs que se tiñan considerado, dada a flexibilidade que aporta ao conxunto non ter definidos pesos fixos para as aportacións de información anterior.

En particular, a inclusión da memoria a longo prazo, $\boldsymbol{k}_t$, da pé a que $\boldsymbol{h}_t$ se centre unicamente na información máis recente, permitindo esquecerse de aportes de información moito anteriores, cuxo gradiente vai sufrindo un esvaecemento progresivo e que serían difíciles de reter.

3.4. Transformers. Modelos baseados na Atención

Non podiamos rematar esta sección sen mencionar os modelos baseados na *atención*. Coa publicación do artigo Vaswani et al. (2017) deuse pé a un novo paradigma de "aprendizaxe" que é a atención, que permite dun xeito eficiente e paralelizable estudar a relación dunha observación con todas as anteriores.

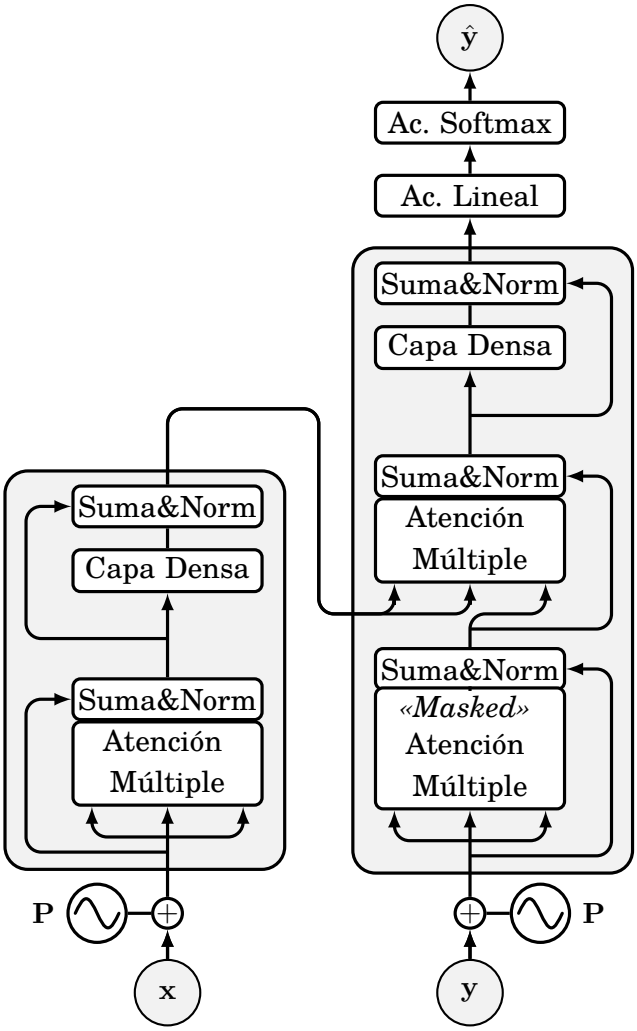

Figura 3.6: Esquema da arquitectura básica dun transformer. Adaptado de Love (2024).

Anos antes do 2017, comezaron a propoñerse métodos que combinaban ideas propias das redes neurais recorrentes, onde a información do pasado se garda en vectores ocultos, con ideas relacionadas co mecanismo de atención. Este fundaméntase no cálculo da relación entre cada par de elementos da secuencia, a cal se consegue codificando cada elemento a través de tres vectores que, operados entre si para dous elementos, determinan o seu grao de relación. Isto conduce a que as relacións entre un conxunto finito de elementos se poidan calcular directamente como produto das matrices formadas ao agrupar os vectores de ditos elementos.

No artigo Vaswani et al. (2017) propúxose a arquitectura do *transformer*, que deu lugar a toda unha familia de modelos que empregaban só o mecanismo de atención e que, polo tanto, deixaban atrás a estrutura recorrente das RNNs.

Na actualidade estes modelos conforman unha das metodoloxías máis empregadas no campo das redes neurais e o aprendizaxe profunda. Circunscribíndonos ás series temporais, existen múltiples adaptacións desta metodoloxía para a predición mostrando resultados moi positivos, especialmente de series moi longas de datos con horizontes de predición tamén grandes, como se pode ver en Zhou et al. (2021).

Capítulo

4 Comparación entre Metodoloxías

Chegados a este punto do traballo, é normal sentir a necesidade de comparar os distintos modelos vistos ata agora, valorando a idoneidade dos mesmos para cada situación concreta. O certo é que non somos os primeiros con tal inquedanza e hai abundante literatura ao respecto. Neste capítulo faremos primeiro un repaso histórico do estado da arte en tanto en canto ás diversas comparativas sistemáticas de metodoloxías de predición (competicións de predición) que se teñen levado a cabo, para logo analizar o rendemento e utilidade práctica dos distintos métodos vistos no traballo á vista dos resultados anteriormente mencionados.

4.1. Introdución Histórica ás Competicións de Predición

Se o lector ten curiosidade sobre o tema, en Hyndman (2020) faise unha retrospectiva breve e completa das competicións de predición previas ao 2020. As competicións de predición nacen nos anos 70 co obxectivo de comparar a cada vez máis ampla variedade de métodos dispoñibles no momento. Segundo veremos ao longo de esta sección, motivaron un cambio de mentalidade respecto do xeito no que se afrontaba a predición de series temporais.

As primeiras competicións, realizadas na universidade de Nottingham, constaban de un ou dous investigadores encargados da tarefa de implementar e contrastar, por eles mesmos, o rendemento dos métodos sobre conxuntos de arredor de 100 series temporais.

Estes estudos precursores sérvennos para identificar as problemáticas que trataron de mellorar todos os que os sucederon. A partir dos anos 80 abríronse á participación de múltiples académicos de diferentes universidades, despexando así as reticencias das primeiras competicións sobre a influenza da destreza dos autores nos resultados. Mentres, como era de esperar, segundo a accesibilidade dos datos foi a máis, a cantidade de series temporais empregadas nas competicións viuse incrementada, aportándolle maior robustez estatística aos resultados.

Estes primeiros estudos viñeron acompañados de polémica. Era a primeira vez que se obtiña unha foto fixa do rendemento de todos os modelos, e certas figuras da época non encaixaron moi ben os resultados, asegurándose de deixalo claro no seu rexistro epistolar. Hai que ter en conta que na época existía a convicción de que a predición de series temporais se baseaba unicamente na procura do modelo subxacente que xeraba aos datos. Isto levaba a que non se asumiran resultados que indicaban, por exemplo, que modelos mixtos, formados pola combinación de varios, obtiveran mellores resultados que os orixinais, ou os que aseguraban que modelos simples obtiñan mellores resultados que outros moito máis complexos.

Estas hostilidades fóronse despexando segundo se sucederon os avances metodolóxicos nas diferentes competicións. Con todo, estes primeiros experimentos chamaron a atención da comunidade, centrando cada vez máis os esforzos na precisión dos modelos máis alá das propiedades matemáticas dos mesmos. Deste xeito, progresivamente dende os anos 90, foise erixindo cada vez máis a "predición" (do inglés: *«forecasting»*) como unha disciplina independente da análise de series temporais.

A raíz da competición M1, publicada en Makridakis, Andersen et al. (1982), na que se consideraron máis de 1000 series, observouse como modelos máis sofisticados non obtiñan mellores resultados respecto de modelos máis simples. Quedou clara a variabilidade nos resultados que introducía a medida de erro empregada. Ademais, comprobouse que o rendemento dos métodos dependía do "horizonte de predición", que nós denotamos como h.

Xa nos anos 2000, os mesmos autores levaron a cabo a competición M3, sobre máis de 3000 series. Nela observouse un mellor rendemento dos modelos ARIMA, que xa existían dende había preto de 30 anos, mais que ata aquel momento non conseguiran atallar o problema do sobreaxuste (do inglés: *«overfitting»*). Nesta competición participou por vez primeira un método baseado en redes neurais, aínda que os seus resultados foron bastante pobres. Na mesma liña, seguiulle a competición M4, levada a cabo en 2020 sobre preto de 100,000 series temporais. Nela consolidouse un paso adiante no campo, acadando resultados moi salientables varias propostas metodolóxicas que incorporaban, como idea fundamental, a combinación nun só algoritmo de varios métodos.

Por último, é xusto mencionar a irrupción de Kaggle, como un plataforma (propiedade de Google) para a publicación de datos e modelos que, dende a súa fundación no 2010, leva acollendo múltiples competicións de predición en varios ámbitos, ademais do das series temporais.

4.2. Análise do Rendemento dos Modelos

Comparación sobre Series Curtas de Datos. Dataset M3

Hai que ter en conta que en datasets como o de M3, é habitual atopar series moi curtas, de aproximadamente entre 20 e 120 observacións, ao tratarse de datos anuais, cuadrimestrais ou mensuais. Isto leva a que, en series tan curtas, o rendemento de modelos baseados en redes neurais resulte bastante pobre, dado que a falta de datos lles impide poder recoñecer os patróns da serie. É precisamente neste tipo de situacións nas que metodoloxías máis simples acadan un mellor rendemento. En particular, como se indica nos resultados de M3, Makridakis e Hibon (2000), modelos máis complexos baseados en modelos ARIMA obteñen resultados lixeiramente peores que modelos máis simples como un alisado exponencial con pendente aditiva atenuada.

Malia todo, non se pode afirmar categoricamente que un método é mellor que outro para calquera dúas series temporais que poidamos considerar. Un dos factores máis sensibles é a magnitude no horizonte de predición, resultando mellor parados os métodos máis elaborados, destacando os modelos ARIMA. Tamén pode chegar a resultar relevante no rendemento dos métodos a orixe da serie temporal que se estea a predicir, sendo dispares os métodos máis precisos segundo as series teñan unha procedencia económica, industrial, demográfica...

Comparación sobre Series Longas de Datos. Dataset de Wikipedia

Ao considerarmos tamén series máis longas de datos, con periodicidades diarias ou mesmo horarias, apreciamos un certo cambio nos resultados obtidos. Fixarémonos agora, por exemplo, na competición *Web Traffic Time Series Forecasting*, organizada por Google en Kaggle, que propón buscar a mellor metodoloxía de predición para un conxunto de series temporais correspondentes cos historiais de visitas de páxinas da Wikipedia (da orde de 700 observacións). Podemos comprobar como a metodoloxía cos mellores resultados emprega un enfoque ad-hoc baseado en redes neurais recorrentes pero que, á vez, emprega ferramentas propias da análise clásica de series temporais, como as autocorrelacións ou os procesos autorregresivos e de medias móbiles para obter características sobre as series que logo poder empregar como entrada na rede neural recorrente. Deste xeito, segundo se explica na descrición da metodoloxía gañadora, as redes neurais recorrentes pódense pensar como unha xeneralización dos modelos ARIMA que, admitindo maior flexibilidade e capacidade de recoñecer patróns, precisa dunha cantidade de datos maior para adestrarse.

Porén, o esvaecemento do gradiente nas redes neurais recorrentes presenta desafíos engadidos segundo a lonxitude das series se incrementa. Solucións como os modelos LSTM ofrecen bos resultados para series non moi longas (entre 100 e 130 observacións), mais de

cara a modelar series aínda máis longas, empregar mecanismos de atención pode ser útiles. Este precisamente é o caso da metodoloxía gañadora da competición mencionada no parágrafo anterior, que implementa un mecanismo de atención que complementa á súa vez o funcionamento dun modelo GRU (semellante a un LSTM).

Comparación sobre Todo Tipo de Series. Dataset M4

Na competición M4 podemos comparar o mellor de ambos mundos, considerando dende series anuais de apenas 13 observacións ata series horarias de máis de 700, segundo se indica no artigo que analiza os resultados da mesma: Makridakis, Spiliotis e Assimakopoulos (2020). Desta forma, os resultados da competición avalían aos métodos segundo a súa idoneidade para seren capaces de realizar predicións consistentemente nunha ampla gama de series temporais. Nos seguintes parágrafos discutimos varias das conclusións que se extraen dos resultados de M4.

Confirmouse, logo de xa ter aparecido indicios en competicións anteriores, o bo rendemento obtido ao combinar distintas metodoloxías de predición. Chegando ao punto de que, para predicións puntuais, só un dos dez mellores métodos non é combinación de outros. Isto reivindica a importancia de coñecer e empregar os distintos métodos dispoñibles, combinándoos para poder aproveitar os beneficios que cada un aporta, despexando quizais a idea de atopar un único método que renda consistentemente mellor que todos os demais. Deste xeito, modelos como o alisado exponencial ou os ARIMA, que se formulan baixo a suposición de estar describindo o "proceso subxacente" que xera os datos, non serían os máis axeitados, e efectivamente, ambos vense superados por combinacións de outros métodos que teoricamente son máis sinxelos. Isto non implica necesariamente que o seu rendemento sexa pobre, e tamén hai que ter en conta que a súa maior interpretabilidade, supón unha vantaxe ao seu favor. Porén, ao definir modelos que xorden como combinación de varias metodoloxías, a selección dos mesmos remata por ser un problema engadido, precisando, en ocasións, que algunha metodoloxía de aprendizaxe automática realice a estimación dos pesos cos que contribúen cada un dos métodos.

En representación deste grupo de metodoloxías podemos destacar o segundo mellor método en canto á precisión puntual, proposto, entre outros, por Pablo Montero Manso, formado na Universidade da Coruña. Este combina sete métodos estatísticos xunto con un de aprendizaxe automática, aos que se lle suma, como vimos antes, outro algoritmo de aprendizaxe automática adestrado para estimar os pesos necesarios para combinar os modelos reducindo o erro de predición.

Unha das maiores novas que deixou M4 foi a obtención de resultados moi prometedores por parte dos chamados métodos híbridos. Estes son métodos que integran de xeito profundo,

non unicamente combinando os resultados, varios modelos estatísticos e de aprendizaxe automática. Por exemplo, o método de predición puntual que mellores resultados acadou, proposto por Slawek Smyl, combina dentro dunha rede neural recorrente fórmulas correspondentes cun modelo de alisado exponencial, de forma que estas quedan completamente integradas na recorrencia da rede, calculándose os parámetros de suavizado correspondentes ao alisado exponencial empregando o mesmo algoritmo de descenso do gradiente co que se optimizan os parámetros da rede neural recorrente.

Respecto da predición dos intervalos de confianza, tamén se rexistrou unha mellora substancial. Unha das principais carencias que presentaban varios dos métodos que, antes de M4, conseguiran boas predicións puntuais, era a de subestimar os intervalos de confianza das súas predicións. Mais, en M4, os métodos de Syml e Montero-Manso, aportaron unha mellora significativa neste aspecto.

Á vista dos resultados anteriores, produciuse un cambio de tendencia no mundo das series temporais, sobre a cuestión de se modelos máis complexos poden realmente levar á obtención de predicións máis precisas. Lembramos que, en M3 e nas súas predecesoras, modelos moi sinxelos conseguiran superar a outros bastante máis complexos. Porén, en M4 unha maior complexidade dos métodos, asociada tamén cun maior custo computacional, si redundou na obtención de mellores resultados.

Outra das novidades técnicas, en termos de rendemento, identificadas en M4, foi a de empregar información de múltiples series temporais para seleccionar a combinación axeitada de métodos necesaria para predicir unha serie en concreto. Isto foi particularmente visible en metodoloxías que combinaran varios métodos. Esta técnica, segundo se indica en Makridakis, Spiliotis e Assimakopoulos (2020), coñécese como: *"enfoque de aprendizaxe cruzada"* (do inglés: *«cross-learning approach»*).

Precisamente pola ausencia do enfoque de aprendizaxe cruzada, poden tratar de explicarse os malos resultados obtidos polos métodos "puros" (que non son combinación de outros) de aprendizaxe automática. Outro dos motivos puido ser a dificultade para atallar o problema do sobreaxuste. Iso súmase ao tamén pobre rendemento dos modelos estatísticos puros respecto dos modelos mixtos e híbridos.

Recapitulando, podemos concluír que o campo da predición de series temporais está nun momento de eclosión. Dende a competición M4, celebrada en 2020, 20 anos despois de M3, xa se teñen celebrado as competicións M5, en 2022, e M6, en 2024. Con todo, queda moita investigación por facer, pero en base á experiencia anterior, podemos comprobar como todas as metodoloxías, independentemente da súa precisión ou utilidade práctica, promoveron o avance da disciplina, axudando a entender mellor o problema que todas elas buscan resolver.

Anexo

A Material Suplementario. Gráficas e Esquemas

Neste anexo incluímos, como material suplementario, un compendio de figuras que complementan os contidos vistos no traballo. Ao longo do mesmo invítase ao lector a consultalas para ilustrar conceptos, poñer exemplos ou visualizar o rendemento das diferentes metodoloxías de predición.

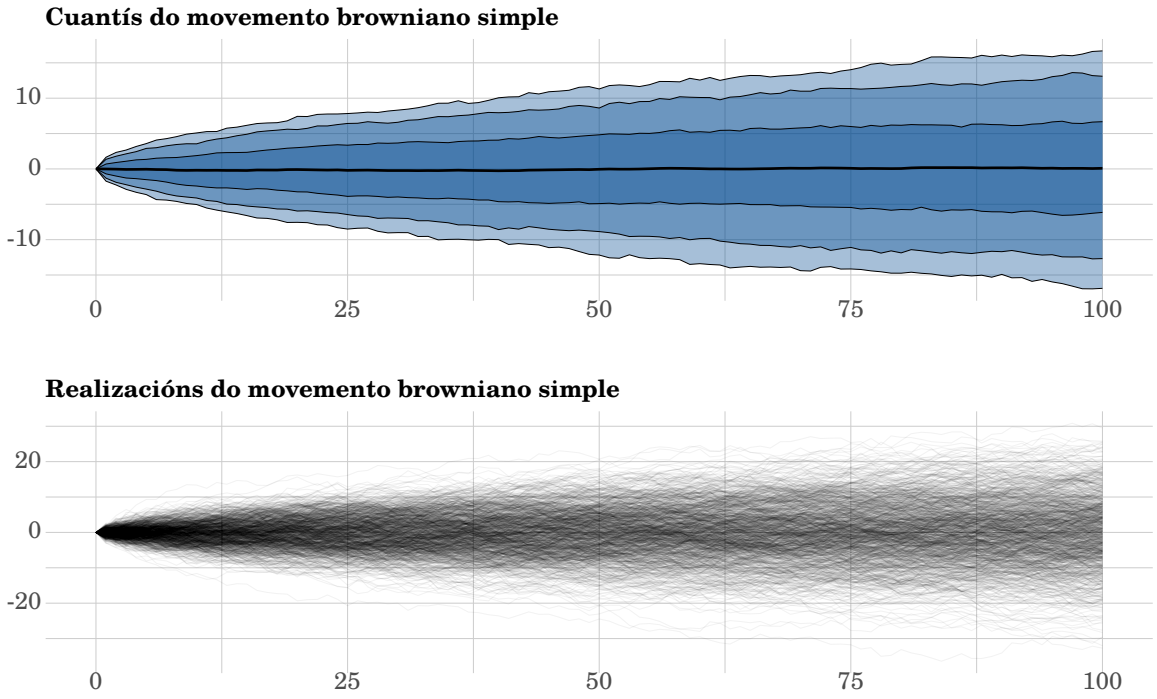

Figura A.1: Representación gráfica dun proceso estocástico, neste caso, un movemento browniano dado por $X_t = X_{t-1} + \varepsilon_t$, con $\varepsilon_t \sim \mathrm{N}(0,1)$. Na primeira gráfica podemos ver os intervalos de confianza para as realizacións para $\alpha = \{0{,}1, 0{,}2, 0{,}5\}$. Na segunda podemos ver a gráfica de 1000 realización do proceso. Elaboración propia con LIB ggplot2 en R.

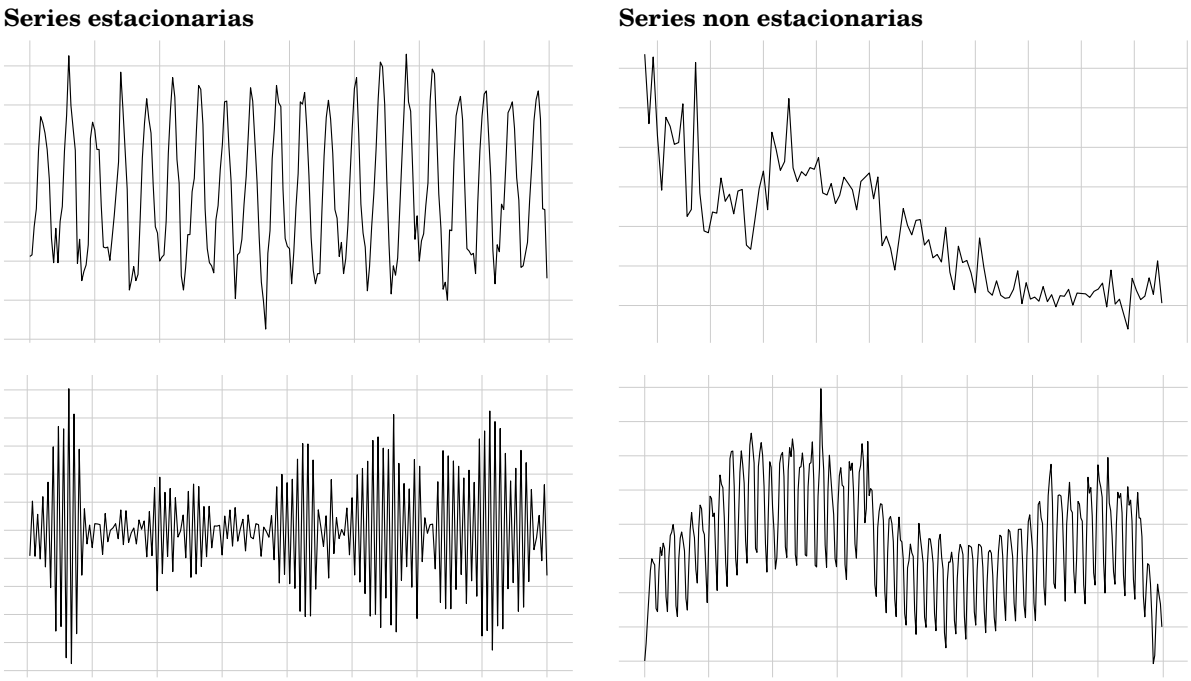

Figura A.2: Exemplos de series temporais estacionarias e non estacionarias. Elaboración propia cos paquetes LIB ggplot2 e LIB fpp3 de R.

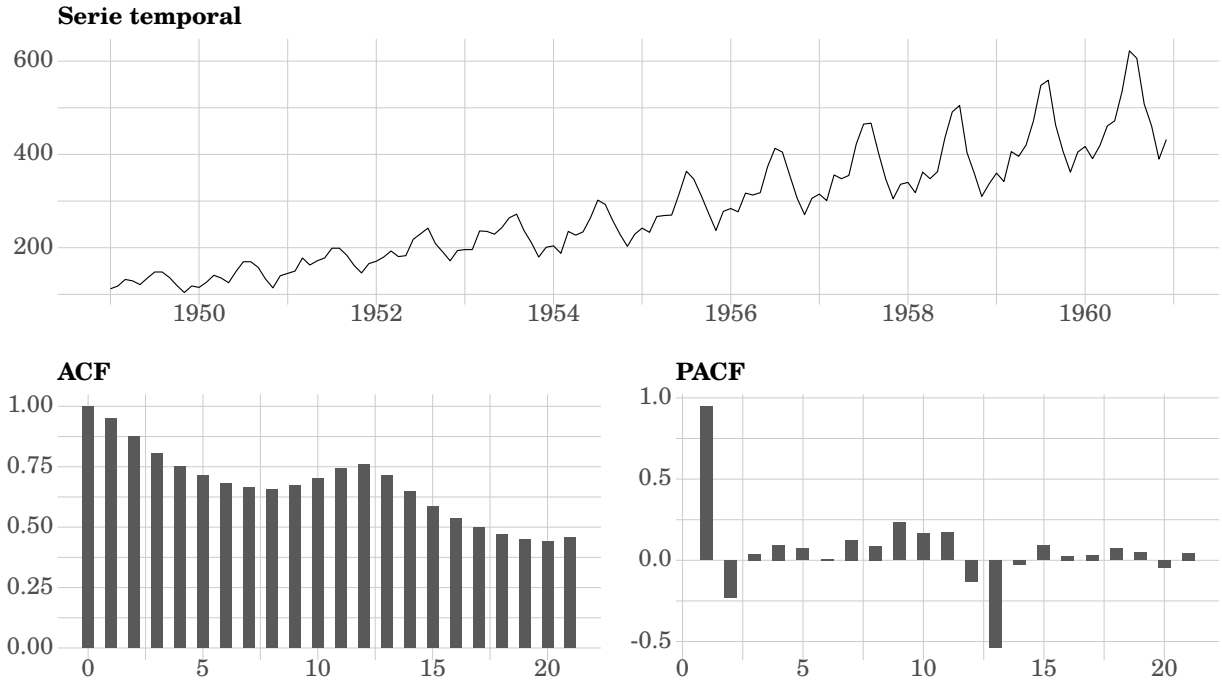

Figura A.3: Representación da función de autocorrelación e da función de autocorrelación parcial da serie temporal dada polo dataset `airpassengers`, introducido en Box, Jenkins et al. (1970):p. 547. Elaboración propia co paquete LIB ggplot2 en R.

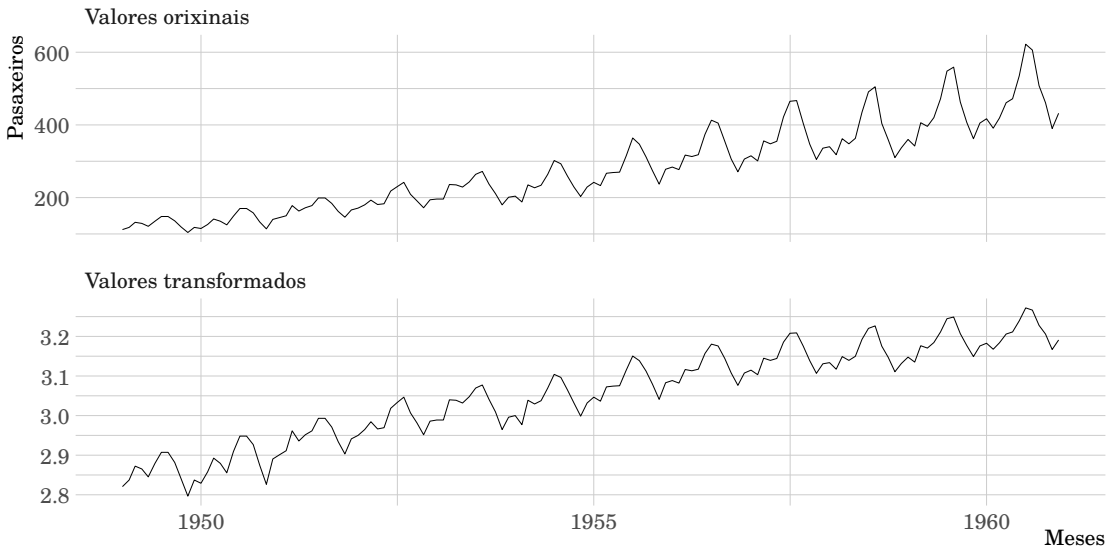

Figura A.4: Diferenza entre a serie de pasaxeiros do dataset `airpassengers` introducido en Box, Jenkins et al. (1970):p. 547, onde se pode observar unha varianza heterocedástica, fronte aos valores homocedásticos obtidos mediante unha transformación Box-Cox con $\lambda = -0{,}24$. Elaboración propia con LIB ggplot2 en R.

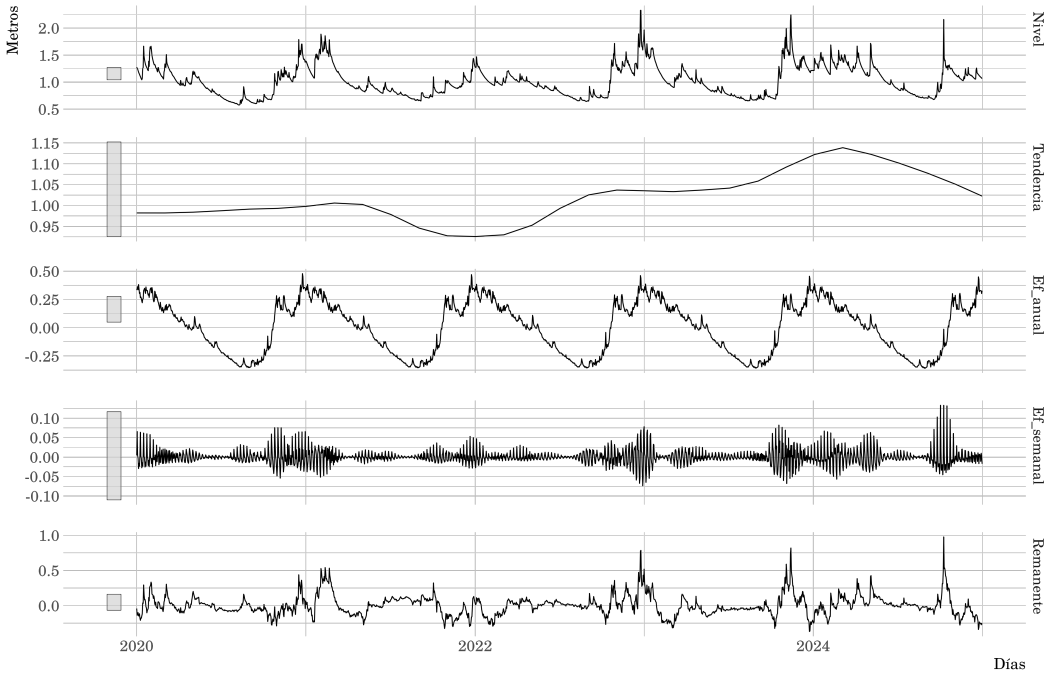

Figura A.5: Representación das compoñentes da descomposición STL do nivel diario do río Sar entre o 2020 e o 2024[1]. De arriba a abaixo: os datos orixinais, a tendencia, as compoñentes estacionais (semanal e mensual) e o remanente. O rectángulo ***gris*** é o mesmo en todas as escalas. Elaboración propia baseada co paquete LIB fabletools de R.

[1] Fonte. Meteogalicia. Nivel histórico diario do río Sar ao seu paso por Santiago. URL: https://servizos.meteogalicia.gal/mgafos/estacionshistorico/historico.action?idEst=140548.

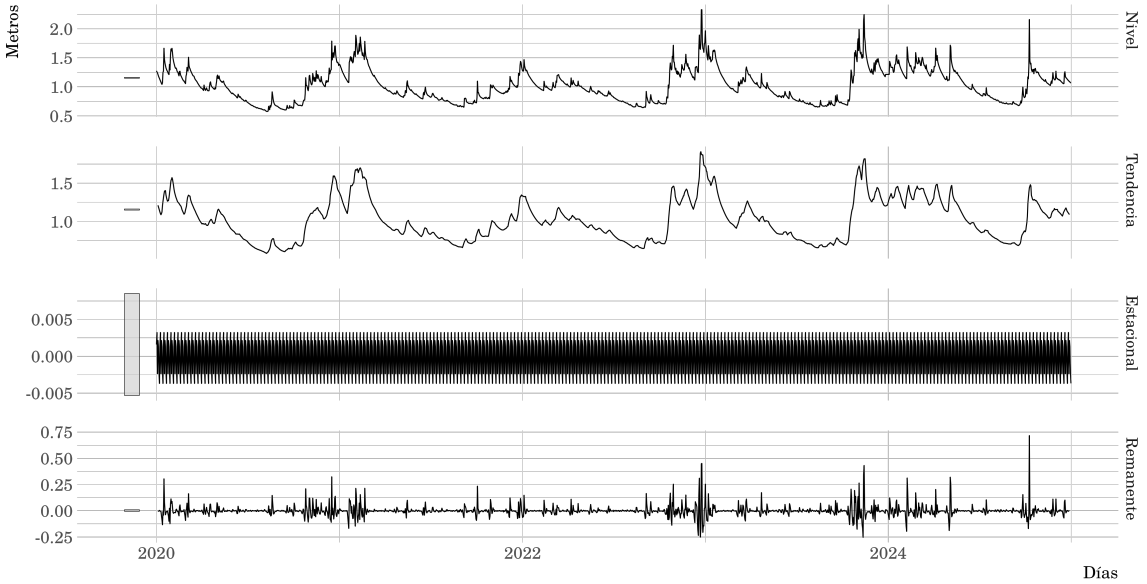

Figura A.6: Representación das compoñentes da descomposición Clásica (aditiva) do nivel diario do río Sar entre o 2020 e o 2024. De arriba a abaixo vemos: os datos orixinais, a tendencia, a compoñente estacional e o remanente. O rectángulo ***gris*** é o mesmo en todas as escalas. Elaboración propia baseada na función `autoplot()` do paquete LIB fabletools de R.

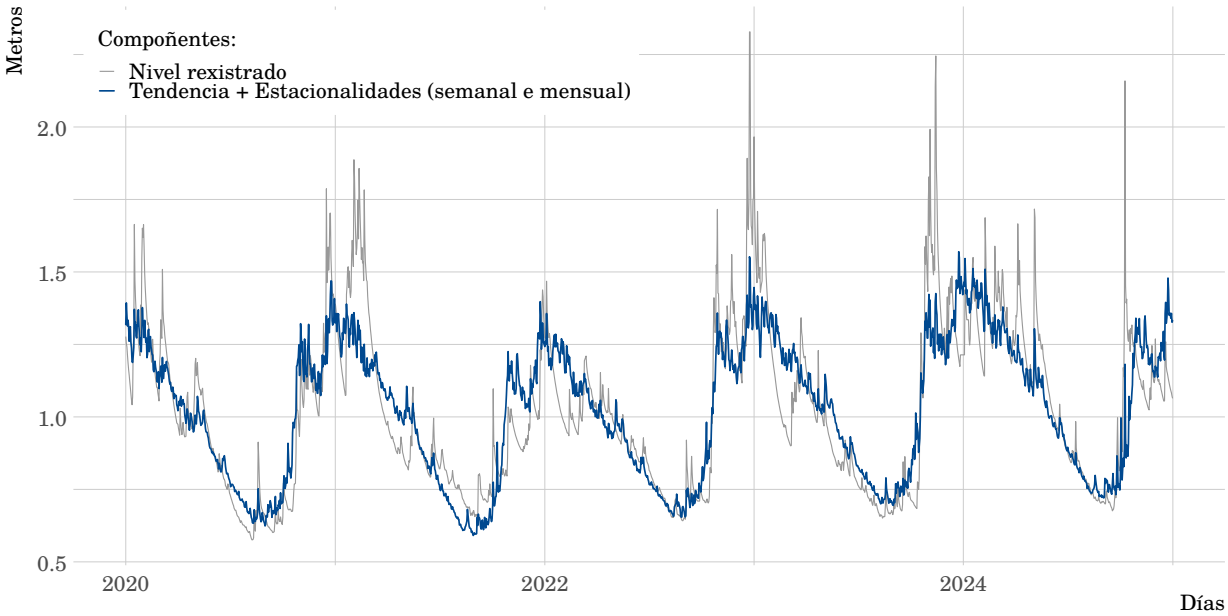

Figura A.7: Representación do nivel diario do río Sar entre o 2020 e o 2024. En ***gris*** o nivel rexistrado e en ***azul*** a suma da tendencia e as dúas compoñentes estacionais (semanal e mensual) froito dunha descomposición STL. Elaboración propia co paquete LIB ggplot2 de R.

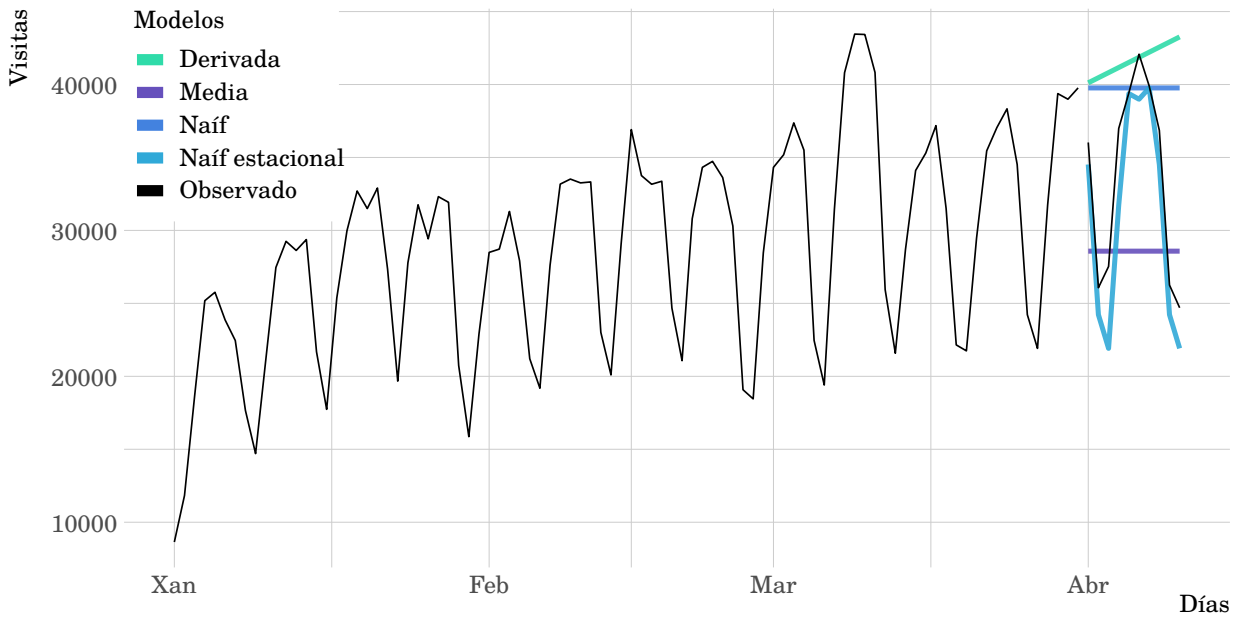

Figura A.8: Comparación das 4 metodoloxías triviais de predición para os datos de visitas da versión online do libro Hyndman e Athanasopoulos (2021). Elaboración propia cos paquetes LIB fable e LIB ggplot2 en R.

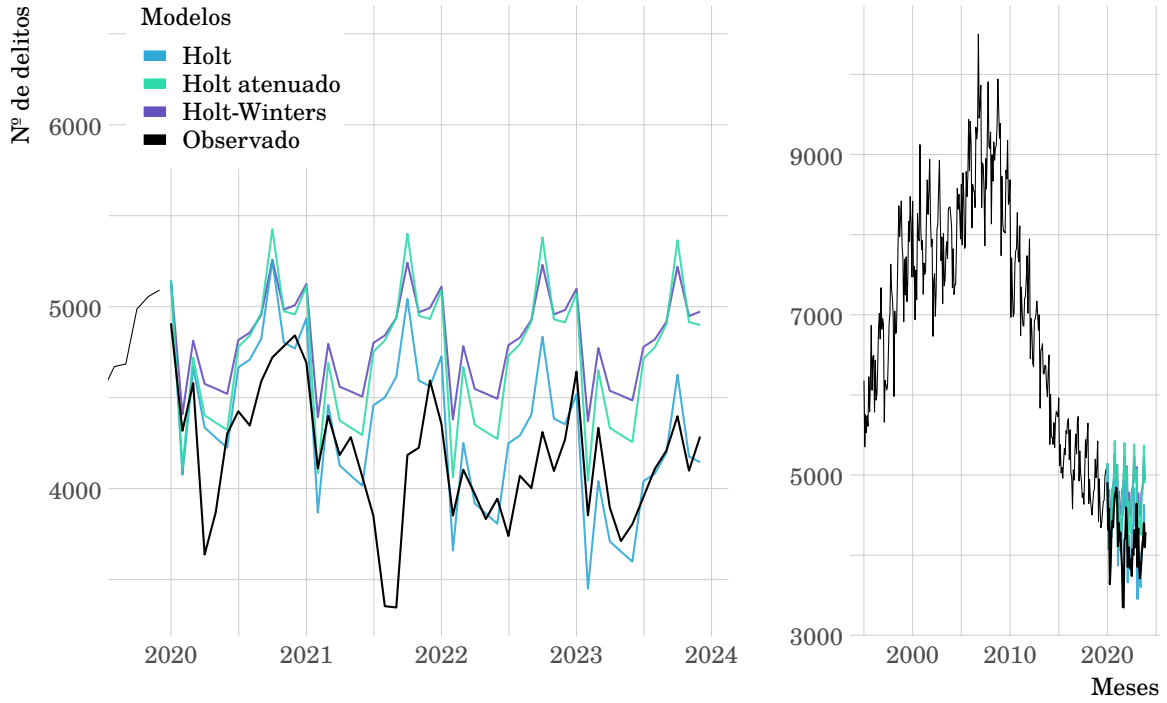

Figura A.9: Exemplo de predición con distintas metodoloxías de alisado exponencial sobre os datos de "danos á propiedade privada" da área metropolitana de Nova York. Á esquerda podemos ver unha ampliación das predicións. Elaboración propia con LIB ggplot2 en R.

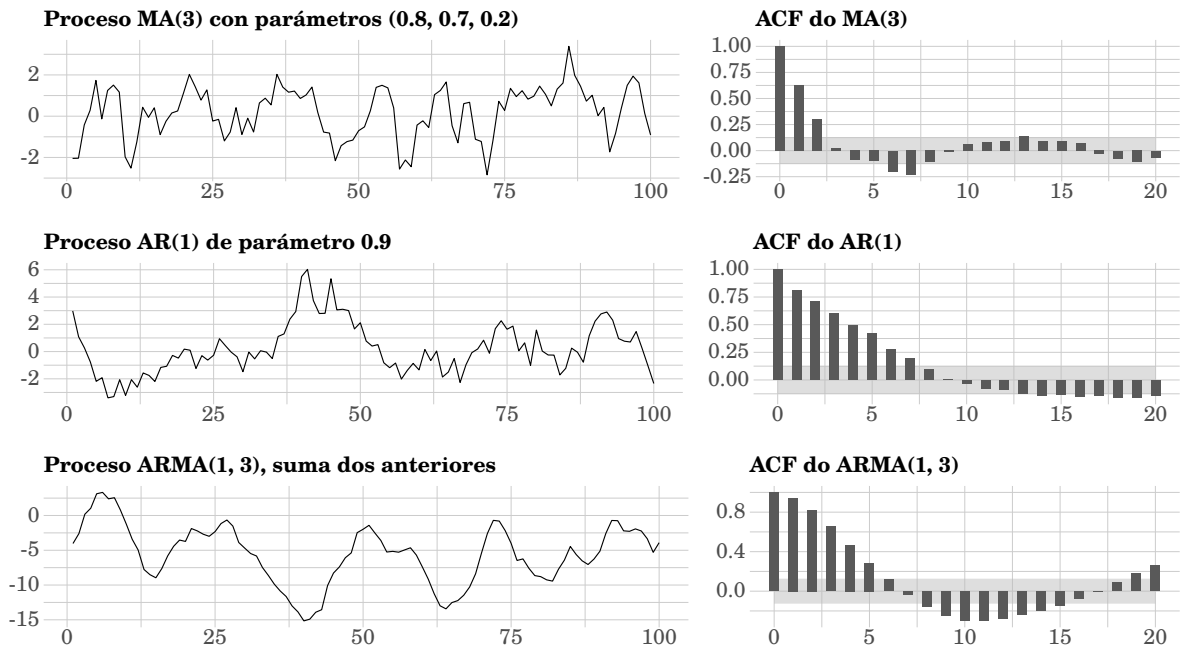

Figura A.10: Exemplos ilustrativos dos procesos MA, AR e ARMA. Á esquerda vemos a representación do proceso en si, mentres que á dereita vemos a función de autocorrelación correspondente. Nesta última, o rectángulo *gris* representa a rexión na que o valor da ACF non é significativo ao 95 %. Elaboración propia con LIB ggplot2 en R.

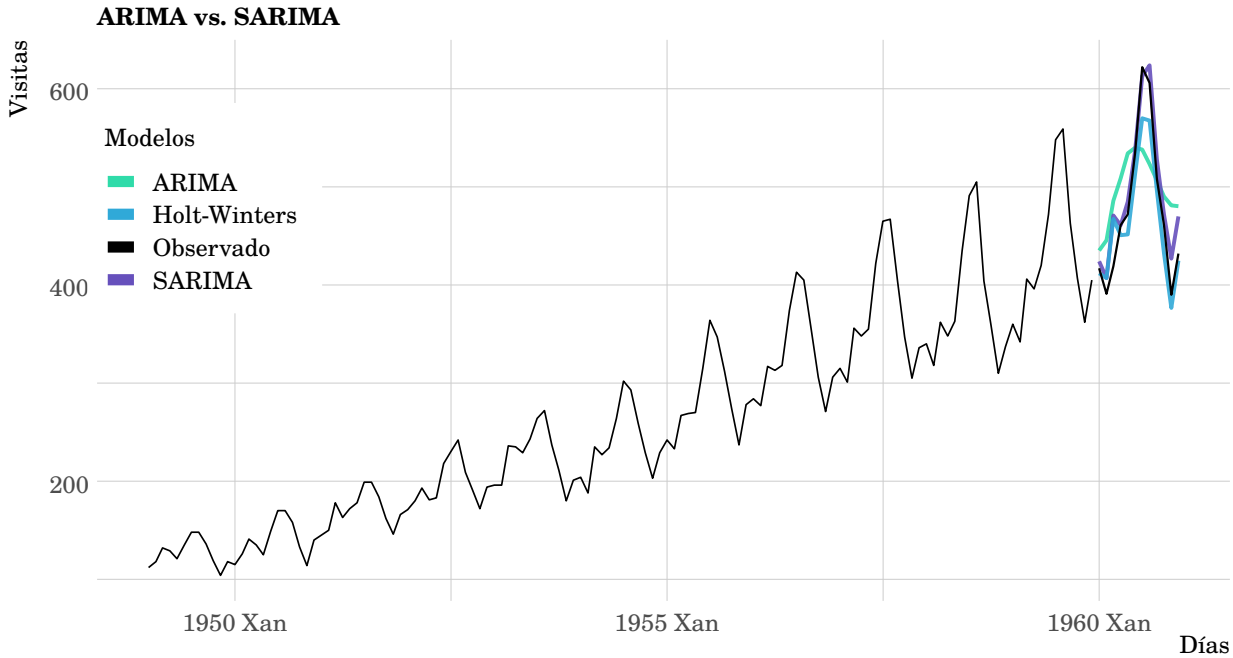

Figura A.11: Exemplos de predición con modelos ARIMA e SARIMA sobre o dataset `airpassengers`, introducido en Box, Jenkins et al. (1970):p. 547. Elaboración propia cos paquetes LIB forecast e LIB ggplot2 de R.

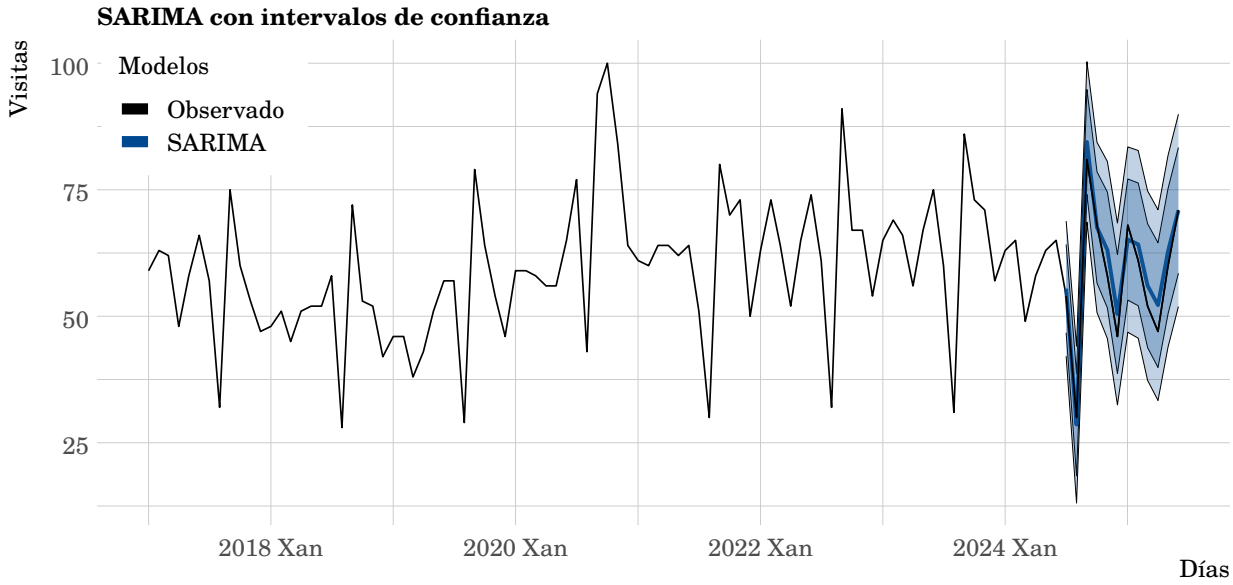

Figura A.12: Exemplo de predición cun modelo SARIMA, xunto cos correspondentes intervalos de predición ao 80 % e 95 %, para os datos medios mensuais de visitas diarias á web da USC co navegador de Google. Elaboración propia con LIB ggplot2 en R.

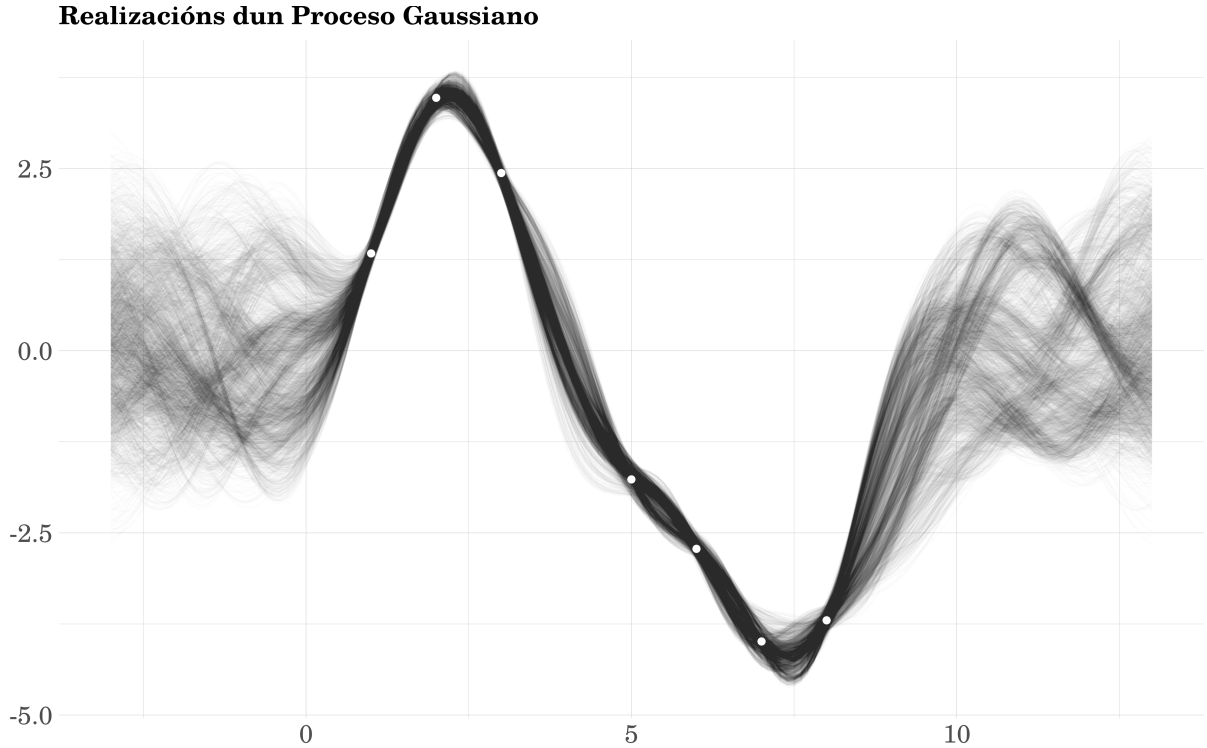

Figura A.13: Exemplo das realizacións dun proceso gaussiano para unha serie temporal con un dato faltante. Os puntos brancos son os datos da serie. As liñas negras son realizacións do proceso gaussiano axustado. Elaboración propia, baseada no recurso web, empregando `Stan` xunto con R.

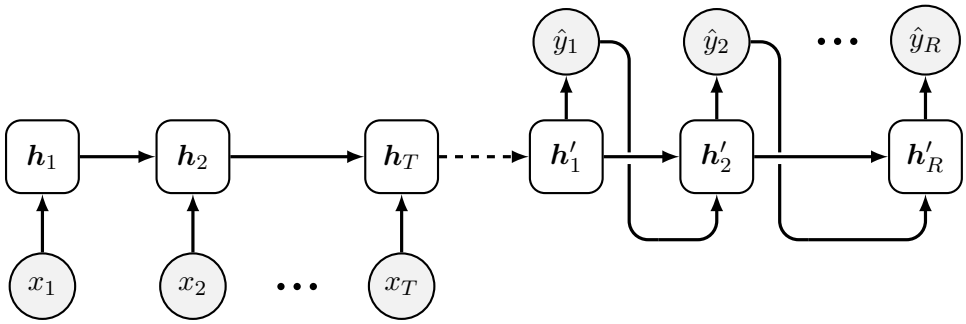

Figura A.14: Esquema da arquitectura *codificador-decodificador* nunha RNN. Adaptación de Love (2024).

Anexo

B Regresión LOESS e Algoritmo de Descomposición STL

Neste apéndice abordamos unha descrición pormenorizada do algoritmo STL para a descomposición de series temporais. Comezamos describindo a regresión LOESS da que logo se valerá o algoritmo para o suavizado das series. Finalmente analizamos a estrutura do algoritmo, especificado na Figura B.1 e no Algoritmo 1.

B.1. Regresión LOESS

É a metodoloxía de regresión na que se fundamenta principalmente o método STL. O termo *LOESS* é un acrónimo, do inglés «*Locally Estimated Scatterplot Smoothing*», que en galego vén a significar "*Suavizado de Diagramas de Dispersión Estimado Localmente*"[I]. É dicir, a partir dun conxunto de datos (x_i, Y_i) permite obter unha curva $g(x)$ que suaviza os datos e está definida $\forall x \in \mathbb{R}$. Deste xeito, temos unha relación entre x e Y dada por: $Y_i = g(x_i) + \varepsilon_i$, na que ε_i son os erros, para os que se supón media cero, independencia e homocedasticidade.

A estimación de g faise localmente e de xeito independente para cada punto. Dado un $x \in \mathbb{R}$, asígnanselle un peso aos $q \in \mathbb{Z}^+$ puntos x_i máis próximos a x[II]. Estes pesos asígnanse empregando unha función de tipo núcleo. Neste caso, introduciremos dúas, a función tricúbica e a bicadrada. Cada unha empregarase nun punto distinto do algoritmo segundo o deseño orixinal do mesmo. Denótanse respectivamente por

$$W_3(u) = \begin{cases} (1-u^3)^3 & \text{se } 0 \leq u < 1, \\ 0 & \text{se } u \geq 1, \end{cases} \qquad \text{e} \qquad W_2(u) = \begin{cases} (1-u^2)^2 & \text{se } 0 \leq u < 1, \\ 0 & \text{se } u \geq 1. \end{cases}$$

Entón, o peso asociado á observación i-ésima respecto de x, onde $\lambda_q(x)$ é a distancia do

[I] Nalgunhas fontes antigas refírense a este tipo de métodos como "*filtro de Savitzky-Golay*" xa que, con outros obxectivos, emprega os mesmos fundamentos e foi proposto 15 anos antes.

[II] En inglés, segundo figura en Cleveland et al. (1990):p. 4, coñécense como «*neighbourhood weights*».

q-ésimo x_i máis afastado de x[III], será o seguinte

$$\omega_i(x) = W_3\left(\frac{|x_i - x|}{\lambda_q(x)}\right),$$

de xeito que as observacións x_i máis próximas a x teñan un peso maior. Unha vez determinados os pesos, procédese coa estimación de $g(x)$. A tal fin, axústase un modelo local polinómico de grao $d = \{1, 2\}$, é dicir, lineal ou cadrático, da forma: $g(x_i) \approx \beta_0^x + \cdots + \beta_d^x (x_i - x)^d$. Ao ser unha aproximación local, os coeficientes $\beta_0^x, \ldots, \beta_d^x$ serán diferentes para cada $x \in \mathbb{R}$.

A estimación dos coeficientes faise cun axuste por mínimos cadrados ponderados, empregando os pesos $\omega_i(x)$ calculados anteriormente. Ao obterse o axuste do polinomio de grao d para os $q + 1$ puntos, estímase o valor de $g(x)$ a partir do seu intercepto, é dicir, $\hat{g}(x) = \hat{\beta}_0$.

É posible introducir pesos externos, ρ_i, no modelo de xeito que se teña en conta, por exemplo, a fiabilidade das observacións. Iso consíguese multiplicando estes pesos polos $\omega_i(x)$ anteriores. En xeral, a introdución deste tipo de pesos permite que a regresión sexa menos dependente de certas observacións sobre as que se ten unha maior incerteza, o que redunda nun axuste máis robusto.

Con todo, a regresión *LOESS* emprega dous parámetros, d e q, cuxos valores dependerán da aplicación que se lle queira dar ao modelo. O parámetro q determina o grao de suavidade do axuste, en particular, se $q \to +\infty$, os pesos $\omega_i(x) \xrightarrow{q \to +\infty} 1$, polo que $\hat{g}(x)$ tendería ao axuste (global) polinómico de grao d obtido por mínimos cadrados. Por outra banda, o parámetro d caracteriza a curvatura local que presentan os datos. Se esta é suave, a elección adoita ser $d = 1$, mais se as flutuacións son máis abruptas, $d = 2$ é o indicado.

B.2. Deseño do algoritmo STL

Estruturalmente, o algoritmo STL susténtase no uso repetido medias móbiles e de regresións LOESS. Na Figura B.1 podemos ver a estrutura do algoritmo STL, correspondéndose esta coa súa definición dada no Algoritmo 1. As etiquetas: ❶, ❷, ❸ e ❹, identifican os mesmo puntos chave en ambas representacións do algoritmo.

Como podemos observar, o algoritmo divídese en dúas seccións: o bucle interno e o externo. No bucle interno ten lugar o proceso de suavizado da tendencia e da compoñente estacional (pasos ❶, ❷ e ❸), é dicir, este é o momento no que se vai extraendo da serie temporal orixinal, en sucesivas iteracións, a variabilidade propia de cada compoñente. No bucle externo (paso ❹) realízase o cálculo dos pesos de robustez[IV], que precisamente permiten garantir que o método sexa robusto.

[III] Se $q > n$, segundo se indica en Cleveland et al. (1990):p. 4, tómase por convenio: $\lambda_q = \frac{q}{n}\lambda_n$.
[IV] Do inglés «*robustness weights*», segundo se menciona en Cleveland et al. (1990):p. 6.

Así mesmo, o algoritmo STL emprega en total 6 parámetros no seu funcionamento, descritos no Cadro B.1. Para algúns a súa elección é inmediata, mentres que para outros resulta máis difícil xa que dependen das características da serie que se estea a descompoñer.

Parámetro	Descrición
n_p	Número de observacións de cada período estacional
n_i	Número de execucións do bucle interno
n_e	Número de execucións do bucle externo
n_l	Parámetro de suavizado do filtro de baixas frecuencias
n_t	Parámetro de suavizado para a tendencia
n_s	Parámetro de suavizado para a compoñente estacional

Cadro B.1: Parámetros que emprega o algoritmo STL.

Notación. Denotamos por $T_t^{(k)}$ e por $S_t^{(k)}$ ao valor da tendencia e da compoñente estacional na iteración k-ésima do bucle interno do algoritmo.
Deterémonos agora nos catro pasos fundamentais, sinalados coas etiquetas ❶, ❷, ❸ e ❹, dos que se compón o algoritmo STL:

❶ Neste paso tómase a serie sen tendencia: $X_t - T_t^k$, e divídese nas súas subseries estacionais, é dicir, as series temporais formadas polos elementos de X_t que pertencen a unha estación concreta[v]. A cada unha destas subseries aplícaselles un suavizado empregando unha regresión LOESS de parámetros $(d, q) = (1, n_s)$, de xeito que se obteñen estimacións para cada punto da subserie incluíndo tanto os datos faltantes como unha observación extra por cada lado. O resultado almacénase no vector C_t^{k+1} de lonxitude $n + 2n_p$.

❷ Neste paso aplícase un filtro de paso baixo que só permite o paso das frecuencias graves, é dicir, que só conserva as variacións de baixa frecuencia que, por deseño, tratamos de agrupar na tendencia. Este filtro está composto a partir de tres medias móbiles e unha regresión LOESS. Ao aplicárllelas a C_t^{k+1} obtemos o vector L_t^{k+1} de lonxitude n. Ao restar ambos vectores no dominio orixinal, obtense o suavizado da compoñente estacional S_t^{k+1}, evitando así que variabilidade pertencente á tendencia poida contaminala.

❸ Ao final de cada iteración do bucle interno ten lugar o suavizado da tendencia, para o que se emprega unha regresión LOESS de parámetros $(d, q) = (1, n_t)$ sobre os últimos datos desestacionalizados: $X_t - S_t^{k+1}$. Co cal, segundo S_t^{k+1} sexa capaz de explicar mellor a variabilidade estacional, T_t^{k+1} tamén poderá extraer a variabilidade da que se corresponde coa tendencia con menos interferencias, o cal á súa vez mellorará o rendemento de S_t^{k+2}, e así sucesivamente.

[v]Por exemplo, nunha serie temporal con datos mensuais, teríamos 12 subseries estacionais, onde os elementos da primeira delas virían dados pola observacións anuais do mes de xaneiro.

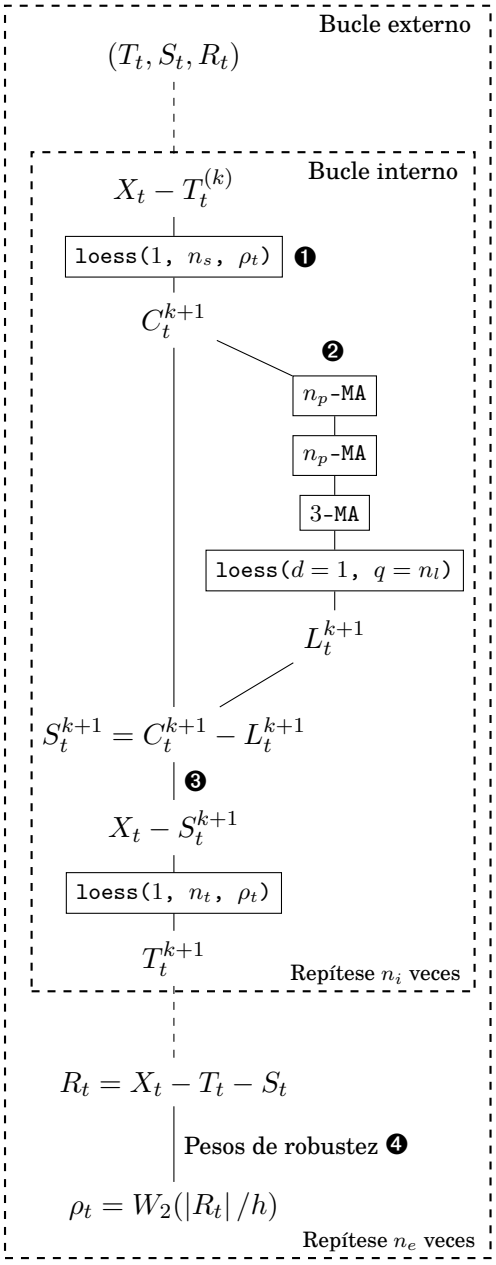

Figura B.1: Esquema do algoritmo STL. Elaboración propia co paquete tikz de LaTeX.

Algoritmo 1 Algoritmo STL.

BUCLE EXTERNO:
for $\hat{k} = 1$ to n_e **do**

BUCLE INTERNO:
for $k = 1$ to n_i **do**

Inicializamos os valores para o suavizado de T_t e S_t dentro do bucle interno:

if $k == 1$ **then**
if $\hat{k} == 1$ **then**
$T_t^k = 0$; $\rho_t = 0$
end if
$T_t^k = T_t$
end if

Aplicamos a regresión LOESS a cada subserie dada polas observacións da estación j-ésima: ❶

for $j = 1$ to n_p **do**
$m = n/n_p$
$C_t^{k+1} = \texttt{LOESS}(X_{j+tm} - T_{j+tm}^k, d = 1, q = n_s)$
end for

Extráese en L_t^{k+1}a variabilidade de baixa frecuencia de C_t^{k+1}, máis propia da tendencia: ❷

$\text{aux} = \texttt{ma}(X_t - T_t^k, d = n_p)$
$\text{aux} = \texttt{ma}(\text{aux}, d = n_p)$
$\text{aux} = \texttt{ma}(\text{aux}, d = 3)$
$L_t^{k+1} = \texttt{LOESS}(\text{aux}, d = 1, q = n_l)$

Actualízanse os novos valores suavizados da tendencia e a compoñente estacional: ❸

$S_t^{k+1} = C_t^{k+1} - L_t^{k+1}$
$T_t^{k+1} = \texttt{LOESS}(X_t - S_t^{k+1}, d = 1, q = n_t)$
end for

Calcúlanse os "pesos de robustez" a partir do valor absoluto dos residuos: ❹

$T_t = T_t^{n_i}$; $S_t = S_t^{n_i}$; $R_t = X_t - T_t - S_t$
$h = \text{mediana}(|R_t|)$
$\rho_t = W_2(|R_t|/h)$
end for

❹ Hai un conxunto non trivial de series temporais para as que o algoritmo ata aquí descrito acada un funcionamento satisfactorio. Mais para series temporais que presenten un comportamento aberrante (non Gaussiano), con flutuacións abruptas e transitorias, é desexable que a descomposición sexa robusta, non dándolle importancia a estas flutuacións.

O xeito que temos de darlle menos importancia a certas observacións é definindo os "*pesos de robustez*", que son menores para observacións que explicamos peor a partir da suma das compoñentes. Estes calcúlanse consecuentemente a partir dos residuos $R_t = X_t - T_t - S_t$, como $\rho_t = W_2(|R_t|/h)$, onde $h = 6 \cdot \text{mediana}(|R_t|)$. Finalmente inclúense como pesos externos nas regresións LOESS da seguinte execución do bucle interno.

Bibliografía

Fontes Bibliográficas Principais

Brockwell, Peter J e Richard A Davis (1991). *Time series: theory and methods*. Springer science & business media.

Bartlett, Peter (2007). "Introduction to Time Series Analysis. Lectures." En: URL: https://www.stat.berkeley.edu/~bartlett/courses/fall2007/.

Hyndman, Rob J e George Athanasopoulos (2021). *Forecasting: principles and practice*. OTexts. URL: https://otexts.com/fpp3/.

Sobre a Descomposición e as Compoñentes das Series Temporais

Cleveland, Robert B et al. (1990). "STL: A seasonal-trend decomposition procedure based on loess". En: *J Off Stat* 6, pp. 3–73. URL: http://bit.ly/stl1990.

Sobre as Características das Series Temporais

Qian, Bo e Khaled Rasheed (2004). "Hurst exponent and financial market predictability". En: *IASTED conference on Financial Engineering and Applications*. Proceedings of the IASTED International Conference. Chicago Cambridge, MA, pp. 203–209. URL: https://c.mql5.com/forextsd/forum/170/hurst_exponent_and_financial_market_predictability.pdf.

Sobre a Análise dos Resultados dos Modelos

Makridakis, Spyros e Michele Hibon (2000). "The M3-Competition: results, conclusions and implications". En: *International journal of forecasting* 16(4), pp. 451–476.

Hanley, James A et al. (2001). "Visualizing the median as the minimum-deviation location". En: *The American Statistician* 55(2), pp. 150–152. URL: http://www.med.mcgill.ca/epidemiology/Joseph/publications/Methodological/median.pdf.

Hyndman, Rob J e Anne B Koehler (2006). "Another look at measures of forecast accuracy". En: *International journal of forecasting* 22(4), pp. 679–688.

Hyndman, Rob J (2020). "A brief history of forecasting competitions". En: *International Journal of Forecasting* 36(1), pp. 7–14.

Makridakis, Spyros, Evangelos Spiliotis e Vassilios Assimakopoulos (2020). "The M4 Competition: 100,000 time series and 61 forecasting methods". En: *International Journal of Forecasting* 36(1), pp. 54–74.

Sobre Procesos Gaussianos

Williams, Christopher KI e Carl Edward Rasmussen (2006). *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA. URL: http://gaussianprocess.org/gpml/chapters/RW.pdf.

Görtler, Jochen, Rebecca Kehlbeck e Oliver Deussen (2019). "A visual exploration of gaussian processes". En: *Distill* 4(4), e17. URL: https://distill.pub/2019/visual-exploration-gaussian-processes/.

Murphy, Kevin P (2023). *Probabilistic machine learning: Advanced topics*. MIT press. URL: https://archive.org/download/pml-book/book2.pdf.

Sobre Modelos baseados en Redes Neurais e Transformers

Bengio, Yoshua, Ian Goodfellow, Aaron Courville et al. (2017). *Deep learning*. Vol. 1. MIT press Cambridge, MA, USA.

Vaswani, Ashish et al. (2017). "Attention is all you need". En: *Advances in neural information processing systems* 30. URL: https://arxiv.org/pdf/2012.07436.

Zhou, Haoyi et al. (2021). "Informer: Beyond efficient transformer for long sequence time-series forecasting". En: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 12, pp. 11106–11115. URL: https://arxiv.org/pdf/2012.07436.

Gráficas

Love, Fraser (2024). *NNTikZ - TikZ Diagrams for Deep Learning and Neural Networks*. GitHub repository. URL: https://github.com/fraserlove/nntikz.

Publicacións Relevantes non Consultadas

Hurst, Harold Edwin (1951). "Long-term storage capacity of reservoirs". En: *Transactions of the American society of civil engineers* 116(1), pp. 770–799.

Box, George EP e David R Cox (1964). "An analysis of transformations". En: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 26(2), pp. 211–243.

Box, George EP, Gwilym M Jenkins et al. (1970). *Time series analysis: forecasting and control*. John Wiley & Sons. URL: https://elib.vku.udn.vn/bitstream/123456789/2536/1/1994.%20Time%20Series%20Analysis-Forecasting%20and%20Control.pdf.

Box, George EP e David A Pierce (1970). "Distribution of residual autocorrelations in autoregressive integrated moving average time series models". En: *Journal of the American statistical Association* 65(332), pp. 1509–1526. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=7cd4e19b3eeecf086574969a2cc9d5a4b987275b.

Ljung, Greta M e George EP Box (1978). "On a measure of lack of fit in time series models". En: *Biometrika* 65(2), pp. 297–303. URL: https://larrylisblog.net/WebContents/Financial%20Models/LjungBox.pdf.

Makridakis, Spyros, Allan Andersen et al. (1982). "The accuracy of extrapolation (time series) methods: Results of a forecasting competition". En: *Journal of forecasting* 1(2), pp. 111–153.

Glosario

Neste glosario recompílanse os termos tanto de uso habitual como esporádico en relación co traballo, xunto cunha pequena definición ou unha referencia á parte do texto na que se tratan.

A | **C** | **E** | **F** | **H** | **I** | **K** | **M** | **N** | **O** | **P** | **R** | **S** | **T** | **V**

A

Atípico Dato cuxo valor dista do que é esperable, podendo non ser representativo.

C

Característica (do inglés: «*features*»), indicadores numéricos que resumen algunha calidade ou aspecto dunha serie temporal.

Ciclos No contexto das series temporais son variacións nos datos que non teñen unha frecuencia fixa. Un exemplo son os ciclos climáticos ou os ciclos económicos.

Criterio de Cauchy Establece que se $\{x_n\}$ é unha sucesión nun espazo de Hilbert $\mathcal{H}$, entón $\{x_n\}$ converse en norma (cuadrática) se, e só se,

$$\|x_n - x_m\| \xrightarrow{n,m\to\infty} 0.$$

E

Erro cadrático medio (MSE) Ver Definición 22.

Espazo de Hilbert Un espazo de Hilbert $\mathcal{H}$ é un espazo vectorial dotado dun produto interior que ademais é completo respecto da métrica que induce dito produto.

Espazo de probabilidade Terna de tres elementos $(\Omega, \mathcal{A}, \mathbb{P})$, onde: Ω é o *espazo mostral*, $\mathcal{A}$ é o *espazo de sucesos* e $\mathbb{P}$ é unha *función de probabilidade*.

Esperanza condicionada Ver Definición 13.

Estacional (En inglés: «*seasonal*») Dise de algo que depende ou varía en función de períodos definidos de tempo (como as estacións). Non confundir con estacionario.

Estacionario (En inglés: «*stationary*») Dise de algo non evoluciona co tempo. Non confundir con estacional.

F

Filtro de paso baixo (En inglés: «*low-pass filter*») Filtro que só permite o paso das frecuencias graves. É un termo de uso habitual no contexto da enxeñería (ver: "filtro paso bajo" no *Diccionario Español de Ingeniería*).

Función característica Para unha variable aleatoria X, é unha función $\varphi_X : \mathbb{R} \to \mathbb{C}$ dada por

$$\begin{aligned}\varphi_X(t) = \mathbb{E}\left(e^{itX}\right) &= \int_{-\infty}^{\infty} e^{itx} dF_X(x) = \int_{-\infty}^{\infty} e^{itx} f_X(x)dx \\ &= \int_{-\infty}^{\infty} \cos(tx) f_X(x)dx + i\int_{-\infty}^{\infty} \operatorname{sen}(tx) f_X(x)dx.\end{aligned}$$

En particular, se X admite función de densidade, a súa función característica correspóndese coa transformada de Fourier da mesma.

Función de autocorrelación Ver Definición 10.

Función de autocorrelación parcial Ver Definición 11.

Función de autocovarianza Ver Definición 6.

Función de tipo núcleo É unha función K par, non negativa e con valores reais, tal que

$$\int_{-\infty}^{\infty} K(u)du = 1.$$

H

Heterocedasticidade Nun modelo de regresión, cualidade que se presenta cando a varianza dos erros non é constante/homoxénea ao longo do dominio. Contraposición de homocedasticidade.

Homocedasticidade Nun modelo de regresión, cualidade que se presenta cando a varianza dos erros é constante/homoxénea ao longo do dominio. Contraposición de heterocedasticidade.

I

Innovación Ver Definición 14.

K

Kernel Un kernel de Mercer é unha función $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, onde $\mathcal{X}$ é un espazo de Hilbert, cumprindo que

$$\sum_{i=1}^{n}\sum_{j=0}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_j)c_i c_j \geq 0,$$

para todo $n \in \mathbb{Z}^+$, $\boldsymbol{x}_\cdot \in \mathcal{X}$ e $c_\cdot \in \mathbb{R}$. A condición é análoga a que a matriz de Gram dada por

$$\mathcal{K} = \begin{pmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & \dots & K(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_n, \boldsymbol{x}_1) & \dots & K(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix},$$

sexa definida positiva para cada conxunto de elementos de $\mathcal{X}$, $\{\boldsymbol{x}_i\}_{i=1}^n$.

M

Matriz de Covarianza Para un vector finito de variables aleatoria, matriz formada polas covarianzas dous a dous dos seus elementos.

Media móbil (En inglés: «*mean average*», MA) No contexto das series temporais, é unha operación coa que se obtén unha nova serie temporal cuxas observacións son medias dun subconxunto dos datos orixinais.

Momentos Para unha función de probabilidade, son medidas cuantitativas que, en conxunto, determinan completamente a súa distribución.

N

Nivel En modelos de alisado exponencial (tratados na Sección 2.2), compoñente dunha serie temporal que expresa a maior parte da variabilidade da mesma de xeito "suave". A súa expresión concreta dependerá do modelo de alisado exponencial empregado.

O

Operador de retardo Ver Definición 5.

P

Pendente En modelos de alisado exponencial (tratados na Sección 2.2), é a compoñente que caracteriza a serie derivada suavizada.

Proceso estocástico Ver Definición 2.

Proceso estocástico gaussiano Ver Definición 38.

Procesos $\mathrm{ARIMA}(p,d,q)$ Ver Definición 36.

Procesos $\mathrm{ARMA}(p,q)$ Ver Definición 32.

Procesos $\mathrm{AR}(p)$ Ver Definición 31.

Procesos $\mathrm{MA}(q)$ Ver Definición 30.

Procesos $\mathrm{SARIMA}(p,d,q) \times (P,D,Q)$ Ver Definición 37.

Procesos Gaussianos Metodoloxía bayesiana e non paramétrica de regresión para series temporais. Ver Sección 3.1.

R

Realización Ver Definición 3.

Ruído branco Ver Definición 29.

S

Serie temporal Ver Definición 1.

Subserie (temporal) estacional Subserie formada polas observacións correspondentes con unha única estación dunha serie estacional, de xeito que o seu número de observacións coincidirá co número de ciclos estacionais da serie orixinal.

T

Tendencia Para unha serie temporal, é a evolución suavizada e a longo prazo.

Transformacións Box-Cox Ver Definición 15.

V

Variable aleatoria Sobre un espazo de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é unha función medible do propio espazo de probabilidade nun espazo medible (S, Σ) (normalmente $\mathbb{R}$). Asocia un valor (real) a cada posible suceso.