Traballo Fin de Mestrado

Selectores bootstrap do parámetro de suavizado para a estimación non paramétrica da función de densidade con datos dependentes

Inés Barbeito Cal

Mestrado en Técnicas Estatísticas

Curso 2015-2016

Proposta de Traballo Fin de Mestrado

Título en galego: Selectores bootstrap do parámetro de suavizado para a estimación non paramétrica da función de densidade con datos dependentes

Título en español: Selectores bootstrap del parámetro de suavizado para la estimación no paramétrica de la función de densidad con datos dependientes

English title: Bootstrap bandwidth selection for non parametric density function estimation with dependent data

Modalidade: A

Autora: Inés Barbeito Cal, Universidade da Coruña

Director: Ricardo Cao Abad, Universidade da Coruña

Breve resumo do traballo:

Abórdase unha revisión bibliográfica sobre as técnicas bootstrap nunha situación de dependencia xeral (bootstrap por bloques, bootstrap estacionario e submostraxe), centrándose no caso particular do bootstrap suavizado con datos dependentes. Posteriormente, estúdanse selectores do parámetro de suavizado, centrándonos nos escollidos mediante bootstrap, para a estimación non paramétrica da función de densidade no contexto de dependencia. Obtense, ademais, unha expresión exacta para o $MISE^*(h)$ baixo dependencia. Finalmente, realízase un estudo de simulación comparativo entre diferentes selectores do parámetro ventá.

O abaixo asinante, Ricardo Cao Abad, Catedrático de Universidade da área de Estatística e Investigación Operativa do Departamento de Matemáticas da Universidade da Coruña, informa de que o Traballo de Fin de Mestrado titulado

Selectores bootstrap do parámetro de suavizado para a estimación non paramétrica da función de densidade con datos dependentes

foi realizado baixo a súa dirección por Dona Inés Barbeito Cal para o Mestrado en Técnicas Estatísticas. Ámbolos dous estiman que o traballo está rematado e dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En A Coruña, a 8 de xaneiro de 2016.

O director: A autora:

Don Ricardo Cao Abad Dona Inés Barbeito Cal

Índice xeral

Resumo

Resumo en galego

Para poder estudar e tratar unha variable aleatoria é necesario coñecer a súa distribución. Ademais, a distribución daquelas variables aleatorias que sexan absolutamente continuas pódese caracterizar pola súa función de densidade. Nembargantes, a maioría das veces que queremos estudar unha característica dunha poboación, o único que temos da mesma é unha mostra. Neste contexto, existen numerosos estudos sobre a estimación non paramétrica da función de densidade, que se pode obter a partir desa mostra. Ademais, é de vital importancia unha boa elección do parámetro ventá, h, que regula o grao de suavización do estimador. Neste traballo, abordaremos un estudo da elección de dito parámetro nun contexto de dependencia.

En primeiro lugar, levamos a cabo unha revisión bibliográfica acerca das técnicas bootstrap válidas en ausencia dunha expresión explícita que modelice a dependencia dunha serie temporal (bootstrap por bloques, bootstrap estacionario e submostraxe). Ademais, estúdanse diferentes selectores de ventá, e establécese unha suavización do bootstrap estacionario de Politis e Romano (1994a). Neste contexto obtense unha expresión exacta para o análogo bootstrap do erro cuadrático medio integrado baixo dependencia. Deste xeito, minimizando o mesmo, pode obterse un parámetro ventá sen necesidade de facer unha aproximación por Monte Carlo. Finalmente, lévase a cabo un estudo de simulación para comparar os distintos selectores de ventá; e ilústrase cunha aplicación práctica dos mesmos, abordando a estimación da densidade cun conxunto de datos reais.

English abstract

In order to study and deal with a random variable, it is necessary to know its distribution. In addition, the distribution of those random variables that are absolutely continuous is characterized by their density function. However, whenever we want to study a characteristic of a population, most of the times the only information we have about it is a sample. In this context, there are numerous studies related to the non parametric estimation of the density function that can be obtained from that sample. Besides, it is of vital importance a good election of the bandwidth parameter, h, which regulates the degree of smoothness of the estimator. In this work, we will carry out a study on the selection of the smoothing parameter in a context of dependence.

First, a bibliographic review is included regarding bootstrap techniques for time series without an explicit expression that models the dependence (moving blocks bootstrap, stationary bootstrap and subsampling). Furthermore, different bandwidth selectors are studied, and a smoothed version of the stationary bootstrap by Politis and Romano (1994a) is established. An exact expression for the bootstrap version of the mean integrated squared error under dependence is obtained in this context. Finally, a simulation study is carried out to compare the different bandwidth selectors. A practical application of these selectors is illustrated with density estimation using a real data set.

Prefacio

A estimación non paramétrica da función de densidade é un tema que se leva estudando ó longo dos últimos cincuenta anos, xa que é unha das formas de caracterizar unha variable aleatoria, foco de estudo da Estatística. Dende os inicios da estimación non paramétrica de curvas co uso do histograma como estimador da función de densidade, pasando polo histograma móbil, ata o estimador da densidade tipo núcleo; un dos principais problemas cos que nos atopamos é a correcta elección do parámetro de suavizado, h, que controla o grao de suavización do estimador. Polo tanto, unha mala escolla do mesmo pode provocar que o estimador da densidade sexa infrasuavizado ou sobresuavizado. Na literatura existente é posible atopar numerosos estudos acerca da selección do parámetro de suavizado baixo o contexto de independencia dos datos empregados.

Na actualidade, sen embargo, moitos conxuntos de datos dependen do tempo, xa que evolucionan ao longo do mesmo. Sen ir máis lonxe, moitos conxuntos de datos relativos á economía ou á sociedade variarán en función do tempo. Abordarase entón neste traballo o estudo de diferentes selectores de ventá cando os datos son dependentes.

Unha das aplicacións máis interesantes da estimación non paramétrica da densidade, ademais de poder coñecer a nivel exploratorio a estructura do conxunto de datos que estamos a tratar, é poder extraer conclusións para unha determinada poboación a partir dunha mostra que sexa representativa da mesma. Isto é de especial relevancia xa que en numerosas ocasións non é posible obter os datos de toda a poboación de estudo. Cobran gran importancia neste ámbito as técnicas de remostraxe, entre as que destaca o bootstrap. Con elas, a partir dunha mostra que sexa representativa da poboación, é posible xerar tantos datos como se queira cunha distribución similar á distribución da poboación orixinal. Centrarémonos neste tipo de técnicas neste traballo, cuxo obxectivo será establecer o parámetro de suavizado, h, para o cal se minimice un criterio de erro a saber, o erro cuadrático medio integrado (ou $MISE(h)$). Abordarase a obtención do selector h dende un punto de vista que difire do habitual, de xeito que se propoñerá unha expresión exacta para a versión bootstrap de dito criterio de erro ($MISE^*(h)$), en vez de tratar de obter unha aproximación do mesmo por Monte Carlo. Desta forma, aforrarase tempo de execución, así como se obterá un selector de ventá máis preciso, sen erros da aproximación de Monte Carlo.

Capítulo 1

Introdución

O principal obxectivo da Inferencia Estatística é intentar extraer conclusións para unha determinada poboación a partir dunha mostra da mesma. Algunhas das súas principais finalidades son: a estimación puntual dun parámetro, a construcción de intervalos de confianza para o mesmo, e a realización de contrastes de hipóteses. Porén, para facer este tipo de tarefas precísase do coñecemento dalgunha das funcións que caracterizan a distribución (como son a función de densidade, ou a función de distribución) que segue a poboación a estudar, que na práctica non sempre resultan coñecidas.

Unha das formas de resolver este problema é supoñer que a función de distribución pertence a unha familia paramétrica e estimar eses parámetros. Non obstante, neste traballo empregaremos un enfoque non paramétrico, no cal estimaremos algunhas das curvas que caracterizan a poboación (en particular, a función de densidade), sen necesidade de supoñer que a función de distribución pertence a unha familia paramétrica.

Consideramos entón o estimador non paramétrico da función de densidade de Parzen-Rosenblatt, que vén dado por:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \tag{1.1}$$

onde K é unha función tipo núcleo ($K_h(u) = \frac{1}{h}K(\frac{u}{h})$), a cal se pode pedir que sexa simétrica en torno ó cero; e $h > 0$, o parámetro de suavizado (ou parámetro ventá) que regula o tamaño do entorno que se usa para levar a cabo a estimación, e será moi importante para unha correcta estimación. Ademais, é habitual esixir que a función núcleo, K, sexa non negativa e a súa integral sexa 1:

$$K(u) \geq 0, \forall u, \int_{-\infty}^{+\infty} K(u)du = 1.$$

Vexamos un pouco máis a fondo as propiedades do estimador de Parzen-Rosenblatt, que empregaremos no Capítulo 4. Para un valor x fixo, o valor esperado de dito estimador será

$$\mathbb{E}(\hat{f}_h(x)) = \mathbb{E}[K_h(x - X_1)] = \int K_h(x - y)f(y)dy = (K_h * f)(x) = \int K(u)f(x - hu)du,$$

onde $*$ denota a convolución. Isto quere dicir que a curva esperada que se obtén co estimador tipo núcleo da densidade de Parzen-Rosenblatt non é a verdadeira densidade f, senón unha versión suavizada da mesma, dada por $(K_h * f)$.

Desenvolvemos cun pouco máis detalle os cálculos de esperanza e varianza de $\hat{f}_h(x)$, para un x fixo, co fin de describir cal é o efecto de h sobre as mesmas. Baixo condicións de regularidade sobre a función de densidade real, f, pódese aproximar empregando un desenvolvemento de Taylor:

$$f(x-hu) = f(x) - huf'(x) + \frac{1}{2}(hu)^2 f''(x) + o(h^2).$$

Supoñendo que $\mu_2(K) = \int u^2 K(u)du < \infty$, entón:

$$\mathbb{E}(\hat{f}_h(x)) = f(x) + \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^2),$$

onde se emprega que K é unha función de densidade simétrica.

No caso da varianza, a expresión resulta como segue

$$\begin{aligned} Var(\hat{f}_h(x)) &= \frac{1}{n}Var(K_h(x-X_1)) \\ &= \frac{1}{n}\left[\mathbb{E}(K_h^2(x-X_1)) - \mathbb{E}^2(K_h(x-X_1))\right] \\ &= \frac{1}{n}\left[\int K_h^2(x-y)f(y)dy - \left(\int K_h(x-y)f(y)dy\right)^2\right] \\ &= \frac{1}{nh}\int K^2(u)f(x-hu)du - \frac{1}{n}\left(\int K(u)f(x-hu)du\right)^2 \\ &= \frac{1}{nh}\int K^2(u)(f(x)+o(1))du - \frac{1}{n}(f(x)+o(1))^2 \\ &= \frac{1}{nh}R(K)f(x) + o((nh)^{-1}), \end{aligned}$$

sendo $R(K) = \int K^2(u)du$.

Polo tanto, para analizar as propiedades do estimador de Parzen-Rosenblatt de forma asintótica, estudaremos que ocorre cando $n \to \infty$, $h \to 0$ e $nh \to \infty$. Temos entón, no caso da esperanza, cando $h \to 0$:

$$\mathbb{E}(\hat{f}_h(x)) - f(x) = \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^2),$$

e no caso da varianza, cando $nh \to \infty$:

$$\mathrm{Var}(\hat{f}_h(x)) = \frac{1}{nh}R(K)f(x) + o((nh)^{-1}).$$

Como observamos, pódese concluir que o sesgo diminúe cando $h \to 0$. É dicir, para valores de h pequenos, teremos estimadores centrados, pero a costa de que incremente a varianza. Serán estimadores infrasuavizados. Por outra banda, se h é grande, redúcese a varianza, pero a expensas de aumentar o sesgo, dando lugar a estimadores sobresuavizados.

En xeral, para avaliar o comportamento dun estimador da densidade (como por exemplo, o de Parzen-Rosenblatt), debemos definir medidas de erro adecuadas ó contexto no que nos atopamos. En particular, poderíamos considerar medidas de erro locais (é dicir, para un x fixo) ou medidas de erro globais (é dicir, para toda a curva estimada). Ademais, estas medidas de erro proporcionarannos funcións obxectivo a minimizar para obter valores óptimos do parámetro de suavizado, h.

Neste momento, introduciremos as principais medidas de erro empregadas na literatura. Posteriormente, veremos as ventás óptimas obtidas dende un punto de vista teórico, baixo hipótese de

independencia.

Sexa $\hat{f}_h$ o estimador tipo núcleo de Parzen-Rosenblatt da función de densidade. Consideraremos $\hat{f}_h(x)$ o estimador puntual para $f(x)$, dado x fixo. Podemos utilizar unha medida de erro local como o erro cuadrático medio (MSE), definido como segue

$$MSE_x(h) = \mathbb{E}(\hat{f}_h(x) - f(x))^2 + \mathrm{Var}(\hat{f}_h(x)) = \mathrm{Sesgo}^2(\hat{f}_h(x)) + \mathrm{Var}(\hat{f}_h(x)).$$

Para poder dar unha expresión completa deste erro local necesitamos empregar as expresións asintóticas explícitas do sesgo e da varianza, obtidas anteriormente. Deste xeito,

$$MSE_x(h) = \frac{1}{nh}R(K)f(x) + \frac{1}{4}h^4\mu_2(K)^2 f''(x)^2 + o((nh)^{-1} + h^4). \tag{1.2}$$

Unha vez obtido un criterio de erro, resulta lóxico preguntarse qué valor de h proporciona o erro mínimo. Sen embargo, minimizar en h o $MSE_x(h)$ non resulta sinxelo, xa que aparecen termos residuais das aproximacións asintóticas que realizamos. Consecuentemente, presentaremos unha versión asintótica do MSE, o erro cuadrático medio asintótico (ou $AMSE$), que será a que minimizaremos:

$$AMSE(h) = \frac{1}{nh}R(K)f(x) + \frac{1}{4}h^4\mu_2(K)^2 f''(x)^2.$$

Logo, sempre e cando $f''(x) \neq 0$, entón o valor de h que minimiza o $AMSE(\hat{f}_h(x))$ será:

$$h_{AMSE} = \left(\frac{R(K)f(x)}{n\mu_2(K)^2 f''(x)^2}\right)^{1/5}.$$

Deste xeito, obtemos unha ventá local, xa que depende do punto x. Por outra banda, non é directamente aplicable na práctica, xa que depende da segunda derivada da densidade e da propia función de densidade no punto x, sendo esta descoñecida.

Unha vez obtidas unha medida de erro local e a expresión dunha ventá óptima, tamén local; obteremos agora unha medida de erro global, que nola proporciona o erro cuadrático medio integrado ($MISE$), que podemos obter de dous xeitos equivalentes:

- Promediando o erro cuadrático integrado (ISE).
- Integrando o erro cuadrático medio (MSE).

Definamos, en primeiro lugar, o erro cuadrático integrado (ou ISE).

$$ISE(X_1, \ldots, X_n; h) = \int \left(\hat{f}_h(x) - f(x)\right)^2 dx,$$

que é unha cantidade aleatoria que depende tanto da mostra considerada como do parámetro de suavizado (h). Se promediamos este erro obtemos o $MISE$, que tamén se pode entender como a integral do MSE:

$$MISE(h) = \mathbb{E}(ISE(X_1, \ldots, X_n; h)) = \mathbb{E}\left(\int \left(\hat{f}_h(x) - f(x)\right)^2 dx\right) = \int MSE_x(h)dx.$$

Neste caso, por medio de desenvolvementos de Taylor pode obterse a seguinte expresión para o $MISE$:

$$MISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2 R(f'') + o((nh)^{-1} + h^4), \tag{1.3}$$

e a súa aproximación asintótica, o erro cuadrático medio integrado asintótico (ou $AMISE$), vén dada por:

$$AMISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f''). \tag{1.4}$$

Se agora minimizamos esta función con respecto a h, obtemos unha ventá óptima:

$$h_{AMISE} = \left(\frac{R(K)}{\mu_2(K)^2R(f'')n}\right)^{1/5}. \tag{1.5}$$

Neste caso, trátase dunha ventá global pero que depende da integral da derivada segunda da función de densidade, que é a función que desexamos estimar. Logo, malia ter unha ventá óptima en sentido teórico, non podemos empregala na práctica.

Dado que as ventás óptimas e asintoticamente óptimas que acabamos de considerar non se poden empregar na práctica, xa que dependen da verdadeira función de densidade (que resulta descoñecida) e a súa derivada segunda; precisaremos dalgún método para construir selectores de ventá a partir dos datos. En concreto, trataremos, primeiramente nun contexto de independencia, os métodos plug-in, validación cruzada e bootstrap. En segundo lugar, presentaremos distintos métodos para o caso no que os datos sexan dependentes. Concretamente, describiremos as metodoloxías plug-in, validación cruzada e, finalmente, proporemos e estudaremos unha versión suavizada do bootstrap estacionario de Politis e Romano (1994a), co fin de obter unha expresión explícita para a versión bootstrap do $MISE$.

Así, no Capítulo 2, presentaremos unha revisión sobre os contextos de dependencia paramétricos e xerais. A continuación, no Capítulo 3, describiremos as ideas básicas da metodoloxía bootstrap nun contexto non paramétrico. En primeiro lugar, trataremos o bootstrap uniforme, suavizado e o método de submostraxe nun contexto de independencia. En segundo lugar, centrarémonos no caso de que os datos sexan dependentes, empregando para iso os contextos de dependencia xeral descritos no Capítulo 2, para así describir a metodoloxía bootstrap por bloques, bootstrap estacionario e o método de submostraxe nun contexto de dependencia. Ademais, presentaranse resultados teóricos que garanten a validez de cada un dos métodos expostos. No Capítulo 4, farase unha revisión dos principais selectores do parámetro ventá nos contextos de independencia e de dependencia. Estudarase con profundidade o bootstrap estacionario suavizado, así como a proposta da expresión explícita para a versión bootstrap do $MISE$. Posteriormente, no Capítulo 5, realizaremos un estudo de simulación co fin de comparar os distintos selectores de ventá estudados no Capítulo 4. Ademais, centrarémonos no selector derivado de minimizar a expresión explícita para o $MISE^*(h)$, co fin de probar o seu bo comportamento na práctica. Da mesma forma, no Capítulo 6 ilustraremos os métodos estudados coa estimación da función de densidade a partir de dous conxuntos de datos reais cos distintos parámetros de suavizado. Finalmente, adicaremos o Capítulo 7 a presentar as conclusións máis relevantes: tanto revisión dos métodos estudados, como a nova proposta e o comportamento da mesma.

Capítulo 2

Datos dependentes

Unha hipótese que xeralmente se asume en numerosas situacións contempladas pola Inferencia Estatística é a independencia das observacións mostrais, é dicir, as variables aleatorias que conforman a mostra, $X_1, X_2, \dots, X_n$, son independentes.

Imaxinemos, sen embargo, unha variable aleatoria X que recolla o número de accidentes na estrada en A Coruña cada ano. Baixo a hipótese de independencia, para calquera momento no tempo a probabilidade de observar un valor en particular non depende do valor observado no ano anterior. Non obstante, o número de accidentes en anos sucesivos aseméllase moito máis que en anos moi distantes. Existen numerosos factores (económicos, sociais, políticos,...) que cambian a medida que o tempo pasa, e que fan que as observacións temporalmente máis achegadas se asemellen máis que aquelas separadas por un longo período de tempo. Logo, neste caso non nos atopamos ante un conxunto de datos independentes e identicamente distribuídos, xa que evolucionan ao longo do tempo. A este tipo de conxuntos de datos denominarémolos series de tempo.

Formalmente, unha serie de tempo é unha colección de observacións dunha variable, X, recollidas secuencialmente ó longo do tempo. Como xa dixemos, éstas non se poderán enmarcar dentro dun contexto de independencia, xa que precisamente ditas observacións dependerán unhas doutras ao longo do tempo.

Consideremos agora un proceso estocástico en tempo discreto e con espacio de estados continuo (normalmente, o conxunto dos números reais), é dicir, un conxunto de variables aleatorias $\{X_t\}_{t\in\mathbb{Z}}$ definidas todas no mesmo espacio de probabilidade, sendo $\mathbb{Z}$ o conxunto dos números enteiros. Supoñeremos ademais que somos capaces de observar parte da súa traxectoria, é dicir, dada unha variable X que foi observada nos instantes $1, 2, \dots, n$, a serie de tempo observada da variable X será representada por $X_1, \dots, X_n$; e será unha realización dun proceso estocástico. Noutras palabras, $X_1, \dots, X_n$ será unha mostra de datos dependentes.

Presentaremos a continuación unha revisión das condicións máis comúns de dependencia, así como dos modelos paramétricos de dependencia, recollidos por Doukhan *et al.* (2010).

2.1. Modelos paramétricos de dependencia

Existen numerosos modelos paramétricos para o tratamento de datos dependentes. Consideraremos, primeiramente, un proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$ xerador da serie de tempo que admite a representación:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t, \tag{2.1}$$

onde $c, \phi_1, \phi_2, ..., \phi_p$ son constantes con a_t independentes de $X_{t-1}, X_{t-2}, ...$ Trátase dun proceso autorregresivo de orde p ($AR(p)$).

En segundo lugar, sexa un proceso $\{X_t\}_{t\in\mathbb{Z}}$ que admite a seguinte representación:

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}, \tag{2.2}$$

onde $c, \theta_1, \theta_2, ..., \theta_q$ son constantes. Neste caso, trátase dun proceso de medias móbiles de orde q ($MA(q)$).

En terceiro lugar, un proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$ que admite a representación:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}, \tag{2.3}$$

onde $c, \phi_1, \phi_2, ..., \phi_p, \theta_1, \theta_2, ..., \theta_q$ son constantes, é coñecido como un proceso $ARMA(p, q)$.

Un proceso $\{X_t\}_{t\in\mathbb{Z}}$ que admite a representación:

$$\phi(B)(1-B)^d X_t = c + \theta(B) a_t, \tag{2.4}$$

onde B denota o operador retardo, definido por $BX_t = X_{t-1}$, sendo

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q,$$

é coñecido como un proceso $ARIMA(p, d, q)$, con d diferenzas regulares.

Finalmente, un proceso $\{X_t\}_{t\in\mathbb{Z}}$ que admite a representación:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = c + \theta(B)\Theta(B^s) a_t, \tag{2.5}$$

onde B^s é o operador retardo estacional definido como $B^s X_t = X_{t-s}$, sendo

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps},$$

$$\Theta(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs},$$

é coñecido como un proceso $ARIMA(p, d, q) \times (P, D, Q)_s$, con d diferenzas regulares e D diferenzas estacionais de período s.

Nas expresións (2.1),(2.2),(2.3),(2.4),(2.5) o conxunto $\{a_t\}_{t\in\mathbb{Z}}$ denota unha colección de variables aleatorias incorreladas, con media 0 e varianza finita σ_a^2, que denominaremos ruído branco. Se este é gaussiano, entón as variables aleatorias que o conforman son *iid*.

2.2. Situacións de dependencia xeral

Noutras ocasións, non se asume ningún tipo de estructura paramétrica sobre o proceso estocástico, senón que se establece algunha condición de dependencia xeral. Consideremos $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estocástico. Normalmente, supoñeremos que dito proceso é estacionario. A continuación, presentaremos varias condicións de dependencia que empregaremos nos Capítulos 3 e 4.

2.2.1. Fortemente mixing (ou α-mixing)

Consideremos $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario. A condición de ser α-mixing, recollida en Doukhan *et al.* (2010), establece que a dependencia entre as variables aleatorias que conforman as observacións da mostra vese atenuada a medida que os seus instantes temporais se distancian, isto é:

$$\sup_{A\in\mathcal{F}_1^n, B\in\mathcal{F}_{n+k}^\infty}|\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)|\leq\alpha_k,$$

con $\alpha_k\to 0$, e sendo $\mathcal{F}_s^t$ a σ-álxebra xerada polas variables aleatorias $X_s,\dots,X_t$.

2.2.2. Uniformemente mixing (ou ϕ-mixing)

Novamente, consideremos $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario. Diremos que $\{X_t\}_{t\in\mathbb{Z}}$ é uniformemente mixing, acorde á literatura existente (Doukhan *et al.*, 2010), se se verifica:

$$|\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)|\leq\phi_k\mathbb{P}(A),\forall A\in\mathcal{F}_1^n,\forall B\in\mathcal{F}_{n+k}^\infty,\ \text{con}\ \phi_k\to 0.$$

Analogamente á definición de fortemente mixing, afirmar que un proceso $\{X_t\}_{t\in\mathbb{Z}}$ sexa uniformemente mixing implica que a dependencia existente entre as variables aleatorias que conforman as observacións mostrais debilítase conforme os seus instantes temporais se afastan cada vez máis.

2.2.3. m-dependente

Sexa $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario. Acorde á definición dada por Billingsley (1968), diremos que o proceso estocástico é m-dependente se para calesquera números enteiros k e l maiores que cero, $(X_1,\dots,X_k)$ e $(X_{k+n},\dots,X_{k+n+l})$ son independentes sempre que $n>m$.

Nótese que nesta terminoloxía unha secuencia independente é 0-dependente. Ademais, se un proceso estocástico é m-dependente tamén será α-mixing con $\alpha_n=0,\forall n>m$, e ϕ-mixing con $\phi_n=0,\forall n>m$.

2.2.4. Ψ dependente

Esta noción de dependencia débil foi introducida por Hwang e Shin (2012), e precisa da previa definición dalgunhas clases de funcións.

Sexa $\mathbb{L}^\infty=\bigcup_{n=1}^{\infty}\mathbb{L}^\infty(\mathbb{R}^n)$ o conxunto de funcións acotadas que toman valores reais no espazo $\mathbb{R}^n$, para algún $n=1,2,\cdots$, e tomemos a función $G:\mathbb{R}^n\to\mathbb{R}$, considerando a norma l_1 para $\mathbb{R}^n$. Definiremos o módulo Lipschitz de G como segue:

$$\text{Lip}(G)=\sup_{x\neq y}\frac{|G(x)-G(y)|}{||x-y||_1}.$$

Ademais, sexa $\mathcal{L}=\bigcup_{n=1}^{\infty}\mathcal{L}_n$, onde $\mathcal{L}_n=\{G\in\mathbb{L}^\infty(\mathbb{R}^n);\text{Lip}(G)<\infty,||G||_\infty\leq 1\}$. Consideremos agora as seguintes dúas funcións:

$$\Psi_1(G,H,n,m)=\min(n,m)\text{Lip}(G)\text{Lip}(H)$$

$$\Psi_2(G,H,n,m)=4(n+m)\text{Lip}(G)\text{Lip}(H),$$

onde G e H son funcións definidas en $\mathbb{R}^n$ e $\mathbb{R}^m$, respectivamente.

Supoñamos $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estocástico estacionario, diremos que é $(\theta,\mathcal{L},\Psi)$-dependente (ou simplemente Ψ-dependente), se existe unha sucesión $\theta=(\theta_r)_{r\in\mathbb{Z}}$ decrecente e tendendo a 0 a medida

que r tende a ∞; e unha función Ψ con argumentos $(G, H, n, m) \in \mathcal{L}_n \times \mathcal{L}_m \times \mathbb{N}^2$ tales que, dada unha n-tupla $(i_1, \ldots, i_n)$ e unha m-tupla $(j_1, \ldots, j_m)$, con $i_1 \leq \cdots \leq i_n < i_n + r \leq j_1 \leq \cdots \leq j_m$, verifícase que:

$$|\mathrm{Cov}(G(X_{i_1}, \ldots, X_{i_n}), H(X_{j_1}, \ldots, X_{j_m}))| \leq \Psi(G, H, n, m)\theta_r.$$

Capítulo 3

Metodoloxía bootstrap nun contexto non paramétrico

Neste capítulo revisarase a metodoloxía bootstrap tanto nun contexto de independencia como de dependencia.

Trataremos, en primeiro lugar, as ideas básicas do método bootstrap nun contexto de independencia, recollidas por Efron e Tibishirani (1986). Consideremos $\vec{X}= (X_1, \dots, X_n)$ unha mostra aleatoria simple procedente dunha poboación con distribución F, descoñecida, e supoñamos que queremos facer inferencia sobre un estatístico $R = R(\vec{X}, F)$. Non obstante, como a distribución F é descoñecida, teremos que estimala dalgún xeito. Un dos métodos que imos empregar é o método bootstrap. Para iso, consideramos $\hat{F}$ unha estimación de F, e condicionalmente a esta mostra aleatoria simple, obteremos unha remostra $\vec{X}^* = (X_1^*, \dots, X_n^*)$ con distribución $\hat{F}$. Consideramos a distribución na remostraxe (distribución bootstrap) do estatístico $R^* = R(\vec{X}^*, \hat{F})$, e aproximaremos a distribución na mostraxe de R pola distribución bootstrap de R^*. Outro problema ó que nos enfrontamos na práctica é que a distribución bootstrap de R^* é poucas veces calculable directamente, entón procederemos por Monte Carlo para poder aproximala.

Distinguiremos a continuación varios casos segundo a estimación que empreguemos para a distribución poboacional: o bootstrap uniforme, o bootstrap suavizado e o método de submostraxe para datos independentes.

3.1. Bootstrap uniforme

Trataremos primeiramente o bootstrap uniforme, no cal o estimador, $\hat{F}$, da distribución descoñecida, F, é a distribución empírica F_n, de modo que $\hat{F} = F_n$. Esta distribución defínese como segue:

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

Deste xeito, seguimos o plan de remostraxe, coñecido como bootstrap uniforme:

1. Para cada $i = 1, \dots, n$, xerar X_i^* a partir da distribución empírica F_n, é dicir, $\mathbb{P}^*(X_i^* = X_j) = \frac{1}{n}, j = 1, \dots, n$.

2. Obter $\vec{X}^* = (X_1^*, \dots, X_n^*)$.

3. Calcular o estatístico $R^* = (\overrightarrow{X}^*, F_n)$.

Como xa dixemos, no caso de non poder obter exactamente a distribución bootstrap de R^*, procederemos a aproximala por Monte Carlo, lanzando unha gran cantidade, B, de réplicas do estatístico na remostraxe R^*. Neste caso, o algoritmo contaría con dous pasos a maiores dos propostos anteriormente, resultando o seguinte plan de remostraxe:

1. Para cada $i = 1, \dots, n$, xerar X_i^* a partir da distribución empírica F_n, é dicir, $\mathbb{P}^*(X_i^* = X_j) = \frac{1}{n}, j = 1, \dots, n$.
2. Obter $\overrightarrow{X}^* = (X_1^*, \dots, X_n^*)$.
3. Calcular o estatístico $R^* = (\overrightarrow{X}^*, F_n)$.
4. Repetir B veces os pasos anteriores para obter as B réplicas bootstrap do estatístico R^*, resultando $R^{*(1)}, \dots, R^{*(B)}$.
5. Empregar estas réplicas bootstrap para aproximar a distribución na mostraxe de R.

3.2. Bootstrap suavizado

No caso de que F sexa unha distribución continua debemos introducir esa información á hora de proceder por bootstrap. Polo tanto, empregaremos un estimador da función de densidade e remostrexaremos del, o que se coñece como bootstrap suavizado.

A partires do estimador da función de densidade de Parzen-Rosenblatt, introducido no Capítulo 1, presentaremos o método bootstrap suavizado, que conta co seguinte plan de remostraxe:

1. A partir da mostra $(X_1, \dots, X_n)$ e empregando un valor $h > 0$ como parámetro de suavizado, calcúlase o estimador de Parzen-Rosenblatt $\hat{f}_h$.
2. Xéranse remostras bootstrap $\overrightarrow{X}^* = (X_1^*, \dots, X_n^*)$ a partir da densidade $\hat{f}_h$.
3. Obtense o estatístico na remostraxe $R^* = R(\overrightarrow{X}^*, \hat{F}_h)$.

O estimador, $\hat{F}_h$, da función de distribución utilizado no paso 3 defínese mediante:

$$\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t)dt = \frac{1}{n}\sum_{i=1}^{n} \mathbb{K}\left(\frac{x - X_i}{h}\right),$$

onde $\mathbb{K}(u) = \int_{-\infty}^{u} K(v)dv$. Este é o estimador chamado empírico suavizado da función de distribución.

Analogamente ó que ocorría co plan de remostraxe do bootstrap uniforme, de non ser posible o cálculo directo da distribución bootstrap do estatístico R^*, procederemos a aproximala por Monte Carlo, é dicir, repetiremos B veces os pasos 1-3 para obter réplicas bootstrap $R^{*(1)}, \dots, R^{*(B)}$; e empregaremos estas réplicas para aproximar a distribución na remostraxe de R^*.

3.3. Método de submostraxe para datos independentes

O método de submostraxe foi proposto por Politis e Romano (1994b). Como veremos na Sección 3.6, existen dúas versións para este método: unha para datos dependentes e outra para datos independentes, esta última descríbese a continuación

Consideremos as observacións $X_1, \dots, X_n$ que proveñen de variables aleatorias *iid* con distribución F, e sexa $\theta = \theta(F)$ un parámetro, $T_n = T_n(X_1, \dots, X_n)$ un estimador do mesmo, e $J_n(\cdot, F)$ a función de distribución na mostraxe de $\tau_n(T_n - \theta)$, é dicir, $J_n(u, F) = \mathbb{P}(\tau_n(T_n - \theta) \leq u)$. Fixemos un enteiro $b < n$, e definamos:

$$S_{n,i} = T_b(Y_i), i = 1, 2, \dots, N$$

sendo $Y_1, Y_2, \dots, Y_N$ todas as $N = \binom{n}{b}$ submostras sen reemprazamento posibles de tamaño b da mostra orixinal.

A función de distribución empírica que debemos empregar como aproximación da distribución na mostraxe de $\tau_n(T_n - \theta)$ será a correspondente ós valores $\tau_b(S_{n,i} - T_n)$, isto é:

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

Presentaremos, en segundo lugar, a metodoloxía bootstrap no contexto de dependencia, tendo en conta as condicións habituais de dependencia tratadas no Capítulo 2. Distínguense dous casos: o caso no cal existe un modelo de dependencia explícita que relaciona o valor actual da serie cos seus valores pasados, e o caso en que só existen condicións xerais de dependencia, no cal nos focalizaremos.

Primeiramente, asumiremos que se verifica a condición α-mixing de dependencia, e describiremos tres plans de remostraxe relativos a este contexto: o bootstrap por bloques, o bootstrap estacionario e o método de submostraxe. Unha revisión máis detallada pode verse en Cao (1999) e en Kreiss e Paparoditis (2011).

3.4. Bootstrap por bloques

Ante a ausencia dunha expresión explícita que modelice a dependencia dos parámetros dunha serie temporal, xorde a primeira proposta: o bootstrap por bloques (ou MBB, do inglés Moving Blocks Bootstrap), proposto por Künsch (1989) e por Liu e Singh (1992). O plan de remostraxe correspondente ó MBB procede como segue:

1. Fixar un enteiro positivo, b (que se corresponde co tamaño do bloque), e tomar k como o menor enteiro maior ou igual que $\frac{n}{b}$.
2. Definir os bloques do seguinte xeito:

   $$B_{i,b} = (X_i, X_{i+1}, \dots, X_{i+b-1}).$$

   Dito doutra forma, definir os bloques B_i, de b valores consecutivos da mostra, comezando na i-ésima observación, $\forall i = 1, 2, \dots, q$, sendo $q = n - b + 1$.
3. Xerar k observacións (é dicir, xerar k bloques), $\xi_1, \xi_2, \dots, \xi_k$, con distribución equiprobable sobre o conxunto de posibles bloques $\{B_1, B_2, \dots, B_q\}$. Nótese que cada ξ_i é un vector b-dimensional $(\xi_{i,1}, \xi_{i,2}, \dots, \xi_{i,b})$.
4. Finalmente, definir $\overrightarrow{X}^*$ como o vector formado polas n primeiras compoñentes de

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

O bootstrap ordinario sería facilmente acadable a partir do bootstrap por bloques: non habería máis que tomar $b = 1$, e por tanto, $k = n$.

3.5. Bootstrap estacionario

O bootstrap estacionario (ou SB, do inglés Stationary Bootstrap) foi proposto por Politis e Romano (1994a) como consecuencia da falta de estacionariedade do bootstrap por bloques. O plan de remostraxe deste método pode presentarse de dúas formas que resultan equivalentes. Ademais, para ambas é necesaria a elección dun parámetro $p \in [0,1]$. Presentamos a continuación a primeira delas:

1. A partir da mostra $(X_1, \dots, X_n)$, construimos a función de distribución empírica F_n. A continuación, xeramos X_1^* a partir de F_n.
2. Unha vez obtido o valor $X_i^* = X_j$, para algún $j \in \{1, 2, \dots, n-1\}$, sendo $i < n$, defínese a seguinte observación bootstrap X_{i+1}^* como segue:

$$X_{i+1}^* = X_{j+1}, \text{ con probabilidade } 1-p$$

$$X_{i+1}^* \text{ é arroxada da función } F_n \text{ con probabilidade } p$$

   No caso de que $j = n$, a observación X_{j+1} reemprazarase por X_1.

Describiremos agora a segunda forma de presentar o plan de remostraxe do bootstrap estacionario:

1. Definimos, a partir da mostra $(X_1, \dots, X_n)$, os bloques circulares como segue:

$$B_{i,b} = (X_i, X_{i+1}, \dots, X_{i+b-1}), b \in \mathbb{N}, i = 1, 2, \dots, n$$

$$\text{sendo } X_t = X_{((t-1) \bmod n)+1}, \text{ se } t > n$$

2. Xeramos realizacións iid, $L_1, L_2, \dots$, con distribución xeométrica de parámetro p, é dicir

$$\mathbb{P}(L_1 = m) = p(1-p)^{m-1}, \forall m \in \mathbb{N}$$

3. Obtemos enterios aleatorios $I_1, I_2, \dots$, con distribución equiprobable sobre o conxunto $\{1, 2, \dots, n\}$.
4. Finalmente, definimos a remostra $(X_1^*, X_2^*, \dots, X_n^*)$ como os n primeiros valores obtidos ó unir os bloques $B_{I_1,L_1}, B_{I_2,L_2}, \dots$

Destacaremos a continuación algúns aspectos relevantes do método bootstrap estacionario. Nótese, en primeiro lugar, que o número mínimo de bloques necesarios, k, no segundo método de remostraxe exposto, coincide co menor enteiro k para o cal $\sum_{i=1}^{k} L_i \geq n$, de modo que o conxunto de bloques $B_{I_1,L_1}, B_{I_2,L_2}, \dots, B_{I_k,L_k}$ teña polo menos n observacións. Ademais, se $p = 1$, obtense o bootstrap clásico.

En segundo lugar, o proceso bootstrap $\{X_i^*\}$ obtido condicionalmente á mostra observada é estacionario. Se ademais non hai datos empatados, este proceso será markoviano. En xeral, tratarase dun proceso de Markov de orde $r+1$, onde $r = \max\{b \in \mathbb{N} / \exists i, j, i \neq j \text{ con } B_{i,b} = B_{j,b}\}$.

En terceiro lugar, podemos xeralizar o segundo dos procedementos explicados para casos nos que a distribución de L_i non sexa xeométrica e a distribución dos I_i non sexa necesariamente equiprobable. Deste xeito, o bootstrap por bloques (MBB) poderá pensarse como un caso particular do bootstrap estacionario xeralizado, sendo:

$$\mathbb{P}(L_1 = m) = \begin{cases} 1 & \text{se } m = b \\ 0 & \text{se } m \neq b \end{cases}$$

$$\mathbb{P}(I_1 = j) = \begin{cases} 1/q & \text{se } j = 1, 2, \dots, q \\ 0 & \text{se } j = q+1, q+2, \dots, n \end{cases}, \text{ con } q = n - b + 1.$$

Sen embargo, haberá que ter coidado na elección das distribucións de forma que o proceso bootstrap continúe sendo estacionario.

Finalmente, tendo en conta que o tamaño medio do bloque no método SB é $\frac{1}{p}$, podemos relacionar o valor de p co tamaño do bloque do método MBB, b, xa que un xoga un papel inverso con respecto ó outro.

3.6. Método de submostraxe para datos dependentes

Como vimos, na Sección 3.3 describíase o método de submostraxe no caso de que estivésemos a traballar con datos independentes. Ademais, Politis e Romano (1994b) propoñen unha forma de proceder con esta metodoloxía no caso de que nos atopemos ante un contexto de dependencia.

Consideremos as observacións $X_1, \dots, X_n$ que proveñen dun proceso estocástico fortemente mixing (ou α-mixing). Ademais, sexa $\theta = \theta(F)$ un parámetro e $T_n = T_n(X_1, \dots, X_n)$ un estimador do mesmo. Denotemos por $J_n(\cdot, F)$ a función de distribución na mostraxe de $\tau_n(T_n - \theta)$. Posteriormente, fixado un enteiro $b < n$, defínese $S_{n,i} = T_b(B_{i,b}), i = 1, 2, \dots, N$, sendo $B_{i,b}$, $i = 1, 2, \dots, N$, todos os posibles bloques de tamaño b; e $N = n - b + 1$.

Ademais, empregaremos como aproximación da distribución na mostraxe de $\tau_n(T_n - \theta)$ a función de distribución empírica correspondente ós valores $\tau_b(S_{n,i} - T_n)$, é dicir:

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

3.7. Resultados teóricos

Detallaremos nesta sección os resultados teóricos que xustifican a validez de cada un dos métodos expostos neste Capítulo 3. Comezaremos polo bootstrap por bloques, para continuar co bootstrap estacionario, e co método de submostraxe nun contexto de independencia e de dependencia. Así mesmo, remataremos cun resultado de bo comportamento asintótico do bootstrap estacionario aplicado ó estimador da función de densidade de Parzen-Rosenblatt, proposto por Hwang e Shin (2012).

3.7.1. Bootstrap por bloques

A continuación, presentaranse dous resultados que proban a validez asintótica da metodoloxía bootstrap por bloques baixo condicións pouco restrictivas sobre o grao de dependencia e o tamaño do bloque. Künsch (1989) establece a proba para xustificar a validez asintótica do MBB no caso da media aritmética. Sen embargo, serán Liu e Singh (1992) quen o fagan para un estatístico xeral. Comezaremos introducindo varias definicións previas á exposición de dita demostración.

Definición 1 *Un funcional $T(\cdot)$ é Frechét diferenciable en F se, para algunha función g_F se verifica*

$$T(G) - T(F) = \int g_F d(G - F) + o(||G - F||),$$

sendo $||\cdot||$ a norma do supremo.

Definición 2 *Un funcional $T(\cdot)$ é dúas veces Frechét diferenciable en F se, para algunhas funcións g_F e $h_F(\cdot,\cdot)$,*

$$T(G)-T(F)=\int g_F(x)d(G(x)-F(x))-\int\int h_F(x,y)d[G(x)-F(x)]d[G(y)-F(y)]+o(||G-F||^2),$$

sendo $||\cdot||$ a norma do supremo.

Neste momento, podemos expoñer a proposta de Liu e Singh (1992), considerando $T_n = T(F_n)$ un estatístico. Asumiremos, ademais, que T é un funcional dúas veces Frechét diferenciable e que o tamaño do bloque $b \to \infty$. Tendo en conta estas hipóteses podemos enunciar o Teorema 1, que pon de manifesto a validez asintótica do MBB.

Teorema 1 *Consideremos $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario e α-mixing. Baixo as condicións anteriores, se $b \to \infty$ e $b \cdot log\ n/\sqrt{n} \to 0$, entón*

$$||\mathbb{P}^*(\sqrt{n}(T(F_n^*)-T(F_n))\le x)-\mathbb{P}(\sqrt{n}(T(F_n)-T(F))\le x)||\to 0,$$

en probabilidade, sendo $||\cdot||$ a norma do supremo.

Posteriormente, Radulovic (1996) será quen de realizar esta proba baixo condicións menos restrictivas. Este autor demostra que sempre que unha sucesión de variables aleatorias fortemente mixing satisfaga o Teorema Central do Límite, dito resultado tamén é válido para a versión bootstrap por bloques. Este enunciado vén reflectido nos Teoremas 2 e 3.

Teorema 2 *Sexa $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario que satisfai a condición de ser α-mixing, tal que verifica que $\sigma_n^{-1}S_n \to N(0,1)$ en distribución, sendo $S_n = X_1 + \cdots + X_n$ e $\sigma_n^2 = Var(S_n)$. Sexa $\sigma_n^{*2} = Var^*(\sum_{i=1}^n X_i^*)$, onde X_i^* son as remostras xeradas polo método MBB cun tamaño de bloque b. Supoñamos ademais que $b/n \to 0$ a medida que $n \to \infty$. Logo, verifícanse as seguintes converxencias:*

1. $H_n^* = \dfrac{1}{\sigma_n^*}\sum_{i=1}^n (X_i^* - \mathbb{E}^*(X_i^*)) \to N(0,1)$ *en probabilidade.*

2. $\tilde{H}_n^* = \dfrac{1}{\sqrt{k}\sigma_b}\sum_{i=1}^n (X_i^* - \mathbb{E}^*(X_i^*)) \to N(0,1)$ *en probabilidade.*

A partir da converxencia dada no epígrafe 1 no Teorema 2, temos que $\|F_{H_n^*} - \Phi\|_\infty \to 0$ en probabilidade, sendo $F_{H_n^*}$ a función de distribución de H_n^*, e Φ a función de distribución normal estándar. Sen embargo, aínda que non é posible enunciar o Teorema 2 nun caso xeral, engadindo unha hipótese adicional relativa á integrabilidade uniforme, podemos obter un resultado similar, tal como se observa no Teorema 3.

Teorema 3 *Sexa $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario que satisfai a condición de ser α-mixing, tal que verifica, sendo $S_n = X_1 + ... + X_n$:*

1. $\dfrac{1}{\sqrt{n}}S_n \to N(0,\sigma^2)$ *en distribución.*

2. $\left\{\left(\dfrac{S_n}{\sqrt{n}}\right)^2\right\}_{n=1}^{\infty}$ *é uniformemente integrable.*

Sexa X_i^* unha remostra xerada polo método MBB, considerando un tamaño de bloque b tal que $b/n \to 0$ cando $n \to \infty$. Entón

$$\hat{H}_n^* = \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i^* - \mathbb{E}(X_i^*)) \to N(0,\sigma^2),$$

en probabilidade.
Se ademais $\{X_t\}_{t\in\mathbb{Z}}$ *verifica:*

$$\frac{\sigma_n^2}{n} \overset{n\to\infty}{\longrightarrow} \sigma^2,\ \sigma^2 > 0, \tag{3.1}$$

sendo $\sigma_n^2 = Var(S_n)$*, entón* $\frac{1}{\sqrt{n}}S_n \to N(0,\sigma^2)$ *equivale a dicir que* $\frac{S_n}{\sigma_n} \to N(0,1)$ *(o que implica a integrabilidade uniforme de 2).*

Logo, sempre e cando (3.1) se verifique, o Teorema 2 e o Teorema 3 serán equivalentes.

3.7.2. Bootstrap estacionario

Nesta sección, exporemos o resultado da validez asintótica do bootstrap estacionario, dado por Politis e Romano (1994a). Consideremos $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario con función de covarianza $R(\cdot)$, verificando:

$$R(0) + \sum_r |rR(r)| < \infty \tag{3.2}$$

e tamén,

$$\sum_{u,v,w} |\kappa_4(u,v,w)| = K < \infty, \tag{3.3}$$

sendo $\kappa_4(u,v,w)$ o cumulante de orde 4 da distribución de $(X_j, X_{j+v}, X_{j+u}, X_{j+u+v+w})$. Baixo estas condicións, enunciemos o Teorema 4.

Teorema 4 *Sexa* $\{X_t\}_{t\in\mathbb{Z}}$ *un proceso estacionario que verifica as condicións (3.2) e (3.3). Asumamos, ademais, que para algún* $d > 0$, $\mathbb{E}|X_i|^{d+2} < \infty$, *e* $\sum_k [\alpha_X(k)]^{d/(2+d)}$*. Entón,* $\sigma_\infty^2 = Var(X_1) + 2\sum_{i=1}^{\infty} cov(X_1, X_{i+1})$ *é finita. Se* $\sigma_\infty > 0$*, tense que*

$$sup_x|\mathbb{P}(\sqrt{n}(\bar{X_n} - \mu) \le x) - \Phi(x/\sigma_\infty)| \to 0,$$

sendo $\Phi(\cdot)$ *a función de distribución normal estándar. Consideremos o método de remostraxe bootstrap estacionario e asumamos agora que* $p \to 0$ *e que* $np \to \infty$*. Logo,*

$$sup_x|\mathbb{P}^*(\sqrt{n}(\bar{X_n}^* - \bar{X_n}) \le x) - \mathbb{P}(\sqrt{n}(\bar{X_n} - \mu) \le x)| \to 0,$$

en probabilidade.

En efecto, estes autores proban que sempre que $p \to 0$ e $np \to \infty$, a distribución de $\sqrt{n}(\bar{X_n} - \mu)$ pódese aproximar pola distribución na remostraxe de $\sqrt{n}(\bar{X_n}^* - \bar{X_n})$. Politis e Romano (1994a) dan tamén unha idea sobre cómo xeralizar este resultado a estatísticos funcionais $T(F_n)$, onde T é un funcional Frechét diferenciable.

3.7.3. Método de submostraxe

Nas Seccións 3.3 e 3.6 descríbese a proposta de Politis e Romano (1994b): o método de submostraxe, cuxa idea clave é proporcionar un método bootstrap que sexa válido baixo condicións minimais, tanto no caso en que os datos sexan independentes como dependentes. Ademais, Politis e Romano (1994b) proporcionan a proba de dita validez en ambos contextos.

Consideremos, primeiramente, o caso en que $(X_1, \dots, X_n)$ sexa unha mostra de n observacións independentes e identicamente distribuídas. A partir da notación empregada nas Seccións 3.3 e 3.6, asumamos:

$$\tau_n(T_n - \theta) \text{ converxe en distribución a } J(\cdot, F) \text{ a medida que } n \to \infty. \tag{3.4}$$

A continuación, enunciaremos o Teorema 5, que xustifica a validez asintótica deste método no contexto de independencia.

Teorema 5 *Asumindo (3.4), consideremos que* $\tau_b/\tau_n \to 0$, $b \to \infty$ *e* $b/n \to 0$ *a medida que* $n \to \infty$. *Sexa* x *un punto de continuidade de* $J(\cdot, F)$. *Entón verifícase:*

1. $L_n(x) \to J(x, F)$ *en probabilidade.*
2. *Se* $J(\cdot, F)$ *é continua, entón*
$$sup_x|L_n(x) - J_n(x, F)| \to 0,$$
en probabilidade.

En segundo lugar, consideremos o contexto de dependencia. Neste caso, partiremos de $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario fortemente mixing. Enunciaremos agora o Teorema 6, onde se garante a validez asintótica do método de submostraxe no caso de datos dependentes.

Teorema 6 *Sexa* $\{X_t\}_{t\in\mathbb{Z}}$ *un proceso estacionario que satisfai a condición de ser* α*-mixing. Asumamos (3.4), que* $\tau_b/\tau_n \to 0$, $b \to \infty$ *e* $b/n \to 0$ *cando* $n \to \infty$. *Consideremos* x *un punto de continuidade de* $J(\cdot, F)$. *Entón as conclusións do Teorema 5 seguen sendo certas.*

Logo, os Teoremas 5 e 6 garanten que, baixo condicións minimais sobre o tamaño do bloque, o método de submostraxe é asintóticamente válido, sempre e cando o estatístico de interese teña distribución límite.

3.7.4. Resultados de bo comportamento asintótico de Hwang e Shin

Asumamos, finalmente, que se verifica a condición de Ψ dependencia descrita no Capítulo 2. Ademais, consideremos o estimador non paramétrico da función de densidade de Parzen-Rosenblatt (1.1), e a metodoloxía relativa ó bootstrap estacionario descrita na Sección 3.5.

Hwang e Shin (2012) proban a validez asintótica do bootstrap estacionario aplicado ó estimador da función de densidade de Parzen-Rosenblatt, $\hat{f}_h(x)$, baixo unha serie de condicións (entre elas, a de Ψ dependencia), e impoñendo unha serie de hipóteses sobre a función de densidade, f. Noutras palabras, proban que, baixo certas condicións, o estimador bootstrap $\hat{f}_h^*(x)$ obtido condicionalmente á mostra considerada $(X_1, \dots, X_n)$ ten a mesma distribución límite que $\hat{f}_h(x)$.

As hipóteses que a función de densidade f debe verificar son as seguintes:

1. Sexa ρ a regularidade da función de densidade f. Definamos $\rho = a + b, \forall a \in \mathbb{N}^+, 0 \leq b < 1$, entón existe unha constante $A > 0$ tal que f é a veces continuamente diferenciable verificando
$$|f^{(a)}(x) - f^{(a)}(y)| \leq A|x - y|^b,$$
onde x, y pertencen a un intervalo compacto arbitrario de $\mathbb{R}$.
2. Asumamos que ρ é un número enteiro, e que a función tipo núcleo K satisfai as seguintes dúas condicións:
$$\begin{gathered} \int x^i K(x)dx = 0, i = 1, \dots, \rho - 1 \\ \int x^\rho K(x)dx \neq 0 \end{gathered} \tag{3.5}$$
3. Sexa $f_{0,n}$ a densidade conxunta de X_0 e X_n, e asumamos que existe algunha constante positiva C tal que para todo enteiro positivo n, se verifique

$$\begin{aligned} ||f||_\infty &\leq C \\ ||f_{0,n}||_\infty &\leq C \end{aligned} \tag{3.6}$$

Politis e Romano (1994a) proban que, condicionalmente ó proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$, o proceso $\{X_t^*\}_{t\in\mathbb{Z}}$ tamén será estacionario. O estimador bootstrap de Parzen-Rosenblatt da función de densidade, obtido a partir dunha remostra bootstrap $(X_1^*, \ldots, X_n^*)$, virá entón dado por:

$$\hat{f}_h^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right). \tag{3.7}$$

Como xa dixemos, Hwang e Shin (2012) proban que, baixo certas condicións, o estimador presentado en (3.7) ten a mesma distribución límite que o estimador de Parzen-Rosenblatt, $\hat{f}_h(x)$. Deste xeito, a distribución de $\sqrt{nh}(\hat{f}_h(x) - f(x))$ pode ser aproximada pola distribución na remostraxe de $\sqrt{nh}(\hat{f}_h^*(x) - \hat{f}_g(x))$, sendo $\hat{f}_g(x)$ un estimador tipo núcleo, e g un parámetro de suavizado que ten maior orde que h. Vexamos baixo que condicións se dá esta propiedade de bo comportamento asintótico.

Teorema 7 *Sexa $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario que satisfai a condición (3.6). Supoñamos que $\{X_t\}_{t\in\mathbb{Z}}$ é $(\theta, \mathcal{L}, \Psi_1)$-dependente, con $\theta_r = \mathcal{O}(r^{-12-\nu})$, ou $(\theta, \mathcal{L}, \Psi_2)$-dependente con $\theta_r = \mathcal{O}(r^{-9-\nu})$, para algún $\nu > 0$. Entón, se $n \to \infty$, temos*

1. *Se $nh^{1+\epsilon} \to \infty$, para algún $\epsilon > 0$ e $\rho = \mathcal{O}(1/\sqrt{n})$, entón para todo $x, z \in \mathbb{R}$ verifícase:*

$$\mathbb{P}^*\left(\sqrt{nh}\left[\hat{f}_h^*(x) - \mathbb{E}^*(\hat{f}_h^*(x))\right] \leq z\right) - \mathbb{P}\left(\sqrt{nh}\left[\hat{f}_h(x) - \mathbb{E}(\hat{f}_h(x))\right] \leq z\right) \longrightarrow 0, \text{ en probabilidade.}$$

2. *Se a regularidade ρ da función de densidade f é un enteiro positivo, e a función núcleo, K, satisfai a condición (3.5), $h = \mathcal{O}(n^{-1/(2\rho+1)})$, $gn^{1/(2\rho+1)} \to \infty$ e $\rho = \mathcal{O}(1/\sqrt{n})$, entón para todo $x, z \in \mathbb{R}$ verifícase:*

$$\mathbb{P}^*\left(\sqrt{nh}\left[\hat{f}_h^*(x) - \hat{f}_g(x)\right] \leq z\right) - \mathbb{P}\left(\sqrt{nh}\left[\hat{f}_h(x) - f(x)\right] \leq z\right) \longrightarrow 0, \text{ en probabilidade.}$$

A partir do Teorema 7, vemos que, en efecto, tendo en conta unha serie de hipóteses sobre a orde dos parámetros ventá g e h considerados, e sobre o parámetro ρ; o sesgo bootstrap do estimador (3.7) resulta despreciable, o que implica un bo comportamento asintótico do mesmo, aínda baixo un método de remostraxe non suavizado.

Capítulo 4

Selección do parámetro de suavizado para datos dependentes

Consideremos o estimador non paramétrico da función de densidade de Parzen-Rosenblatt e as súas propiedades asintóticas definidas no Capítulo 1. Como vimos, para valores de h pequenos, teríamos estimadores centrados, pero a varianza incrementaríase. Sen embargo, para valores de h grandes, diminuiría a varianza, pero a costa de aumentar o sesgo. Polo tanto, un dos problemas centrais da estatística non paramétrica será atopar o valor do parámetro de suavizado, h, que proporciona o balance óptimo entre o sesgo e a varianza.

Como vimos no Capítulo 1, para obter o valor óptimo do parámetro de suavizado, h, así como para avaliar o comportamento dun estimador da función de densidade, defínense varias medidas de erro. Neste capítulo focalizarémonos principalmente no erro cuadrático medio integrado ($MISE$). Ademais, proporcionaremos, a partir destas medidas de erro, distintos selectores de ventá construidos a partir dos datos, primeiro nun contexto de independencia e logo nun contexto de dependencia. Concretamente, describiremos os métodos de validación cruzada, plug-in, bootstrap e bootstrap estacionario suavizado. Finalmente, obteremos unha expresión explícita para a versión bootstrap do $MISE$.

4.1. Plug-in

Comezaremos dando unha descrición da obtención da ventá plug-in nun contexto de independencia, para logo traballar baixo a hipótese de dependencia. Así, considerando a ventá $AMISE$ (1.5), Seather e Jones (1991) propoñen empregar o estimador tipo núcleo de Parzen-Rosenblatt para estimar, de xeito non paramétrico, $R(f'') = \int f''(x)^2 dx$. Se denotamos $\Psi_r = \mathbb{E}(f^{(r)}(X))$, e empregando o método de integración por partes, temos que:

$$\int f''(x)^2 dx = \int f^{(4)}(x) f(x) dx = \Psi_4.$$

Polo tanto, o problema radica en estimar Ψ_4. Xeralmente, unha aproximación de $\Psi_r = \mathbb{E}(f^{(r)}(X)) = \int f^{(r)}(x) f(x) dx$ vén dada por:

$$\hat{\Psi}_r = \frac{1}{n} \sum_{i=1}^{n} \hat{f}^{(r)}(X_i),$$

sendo $\hat{f}^{(r)}$ un estimador tipo núcleo da densidade da r-ésima derivada. Chegados a este punto, precisaríamos dunha ventá g para obter esta última estimación, de xeito que, considerando un núcleo

L:

$$\hat{f}^{(4)}(X_i) = \frac{1}{n}\sum_{j=1}^{n} L_g^{(4)}(X_i - X_j),$$

onde poderíamos obter a ventá g empregando argumentos similares, precisando de $\hat{\Psi}_6$, e así sucesivamente. Seather e Jones (1991) suxiren asumir, nun destes pasos, que f é normal, con desviación típica σ, resultando:

$$\Psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\sqrt{\pi}}.$$

Procederemos neste caso estimando Ψ_r, empregando $\hat{\sigma}$ e iterando cara atrás.

Unha vez descrita a metodoloxía plug-in baixo a hipótese de independencia, focalicémonos na descrición deste método nun contexto de dependencia, proposto por Hall *et al.* (1995). Consideremos $R(f(x)) = \int f(x)^2 dx$, $R(f''(x)) = \int f''(x)^2 dx$, $R(f'''(x)) = \int f'''(x)^2 dx$ e $\mu_k = \int z^k K(z) dz$. A continuación, presentamos a expresión para o $AMISE$ con datos dependentes que desenvolveron Hall *et al.* (1995), supoñendo que f é, como mínimo, seis veces diferenciable:

$$AMISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2^2 R(f'') - h^6\frac{1}{24}\mu_2\mu_4 R(f''') - \frac{1}{n}\left(2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\int g_i(x,x)dx - R(f)\right), \tag{4.1}$$

sendo $g_i(x_1, x_2) = f_i(x_1, x_2) - f(x_1)f(x_2)$, e f_i a densidade de (X_j, X_{i+j}).

Hall *et al.* (1995) propoñen que o parámetro de suavizado óptimo, h, será aquel que minimice a expresión do $AMISE$ (4.1). Así, temos que o selector de ventá empírico polo método plug-in, $\hat{h}$, sería

$$\hat{h} = \left(\frac{\hat{J}_1}{n}\right)^{1/5} + \hat{J}_2\left(\frac{\hat{J}_1}{n}\right)^{3/5},$$

onde $\hat{J}_1$ é o estimador de $J_1 = \dfrac{R(K)}{\mu_2^2 R(f'')}$, e $\hat{J}_2$ é o estimador de $J_2 = \dfrac{\mu_4 R(f''')}{20\mu_2 R(f'')}$.

Seguidamente, propoñen reemprazar directamente as cantidades $R(f'')$ e $R(f''')$ polos seus respectivos estimadores, $\hat{I}_2$ e $\hat{I}_3$, que se obteñen como segue:

$$\hat{I}_k = 2\hat{\theta}_{1k} - \hat{\theta}_{2k}, k = 2, 3$$

sendo $\hat{\theta}_{1k}$ e $\hat{\theta}_{2k}$ os respectivos estimadores de $\theta_{1k} = \int \left(\mathbb{E}(\hat{f}_1)\right) f^{(k)} dx, \theta_{2k} = \int \left(\mathbb{E}(\hat{f}_1^{(k)})\right)^2, k = 1, 2, 3.$

Ademais, $\hat{f}_1$ é o estimador non paramétrico da densidade de Parzen-Rosenblatt obtido cun núcleo K_1 e un parámetro ventá h_1. Téñense as seguintes expresións para $\hat{\theta}_{1k}$ e $\hat{\theta}_{2k}$:

$$\hat{\theta}_{1k} = 2\left(n(n-1)h_1^{2k+1}\right)^{-1}\sum_{1\le i<j\le n}\sum K_1^{(2k)}\left(\frac{X_i - X_j}{h_1}\right),$$

$$\hat{\theta}_{2k} = 2\left(n(n-1)h_1^{2(k+1)}\right)^{-1}\sum_{1\le i<j\le n}\sum \int K_1^{(k)}\left(\frac{x - X_i}{h_1}\right) K_1^{(k)}\left(\frac{x - X_j}{h_1}\right) dx.$$

Do mesmo xeito, Hall *et al.* (1995) proban que, baixo certas condicións impostas sobre o proceso estacionario, pedindo a diferenciabilidade do núcleo K_1 considerado, e escollendo a ventá h_1 tal que $n^{-1/(4k+1)} \le h_1 \le 1$; o método plug-in para a obtención do parámetro de suavizado h nun contexto de dependencia é consistente, xa que aínda que os estimadores $\hat{\theta}_{1k}$ e $\hat{\theta}_{2k}$ presentan un pequeno sesgo, este é despreciable.

4.2. Validación cruzada

Novamente, comezaremos describindo a metodoloxía de validación cruzada baixo a hipótese de independencia, plantexada por Bowman (1984). Posteriormente, centrarémonos no caso de traballar con datos dependentes, seguindo a proposta de Cox e Kim (1997).

Consideremos o caso *iid*. Este método afronta o problema de selección de ventá dende unha perspectiva diferente. En lugar de basearse nas expresións de ventá óptimas no senso do $AMISE$, parte directamente de aproximar as medidas de erro. Primeiramente, o erro cuadrático integrado (ISE) de $\hat{f}_h$ como estimador de f é:

$$ISE(X_1, \dots, X_n; h) = \int \left(\hat{f}(x) - f(x)\right)^2 dx,$$

que se pode reescribir como segue:

$$\begin{aligned} ISE(h) &= \int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx + \int f^2(x)dx \\ &= R(\hat{f}_h) - 2\int \hat{f}_h(x)f(x)dx + R(f), \end{aligned}$$

onde o último sumando non depende de h. Así, a ventá óptima para o ISE resulta:

$$\begin{aligned} h_{ISE} &= \text{arg mín}_h \left(\int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx\right) \\ &= \text{arg mín}_h \left(R(\hat{f}_h) - 2\int \hat{f}_h(x)f(x)dx\right), \end{aligned}$$

onde h aparece dentro do estimador tipo núcleo $\hat{f}_h$. Chegados a este punto, o método de validación cruzada trata de aproximar os dous sumandos que compoñen a expresión a minimizar na ventá. No caso do primeiro sumando, reescribirase do seguinte xeito:

$$R(\hat{f}_h) = \int \hat{f}_h^2(x)dx = \frac{1}{n^2 h}\sum_{i,j} K * K\left(\frac{X_i - X_j}{h}\right),$$

dependendo unicamente da ventá h e da función tipo núcleo escollida. Notemos, para o segundo sumando, que:

$$\int \hat{f}_h(x)f(x)dx = \mathbb{E}\left(\hat{f}_h(Y)\Big|_{X_1,\dots,X_n}\right),$$

onde Y é unha variable aleatoria con densidade f e independente de $X_1, \dots, X_n$. Polo tanto, condicionando á mostra $X_1, \dots, X_n$, poderíamos aproximar esta cantidade por:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_h^{-i}(X_i),$$

sendo $\hat{f}_h^{-i}(\cdot)$ o estimador tipo núcleo de Parzen-Rosenblatt da densidade obtido con toda a mostra excepto o dato i-ésimo.

Logo, o método de validación cruzada baséase en obter o parámetro de suavizado, h, tal que minimiza a función de validación cruzada:

$$CV(h) = R(\hat{f}_h) - \frac{2}{n}\sum_{i=1}^{n} f_h^{-i}(X_i),$$

e así definir $h_{CV} = \arg\,\text{mín}_h CV(h)$.

Nun contexto de dependencia, Cox e Kim (1997) propoñen unha modificación para o método de validación cruzada, debido a que este último produce estimadores sesgados. Xorde así a versión «leave-$(2l+1)$-out» do método de validación cruzada (método de validación cruzada modificado). Definimos

$$CV_l(h) = \int \hat{f}^2(x)dx - \frac{2}{n}\sum_{j=1}^{n} \hat{f}_l^j(X_j),$$

sendo

$$\hat{f}_l^j(x) = \frac{1}{n_l}\sum_{i:|j-i|>l}^{n} \frac{1}{h}K\left(\frac{x-X_i}{h}\right),$$

e l é unha sucesión de enteiros positivos chamados a secuencia «leave-out». Pola súa banda, n_l é escollido tal que:

$$nn_l = \#\{(i,j) : |i-j| > l\}.$$

Deste xeito, o método de validación cruzada modificado (MCV) seleccionará o mínimo de $CV_l(h)$, así:

$$h_{MCV} = \arg\,\text{mín}_h CV_l(h).$$

Finalmente, Cox e Kim (1997) proban a consistencia do método de validación cruzada modificado baixo certas condicións sobre o proceso estacionario considerado.

4.3. Bootstrap estacionario suavizado

Nesta sección estudarase unha versión suavizada do método bootstrap estacionario proposto por Politis e Romano (1994a). O propósito neste caso é presentar un selector de ventá para a estimación da densidade nun contexto de dependencia. Para iso, obterase unha expresión exacta para a versión bootstrap do $MISE$. Este é un método novo presentado neste traballo fin de mestrado.

Consideramos en primeiro lugar o caso *iid*. A continuación, presentaremos a metodoloxía bootstrap existente na literatura para seleccionar o parámetro de suavizado, h, con datos independentes. A idea básica consiste en deseñar un plan de remostraxe, do tipo bootstrap suavizado, para estimar o erro cuadrático medio integrado ($MISE$). Seguiremos a proposta de Cao (1993), que procede deste xeito:

1. A partir da mostra $(X_1, \ldots, X_n)$ *iid*, e empregando unha ventá piloto g, calcúlase o estimador da función de densidade de Parzen-Rosenblatt, $\hat{f}_g$.

2. Xéranse remostras bootstrap $(X_1^*, \ldots, X_n^*)$ a partir da densidade $\hat{f}_g$.

3. Para cada $h > 0$, obtense o análogo bootstrap do estimador de Parzen-Rosenblatt

$$\hat{f}_h^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i^*}{h}\right).$$

4. Constrúese a versión bootstrap do $MISE$, resultando:

$$MISE^*(h) = \int \mathbb{E}^*\left[\left(\hat{f}_h^*(x) - \hat{f}_g^*(x)\right)^2\right]dx.$$

5. Minimízase o $MISE^*(h)$ en $h > 0$, e obtense o selector bootstrap:

$$h_{MISE}^* = \arg\,\text{mín}_{h>0} MISE^*(h).$$

Neste contexto pode atoparse unha expresión pechada para o $MISE^*(h)$, sen necesidade de facer Monte Carlo:

$$MISE^*(h) = \int \left(K_h * \hat{f}_g(x) - \hat{f}_g(x)\right)^2 dx + \frac{R(K)}{nh} - \frac{1}{n}\int K_h * \hat{f}_g(x)^2 dx.$$

Esta expresión pode escribirse de xeito máis explícito da seguinte forma:

$$\begin{aligned} MISE^*(h) &= \frac{1}{n^2}\sum_{i,j=1}^{n}\left[(K_h * K_g - K_g) * (K_h * K_g - K_g)\right](X_i - X_j) \\ &\quad + \frac{R(K)}{nh} - \frac{1}{n^3}\sum_{i,j=1}^{n}\left[(K_h * K_g) * (K_g * K_g)\right](X_i - X_j). \end{aligned}$$

Consideremos agora o contexto de dependencia, e sexa un proceso estacionario α-mixing, do cal observamos unha mostra aleatoria $X_1, X_2, \ldots, X_n$. Centrarémonos no problema da estimación da función de densidade marxinal f das X_i, asumindo a súa existencia. Novamente, consideraremos o estimador non paramétrico tipo núcleo da densidade de Parzen-Rosenblatt dado en (1.1). A continuación, desenvolverase un mecanismo bootstrap deseñado para seleccionar o parámetro de suavizado, h. Este plan de remostraxe é unha versión suavizada do bootstrap estacionario, proposto por Politis e Romano (1994a).

Sexa $X_1, X_2, \ldots, X_n$ a mostra observada, e g unha ventá piloto. Entón, o bootstrap estacionario suavizado que propoñemos procede como segue:

1. Arroxamos $X_1^{*(d)}$ a partir da función de distribución empírica da mostra, F_n. Esto será equivalente a fixar $\mathbb{P}^*\left(X_1^{*(d)} = X_i\right) = \frac{1}{n}$, $\forall i = 1, 2, \ldots, n$.

2. Definimos $X_1^* = X_1^{*(d)} + gU_1^*$, onde U_1^* é arroxada con densidade K e independentemente de $X_1^{*(d)}$.

3. Asumamos que xa temos xeradas as observacións $X_1^*, \ldots, X_i^*$ (e, consecuentemente, $X_1^{*(d)}, \ldots, X_i^{*(d)}$). Consideremos o índice j, para o cal $X_i^{*(d)} = X_j$. Definimos entón unha variable aleatoria binaria auxiliar I_{i+1}^*, tal que $\mathbb{P}^*\left(I_{i+1}^* = 1\right) = 1 - p$ e $\mathbb{P}^*\left(I_{i+1}^* = 0\right) = p$. Asignemos $X_{i+1}^{*(d)} = X_{(j \bmod n)+1}$ sempre e cando $I_{i+1}^* = 1$ e fixaremos a distribución empírica para $X_{i+1}^{*(d)}|_{I_{i+1}^*=0}$.

4. Unha vez que foi arroxada $X_{i+1}^{*(d)}$, definimos $X_{i+1}^* = X_{i+1}^{*(d)} + gU_{i+1}^*$, onde, de novo, U_{i+1}^* é xerada a partir da densidade K e independentemente de $X_{i+1}^{*(d)}$.

Este plan de remostraxe é esencialmente unha versión suavizada do bootstrap estacionario descrito na Sección 3.5, que depende dun parámetro $p \in (0, 1)$ e da ventá piloto g.

Observación 1 *O proceso estocástico bootstrap $X_1^{*(d)}, X_2^{*(d)}, \ldots, X_n^{*(d)}$ é estacionario se non hai empates na mostra, é dicir, normalmente é estacionario. Esto implica que todas estas variables aleatorias bootstrap teñen como función de distribución F_n.*

Observación 2 *O proceso estocástico bootstrap $X_1^*, X_2^*, \ldots, X_n^*$ é estacionario e a densidade bootstrap de cada X_i^* é $\hat{f}_g$, o estimador tipo núcleo de Parzen-Rosenblatt obtido a partir da mostra $X_1, X_2, \ldots, X_n$, e empregando unha ventá piloto g. Esta é unha consecuencia inmediata do feito de que $X_i^{*(d)}$ son*

variables aleatorias discretas bootstrap, con función de distribución común F_n:

$$\begin{aligned}
\mathbb{P}^*\left(X_i^* \leq x\right) &= \mathbb{P}^*\left(X_{i+1}^{*(d)} + gU_{i+1}^* \leq x\right) \\
&= \sum_{j=1}^{n} \mathbb{P}^*\left(X_{i+1}^{*(d)} + gU_{i+1}^* \leq x \Big|_{X_{i+1}^{*(d)}=X_j}\right) \mathbb{P}^*\left(X_{i+1}^{*(d)} = X_j\right) \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}^*\left(U_{i+1}^* \leq \frac{x - X_j}{g}\Big|_{X_{i+1}^{*(d)}=X_j}\right) \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{K}\left(\frac{x - X_j}{g}\right),
\end{aligned}$$

onde $\mathbb{K}$ é a función de distribución asociada a K. Derivando a anterior expresión con respecto a x, obtemos a densidade bootstrap de X_i^*:

$$\frac{1}{n}\sum_{j=1}^{n} \frac{1}{g}\mathbb{K}'\left(\frac{x - X_j}{g}\right) = \frac{1}{ng}\sum_{j=1}^{n} K\left(\frac{x - X_j}{g}\right) = \hat{f}_g\left(x\right).$$

Observación 3 *Como ocorre no caso non suavizado, o parámetro p dá liberdade ó método bootstrap. O plan de remostraxe resultante abrangue dende un mecanismo bootstrap suavizado clásico ($p = 1$), descrito por Cao (1993), ata unha permutación circular suavizada da mostra ($p = 0$).*

4.3.1. Expresión exacta para o $MISE\left(h\right)$

Como xa vimos, unha medida global do erro cometido polo estimador tipo núcleo da densidade é o seu erro cuadrático medio integrado, dado por

$$MISE\left(h\right) = \mathbb{E}\left[\int \left(\hat{f}_h\left(x\right) - f\left(x\right)\right)^2 dx\right] = B\left(h\right) + V\left(h\right),$$

onde

$$\begin{aligned}
B\left(h\right) &= \int \left[\mathbb{E}\left(\hat{f}_h\left(x\right)\right) - f\left(x\right)\right]^2 dx, \text{ e} \\
V\left(h\right) &= \int Var\left(\hat{f}_h\left(x\right)\right) dx
\end{aligned}$$

son o sesgo ao cadrado integrado e a varianza integrada do estimador, respectivamente.

No caso *iid*, é sinxelo obter unha expresión exacta para o $MISE\left(h\right)$, empregando que

$$\begin{aligned}
B\left(h\right) &= \int \left[\mathbb{E}\left(K_h\left(x-X_1\right)\right)-f\left(x\right)\right]^2 dx = \int \left[\int K_h\left(x-y\right) f\left(y\right) dy - f\left(x\right)\right]^2 dx \\
&= \int \left(K_h * f\left(x\right) - f\left(x\right)\right)^2 dx, \text{ e} \\
V\left(h\right) &= n^{-1}\int Var\left(K_h\left(x-X_1\right)\right) dx \\
&= n^{-1}\int \left\{\mathbb{E}\left(K_h\left(x-X_1\right)^2\right) - \left[\mathbb{E}\left(K_h\left(x-X_1\right)\right)\right]^2\right\} dx \\
&= n^{-1}h^{-2}\int\int K\left(\frac{x-y}{h}\right)^2 f\left(y\right) dy\, dx - n^{-1}\int \left[\int K_h\left(x-y\right) f\left(y\right) dy\right]^2 dx \\
&= n^{-1}h^{-1}\int\int K\left(u\right)^2 f\left(x-hu\right) du\, dx - n^{-1}\int \left(K_h * f\left(x\right)\right)^2 dx \\
&= n^{-1}h^{-1}R\left(K\right) - n^{-1}\int \left(K_h * f\left(x\right)\right)^2 dx,
\end{aligned}$$

sendo $*$ o operador convolución.

No contexto de dependencia, os mesmos cálculos son válidos para $B\left(h\right)$, mentras que $V\left(h\right)$ pode ser obtida tendo en conta todas as posibles covarianzas. En primeiro lugar, é útil considerar a estacionariedade do proceso, ademais dos termos que compoñen a varianza integrada no caso *iid* (denotémolo $V_0\left(h\right)$):

$$\begin{aligned}
V\left(h\right) &= n^{-1}\int Var\left(K_h\left(x-X_1\right)\right) dx + n^{-2}\sum_{i\neq j}\int Cov\left(K_h\left(x-X_i\right), K_h\left(x-X_j\right)\right) dx \\
&= V_0\left(h\right) + 2n^{-2}\sum_{i<j}\int Cov\left(K_h\left(x-X_i\right), K_h\left(x-X_j\right)\right) dx \\
&= V_0\left(h\right) + 2n^{-2}\sum_{i<j}\int Cov\left(K_h\left(x-X_1\right), K_h\left(x-X_{j-i+1}\right)\right) dx \\
&= V_0\left(h\right) + 2n^{-2}\sum_{\ell=1}^{n-1}\left(n-\ell\right)\int Cov\left(K_h\left(x-X_1\right), K_h\left(x-X_{\ell+1}\right)\right) dx.
\end{aligned}$$

Agora,

$$\begin{aligned}
&\int Cov\left(K_h\left(x-X_1\right), K_h\left(x-X_{\ell+1}\right)\right) dx \\
&= \int \mathbb{E}\left(K_h\left(x-X_1\right) K_h\left(x-X_{\ell+1}\right)\right) dx - \int \left[\mathbb{E}\left(K_h\left(x-X_1\right)\right)\right]^2 dx \\
&= \int\int\int K_h\left(x-y\right) K_h\left(x-z\right) f\left(y\right) f_\ell\left(z|y\right) dx\, dy\, dz - \int \left(K_h * f\left(x\right)\right)^2 dx \\
&= \int\int K_h\left(x-y\right) f\left(y\right) K_h * f_\ell\left(\bullet|y\right)\left(x\right)\, dx\, dy - \int \left(K_h * f\left(x\right)\right)^2 dx,
\end{aligned}$$

sendo $f_\ell\left(z|y\right)$ a densidade condicional de $X_{\ell+1}$ sabendo que $X_1 = y$, avaliada no punto z.

Consecuentemente, finalizamos cunha expresión exacta para o erro cuadrático medio integrado do estimador de Parzen-Rosenblatt para un proceso estacionario:

$$\begin{aligned} MISE\left(h\right) &= B\left(h\right)+V\left(h\right), \text{ onde} \\ B\left(h\right) &= \int \left(K_h * f\left(x\right)-f\left(x\right)\right)^2 dx, \text{ e} \\ V\left(h\right) &= n^{-1}h^{-1}R\left(K\right)-\int \left(K_h * f\left(x\right)\right)^2 dx \\ &\quad +2n^{-2}\sum_{\ell=1}^{n-1}\left(n-\ell\right)\int\int K_h\left(x-y\right)f\left(y\right)K_h * f_\ell\left(\bullet|y\right)\left(x\right)\, dx\, dy. \end{aligned}$$

Como cabía esperar, esta fórmula non depende unicamente da función tipo núcleo, o parámetro de suavizado e a densidade marxinal, f, do proceso; senón que tamén depende das densidades condicionais, f_ℓ, que caracterizan a dependencia.

4.3.2. $MISE^*\left(h\right)$: a versión bootstrap estacionaria suavizada do $MISE\left(h\right)$

Consideremos o caso *iid* e o plan de remostraxe bootstrap suavizado estándar descrito na Sección 3.2 (que corresponde con tomar $p=1$ no plan de remostraxe bootstrap estacionario suavizado). É sinxelo obter unha expresión exacta para

$$MISE^*\left(h\right)=\mathbb{E}^*\left[\int\left(\hat{f}_h^*\left(x\right)-\hat{f}_g\left(x\right)\right)^2 dx\right],$$

a versión bootstrap do $MISE\left(h\right)$, descrita por Cao (1993). A expresión resultante é

$$\begin{aligned} MISE^*\left(h\right) &= B^*\left(h\right)+V^*\left(h\right), \text{ onde} \\ B^*\left(h\right) &= n^{-2}\sum_{i,j=1}^{n}\left[\left(K_h * K_g-K_g\right) * \left(K_h * K_g-K_g\right)\right]\left(X_i-X_j\right), \text{ e} \\ V^*\left(h\right) &= n^{-1}h^{-1}R\left(K\right)-n^{-3}\sum_{i,j=1}^{n}\left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right]\left(X_i-X_j\right). \end{aligned}$$

Esta expresión é moi útil dado que non é preciso o uso de Monte Carlo.

De novo, atopámonos co problema da elección óptima da ventá piloto g, que vén ligado á da estimación óptima da curvatura da función de densidade. Así, unha boa elección de g é a que minimiza:

$$\mathbb{E}\left[\left(\int \hat{f}_g''(x)^2dx-\int f''(x)^2dx\right)^2\right],$$

sendo o valor asintótico de dita ventá g:

$$g_0=\left(\frac{\int K''(t)^2dt}{nd_K\int f^{(3)}(x)^2dx}\right)^{1/7}.$$

A continuación, obterase unha expresión exacta para $MISE^*\left(h\right)$ no caso de dependencia, tendo en conta a estacionariedade do proceso. Consideremos unha mostra aleatoria que proveña dun proceso estacionario, $X_1, X_2, \ldots, X_n$, e a versión suavizada do bootstrap estacionario, $\hat{f}_h^*\left(x\right)$, do estimador tipo núcleo da densidade de Parzen-Rosenblatt, dado por

$$\hat{f}_h^*\left(x\right)=\frac{1}{n}\sum_{i=1}^{n}K_h\left(x-X_i^*\right),$$

onde $X_1^*, X_2^*, \ldots, X_n^*$ é unha remostra bootstrap estacionaria suavizada, tomando unha ventá piloto g. O erro cuadrático medio bootstrap vén dado por

$$\begin{aligned} & \mathbb{E}^* \left[\int \left(\hat{f}_h^*(x) - \hat{f}_g(x) \right)^2 dx \right] \\ = & \int \left[\mathbb{E}^* \left(\hat{f}_h^*(x) \right) - \hat{f}_g(x) \right]^2 dx + \int Var^* \left(\hat{f}_h^*(x) \right) dx \\ = & B^*(h) + V^*(h), \end{aligned}$$

onde

$$\begin{aligned} B^*(h) &= \int \left[\mathbb{E}^* \left(\hat{f}_h^*(x) \right) - \hat{f}_g(x) \right]^2 dx, \text{ e} \\ V^*(h) &= \int Var^* \left(\hat{f}_h^*(x) \right) dx. \end{aligned}$$

Se continuamos cos cálculos, obtemos

$$\begin{aligned} B^*(h) &= \int \left[\int K_h(x-y) \hat{f}_g(y) dy - \hat{f}_g(x) \right]^2 dx \\ &= \int \left[\int K_h(x-y) \frac{1}{n} \sum_{i=1}^n K_g(y - X_i) dy - \frac{1}{n} \sum_{i=1}^n K_g(x - X_i) \right]^2 dx \\ &= n^{-2} \int \left[\sum_{i=1}^n \left(\int K_h(x-y) K_g(y - X_i) dy - K_g(x - X_i) \right) \right]^2 dx \\ &= n^{-2} \int \left[\sum_{i=1}^n \left(\int K_h(x - X_i - u) K_g(u) du - K_g(x - X_i) \right) \right]^2 dx \\ &= n^{-2} \int \left[\sum_{i=1}^n \left(K_h * K_g(x - X_i) - K_g(x - X_i) \right) \right]^2 dx \\ &= n^{-2} \sum_{i,j=1}^n \int \left[(K_h * K_g - K_g)(x - X_i) \right] \left[(K_h * K_g - K_g)(x - X_j) \right] dx \\ &= n^{-2} \sum_{i,j=1}^n \int \left[(K_h * K_g - K_g)(-v) \right] \left[(K_h * K_g - K_g)(X_i - X_j - v) \right] dx \\ &= n^{-2} \sum_{i,j=1}^n \int \left[(K_h * K_g - K_g)(v) \right] \left[(K_h * K_g - K_g)(X_i - X_j - v) \right] dx \\ &= n^{-2} \sum_{i,j=1}^n \left[(K_h * K_g - K_g) * (K_h * K_g - K_g) \right] (X_i - X_j), \end{aligned}$$

que, como era esperable, coincide coa expresión para o sesgo bootstrap integrado no caso *iid*.

A varianza bootstrap integrada, en cambio, resulta:

$$\begin{aligned} V^*(h) &= \int Var^*\left(n^{-1}\sum_{i=1}^{n} K_h(x-X_i^*)\right)dx \\ &= n^{-1}\int Var^*\left(K_h(x-X_1^*)\right)dx \\ &\quad +n^{-2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\int Cov^*\left(K_h(x-X_i^*), K_h\left(x-X_j^*\right)\right)dx. \end{aligned}$$

Cómpre ter en conta que

$$\begin{aligned} Cov^*\left(K_h(x-X_i^*), K_h(x-X_j^*)\right) &= \mathbb{E}^*\left[Cov^*\left(K_h(x-X_i^*), K_h(x-X_j^*)\big|_{B_{ij}^*}\right)\right] \\ &+ Cov^*\left[\mathbb{E}^*\left(K_h(x-X_i^*)|_{B_{ij}^*}\right), \mathbb{E}^*\left(K_h(x-X_j^*)\big|_{B_{ij}^*}\right)\right], \end{aligned}$$

mais este segundo sumando é cero xa que

$$\mathbb{E}^*\left[K_h(x-X_i^*)|_{B_{ij}^*}\right] = \mathbb{E}^*\left[K_h\left(x-X_j^*\right)\big|_{B_{ij}^*}\right] = \mathbb{E}^*\left[K_h(x-X_i^*)\right] = \left(K_h * \hat{f}_g\right)(x),$$

sendo B_{ij}^* o indicador de que X_i^* e X_j^* pertenzan ó mesmo bloque bootstrap, é dicir:

$$B_{ij}^* = I_{i+1}^* \cdot I_{i+2}^* \cdot \cdots \cdot I_j^*.$$

É claro que estes indicadores bootstrap teñen distribución de Bernoulli de tal xeito que

$$\mathbb{P}^*\left(B_{ij}^* = 1\right) = (1-p)^{j-i}.$$

Logo, retomando os cálculos da varianza bootstrap integrada, temos que:

$$\begin{aligned} V^*(h) &= n^{-1}\int\left\{\mathbb{E}^*\left(K_h(x-X_1^*)^2\right) - \left[\mathbb{E}^*\left(K_h(x-X_1^*)\right)\right]^2\right\}dx \\ &\quad +2n^{-2}\sum_{\substack{i,j=1\\ i<j}}^{n}\int \mathbb{E}^*\left[Cov^*\left(K_h(x-X_i^*), K_h\left(x-X_j^*\right)\big|_{B_{ij}^*}\right)\right]dx. \end{aligned}$$

O primeiro termo na varianza bootstrap integrada é exactamente o mesmo que no caso iid, o que conduce a

$$\begin{aligned} & n^{-1}\int\left\{\mathbb{E}^*\left(K_h(x-X_1^*)^2\right) - \left[\mathbb{E}^*\left(K_h(x-X_1^*)\right)\right]^2\right\}dx \\ = \;& n^{-1}h^{-1}R(K) - n^{-3}\sum_{i,j=1}^{n}\left[(K_h * K_g) * (K_h * K_g)\right](X_i - X_j). \end{aligned}$$

Neste momento, centrarémonos no termo onde aparecen as covarianzas. Primeiramente, usaremos a distribución bootstrap de Bernoulli que segue B_{ij}^*, e o feito de que X_i^* e X_j^* son independentes condicionalmente a que $B_{ij}^* = 0$. Ademais, se $X_i^{*(d)} = X_k$ e $B_{ij}^* = 1$, entón tense que $X_j^{*(d)} = X_{[(k+j-i-1)\bmod n]+1}$. Tendo en conta o anterior, temos que:

$$
\begin{aligned}
& \mathbb{E}^*\left[Cov^*\left(K_h\left(x-X_i^*\right),K_h\left(x-X_j^*\right)\right)\big|_{B_{ij}^*}\right] \\
= \;& (1-p)^{j-i}\cdot Cov^*\left(K_h\left(x-X_i^*\right),K_h\left(x-X_j^*\right)\big|_{B_{ij}^*=1}\right) \\
= \;& (1-p)^{j-i}\cdot\left[\mathbb{E}^*\left(K_h\left(x-X_i^*\right)\cdot K_h\left(x-X_j^*\right)\big|_{B_{ij}^*=1}\right)\right. \\
& \left.-\mathbb{E}^*\left(K_h\left(x-X_i^*\right)|_{B_{ij}^*=1}\right)\cdot\mathbb{E}^*\left(K_h\left(x-X_j^*\right)\big|_{B_{ij}^*=1}\right)\right] \\
= \;& (1-p)^{j-i}\cdot\left[\mathbb{E}^*\left[\mathbb{E}^*\left(K_h\left(x-X_i^*\right)\cdot K_h\left(x-X_j^*\right)\big|_{U_i^*,U_j^*,B_{ij}^*=1}\right)\Big|_{B_{ij}^*=1}\right]\right. \\
& \left.-\left[\mathbb{E}^*\left(K_h\left(x-X_1^*\right)\right)\right]^2\right] \\
= \;& (1-p)^{j-i}\cdot\left[\mathbb{E}^*\left[\frac{1}{n}\sum_{k=1}^{n}K_h\left(x-X_k-gU_i^*\right)\cdot K_h\left(x-X_{\lceil(k+j-i-1)\bmod n\rceil+1}-gU_j^*\right)\right]\right. \\
& \left.-\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}^*\left(K_h\left(x-X_k-gU_1^*\right)\right)\right]^2\right] \\
= \;& (1-p)^{j-i}\cdot\left[\frac{1}{n}\sum_{k=1}^{n}\int\int K_h\left(x-X_k-gu\right)\right. \\
& \cdot K_h\left(x-X_{\lceil(k+j-i-1)\bmod n\rceil+1}-gv\right)\cdot K\left(u\right)\cdot K\left(v\right)\,du\,dv \\
& \left.-\left(\frac{1}{n}\sum_{k=1}^{n}\int K_h\left(x-X_k-gu\right)\cdot K\left(u\right)\,du\right)^2\right] \\
= \;& (1-p)^{j-i}\cdot\left[\frac{1}{n}\sum_{k=1}^{n}\int\int K_h\left(x-X_k-s\right)\right. \\
& \cdot K_h\left(x-X_{\lceil(k+j-i-1)\bmod n\rceil+1}-t\right)\cdot K_g\left(s\right)\cdot K_g\left(t\right)\,ds\,dt \\
& \left.-\left(\frac{1}{n}\sum_{k=1}^{n}\int K_h\left(x-X_k-s\right)\cdot K_g\left(s\right)\,ds\right)^2\right] \\
= \;& (1-p)^{j-i}\cdot\left[\frac{1}{n}\sum_{k=1}^{n}K_h*K_g\left(x-X_k\right)\cdot K_h*K_g\left(x-X_{\lceil(k+j-i-1)\bmod n\rceil+1}\right)\right. \\
& \left.-\left[\frac{1}{n}\sum_{k=1}^{n}K_h*K_g\left(x-X_k\right)\right]^2\right].
\end{aligned}
$$

Consecuentemente, o segundo termo da expresión da varianza bootstrap integrada será

$$
2n^{-2}\sum_{\substack{i,j=1\\ i<j}}^{n}\int E^*\left[Cov^*\left(K_h\left(x-X_i^*\right),K_h\left(x-X_j^*\right)\big|_{B_{ij}^*}\right)\right]dx
$$

$$
\begin{aligned}
&= 2n^{-2}\sum_{\substack{i,j=1\\ i<j}}^{n}(1-p)^{j-i}\cdot \\
&\quad\left[n^{-1}\sum_{k=1}^{n}\int K_h * K_g\,(x-X_k)\cdot K_h * K_g\left(x-X_{\lceil (k+j-i-1)\bmod n\rceil+1}\right)dx\right. \\
&\quad\left.-n^{-2}\sum_{k,\ell=1}^{n}\int K_h * K_g\,(x-X_k)\cdot K_h * K_g\,(x-X_\ell)\,dx\right] \\
&= 2n^{-3}\sum_{\substack{i,j=1\\ i<j}}^{n}(1-p)^{j-i}\cdot\sum_{k=1}^{n}\left[(K_h*K_g)*(K_h*K_g)\right]\left(X_k-X_{\lceil (k+j-i-1)\bmod n\rceil+1}\right) \\
&\quad -2n^{-4}\sum_{\substack{i,j=1\\ i<j}}^{n}(1-p)^{j-i}\cdot\sum_{k,\ell=1}^{n}\left[(K_h*K_g)*(K_h*K_g)\right](X_k-X_\ell) \\
&= 2n^{-3}\sum_{\ell=1}^{n-1}(n-\ell)(1-p)^{\ell}\cdot\sum_{k=1}^{n}\left[(K_h*K_g)*(K_h*K_g)\right]\left(X_k-X_{\lceil (k+\ell-1)\bmod n\rceil+1}\right) \\
&\quad -2n^{-4}\left(\sum_{\ell=1}^{n-1}(n-\ell)(1-p)^{\ell}\right)\cdot\sum_{k,\ell=1}^{n}\left[(K_h*K_g)*(K_h*K_g)\right](X_k-X_\ell) \\
&= 2n^{-3}\sum_{\ell=1}^{n-1}\sum_{k=1}^{n}(n-\ell)(1-p)^{\ell}\left[(K_h*K_g)*(K_h*K_g)\right]\left(X_k-X_{\lceil (k+\ell-1)\bmod n\rceil+1}\right) \\
&\quad -2n^{-4}\left(n\frac{1-p-(1-p)^n}{p}-\frac{(n-1)(1-p)^{n+1}-n(1-p)^n+1-p}{p^2}\right) \\
&\quad\cdot\sum_{k,\ell=1}^{n}\left[(K_h*K_g)*(K_h*K_g)\right](X_k-X_\ell)\,.
\end{aligned}
$$

Para probar a última igualdade cómpre calcular a suma $\sum_{\ell=1}^{n-1}(n-\ell)(1-p)^{\ell}$, obtendo:

$$
\begin{aligned}
\sum_{\ell=1}^{n-1}(n-\ell)(1-p)^{\ell} &= n\sum_{\ell=1}^{n-1}(1-p)^{\ell}-\sum_{\ell=1}^{n-1}\ell(1-p)^{\ell} \\
&= n\frac{(1-p)^{n-1}(1-p)-(1-p)}{(1-p)-1}-(1-p)\sum_{\ell=1}^{n-1}\ell(1-p)^{\ell-1} \\
&= n\frac{1-p-(1-p)^n}{p}-(1-p)\sum_{\ell=1}^{n-1}\ell(1-p)^{\ell-1}. \qquad (4.2)
\end{aligned}
$$

Para calcular o valor de $\sum_{\ell=1}^{n-1}\ell(1-p)^{\ell-1}$ consideremos a función de r:

$$g(r)=\sum_{\ell=1}^{n-1}\ell r^{\ell-1}.$$

Obviamente $g(r)$ é a derivada en r da función:

$$G(r) = \sum_{\ell=1}^{n-1} r^{\ell} = \frac{r^{n-1}r - r}{r-1} = \frac{r^n - r}{r-1}.$$

Así pois:

$$\begin{aligned} g(r) = \frac{dG(r)}{dr} &= \frac{(nr^{n-1}-1)(r-1) - (r^n - r)1}{(r-1)^2} \\ &= \frac{(nr^{n-1}-1)(r-1) - (r^n - r)}{(r-1)^2} \\ &= \frac{(n-1)r^n - nr^{n-1} + 1}{(r-1)^2}. \end{aligned}$$

Co cal:

$$\begin{aligned} \sum_{\ell=1}^{n-1} \ell(1-p)^{\ell-1} = g(1-p) &= \frac{(n-1)(1-p)^n - n(1-p)^{n-1} + 1}{(1-p-1)^2} \\ &= \frac{(n-1)(1-p)^n - n(1-p)^{n-1} + 1}{p^2}. \end{aligned}$$

Utilizando isto na expresión (4.2), obtense:

$$\sum_{\ell=1}^{n-1}(n-\ell)(1-p)^{\ell} = n\frac{1-p-(1-p)^n}{p} - \frac{(n-1)(1-p)^{n+1} - n(1-p)^n + 1 - p}{p^2},$$

como queríamos demostrar.

Finalmente, xuntando todos estes termos, obtemos a expresión exacta para o $MISE^*(h)$ nun contexto de dependencia:

$$\begin{aligned} MISE^*(h) &= n^{-2}\sum_{i,j=1}^{n}\left[(K_h * K_g - K_g) * (K_h * K_g - K_g)\right](X_i - X_j) \\ &\quad + n^{-1}h^{-1}R(K) - n^{-3}\sum_{i,j=1}^{n}\left[(K_h * K_g) * (K_h * K_g)\right](X_i - X_j) \\ &\quad + 2n^{-3}\sum_{\ell=1}^{n-1}\sum_{k=1}^{n}(n-\ell)(1-p)^{\ell}\left[(K_h * K_g) * (K_h * K_g)\right]\left(X_k - X_{\lceil (k+\ell-1) \bmod n \rceil + 1}\right) \\ &\quad - 2n^{-4}\left(n\frac{1-p-(1-p)^n}{p} - \frac{(n-1)(1-p)^{n+1} - n(1-p)^n + 1 - p}{p^2}\right) \\ &\quad \cdot \sum_{k,\ell=1}^{n}\left[(K_h * K_g) * (K_h * K_g)\right](X_k - X_{\ell}) \end{aligned}$$

$$
\begin{aligned}
= \quad & n^{-1}h^{-1}R\left(K\right) \\
& + \left[\frac{n-1}{n^3} - 2\frac{1-p-\left(1-p\right)^n}{pn^3} + 2\frac{\left(n-1\right)\left(1-p\right)^{n+1} - n\left(1-p\right)^n + 1 - p}{p^2n^4}\right] \\
& \cdot \sum_{i,j=1}^{n} \left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right]\left(X_i - X_j\right) \\
& -2n^{-2}\sum_{i,j=1}^{n}\left(K_h * K_g * K_g\right)\left(X_i - X_j\right) \\
& +n^{-2}\sum_{i,j=1}^{n}\left(K_g * K_g\right)\left(X_i - X_j\right) \\
& +2n^{-3}\sum_{\ell=1}^{n-1}\left(n-\ell\right)\left(1-p\right)^{\ell}\sum_{k=1}^{n}\left[\left(K_h * K_g\right) * \left(K_h * K_g\right)\right]\left(X_k - X_{\lceil (k+\ell-1) \bmod n \rceil + 1}\right).
\end{aligned}
$$

A partir desta expresión obsérvase que o cálculo da función $MISE^*(h)$ para cada ventá h implica o cálculo de catro dobres bucles en índices que varían dende 1 ata n. Como consecuencia, o número de operacións necesarias é de orde n^2. Por conseguinte, o tempo de cálculo non debe resultar excesivo, agás para tamaños mostrais, n, extremadamente grandes.

Capítulo 5

Simulacións

Co fin de comprobar na práctica o bo comportamento dos selectores do parámetro de ventá proporcionados polos métodos presentados no Capítulo 4, e en particular, para comprobar a eficacia do método bootstrap estacionario suavizado empregando a expresión exacta para o $MISE^*(h)$, desenvolveremos un estudo de simulación neste capítulo. Este estudo de simulación será similar ó levado a cabo por Cao *et al.* (1993).

Este capítulo está organizado en dúas seccións. Por unha parte, describiranse os modelos que se van simular indicando, entre outras cousas, as densidades poboacionais e o tipo de dependencia. Posteriormente, procederase ó correspondente estudo de simulación no que se compararán empiricamente os diferentes selectores de ventá, así como o tempo de CPU que precisa cada método. Finalmente, focalizarémonos na comprobación do bo comportamento (na práctica) do parámetro h obtido a partir da expresión exacta para o $MISE^*(h)$, e consecuentemente, no método bootstrap estacionario suavizado.

5.1. Modelos simulados

A continuación describiremos as diferentes funcións de densidade que empregaremos para mostrar os resultados prácticos para cada situación.

- Un modelo autorregresivo de orde 2 dado por:

$$X_t = -0{,}9X_{t-1} - 0{,}2X_{t-2} + a_t, \tag{5.1}$$

  sendo a_t variables aleatorias *iid* con distribución común $N(0,1)$ independentes de $X_{t-1}, X_{t-2}, ...$

- Un modelo de medias móbiles de orde 2 que vén dado por:

$$X_t = a_t - 0{,}9a_{t-1} + 0{,}2a_{t-2}, \tag{5.2}$$

  onde a_t son variables aleatorias *iid* con distribución común normal estándar.

- A mostra, neste caso, provén dun modelo $AR(1)$ con ruído normal, dado por:

$$X_t = \phi X_{t-1} + (1-\phi^2)^{1/2} a_t, \tag{5.3}$$

  sendo a_t variables aleatorias independentes con distribución común, dada pola normal estándar. Ademais, a_t é independente de $X_{t-1}, X_{t-2}, ...$ A autocorrelación será fixada para os valores $\phi = 0, \pm 0{,}3, \pm 0{,}6, \pm 0{,}9$. Deste xeito, como veremos máis adiante, a distribución marxinal das X_t é sempre unha $N(0,1)$ independentemente do valor de ϕ elixido.

- Un modelo $AR(1)$ con erro que ten distribución mixtura dunha exponencial e dexenerada en cero:

$$X_t = \phi X_{t-1} + a_t. \tag{5.4}$$

  Neste caso, a_t son variables aleatorias, independentes e identicamente distribuídas, con distribución dada a partir do condicionamento dunha variable I_t discreta:

$$\mathbb{P}(I_t = 1) = \phi, \mathbb{P}(I_t = 2) = 1 - \phi, \text{ con}$$

$$a_t|_{I_t=1} \stackrel{d}{=} 0 \text{ (constante)}, a_t|_{I_t=2} \stackrel{d}{=} \exp(1).$$

  Consideraremos $\phi = 0,0{,}3, 0{,}6, 0{,}9$.

- Un modelo $AR(1)$ formalmente definido mediante a mesma expresión (5.5) que o modelo (5.4):

$$X_t = \phi X_{t-1} + a_t, \tag{5.5}$$

  mais a distribución da variable aleatoria do erro é diferente:

$$\mathbb{P}(I_t = 1) = \phi^2, \mathbb{P}(I_t = 2) = 1 - \phi^2, \text{ con}$$

$$a_t|_{I_t=1} \stackrel{d}{=} 0 \text{ (constante)}, a_t|_{I_t=2} \stackrel{d}{=} Dexp(1),$$

  sendo $Dexp(\lambda)$ a distribución dobre exponencial de parámetro λ, é dicir, a que ten por función de densidade $g(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$. Os valores que empregaremos para ϕ serán $\phi = 0, \pm 0{,}3, \pm 0{,}6, \pm 0{,}9$.

- Unha mixtura de dous modelos $AR(1)$ con ruído normal:

$$X_t = \begin{cases} X_t^{(1)} & \text{con probabilidade } 1/2 \\ X_t^{(2)} & \text{con probabilidade } 1/2 \end{cases}, \tag{5.6}$$

  sendo $X_t^{(j)} = (-1)^{j+1} + 0{,}5X_{t-1}^{(j)} + a_t^{(j)}$ con $j = 1, 2$, para todo $t \in \mathbb{Z}$, e $a_t^{(j)} \stackrel{d}{=} N(0, 0{,}6)$ independentes.

É sinxelo ver cal é a distribución que seguirá cada un dos modelos que acabamos de presentar, describindo a súa obtención a continuación.

- O modelo autorregresivo dado en (5.1) ten, como dixemos, ruído normal, de xeito que a_t son variables aleatorias *iid* con distribución normal estándar e independentes de $X_{t-1}, X_{t-2}, \ldots$. Polo tanto, a distribución marxinal de X_t é normal. Bastará entón con calcular a partir de (5.1) a esperanza e a varianza de X_t. Consideremos, por xeralizar, o modelo $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t$, sendo a_t variables aleatorias *iid* con distribución $N(0, \sigma_a^2)$, e vexamos cal é a esperanza de X_t: $\mathbb{E}(X_t) = \mu_X, \forall t \in \mathbb{Z}$.

$$\begin{aligned} \mu_X = \mathbb{E}(X_t) &= \mathbb{E}\left(\phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t\right) \\ &= \phi_1 \mathbb{E}(X_{t-1}) + \phi_2 \mathbb{E}(X_{t-2}) + \mathbb{E}(a_t) \\ &= \phi_1 \mathbb{E}(X_t) + \phi_2 \mathbb{E}(X_t) \\ &= \phi_1 \mu_X + \phi_2 \mu_X, \end{aligned}$$

polo que $(1-\phi_1-\phi_2)\mu_X = 0$, e consecuentemente, se $\phi_1+\phi_2 \neq 1$, entón $\mu_X = 0$. Igualmente, grazas á estacionariedade, resulta doado atopar a varianza de X_t: $\sigma^2(X_t) = \mathrm{Var}(X_t) = \sigma_X^2, \forall t \in \mathbb{Z}$:

$$\begin{aligned}
\sigma_X^2 = \mathrm{Var}(X_t) &= \mathrm{Var}\left(\phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t\right) \\
&= \mathrm{Var}(\phi_1 X_{t-1} + \phi_2 X_{t-2}) + \sigma_a^2 \\
&= \phi_1^2 \mathrm{Var}(X_{t-1}) + \phi_2^2 \mathrm{Var}(X_{t-2}) + 2\phi_1\phi_2 \mathrm{Cov}(X_{t-1}, X_{t-2}) + \sigma_a^2 \\
&= \phi_1^2 \mathrm{Var}(X_t) + \phi_2^2 \mathrm{Var}(X_t) + 2\phi_1\phi_2 \mathrm{Cov}(X_{t-1}, X_{t-2}) + \sigma_a^2 \qquad (5.7) \\
&= \phi_1^2 \mathrm{Var}(X_t) + \phi_2^2 \mathrm{Var}(X_t) + 2\phi_1\phi_2 \frac{\phi_1}{1-\phi_2} \mathrm{Var}(X_t) + \sigma_a^2 \\
&= \left(\phi_1^2 + \phi_2^2 + 2\phi_1\phi_2 \frac{\phi_1}{1-\phi_2}\right) \mathrm{Var}(X_t) + \sigma_a^2 \\
&= \left(\phi_1^2 + \phi_2^2 + 2\phi_1\phi_2 \frac{\phi_1}{1-\phi_2}\right) \sigma_X^2 + \sigma_a^2,
\end{aligned}$$

de onde se deduce que

$$\sigma_X^2 = \frac{\sigma_a^2}{\left(1 - \phi_1^2 - \phi_2^2 - 2\phi_1\phi_2 \frac{\phi_1}{1-\phi_2}\right)}. \qquad (5.8)$$

En (5.7) úsase que

$$\begin{aligned}
\mathrm{Cov}(X_{t-1}, X_{t-2}) &= \mathbb{E}(X_{t-1} \cdot X_{t-2}) - \mathbb{E}(X_{t-1})\mathbb{E}(X_{t-2}) \\
&= \mathbb{E}(X_{t-1} \cdot X_{t-2}) \\
&= \mathbb{E}(X_t \cdot X_{t+1}) \\
&= \mathbb{E}(X_t(\phi_1 X_t + \phi_2 X_{t-1} + a_{t+1})) \\
&= \phi_1 \mathbb{E}(X_t^2) + \phi_2 \mathbb{E}(X_t \cdot X_{t-1}),
\end{aligned}$$

polo que $\mathbb{E}(X_t \cdot X_{t-1}) = \phi_1\sigma_X^2 + \phi_2\mathbb{E}(X_t \cdot X_{t-1})$. Logo $\mathbb{E}(X_t \cdot X_{t-1}) = \dfrac{\phi_1}{1-\phi_2}\sigma_X^2$.

Así, obter a varianza de X_t non ten máis que sustituir en (5.8) polos correspondentes valores de $\phi_1 = -0{,}9$ e $\phi_2 = -0{,}2$, tendo en conta que $\sigma_a^2 = 1$. Deste xeito, resulta que a distribución marxinal das X_t é $X_t \overset{d}{=} N(0, 0{,}42)$.

- Procederemos de xeito análogo para coñecer a distribución do modelo (5.2). Consideremos, a modo de xeralización, un modelo de medias móbiles tal que $X_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2}$, onde a_t son variables aleatorias con distribución normal estándar, e a_s independente de a_t se $s \neq t$. Neste caso, os cálculos da esperanza e varianza de X_t resultan máis sinxelos, xa que cada X_t é suma de variables aleatorias normais independentes con media:

$$\begin{aligned}
\mu_X = \mathbb{E}(X_t) &= \mathbb{E}(a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2}) \\
&= \mathbb{E}(a_t) + \theta_1 \mathbb{E}(a_{t-1}) + \theta_2 \mathbb{E}(a_{t-2}) \\
&= \mathbb{E}(a_t) + \theta_1 \mathbb{E}(a_t) + \theta_2 \mathbb{E}(a_t) \\
&= (1 + \theta_1 + \theta_2)\mathbb{E}(a_t) \\
&= 0.
\end{aligned}$$

e varianza:

$$\begin{aligned}\sigma_X^2 = \mathrm{Var}(X_t) &= \mathrm{Var}(\theta_1 a_{t-1} + \theta_2 a_{t-2}) + \mathrm{Var}(a_t) \\ &= \theta_1^2 \mathrm{Var}(a_{t-1}) + \theta_2^2 \mathrm{Var}(a_{t-2}) + \mathrm{Var}(a_t) \\ &= (\theta_1^2 + \theta_2^2 + 1)\mathrm{Var}(a_t) \\ &= \theta_1^2 + \theta_2^2 + 1.\end{aligned}$$

Como consecuencia, a distribución marxinal das X_t é $X_t \overset{d}{=} N\left(0, \theta_1^2 + \theta_2^2 + 1\right)$, sendo $\theta_1 = -0{,}9$ e $\theta_2 = 0{,}2$, é dicir, $X_t \overset{d}{=} N(0, 1{,}85)$.

- En (5.3) temos de novo un modelo autorregresivo de orde 1, con ruído normal, sendo a_t variables aleatorias *iid* con distribución normal estándar, e a_t independente de X_{t-1}, X_{t-2}.... Logo, é sinxelo demostrar que a distribución marxinal das X_t é tamén normal, con media:

$$\begin{aligned}\mu_X = \mathbb{E}(X_t) &= \mathbb{E}\left(\phi X_{t-1} + (1-\phi^2)^{1/2} a_t\right) \\ &= \phi \mathbb{E}(X_{t-1}) + (1-\phi^2)^{1/2}\mathbb{E}(a_t) \\ &= \phi \mathbb{E}(X_t) + (1-\phi^2)^{1/2}\mathbb{E}(a_t) \\ &= \phi \mathbb{E}(X_t) \\ &= \phi \mu_X,\end{aligned}$$

logo $(1-\phi)\mu_X = 0$, e sempre que $\phi \neq 1$, tense que $\mu_X = 0$. Obteñamos agora a varianza de X_t.

$$\begin{aligned}\sigma_X^2 = \mathrm{Var}(X_t) &= \mathrm{Var}\left(\phi X_{t-1}\right) + \mathrm{Var}\left((1-\phi^2)^{1/2} a_t\right) \\ &= \phi^2 \mathrm{Var}\left(X_{t-1}\right) + (1-\phi^2)\mathrm{Var}\left(a_t\right) \\ &= \phi^2 \mathrm{Var}\left(X_t\right) + (1-\phi^2) \\ &= \phi^2 \sigma_X^2 + (1-\phi^2),\end{aligned}$$

e consecuentemente, $(1-\phi^2)\sigma_X^2 = 1-\phi^2$, de onde concluímos que $\sigma_X^2 = 1$. É dicir, a distribución marxinal das X_t é $X_t \overset{d}{=} N(0,1)$.

- O modelo dado en (5.4) é un proceso autorregresivo de orde 1, con erro que posúe estructura exponencial. Aplicando a ecuación (5.4) recursivamente k veces (sendo k un número natural), obtense:

$$\begin{aligned}X_t &= \phi X_{t-1} + a_t = \phi^2 X_{t-2} + \phi a_{t-1} + a_t = \cdots \\ &= \phi^{k+1} X_{t-k-1} + \phi^k a_{t-k} + \phi^{k-1} a_{t-k+1} + \cdots + \phi a_{t-1} + a_t.\end{aligned} \tag{5.9}$$

Denotando por φ_X e φ_a as funcións características das variables aleatorias X_s e a_s; isto é, $\varphi_X(t) = \mathbb{E}\left(e^{itX_s}\right)$ e $\varphi_a(t) = \mathbb{E}\left(e^{ita_s}\right)$, tense, a partir da expresión (5.9):

$$\varphi_X(t) = \varphi_X\left(\phi^{k+1}t\right) \cdot \varphi_a\left(\phi^k t\right) \cdot \varphi_a\left(\phi^{k-1}t\right) \cdot \cdots \cdot \varphi_a\left(\phi t\right) \cdot \varphi_a\left(t\right),$$

é dicir,

$$\varphi_X(t) = \varphi_X\left(\phi^{k+1}t\right) \cdot \prod_{j=0}^{k} \varphi_a(\phi^j t). \tag{5.10}$$

Por outra banda, é ben coñecido que a función característica de $Y \stackrel{d}{=} \exp(\lambda)$ vén dada por $\varphi_Y(t) = \left(1 - \frac{it}{\lambda}\right)^{-1}$. Como consecuencia, a función característica de a_s é:

$$\begin{aligned}
\varphi_a(t) &= \mathbb{E}\left(e^{ita_s}\right) = \mathbb{E}\left[\mathbb{E}\left(e^{ita_s}\big|_{I_s}\right)\right] \\
&= \mathbb{P}\left(I_s = 1\right) \cdot \mathbb{E}\left(e^{ita_s}\big|_{I_s=1}\right) + \mathbb{P}\left(I_s = 2\right) \cdot \mathbb{E}\left(e^{ita_s}\big|_{I_s=2}\right) \\
&= \phi e^{it0} + (1-\phi) \cdot \left(1 - \frac{it}{1}\right)^{-1} = \phi + (1-\phi) \cdot (1-it)^{-1} \\
&= (1-it)^{-1}\left[\phi \cdot (1-it) + 1 - \phi\right] = (1-it)^{-1} \cdot (1 - i\phi t) \\
&= \frac{1 - i\phi t}{1 - it}.
\end{aligned} \tag{5.11}$$

Utilizando a expresión (5.11) en (5.10) obtense:

$$\begin{aligned}
\varphi_X(t) &= \varphi_X\left(\phi^{k+1}t\right) \cdot \prod_{j=0}^{k}\left(\frac{1 - i\phi^{j+1}t}{1 - i\phi^j t}\right) = \varphi_X\left(\phi^{k+1}t\right) \cdot \left(\frac{\prod_{j=0}^{k}\left(1 - i\phi^{j+1}t\right)}{\prod_{j=0}^{k}\left(1 - i\phi^j t\right)}\right) \\
&= \varphi_X\left(\phi^{k+1}t\right) \cdot \left(\frac{\left(1 - i\phi^{k+1}t\right)\prod_{j=1}^{k}\left(1 - i\phi^j t\right)}{(1-it)\prod_{j=1}^{k}\left(1 - i\phi^j t\right)}\right) \\
&= \varphi_X\left(\phi^{k+1}t\right) \cdot \frac{1 - i\phi^{k+1}t}{1 - it},
\end{aligned}$$

ou, o que é o mesmo,

$$\varphi_X(t) = \varphi_X\left(\phi^{k+1}t\right) \cdot \frac{1 - i\phi^{k+1}t}{1 - it}. \tag{5.12}$$

Dado que $0 < \phi < 1$, tomando límites cando $k \longrightarrow \infty$ e tendo en conta que calquera función característica, $\varphi(t)$, é continua e verifica que $\varphi(0) = \mathbb{E}\left[e^{i \cdot 0 \cdot Y}\right] = 1$, a expresión (5.12) permite deducir:

$$\begin{aligned}
\varphi_X(t) &= \lim_{k\to\infty}\left[\varphi_X\left(\phi^{k+1}t\right) \cdot \frac{1 - i\phi^{k+1}t}{1 - it}\right] = \lim_{k\to\infty}\varphi_X\left(\phi^{k+1}t\right) \cdot \frac{\lim_{k\to\infty}\left(1 - i\phi^{k+1}t\right)}{1 - it} \\
&= \varphi_X\left(\lim_{k\to\infty}\phi^{k+1}t\right) \cdot \frac{1 - i\left(\lim_{k\to\infty}\phi^{k+1}\right)t}{1 - it} = \varphi_X(0) \cdot \frac{1 - i \cdot 0 \cdot t}{1 - it} \\
&= (1-it)^{-1},
\end{aligned}$$

que é a función característica dunha exp(1). Así pois, baixo o modelo (5.4), a distribución marxinal da X_t é unha exp(1).

- Neste caso, o modelo autorregresivo dado en (5.5) posúe estructura doble exponencial. Denotando por Z unha variable aleatoria con distribución $Dexp(\lambda)$ e definindo $\mathbb{P}(S = 1) = \mathbb{P}(S = -1) = \frac{1}{2}$,

e Y unha variable aleatoria $\exp(\lambda)$, independente de S, pode demostrarse que $Z = S \cdot Y$, de xeito que a función característica de Z é:

$$\begin{aligned}
\varphi_Z(t) &= \mathbb{E}\left(e^{itZ}\right) = \mathbb{E}\left[\mathbb{E}\left(e^{itZ}\big|_S\right)\right] \\
&= \mathbb{E}\left(e^{itZ}\big|_{S=1}\right)\cdot\mathbb{P}(S=1) + \mathbb{E}\left(e^{itZ}\big|_{S=-1}\right)\cdot\mathbb{P}(S=-1) \\
&= \frac{1}{2}\cdot\mathbb{E}\left(e^{itY}\right) + \frac{1}{2}\cdot\mathbb{E}\left(e^{it(-Y)}\right) = \frac{1}{2}\cdot\varphi_Y(t) + \frac{1}{2}\cdot\varphi_Y(-t) \\
&= \frac{1}{2}\cdot\left(1-\frac{it}{\lambda}\right)^{-1} + \frac{1}{2}\cdot\left(1+\frac{it}{\lambda}\right)^{-1} = \frac{1}{2}\cdot\frac{1+\frac{it}{\lambda}+1-\frac{it}{\lambda}}{\left(1-\frac{it}{\lambda}\right)\left(1+\frac{it}{\lambda}\right)} \\
&= \left(1-\frac{it}{\lambda}\right)^{-1}\cdot\left(1+\frac{it}{\lambda}\right)^{-1} = \left[1-\left(\frac{it}{\lambda}\right)^2\right]^{-1} \\
&= \left(1+\frac{t^2}{\lambda^2}\right)^{-1}.
\end{aligned}$$

Así pois, no caso concreto $\lambda = 1$, obtense que:

$$\varphi_Z(t) = (1+t^2)^{-1}. \tag{5.13}$$

Como consecuencia da distribución que segue o ruído a_t do modelo (5.5) e de (5.13), obtense:

$$\begin{aligned}
\varphi_a(t) &= \phi^2 e^{it0} + (1-\phi^2)\cdot(1+t^2)^{-1} = \phi^2 + (1-\phi^2)\cdot(1+t^2)^{-1} \\
&= (1+t^2)^{-1}\cdot\left[\phi^2(1+t^2) + 1 - \phi^2\right] \\
&= \frac{1+\phi^2 t^2}{1+t^2}.
\end{aligned} \tag{5.14}$$

Neste modelo, a expresión (5.10) volve ser certa mais cun valor distinto para a función característica $\varphi_a(t)$, que agora é a expresión (5.14). Usando ámbalas dúas obtense:

$$\begin{aligned}
\varphi_X(t) &= \varphi_X\left(\phi^{k+1}t\right)\cdot\prod_{j=0}^{k}\left(\frac{1+\phi^2\phi^{2j}t^2}{1+\phi^{2j}t^2}\right) \\
&= \varphi_X\left(\phi^{k+1}t\right)\cdot\prod_{j=0}^{k}\left(\frac{1+\phi^{2(j+1)}t^2}{1+\phi^{2j}t^2}\right) \\
&= \varphi_X\left(\phi^{k+1}t\right)\cdot\frac{1+\phi^{2k+2}t^2}{1+t^2}.
\end{aligned} \tag{5.15}$$

Tomando límites cando $k \to \infty$ na expresión (5.15), temos que:

$$\varphi_X(t) = \varphi_X(0)\cdot\frac{1}{1+t^2} = \frac{1}{1+t^2} = (1+t^2)^{-1},$$

que é a función característica dunha $Dexp(1)$. Co cal, a distribución marxinal das X_t é unha $Dexp(1)$.

- Por último, o modelo dado en (5.6) é unha mixtura de dous modelos autorregresivos de orde 1 con ruído normal. É ben coñecido que, como o ruído $a_t^{(j)}$ ten distribución normal, tamén ten

distribución normal marxinal a variable aleatoria $X_t^{(j)}$. A súa media e varianza poden calcularse sinxelamente no caso de estacionariedade:

$$\begin{aligned}\mu_X^{(j)} = \mathbb{E}\left(X_t^{(j)}\right) &= (-1)^{j+1} + 0{,}5\mathbb{E}\left(X_{t-1}^{(j)}\right) + \mathbb{E}\left(a_t^{(j)}\right) \\ &= (-1)^{j+1} + 0{,}5\mu_X^{(j)},\end{aligned}$$

o que implica que $0{,}5\mu_X^{(j)} = (-1)^{(j+1)}$. Logo, concluímos que $\mu_X^{(j)} = 2 \cdot (-1)^{j+1}$, de onde se segue que $\mu_X^{(1)} = 2$ e $\mu_X^{(2)} = -2$.

Por outra banda, no caso da varianza:

$$\begin{aligned}\sigma_X^{2(j)} = \mathrm{Var}\left(X_t^{(j)}\right) &= \mathrm{Var}\left((-1)^{j+1} + 0{,}5X_{t-1}^{(j)} + a_t^{(j)}\right) \\ &= 0{,}5^2\mathrm{Var}\left(X_{t-1}^{(j)}\right) + \mathrm{Var}\left(a_t^{(j)}\right) \\ &= 0{,}25\sigma_X^{2(j)} + 0{,}6,\end{aligned}$$

polo que $\sigma_X^{2(j)} = 0{,}8$, para $j = 1, 2$.

Como consecuencia, a distribución marxinal de $X_t^{(1)}$ é unha $N(2, 0{,}8)$, a distribución marxinal de $X_t^{(2)}$ é unha $N(-2, 0{,}8)$, e a distribución marxinal da X_t é a seguinte mixtura de normais:

$$\frac{1}{2}N\left(2, 0{,}8\right) + \frac{1}{2}N\left(-2, 0{,}8\right).$$

5.2. Resultados de simulación

Unha vez que describimos os modelos que imos empregar no seguinte estudo de simulación, procederemos a comezar co mesmo. En primeiro lugar, obteremos, como xa dixemos, mediante cada un dos métodos descritos no Capítulo 4, os parámetros de ventá para cada unha das densidades mostradas anteriormente. Denotaremos por h_{PI} a ventá obtida mediante o método plug-in, h_{MCV} a ventá obtida mediante o método de validación cruzada modificado, h_{MC}^{BOOT} a ventá obtida mediante o método bootstrap estacionario suavizado empregando Monte Carlo para aproximar o $MISE^*(h)$, por h_{BOOT} a ventá obtida mediante o método bootstrap estacionario suavizado empregando a expresión exacta para o $MISE^*(h)$, e por h_{MISE}, a obtida minimizando o $MISE(h)$, facilmente calculable debido a que coñecemos a densidade real da serie temporal a tratar. Ademais, denotaremos por CPU_{PI}, CPU_{MCV}, CPU_{MC}^{BOOT}, CPU_{BOOT}; os tempos de computación requeridos para obter os parámetros ventá polos métodos plug-in, validación cruzada modificada, bootstrap estacionario suavizado empregando Monte Carlo para a aproximación do $MISE^*(h)$, e bootstrap estacionario suavizado empregando a expresión exacta para o mesmo; respectivamente. Estes tempos virán dados en segundos.

Ademais, presentaremos os resultados das simulacións

$$\begin{gathered}\frac{h_{PI} - h_{MISE}}{h_{MISE}} \\ \frac{h_{MCV} - h_{MISE}}{h_{MISE}} \\ \frac{h_{BOOT} - h_{MISE}}{h_{MISE}}\end{gathered} \tag{5.16}$$

para comprobar o bo funcionamento da expresión exacta desenvolta para o $MISE^*(h)$ no Capítulo 4, así como das ventás plug-in e de validación cruzada modificada. Para iso, veremos se os resultados

das expresións (5.16), aproximadas por simulación, son próximas a cero. Nestas expresións poñemos en relación o selector de ventá correspondente, do cal queremos ver o seu bo comportamento a nivel práctico; co selector de ventá h_{MISE}. Logo, é claro que se este resultado é próximo a cero, implicará que ambos selectores son similares.

De xeito análogo, presentaremos tamén os resultados das expresións (5.17) aproximadas por simulación, nas cales poñemos en relación $MISE(h_{MISE})$ cos criterios de erro $MISE(h_{PI}), MISE(h_{MCV})$ e $MISE(h_{BOOT})$, é dicir, o erro que cometemos ó estimar a función de densidade cos selectores h_{PI}, h_{MCV} e h_{BOOT}, respectivamente. Novamente, se este resultado é próximo a cero, implicará que o erro cometido ó estimar a función de densidade co parámetro de suavizado correspondente será próximo ó erro cometido co parámetro h_{MISE}, que é o mínimo erro posible.

$$\begin{aligned} &\frac{MISE(h_{PI}) - MISE(h_{MISE})}{MISE(h_{MISE})} \\ &\frac{MISE(h_{MCV}) - MISE(h_{MISE})}{MISE(h_{MISE})} \\ &\frac{MISE(h_{BOOT}) - MISE(h_{MISE})}{MISE(h_{MISE})} \end{aligned} \tag{5.17}$$

Procederemos como segue para escoller o valor óptimo do parámetro de suavizado:

1. Consideraremos un conxunto de catro ventás equiespaciadas entre os valores 0,01 e 10.
2. Deste xeito, mediante cada método, escolleremos a ventá máis axeitada entre esas catro dadas, é dicir, a que presenta o menor valor de todas elas para a función que se está a minimizar. Denotémola por h_{OPT_1}.
3. Posteriormente, consideraremos entre o conxunto de catro ventás iniciais, a anterior e a posterior a h_{OPT_1}.
4. Neste momento, construiremos un conxunto de catro ventás equiespaciadas entre estes dous valores seleccionados.
5. Finalmente, repetiremos os pasos 2-4 dez veces, sendo o parámetro de ventá óptimo o resultante na última etapa.

Esta será a forma de obter a ventá h para os métodos de validación cruzada modificado e ambos bootstraps estacionarios suavizados (aproximando o $MISE^*(h)$ por Monte Carlo ou empregando a expresión exacta para o mesmo). En cambio, o método plug-in obtén de xeito directo a ventá h, como xa explicamos na Sección 4.1. Consideraremos ademais 50 mostras aleatorias de tamaño $n = 100$ cada unha. No caso do uso de Monte Carlo, empregaremos $B = 100$ simulacións. Sería recomendable que o número de mostras e o número de réplicas bootstrap foran máis grandes, por exemplo 1000 mostras e 1000 réplicas bootstrap. Mais debido ó elevado tempo de computación necesario mantivéronse estes valores. De todos xeitos, mostraremos ao final deste capítulo os resultados das simulacións que describiremos ao longo do mesmo, considerando 500 mostras de tamaño $n = 100$ no caso concreto do modelo (5.3) con $\phi = 0{,}9$, co fin ver os cambios que se producen con respecto ós resultados con só 50 mostras.

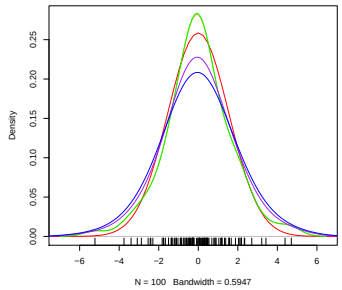
Finalmente, debemos determinar o parámetro que regula a dependencia no caso da validación cruzada modificada, l, e no caso do bootstrap estacionario suavizado, p. Por unha banda, tomaremos un valor $p = 0{,}05$. Por outro lado, l tomará dous valores dependendo do caso: $l = 5$ se a correlación é alta, é dicir, nos modelos (5.1),(5.2),(5.6); para valores $\phi = \pm 0{,}9, \pm 0{,}6$ dos modelos (5.3), (5.5); e por último, para valores $\phi = 0{,}6, 0{,}9$ do modelo (5.4); mentres que nos casos restantes empregaremos un valor $l = 2$. Cómpre dicir que se probaron outros valores para p e l, mais estes resultaron os máis

axeitados (dentro dos que se probaron).

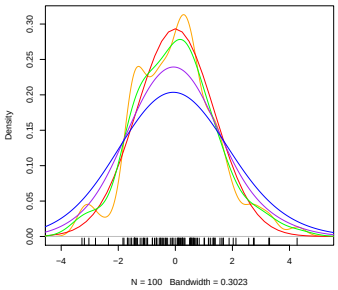
Posteriormente, mostramos o gráfico das estimacións das densidades de cada modelo con cada un dos selectores de parámetro ventá a partir dunha mostra concreta de tamaño $n = 100$ (Figura 5.1). O significado das cores que empregamos nesta figura é a seguinte:

- A liña azul representa a estimación da densidade coa ventá obtida polo método plug-in.
- A liña violeta, a estimación da mesma coa ventá seleccionada polo método de validación cruzada modificado.
- A liña verde, a estimación da densidade tras usar o método bootstrap estacionario suavizado empregando a aproximación por Monte Carlo do $MISE^*(h)$.
- A liña laranxa, a estimación da densidade cun parámetro ventá seleccionado polo método bootstrap suavizado estacionario empregando a expresión exacta para o $MISE^*(h)$.
- A liña vermella, a densidade marxinal de X_t.

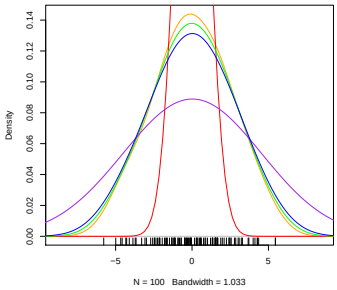
Deste xeito, a partir da Figura 5.1, podemos facer unha primeira análise e comparación do comportamento de cada un dos selectores de ventá explicados no Capítulo 4.



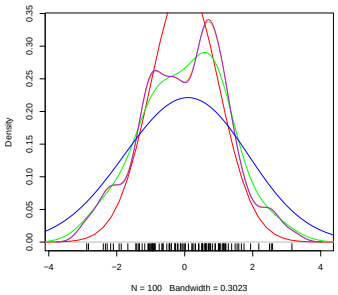
(a) Modelo (5.1)



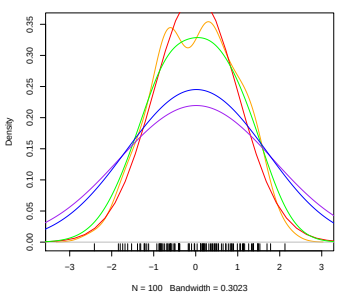
(b) Modelo (5.2)



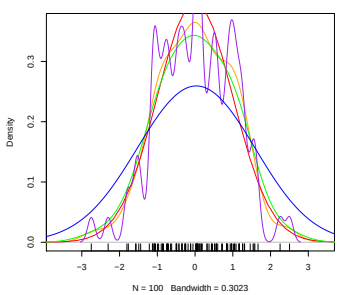
(c) Modelo (5.3), $\phi = -0{,}9$



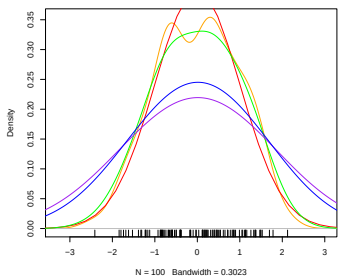
(d) Modelo (5.3), $\phi = -0{,}6$



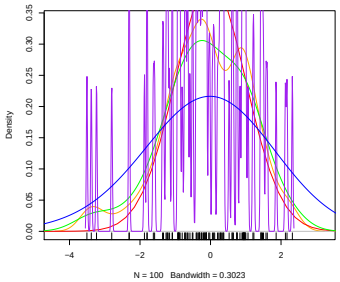
(e) Modelo (5.3), $\phi = -0{,}3$



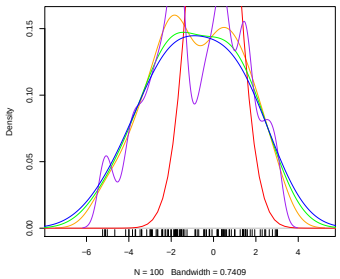
(f) Modelo (5.3), $\phi = 0$



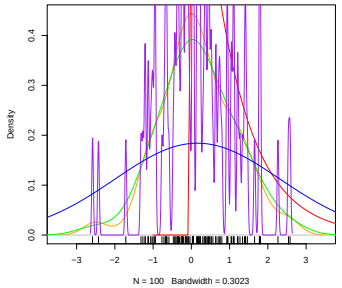
(g) Modelo (5.3), $\phi = 0{,}3$



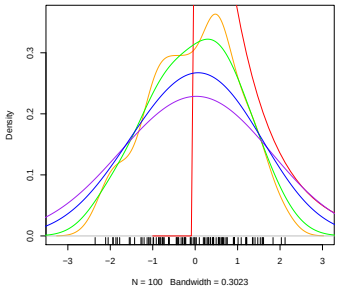
(h) Modelo (5.3), $\phi = 0{,}6$



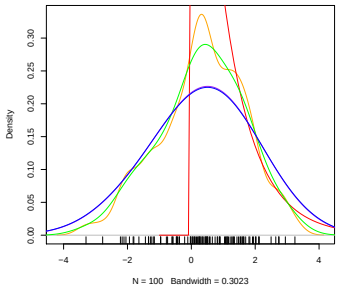
(i) Modelo (5.3), $\phi = 0{,}9$



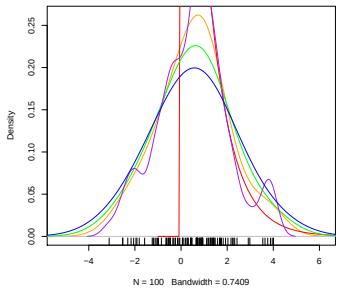
(l) Modelo (5.4), $\phi = 0$



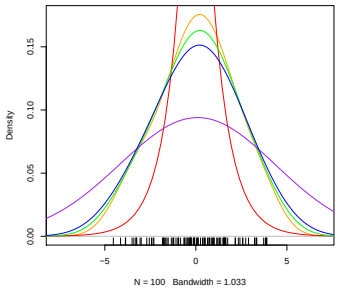
(m) Modelo (5.4), $\phi = 0{,}3$



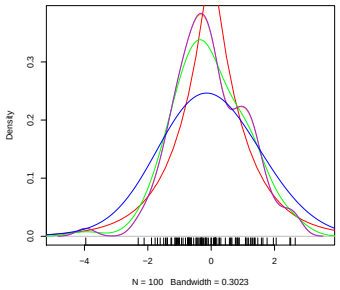
(n) Modelo (5.4), $\phi = 0{,}6$



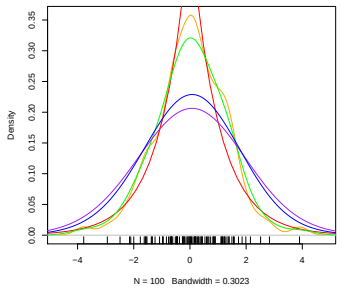
(ñ) Modelo (5.4), $\phi = 0{,}9$



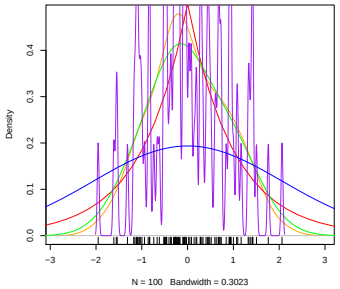
(o) Modelo (5.5), $\phi = -0{,}9$



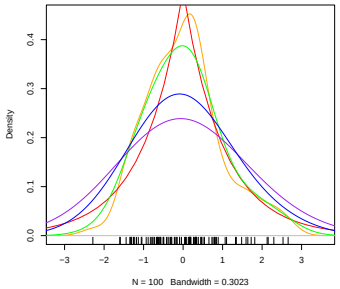
(p) Modelo (5.5), $\phi = -0{,}6$



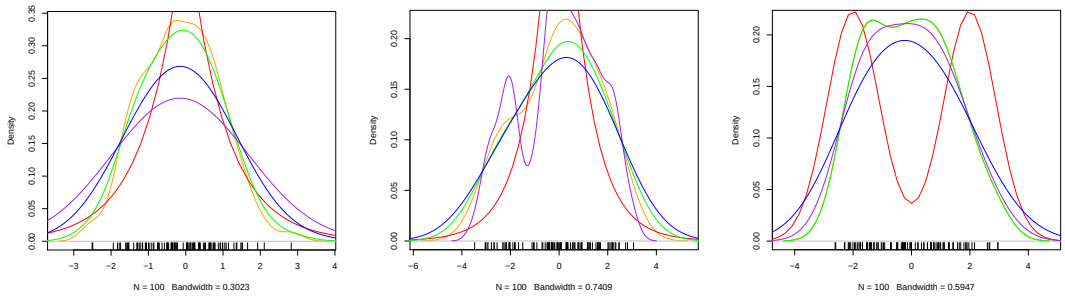
(q) Modelo (5.5), $\phi = -0{,}3$



(r) Modelo (5.5), $\phi = 0$



(s) Modelo (5.5), $\phi = 0{,}3$



(t) Modelo (5.5), $\phi = 0{,}6$ (u) Modelo (5.5), $\phi = 0{,}9$ (v) Modelo (5.6)

Figura 5.1: Estimación da función de densidade para os modelos presentados neste capítulo empregando as ventás obtidas para unha mostra concreta de tamaño $n = 100$.

En primeiro lugar, temos que h_{BOOT} e h_{MC}^{BOOT} proporcionan unha boa estimación da función de densidade en calquera dos casos expostos. Sen embargo, cabe destacar que, no caso do parámetro h_{MC}^{BOOT}, precisaríamos un maior número de réplicas B para diminuir o erro de aproximación de Monte Carlo, o que requeriría máis tempo de computación para a obtención dun resultado. Por outra banda, é apreciable que aínda que en moitos casos h_{PI} e h_{MCV} proporcionan tamén unha boa estimación da mesma; h_{PI} tende á sobresuavización da curva, é dicir, o método plug-in tende a proporcionar valores para o parámetro ventá altos. Tal é o caso das Figuras 5.1l ou 5.1r. Pola contra, h_{MCV} tende a infrasuavizar a curva (obtendo por tanto estimacións rugosas), é dicir, o método de validación cruzada modificado proporciona valores para o parámetro ventá baixos. Tal é o caso das Figuras 5.1f, 5.1h, 5.1l ou 5.1r.

A continuación presentaremos o Cadro 5.1, onde se recollen os promedios dos parámetros ventá obtidos empiricamente para cada un dos métodos vistos no Capítulo 4, empregando como xa dixemos, 50 mostras de tamaño $n = 100$. De seguido, recollemos os tempos de computación requeridos para cada cálculo no Cadro 5.2. Finalmente, expoñemos os resultados das expresións (5.16) e (5.17) aproximadas por simulación nas Figuras 5.2 e 5.3, respectivamente. Presentarémolo en forma de diagrama de caixas ou boxplot, xa que con este diagrama é sinxelo visualizar a boa escolla dos parámetros de ventá $h_{PI}, h_{MCV}, h_{BOOT}$, así como proceder á súa comparación: bastará con ver que a mediana dos datos é próxima a cero; e o primeiro e terceiro cuantiles non se alonxan moito do mesmo.

Novamente, atendendo ó Cadro 5.1, vemos que o selector de ventá que ofrece mellores estimacións da densidade é o selector h_{BOOT}, xa que na maioría dos casos, este é o máis próximo ó parámetro h_{MISE}, en promedio. Por outra banda, é claro que precisaríamos máis réplicas B para obter unha mellor aproximación por Monte Carlo do parámetro h_{MC}^{BOOT}. Ademais, cabe destacar que o selector que ofrece valores máis altos do parámetro ventá é o obtido polo método plug-in, o que conlevará a estimacións da densidade sobresuavizadas. Finalmente, en canto ós selectores de ventá obtidos polo método de validación cruzada modificada, estes tamén se achegan, na maioría dos casos, ó parámetro h_{MISE}. Sen embargo, para poder facer unha análise rigurosa destes parámetros de ventá, debemos ter en conta a variabilidade dos resultados obtidos ó simular as 50 mostras de tamaño $n = 100$.

		h_{PI}	h_{MCV}	h_{MC}^{BOOT}	h_{BOOT}	h_{MISE}
Modelo (5.1)		1,26062	0,56969	0,68727	0,541636	0,66134
Modelo (5.2)		1,74056	0,93806	0,61028	0,4745	0,58126
Modelo (5.3)	$\phi = -0{,}9$	1,79047	1,1123	1,05367	0,7351	0,68466
	$\phi = -0{,}6$	1,28125	0,63492	0,56448	0,44202	0,36601
	$\phi = -0{,}3$	1,11269	0,82036	0,4836	0,36818	0,39438
	$\phi = 0$	1,16547	0,65645	0,45631	0,34133	0,44332
	$\phi = 0{,}3$	1,21478	0,37137	0,47386	0,34349	0,43033
	$\phi = 0{,}6$	1,419868	1,36248	0,54012	0,41517	0,41149
	$\phi = 0{,}9$	1,30183	1,66049	0,90774	0,68461	0,81211
Modelo (5.4)	$\phi = 0$	1,94036	0,61007	0,46411	0,32919	0,38285
	$\phi = 0{,}3$	1,17049	0,77094	0,48749	0,35583	0,40088
	$\phi = 0{,}6$	1,16458	0,53756	0,56643	0,41257	0,46216
	$\phi = 0{,}9$	1,38879	1,0191	0,9572	0,68743	0,56535
Modelo (5.5)	$\phi = -0{,}9$	1,71921	0,90486	1,06939	0,78667	0,78549
	$\phi = -0{,}6$	1,17844	1,31982	0,57715	0,43271	0,41723
	$\phi = -0{,}3$	1,13417	0,38504	0,47386	0,36536	0,4274
	$\phi = 0$	1,88356	0,40866	0,47386	0,34998	0,46573
	$\phi = 0{,}3$	1,20389	0,65209	0,47775	0,35432	0,44884
	$\phi = 0{,}6$	1,02351	1,27743	0,53915	0,42123	0,48414
	$\phi = 0{,}9$	1,36983	1,18068	0,94551	0,66117	0,79365
Modelo (5.6)		1,31741	1,16822	0,68922	0,53059	2,26408

Cadro 5.1: Promedios dos parámetros ventá obtidos por simulación para os modelos presentados anteriormente, empregando os métodos descritos no Capítulo 4.

No Cadro 5.2 exporemos, como xa dixemos, o tempo (en segundos) que precisa a CPU para obter cada un dos resultados. Podemos destacar que o método de selección do parámetro de suavizado que é máis custoso (en termos de tempo) é, con diferenza, o método de validación cruzada modificado.

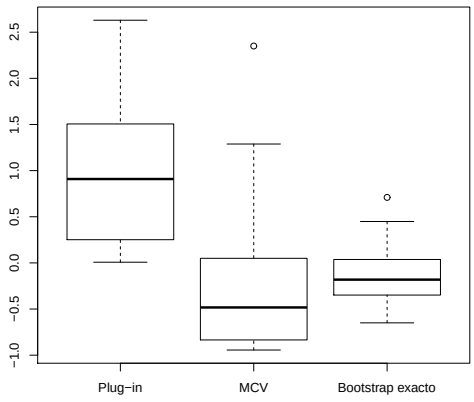
Por outra banda, o método plug-in e o método de selección bootstrap empregando a expresión exacta para o $MISE^*(h)$ son similares en tempo de execución. Sen embargo, na maioría dos casos, é o método de selección bootstrap empregando a expresión exacta para o $MISE^*(h)$ o que precisa menor tempo de execución para acadar o resultado. Finalmente, vemos que o método de selección bootstrap empregando a aproximación por Monte Carlo é tamén bastante custoso. Ademais, como xa dixemos, sería convinte empregar máis réplicas B para acadar menor erro de aproximación Monte Carlo, o que derivaría nun incremento de CPU_{MC}^{BOOT}.

		CPU_{PI}	CPU_{MCV}	CPU_{MC}^{BOOT}	CPU_{BOOT}
Modelo (5.1)		146,18	1390,92	423,83	199,99
Modelo (5.2)		204,18	1762,67	352,186	196,87
	$\phi = -0{,}9$	234,06	1816	437,1216	211,78
	$\phi = -0{,}6$	252,98	1282,01	520,75	222,71
	$\phi = -0{,}3$	263,51	1334,5	367,98	198,97
Modelo (5.3)	$\phi = 0$	286,96	1928,31	308,41	194,99
	$\phi = 0{,}3$	288,69	1488,5	392,09	238,66
	$\phi = 0{,}6$	293,13	1475,9	398,08	230,53
	$\phi = 0{,}9$	309,59	1558,4	413,29	287,8
	$\phi = 0$	322,32	1605,75	424,48	298,98
Modelo (5.4)	$\phi = 0{,}3$	331,81	1666,6	435,43	301,01
	$\phi = 0{,}6$	352,68	1771,5	456,24	296,25
	$\phi = 0{,}9$	363,43	1723,03	466,86	301,93
	$\phi = -0{,}9$	373,43	1918,15	499,15	305,76
	$\phi = -0{,}6$	400,23	2012,04	504,94	344,59
	$\phi = -0{,}3$	410,6	2085	518,01	370,35
Modelo (5.5)	$\phi = 0$	528,77	2615,45	617,4	458,39
	$\phi = 0{,}3$	524,44	2550,45	611,19	444,28
	$\phi = 0{,}6$	525,36	2590,2	613,33	440,23
	$\phi = 0{,}9$	527,74	2606,45	614,95	436,19

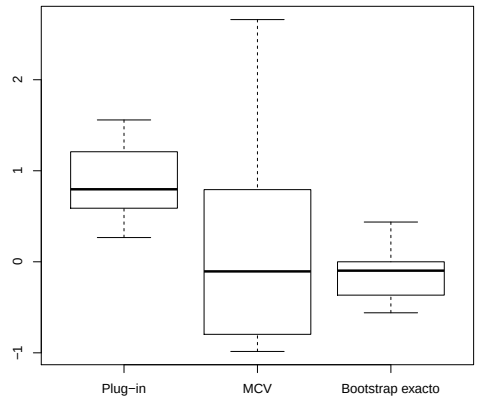
	CPU_{PI}	CPU_{MCV}	CPU_{MC}^{BOOT}	CPU_{BOOT}
Modelo (5.6)	580,02	2912,8	683,58	490,43

Cadro 5.2: Tempos de computación (en segundos) requeridos para obter os resultados do Cadro 5.1.

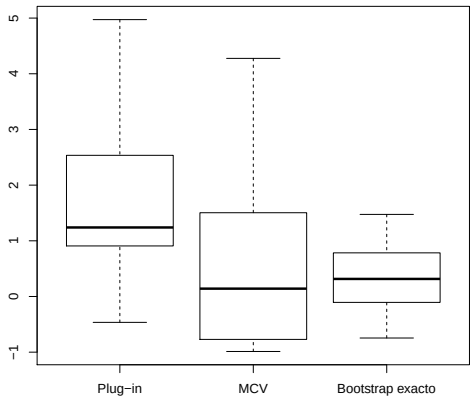
Seguidamente, presentamos os resultados das simulacións feitas a partir das expresións (5.16) e (5.17) por medio do boxplot. Nelas poderemos apreciar a variabilidade dos mesmos, ademais das medianas correspondentes. Posteriormente, compararemos os distintos métodos presentados no Capítulo 4 para a obtención do parámetro ventá. Comezaremos coa simulación a partir da expresión (5.16), reflexada na Figura 5.2.



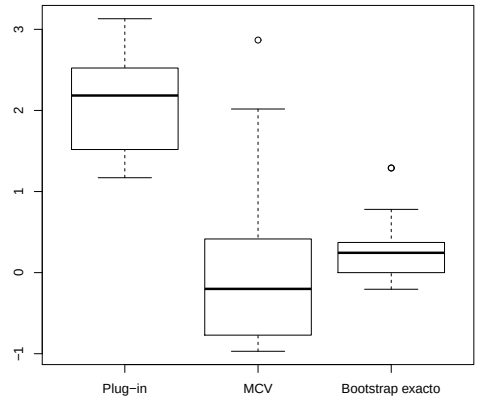
(a) Modelo (5.1)



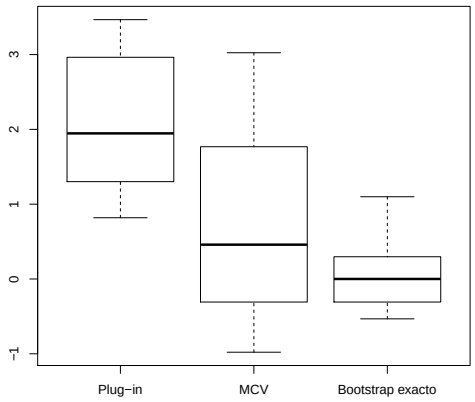
(b) Modelo (5.2)



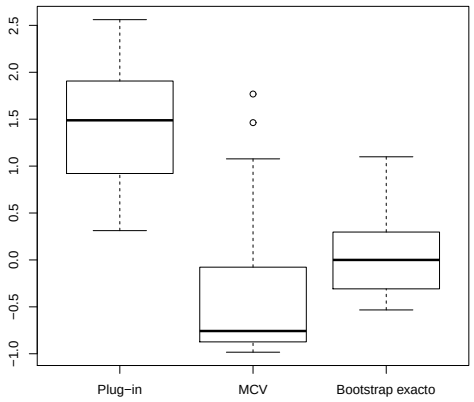
(c) Modelo (5.3), $\phi = -0{,}9$



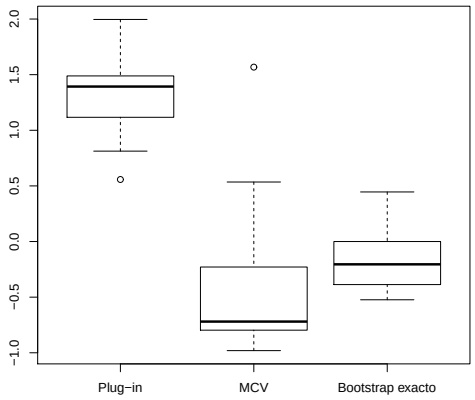
(d) Modelo (5.3), $\phi = -0{,}6$



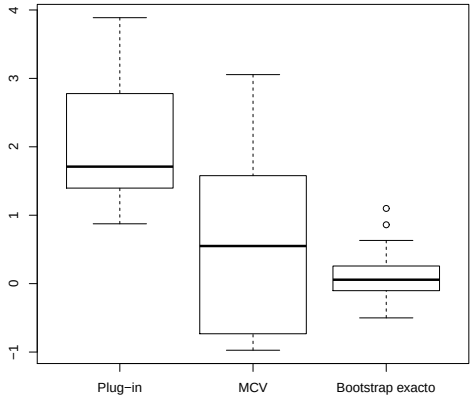
(e) Modelo (5.3), $\phi = -0{,}3$



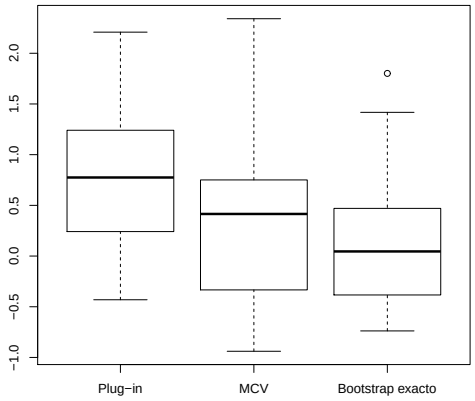
(f) Modelo (5.3), $\phi = 0$



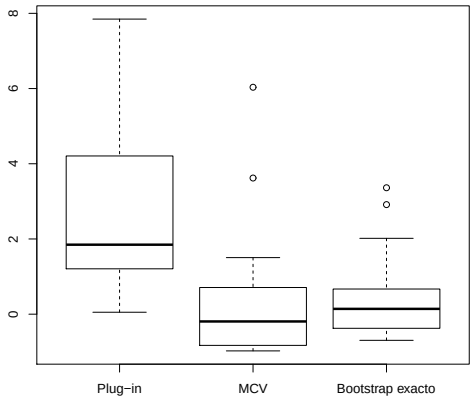
(g) Modelo (5.3), $\phi = 0{,}3$



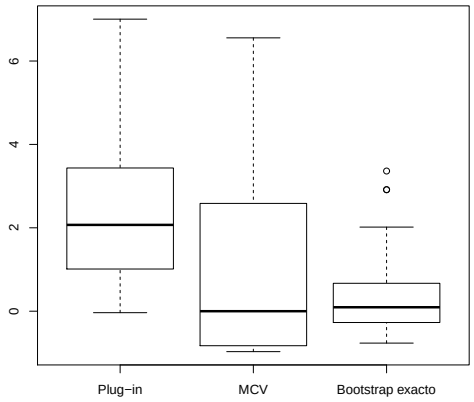
(h) Modelo (5.3), $\phi = 0{,}6$



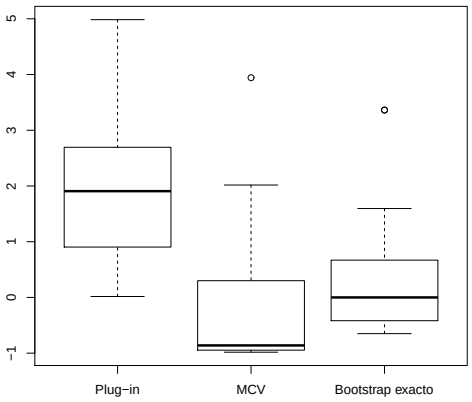
(i) Modelo (5.3), $\phi = 0{,}9$



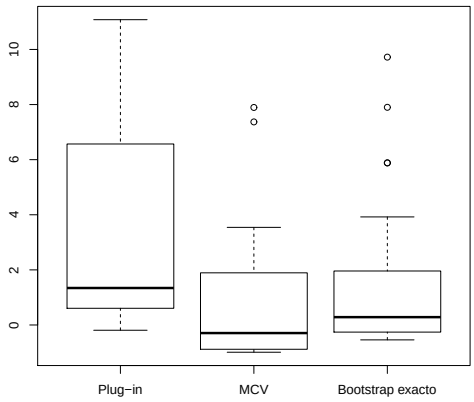
(l) Modelo (5.4), $\phi = 0$



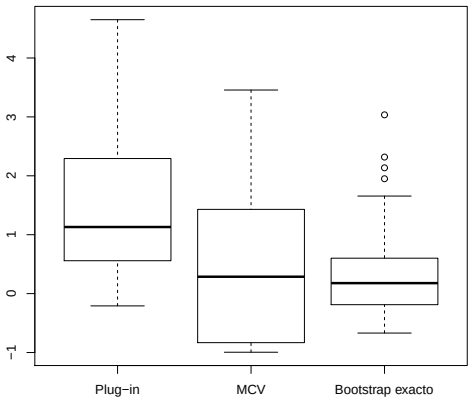
(m) Modelo (5.4) con $\phi = 0{,}3$



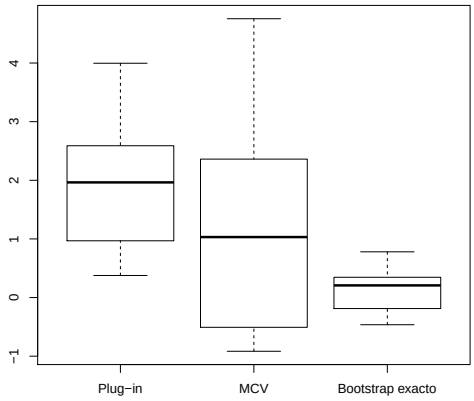
(n) Modelo (5.4) con $\phi = 0{,}6$



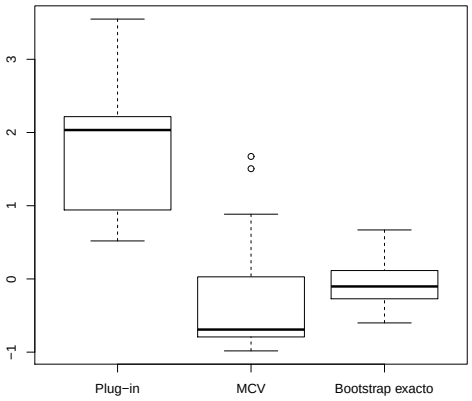
(ñ) Modelo (5.4), $\phi = 0{,}9$



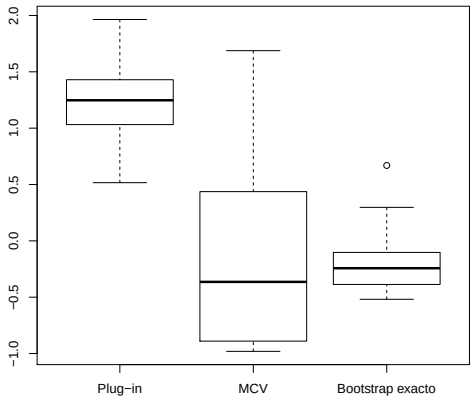
(o) Modelo (5.5), $\phi = -0{,}9$



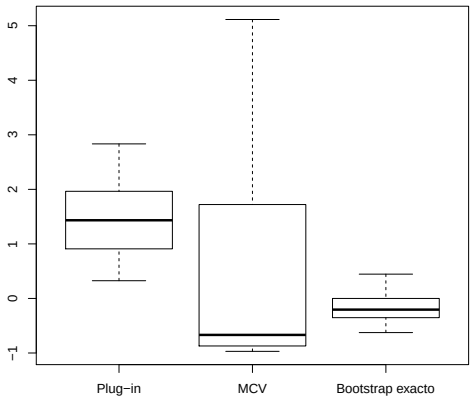
(p) Modelo (5.5), $\phi = -0{,}6$



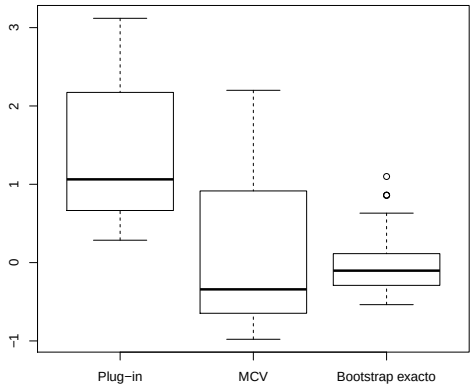
(q) Modelo (5.5), $\phi = -0{,}3$



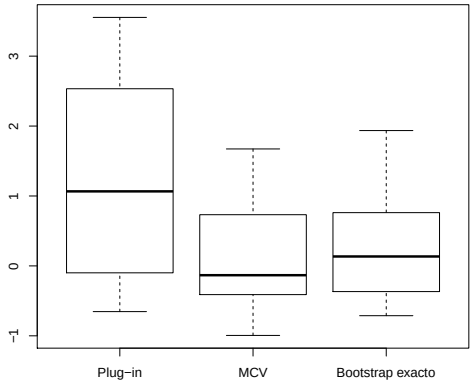
(r) Modelo (5.5), $\phi = 0$



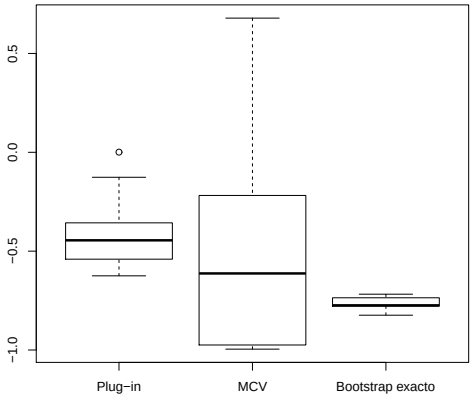
(s) Modelo (5.5), $\phi = 0{,}3$



(t) Modelo (5.5), $\phi = 0{,}6$



(u) Modelo (5.5), $\phi = 0{,}9$

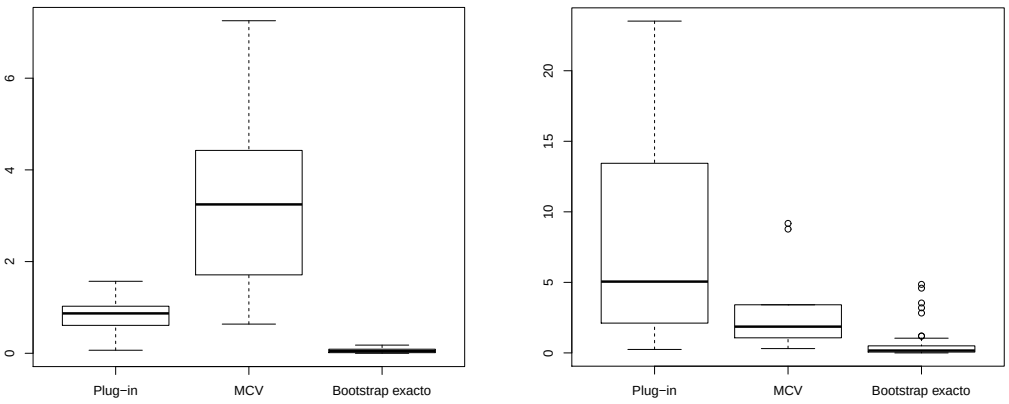


(v) Modelo (5.6)

Figura 5.2: Diagramas de caixas e bigotes para a expresión (5.16) aproximada por simulación.

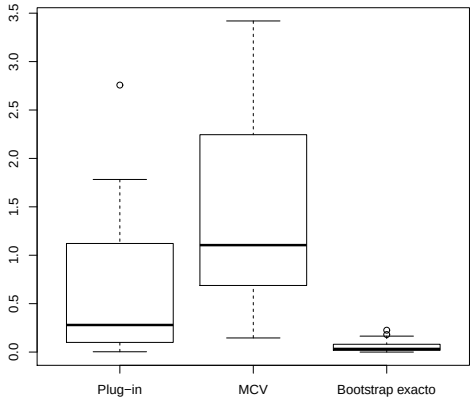
Como podemos ver a partir da Figura 5.2, as medianas dos diagramas de caixas obtidos cos resultados da expresión (5.16) que pon en relación h_{BOOT} con h_{MISE}, na maioría dos casos son próximas a cero. Cabe destacar ademais, que en ningún caso o primeiro e terceiro cuantil están notablemente alonxados deste valor, sendo este selector o que ten menor variabilidade. Deducimos entón, empiricamente, que o selector de ventá h_{BOOT} ten un bo comportamento. Isto é así xa que ó ser as medianas obtidas na Figura 5.2 próximas a cero, temos que h_{BOOT} será próximo a h_{MISE}, que como xa dixemos, é obtido minimizando o $MISE(h)$. Por outra banda, en xeral, o selector de ventá que peores resultados obtén atendendo á mediana dos mesmos, é o obtido co método plug-in. Estas atópanse en torno ó valor 1 ou 2 na maioría dos casos, o que nos indica que o selector h_{PI} é maior que o selector h_{MISE}, o que conlevaría a unha estimación da densidade sobresuavizada, tal e como víamos que ocorría na Figura 5.1 para unha mostra concreta de tamaño $n = 100$. En canto ó selector obtido polo método de validación cruzada modificada, aínda que as medianas dos resultados de simulación son próximos a 0, podemos ver que é o que maior variabilidade presenta. Finalmente, destacaremos que o caso no que peores resultados se obteñen é co modelo (5.6). En primeiro lugar, aínda que o selector de ventá h_{BOOT} apenas presenta variabilidade, a mediana é negativa, o que nos indica que o parámetro ventá h_{BOOT} é menor que o parámetro h_{MISE}, como tamén ocorría cos promedios deste modelo obtidos no Cadro 5.1. Analogamente, o mesmo ocorre cos outros dous selectores de ventá h_{PI} e h_{MCV}, sendo este último o que maior variabilidade presenta, novamente.

Posteriormente, presentamos na Figura 5.3 os boxplots relativos ós resultados da simulación feita a partir da expresión (5.17), nos que poñemos en relación os criterios de erro $MISE(h_{PI})$, $MISE(h_{MCV})$ e $MISE(h_{BOOT})$ co erro producido ó estimar a densidade co parámetro h_{MISE}, $MISE(h_{MISE})$. Vemos a partir desta figura, que os valores da expresión (5.17) relativas ó selector h_{BOOT} son as menores. Logo, deducimos un bo comportamento na práctica do selector de ventá h_{BOOT}, ó ser o erro que comete este ó estimar a función de densidade próximo ó erro que comete o selector h_{MISE}, obtido tras minimizar $MISE(h)$. Temos, ademais, que tanto o erro cometido polo selector h_{PI} como polo selector h_{MCV} son superiores ao erro cometido por h_{BOOT} na maioría dos casos, presentando ambos moita variabilidade (tales son os casos das Figuras 5.3e, 5.3g ou 5.3o).

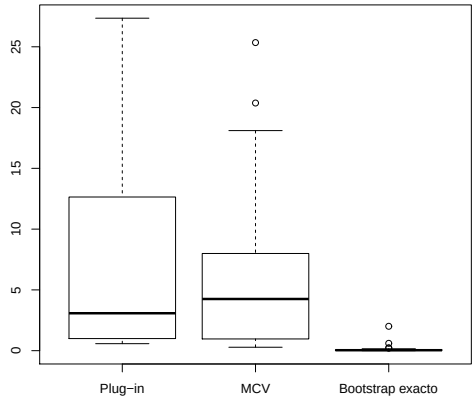


(a) Modelo (5.1)

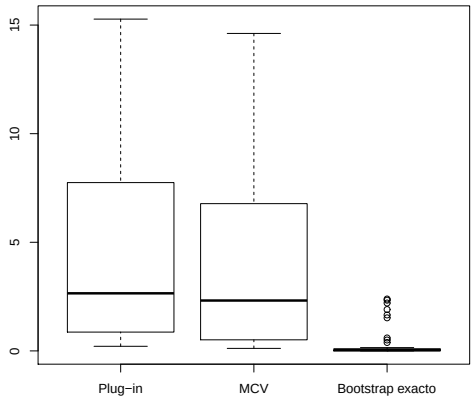
(b) Modelo (5.2)



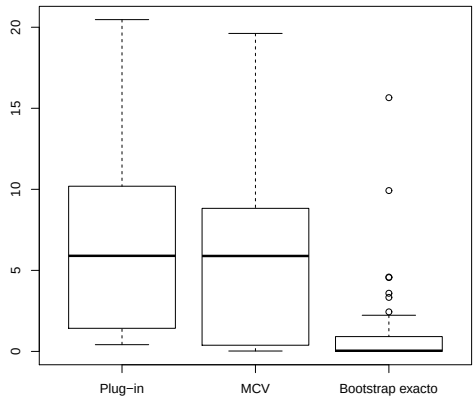
(c) Modelo (5.3), $\phi = -0{,}9$



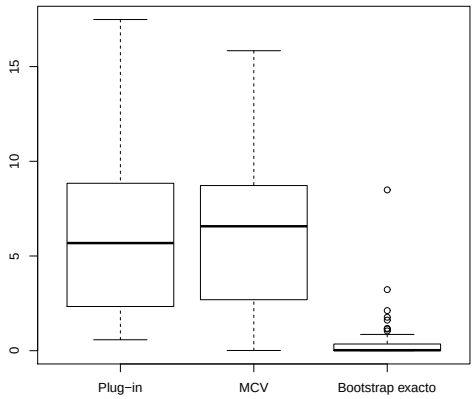
(d) Modelo (5.3), $\phi = -0{,}6$



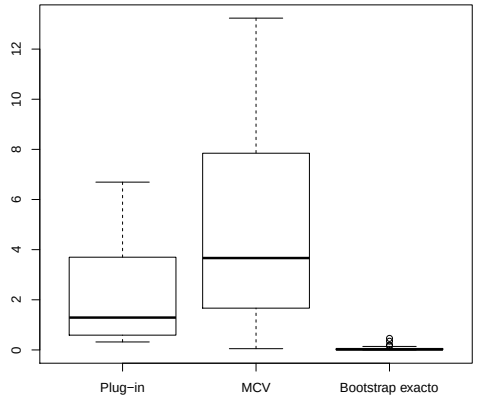
(e) Modelo (5.3), $\phi = -0{,}3$



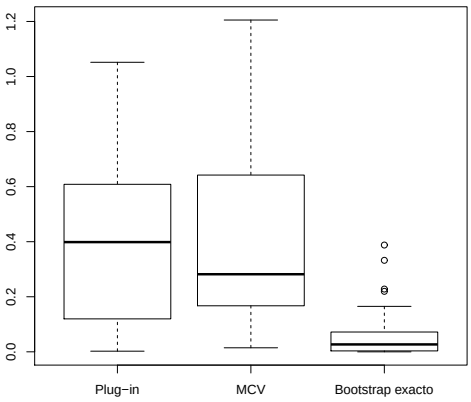
(f) Modelo (5.3), $\phi = 0$



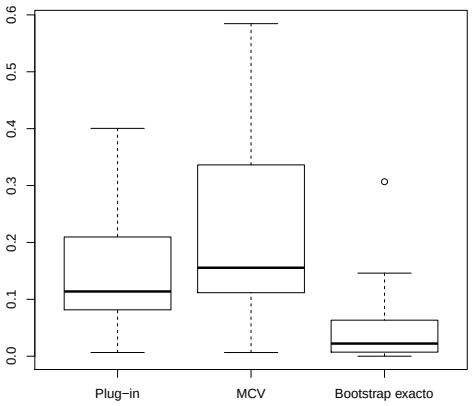
(g) Modelo (5.3), $\phi = 0{,}3$



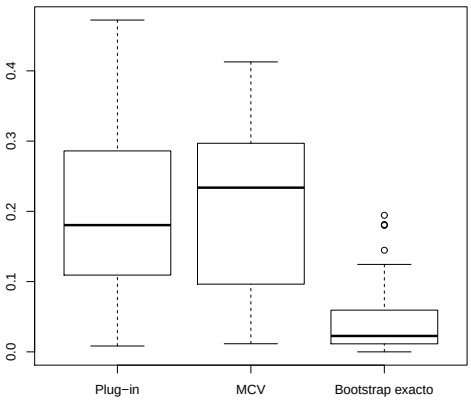
(h) Modelo (5.3), $\phi = 0{,}6$



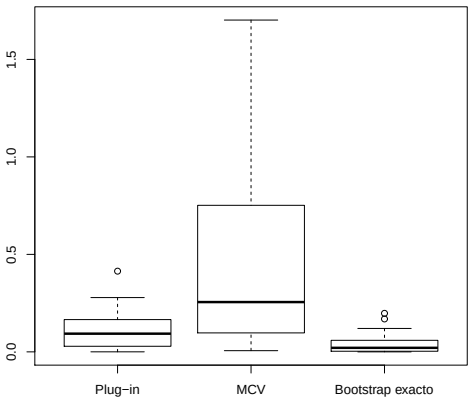
(i) Modelo (5.3), $\phi = 0{,}9$



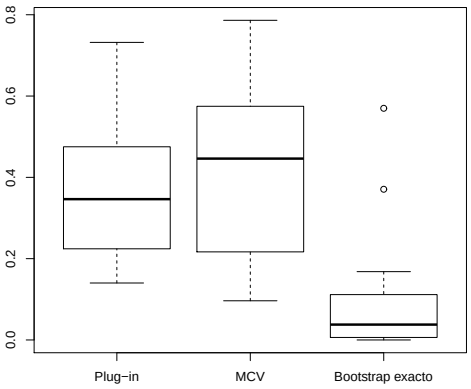
(l) Modelo (5.4), $\phi = 0$



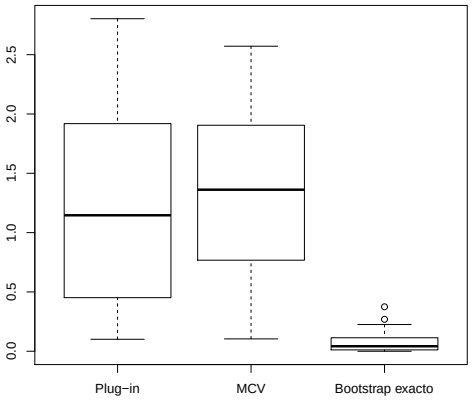
(m) Modelo (5.4), $\phi = 0{,}3$



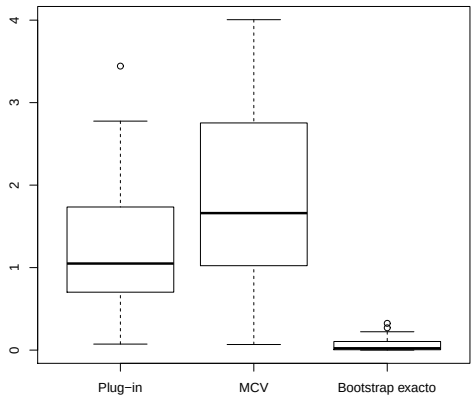
(n) Modelo (5.4), $\phi = 0{,}6$



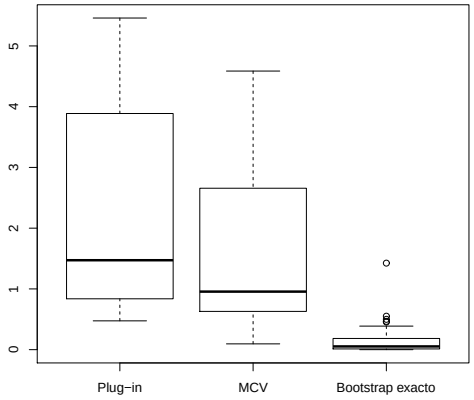
(ñ) Modelo (5.4), $\phi = 0{,}9$



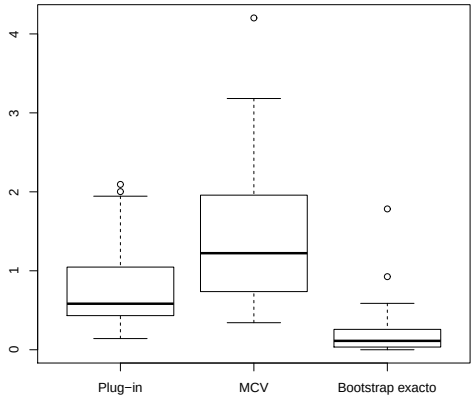
(o) Modelo (5.5), $\phi = -0{,}9$



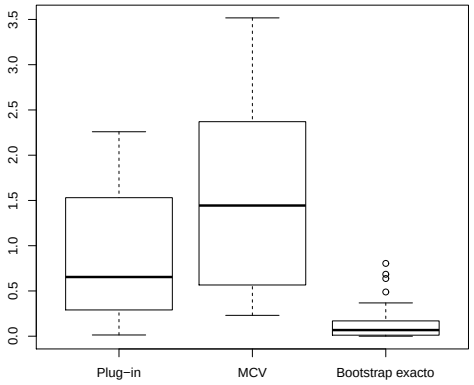
(p) Modelo (5.5), $\phi = -0{,}6$



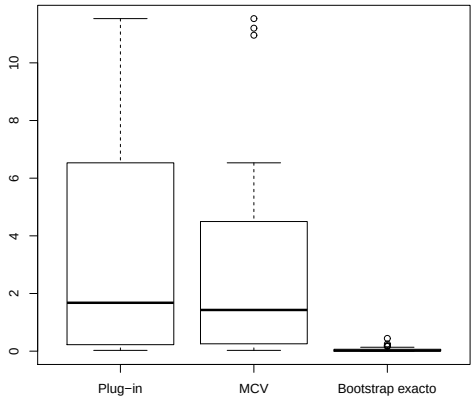
(q) Modelo (5.5), $\phi = -0{,}3$



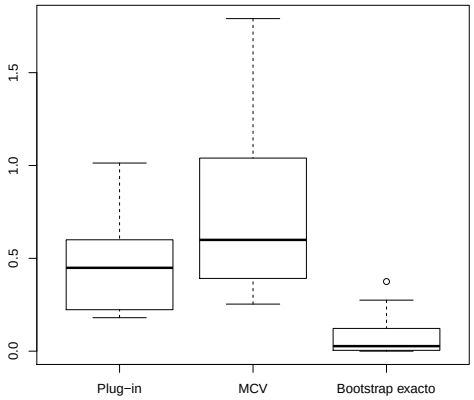
(r) Modelo (5.5), $\phi = 0$



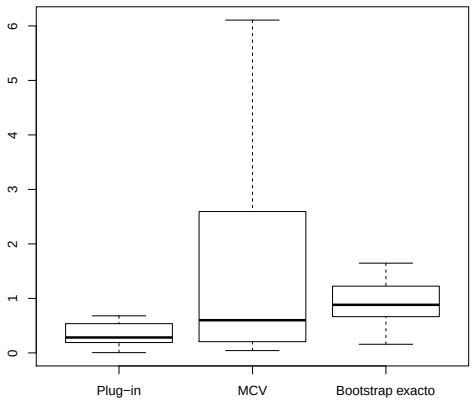
(s) Modelo (5.5), $\phi = 0{,}3$



(t) Modelo (5.5), $\phi = 0{,}6$



(u) Modelo (5.5), $\phi = 0{,}9$



(v) Modelo (5.6)

Figura 5.3: Diagramas de caixas e bigotes para a expresión (5.17) aproximada por simulación.

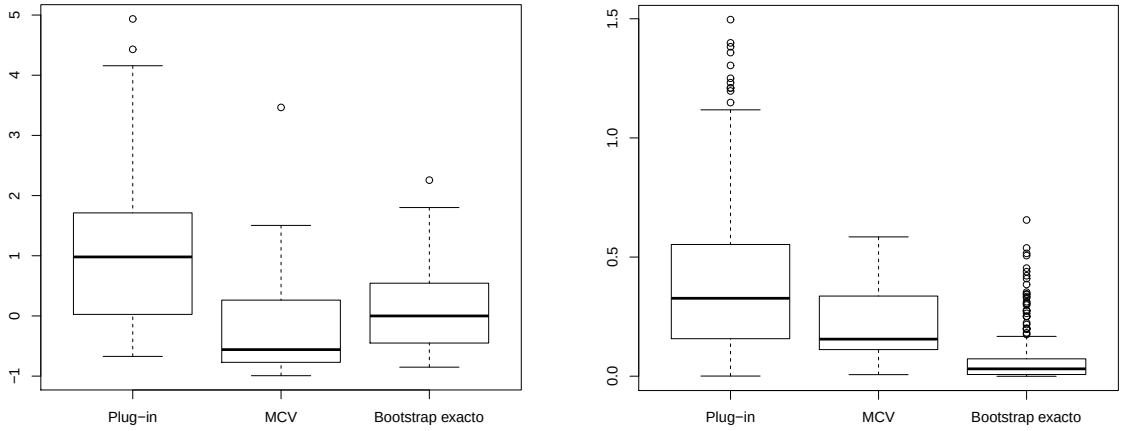
Rematamos este estudo de simulación presentando no Cadro 5.3 os promedios dos parámetros ventá h_{PI}, h_{MCV}, h_{BOOT} e h_{MISE} no caso concreto do modelo (5.3) considerando $\phi = 0{,}9$. Mais desta vez, tomamos 500 mostras de tamaño $n = 100$, para así ver os cambios que se producen con respecto ós resultados con só 50 mostras. Como xa dixemos, sería recomendable que o número de mostras empregadas fosen, para todos os modelos que analizamos, maior que 50 (por exemplo, 500, como neste caso). Non obstante, debido ó elevado tempo de execución non se engadiron os resultados de todos os modelos con 500 mostras. De seguido, reflexamos os tempos de computación requeridos para obter estes resultados (Cadro 5.4), e por último, mostramos as expresións (5.16) e (5.17) aproximadas por simulación para este caso concreto.

h_{PI}	h_{MCV}	h_{BOOT}	h_{MISE}
2,72222	0,628907	0,67586	0,72298

Cadro 5.3: Promedios dos parámetros ventá obtidos por simulación para o modelo (5.3) con $\phi = 0{,}9$, empregando os métodos descritos no Capítulo 4 e considerando 500 mostras de tamaño $n = 100$.

CPU_{PI}	CPU_{MCV}	CPU_{BOOT}
3096,2	27989,89	3119,79

Cadro 5.4: Tempos de computación (en segundos) requeridos para obter os resultados do Cadro 5.3.



(a) Expresión (5.16) aproximada por simulación.

(b) Expresión (5.17) aproximada por simulación.

Figura 5.4: Boxplots para as expresións (5.16) e (5.17) aproximadas por simulación considerando o modelo (5.3) con $\phi = 0{,}9$ e 500 mostras de tamaño $n = 100$.

Se comparamos os valores obtidos no Cadro 5.3 cos resultados para o modelo (5.3) con $\phi = 0{,}9$ do Cadro 5.1, vemos que, considerando 500 mostras, os promedios dos selectores h_{MCV} e h_{BOOT} se achegan máis a h_{MISE}. Sen embargo, notemos que no caso do obtido por bootstrap, ambos valores, con 50 e con 500 mostras, son moi semellantes. Atendendo ó parámetro h_{PI}, este proporciona unha estimación da densidade moito máis suave no caso de empregar 500 mostras. Se agora temos en conta o Cadro 5.4, vemos novamente que, con moita diferenza, é o selector h_{MCV} o que precisa máis tempo de CPU para proporcionar un resultado. Esta diferenza vese agora máis acentuada debido ó aumento de mostras. Por outra banda, a Figura 5.4a suxire que o selector h_{BOOT} é o que máis se achega ó parámetro h_{MISE}, como xa podíamos ver na Figura 5.2i. Mais no caso do selector de ventá obtido por validación cruzada modificada, vemos que a súa mediana é negativa, o que indica que, en moitas das 500 mostras, o selector h_{MCV} proporciona unha estimación infrasuavizada da función de densidade. Pola contra, como xa víamos na Figura 5.2i, o parámetro h_{PI} proporciona estimacións sobresuavizadas. Finalmente, de xeito análogo ó que ocorría no caso de empregar só 50 mostras (Figura 5.3l), observamos na Figura 5.4b que o selector que menor erro comete na estimación da densidade é h_{BOOT}, seguido por h_{MCV}. Nembargantes, cabe destacar que a variabilidade no caso do parámetro h_{MCV} vese reducida.

Concluímos entón con este estudo de simulación que o selector de ventá h_{BOOT} é co que se obteñen mellores resultados de bo comportamento a nivel práctico, xa que é, na maioría dos casos expostos, o que proporciona estimadores da función da densidade máis próximos á curva real. Ademais, é o que presenta un menor tempo de CPU.

Capítulo 6

Aplicación a datos reais

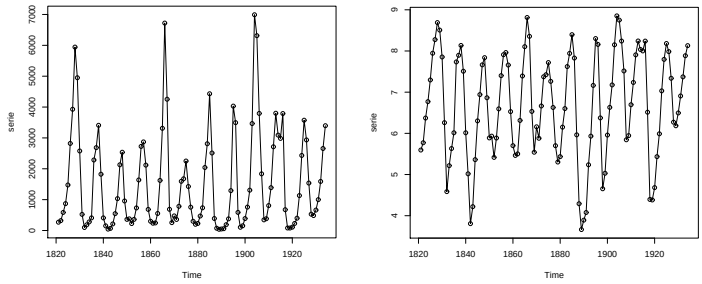
Neste capítulo aplicarase a suavización estacionaria do bootstrap de Politis e Romano (1994a) e a elección do parámetro de suavizado h empregando a expresión exacta para o $MISE^*(h)$ proposta no Capítulo 4 (h_{BOOT}), sobre dous conxuntos de datos reais. Ademais, compararemos os resultados obtidos con este selector co de validación cruzada modificado (h_{MCV}) de Cox e Kim (1997); e co plug-in (h_{PI}) proposto por Hall *et al.* (1995).

A continuación, podemos ver que o capítulo está dividido en dúas seccións. Na primeira delas, presentaremos os conxuntos de datos reais aos que lles aplicaremos, na segunda sección, os selectores do parámetro de suavizado estudados no Capítulo 4.

6.1. Presentación dos conxuntos de datos

En primeiro lugar, empregaremos o conxunto de datos `lynx`, recollidos na librería `datasets` do software estatístico `R`. Neste conxunto de datos recóllense o número de linces anuais cazados entre 1821 e 1934 en Canadá. É sinxelo ver que os datos da mostra dependen claramente do tempo, xa que o número de linces cazados variará segundo a evolución do tempo. Tentaremos modelizar esta serie de tempo de xeito paramétrico, para logo proceder dun xeito non paramétrico. Así, propoñeremos un modelo paramétrico para dita serie, que está constituida de 114 observacións.

A continuación, representaremos o gráfico secuencial da serie de tempo relativa ós datos `lynx` e da serie transformada pola función logarítmica, log(`lynx`).



(a) Serie `lynx`. (b) Serie log(`lynx`).

Figura 6.1: Gráficos secuenciais das series `lynx` e log(`lynx`).

A partir da Figura 6.1a podemos ver un gráfico que presenta estacionalidade e variabilidade. Polo tanto, cómpre aplicar a función logaritmo para ver se esta variabilidade se estabiliza, feito que podemos comprobar na Figura 6.1b. Logo, a partir de agora, traballaremos coa serie transformada log(`lynx`).

Unha vez que diferenciamos a serie log(`lynx`) regularmente para eliminar a tendencia, vemos a partir da función de autocorrelación simple (ou fas) na mostraxe (Figura 6.2) que esta serie non é estacionaria, xa que presenta estacionalidade de período 12 (é dicir, a compoñente estacional aparece cada 12 retardos). Ademais, existe unha forte correlación positiva nos retardos estacionais, que converxen lentamente a 0.

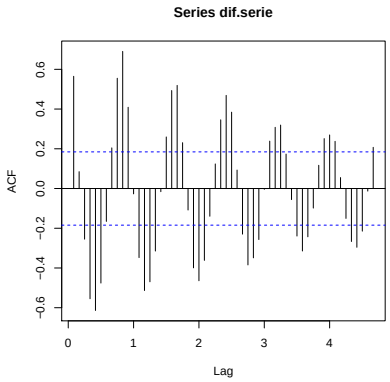


Figura 6.2: fas na mostraxe da serie log(`lynx`) cunha diferencia regular.

Necesitamos, por tanto, diferenciar de novo, esta vez estacionalmente, con período 12. Se representamos o gráfico secuencial e a fas na mostraxe da serie transformada diferenciada dúas veces, podemos ver que xa é estacionaria. Por tanto, estamos no momento de representar as funcións de autocorrelación simple e parcial (fas e fap, respectivamente) na mostraxe da serie diferenciada dúas veces para poder propoñer un modelo paramétrico para a serie temporal.

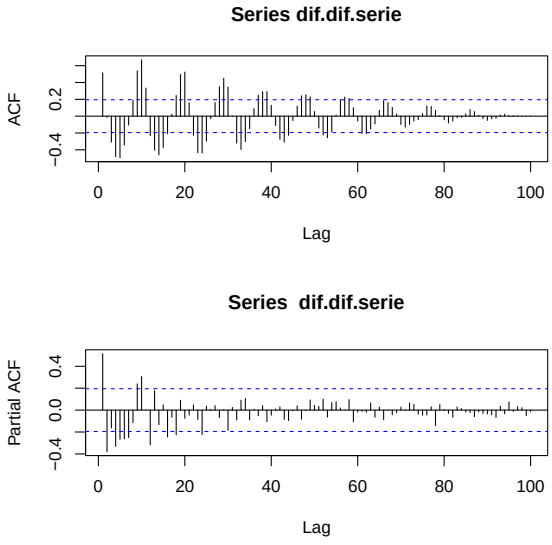


Figura 6.3: fas e fap na mostraxe da serie log(`lynx`) diferenciada dúas veces.

A Figura 6.3 suxire que a serie transformada foi xerada por un proceso ARIMA$(6,1,0)\times(1,1,0)_{12}$, xa que os retardos regulares da fap na mostraxe anúlanse a partir do sexto retardo. Por outra banda, na fap da parte estacional vemos que todos os retardos resultan non significativos despois do primeiro retardo estacional. Sen embargo, as innovacións $\{a_t\}_{t\in\mathbb{Z}}$ deste modelo non presentan incorrelación, polo que este modelo non resulta válido para xerar a serie de tempo ó non verificar a hipótese de que $\{a_t\}_{t\in\mathbb{Z}}$ sexa ruído branco.

Consecuentemente, comprobamos a través da función `auto.arima` do paquete `forecast` do software estatístico `R` se existe algunha outra tentativa para modelar a serie. En efecto, esta función propón un ARIMA$(2,0,3)\times(0,0,1)_{12}$, polo que será a que empregaremos para modelar a serie log(`lynx`). Este modelo pódese representar como:

$$(1-\phi_1B-\phi_2B^2)Y_t = c+(1+\theta_1B+\theta_2B^2+\theta_3B^3)(1+\Theta_1B^{12})a_t, \tag{6.1}$$

sendo $Y_t=\log(X_t)$; X_t, o número de linces cazados no ano t; c, a constante; B, o operador retardo definido como $BY_t=Y_{t-1}$; e B^s, o operador retardo estacional, que verifica $B^sa_t=a_{t-s}$. Outra forma de presentar este modelo será:

$$Y_t = c+\phi_1Y_{t-1}+\phi_2Y_{t-2}+a_t+\theta_1a_{t-1}+\theta_2a_{t-2}+\theta_3a_{t-3}+\Theta_1a_{t-12}+\theta_1\Theta_1a_{t-13}+\theta_2\Theta_1a_{t-14}+\theta_3\Theta_1a_{t-15}.$$

Comprobamos finalmente se o modelo (6.1) verifica a hipótese de que as innovacións a_t son ruído branco. Efectivamente, se realizamos un test de independencia ás innovacións coa función `Box.test`, obtemos un p-valor de 0,3704, polo que non existen evidencias significativas a ningún nivel de significación habitual para rexeitar a hipótese nula de independencia. Se realizamos agora un test para corroborar que as innovacións teñen media cero (`t.test`), obtemos un p-valor de 0,9748, polo que novamente non existen evidencias significativas a ningún nivel de significación habitual para rexeitar a hipótese nula de que a media sexa cero. Por outra banda, non existen evidencias significativas para rexeitar a hipótese nula de normalidade de a_t, xa que realizando o test `jarque.bera.test`, obtemos un p-valor de 0,7022. Ademais, atendendo á Figura 6.4, temos a partir da fas na mostraxe de a_t que as innovacións son incorreladas, xa que o único retardo que sae das bandas de confianza pode ser debido á aleatoriedade.

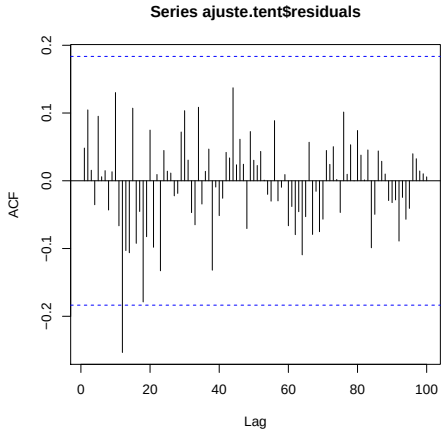


Figura 6.4: fas na mostraxe das innovacións para o modelo (6.1).

Polo tanto, o modelo proposto en (6.1) con constante non nula e innovacións gaussianas resulta axeitado como xerador da serie log(`lynx`).

Finalmente, a partir do modelo (6.1), tentaremos aproximar a distribución marxinal da serie log(`lynx`). Dado que as innovacións $\{a_t\}_{t\in\mathbb{Z}}$ son variables aleatorias *iid* que seguen unha distribución $N(0,\sigma_a^2)$, e son independentes de $Y_{t-1}, Y_{t-2}, ...$, entón a distribución marxinal de Y_t será normal, con esperanza $\mu_Y = \mathbb{E}(Y_t), \forall t \in \mathbb{Z}$:

$$\begin{aligned}
\mu_Y &= c + \phi_1\mathbb{E}(Y_{t-1}) + \phi_2\mathbb{E}(Y_{t-2}) + \mathbb{E}(a_t) + \theta_1\mathbb{E}(a_{t-1}) + \theta_2\mathbb{E}(a_{t-2}) + \theta_3\mathbb{E}(a_{t-3}) \\
&+ \Theta_1\mathbb{E}(a_{t-12}) + \theta_1\Theta_1\mathbb{E}(a_{t-13}) + \theta_2\Theta_1\mathbb{E}(a_{t-14}) + \theta_3\Theta_1\mathbb{E}(a_{t-15}) \\
&= c + \phi_1\mathbb{E}(Y_t) + \mathbb{E}(Y_t) \\
&= c + \phi_1\mu_Y + \phi_2\mu_Y. \qquad (6.2)
\end{aligned}$$

Logo, $\mu_Y = \dfrac{c}{1-\phi_1-\phi_2}$. Por outra banda, a varianza, $\sigma_Y^2 = \mathrm{Var}(Y_t), \forall t \in \mathbb{Z}$, virá dada por:

$$\begin{aligned}
\sigma_Y^2 &= \phi_1^2\mathrm{Var}(Y_{t-1}) + \phi_2^2\mathrm{Var}(Y_{t-2}) + 2\phi_1\phi_2\mathrm{Cov}(Y_{t-1}, Y_{t-2}) + \mathrm{Var}(a_t) + \theta_1^2\mathrm{Var}(a_{t-1}) \\
&+ \theta_2^2\mathrm{Var}(a_{t-2}) + \theta_3^2\mathrm{Var}(a_{t-3}) + \Theta_1^2\mathrm{Var}(a_{t-12}) + \theta_1^2\Theta_1^2\mathrm{Var}(a_{t-13}) \\
&+ \theta_2^2\Theta_1^2\mathrm{Var}(a_{t-14}) + \theta_3^2\Theta_1^2\mathrm{Var}(a_{t-15}) \\
&= \phi_1^2\mathrm{Var}(Y_t) + \phi_2^2\mathrm{Var}(Y_t) + 2\phi_1\phi_2\mathrm{Cov}(Y_{t-1}, Y_{t-2}) \\
&+ \left[1 + \theta_1^2 + \theta_2^2 + \theta_3^2 + \Theta_1^2 + \theta_1^2\Theta_1^2 + \theta_2^2\Theta_1^2 + \theta_3^2\Theta_1^2\right]\mathrm{Var}(a_t) \\
&= \phi_1^2\sigma_Y^2 + \phi_2^2\sigma_Y^2 + 2\phi_1\phi_2\mathrm{Cov}(Y_{t-1}, Y_{t-2}) \\
&+ \left[1 + \theta_1^2 + \theta_2^2 + \theta_3^2 + \Theta_1^2 + \theta_1^2\Theta_1^2 + \theta_2^2\Theta_1^2 + \theta_3^2\Theta_1^2\right]\cdot\sigma_a^2. \qquad (6.3)
\end{aligned}$$

Mais cómpre ter en conta que, se denotamos $\Upsilon_1 = \mathrm{Cov}(Y_{t-1}, Y_{t-2})$, e empregamos un argumento similar ao que usamos na expresión (5.7), entón

$$\begin{aligned}
\Upsilon_1 &= \mathbb{E}(Y_{t-1}\cdot Y_{t-2}) - \mathbb{E}(Y_{t-1})\cdot\mathbb{E}(Y_{t-2}) = \mathbb{E}(Y_{t-1}\cdot Y_{t-2}) - \mu_Y^2 \\
&= c\cdot\mathbb{E}(Y_{t-2}) + \phi_1\mathbb{E}(Y_{t-2}^2) + \phi_2\mathbb{E}(Y_{t-2}\cdot Y_{t-3}) - \mu_Y^2 \\
&= (1-\phi_1-\phi_2)\cdot\mu_Y\cdot\mathbb{E}(Y_{t-2}) + \phi_1\mathbb{E}(Y_{t-2}^2) + \phi_2\mathbb{E}(Y_{t-2}\cdot Y_{t-3}) - \mu_Y^2 \\
&= (1-\phi_1-\phi_2)\cdot\mu_Y^2 + \phi_1\mathbb{E}(Y_{t-2}^2) + \phi_2\mathbb{E}(Y_{t-2}\cdot Y_{t-3}) - \mu_Y^2 \\
&= \phi_1(\mathbb{E}(Y_{t-2}^2) - \mu_Y^2) + \phi_2(\mathbb{E}(Y_{t-2}\cdot Y_{t-3}) - \mu_Y^2) \\
&= \phi_1\sigma_Y^2 + \phi_2\Upsilon_1.
\end{aligned}$$

É dicir, $\Upsilon_1 = \phi_1\sigma_Y^2 + \phi_2\Upsilon_1$, co cal $(1-\phi_2)\Upsilon_1 = \phi_1\sigma_Y^2$, e obtense, sempre e cando $\phi_2 \neq 1$:

$$\Upsilon_1 = \frac{\phi_1}{1-\phi_2}\sigma_Y^2. \qquad (6.4)$$

Así pois, usando (6.3) e (6.4), tense:

$$\sigma_Y^2 = \frac{\left[1 + \theta_1^2 + \theta_2^2 + \theta_3^2 + \Theta_1^2 + \theta_1^2\Theta_1^2 + \theta_2^2\Theta_1^2 + \theta_3^2\Theta_1^2\right]\cdot\sigma_a^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1\phi_2\left(\dfrac{\phi_1}{1-\phi_2}\right)}. \qquad (6.5)$$

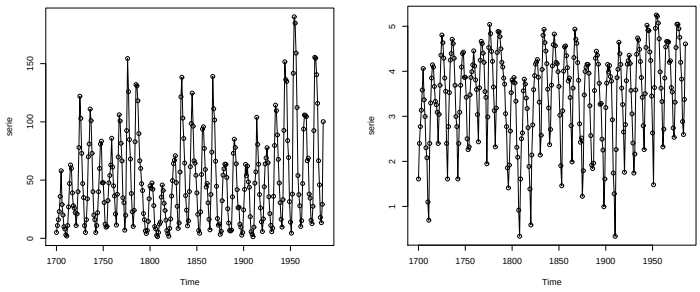
Ademais, a función `auto.arima` proporciona as estimacións dos parámetros do modelo (6.1), resultando:

$$\hat\phi_1 = 1{,}57, \hat\phi_2 = -0{,}962, \hat\theta_1 = -0{,}5086, \hat\theta_2 = -0{,}1218, \hat\theta_3 = 0{,}5677, \hat\Theta_1 = -0{,}3015, \qquad (6.6)$$
$$\hat\mu_Y = 6{,}6643, \hat\sigma_a^2 = 0{,}2056.$$

Empregando (6.2), (6.5) e (6.6), obtemos que a estimación da distribución marxinal de Y_t é $N(6{,}6643, 12{,}27911)$.

En segundo lugar, analizaremos o conxunto de datos `sunspot.year`, proporcionado pola librería `datasets` de `R`. Este conxunto de datos achéganos información sobre o número de manchas solares entre 1700 e 1988. Estes datos recolléronse de forma anual. Novamente, esta mostra depende do tempo, xa que os datos de manchas solares variarán conforme o fan os anos.

Procederemos de xeito análogo a como fixemos co outro conxunto de datos, é dicir, trataremos de identificar un modelo paramétrico que xerase a serie de tempo `sunspot.year`, que está constituida por 289 observacións. Comezaremos presentando o seu gráfico secuencial.



(a) Serie `sunspot.year`. (b) Serie log(`sunspot.year`).

Figura 6.5: Gráficos secuenciais das series `sunspot.year` e log(`sunspot.year`).

Como podemos ver, a partir da Figura 6.5a, para comezar a traballar coa mostra precisamos realizar algunhas modificacións. A primeira delas será eliminar os valores que sexan cero, é dicir, os anos nos que o número de manchas solares identificadas é nulo. Seguidamente, ante a presenza de variabilidade na Figura 6.5a, aplicamos a función logaritmo á serie `sunspot.year`, obtendo log(`sunspot.year`), na Figura 6.5b. Ademais, detectamos presenza de estacionalidade.

Logo de diferenciar regularmente a serie log(`sunspot.year`) para eliminar a tendencia, representamos a fas na mostraxe desta serie para comprobar se é estacionaria.

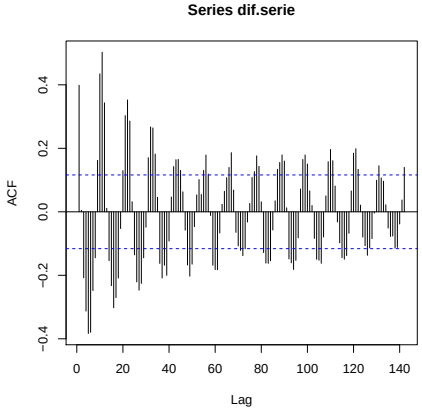


Figura 6.6: fas na mostraxe da serie log(`sunspot.year`) diferenciada regularmente.

Á vista da Figura 6.6, deducimos que esta serie non é estacionaria, xa que presenta estacionalidade de período 12, converxendo lentamente a 0 os retardos estacionais, polo que se evidencia unha

forte correlación positiva. É por iso que precisaremos diferenciar de novo a serie log(`sunspot.year`) estacionalmente con período 12. Se representamos o gráfico secuencial e a fas na mostraxe da serie transformada polo logaritmo diferenciada dúas veces, podemos ver que xa conseguimos a súa estacionariedade. Logo, é hora de representar a fas e a fap na mostraxe de dita serie para propoñer un modelo.

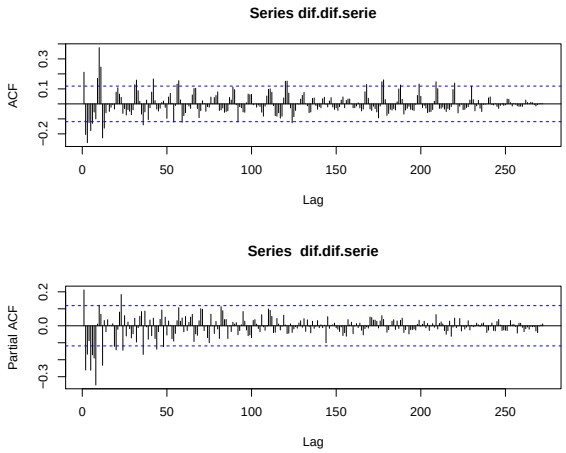


Figura 6.7: fas e fap na mostraxe da serie log(`sunspot.year`) diferenciada dúas veces.

A Figura 6.7 suxire que a serie transformada foi xerada por un proceso ARIMA$(0,1,6)\times(1,1,0)_{12}$, xa que os retardos regulares da fas mostral anúlanse a partir do sexto retardo. Ademais, na fap empírica da parte regular aparecen moitos retardos non nulos. Por outro lado, na fap da parte estacional temos que ningún retardo é significativo agás o primeiro retardo estacional, e a fas da parte estacional presenta moitos retardos non nulos. Non obstante, as innovacións a_t deste modelo presentan correlación e non son independentes, polo que este proceso non resulta axeitado para xerar a serie de tempo log(`sunspot.year`).

Novamente, comprobamos coa función `auto.arima` se hai algunha tentativa que puidese xerar esta serie de tempo. Esta función propón o modelo ARIMA$(4,1,4)\times(0,1,1)_{12}$ como posible xerador da serie `sunspot.year`. Este modelo pódese representar como:

$$(1-\phi_1B-\phi_2B_2-\phi_2B^3-\phi_4B^4)(1-B)(1-B^{12})Y_t = c+(1+\theta_1B+\theta_2B^2+\theta_3B^3+\theta_4B^4)(1+B^{12}\Theta_1)a_t, \tag{6.7}$$

onde $Y_t=\log(X_t)$; X_t, o número de manchas solares detectadas no ano t; c a constante; e $BY_t = Y_{t-1}, B^sY_t = Y_{t-s}, Ba_t = a_{t-1}, B^sa_t = a_{t-s}$. Outra forma de presentar este modelo será:

$$\begin{aligned} Y_t &= c+(1+\phi_1)Y_{t-1}+(\phi_2-\phi_1)Y_{t-2}+(\phi_3-\phi_2)Y_{t-3}+(\phi_4-\phi_3)Y_{t-4}-\phi_4Y_{t-5} \\ &+ Y_{t-12}-(1+\phi_1)Y_{t-13}+(\phi_1-\phi_2)Y_{t-14}(\phi_2-\phi_3)Y_{t-15}+(\phi_3-\phi_4)Y_{t-16}+\phi_4Y_{t-17} \\ &+ a_t+\theta_1a_{t-1}+\theta_2a_{t-2}+\theta_3a_{t-3}+\theta_4a_{t-4}+\Theta_1a_{t-12}+\theta_1\Theta_1a_{t-13}+\theta_2\Theta_1a_{t-14} \\ &+ \theta_3\Theta_1a_{t-15}+\theta_4\Theta_1a_{t-16}. \end{aligned}$$

As estimacións dos parámetros do modelo (6.7) que proporciona a función `auto.arima` son as seguintes (considerando a constante nula):

$$\hat{\phi}_1 = 2{,}466, \hat{\phi}_2 = -3{,}2092, \hat{\phi}_3 = 2{,}2679, \hat{\phi}_4 = -0{,}8568,$$
$$\hat{\theta}_1 = -2{,}4807, \hat{\theta}_2 = 2{,}9486, \hat{\theta}_3 = -1{,}9249, \hat{\theta}_4 = 0{,}5612,$$

$$\hat{\Theta}_1 = -1, \hat{\sigma}_a^2 = 0{,}219.$$

Finalmente, realizaremos a diagnose do modelo (6.7), é dicir, comprobaremos se as innovacións $\{a_t\}_{t\in\mathbb{Z}}$ son, en efecto, ruído branco. Se realizamos o test de independencia `Box.test`, obtemos un p-valor de 0,9293, polo que non existen evidencias significativas para rexeitar a hipótese nula de independencia das innovacións. Seguidamente, se realizamos un test para ver se a_t ten media cero (`t.test`), obtemos un p-valor de 0,4841, polo que, de novo, non temos probas significativas para rexeitar a hipótese nula de que a media sexa cero. Por outra banda, se realizamos un test de normalidade (`jarque.bera.test`), obtemos un p-valor de $1{,}163 \times 10^{-12}$, logo atopamos probas significativas para rexeitar a hipótese nula de normalidade a calquera nivel de significación habitual. Para rematar, vexamos se as innovacións son incorreladas.

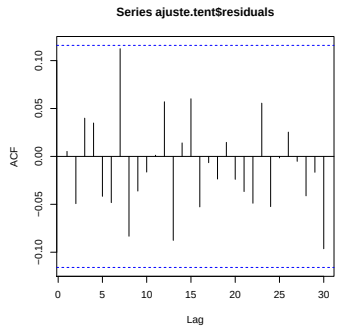


Figura 6.8: fas na mostraxe das innovacións para o modelo (6.7).

Atendendo á Figura 6.8, concluímos que as innovacións son incorreladas. Polo tanto, o modelo (6.7) con constante nula e innovacións non gaussianas resulta axeitado como xerador da serie log(`sunspot.year`).

6.2. Resultados

Tendo en conta as conclusións extraídas para ambos conxuntos de datos reais na Sección 6.1, os dous modelos paramétricos propostos parecen unha boa escolla. Sen embargo, é certo que poderían existir outros modelos paramétricos distintos dos propostos que proporcionasen un bo axuste. Un dos principais problemas de empregar un procedemento paramétrico para coñecer a distribución marxinal da serie temporal é que, ademais de ter que seleccionar un proceso para modelizar a dependencia, e da necesidade de asumir un modelo paramétrico para a distribución das innovacións $\{a_t\}_{t\in\mathbb{Z}}$, habería que corroborar se esta elección é a mellor.

Por outra banda, se empregásemos un procedemento non paramétrico, evitaríamos este tipo de problemas que acabamos de plantexar, xa que poderíamos estimar o modelo sen ter en conta se o modelo paramétrico que estamos a empregar é correcto ou non. Ademais, no caso do modelo (6.7), as innovacións non resultaban gaussianas, polo que non podíamos obter a distribución marxinal da serie temporal ó non coñecer a distribución de $\{a_t\}_{t\in\mathbb{Z}}$. Esta sería unha cuestión que empregando un procedemento non paramétrico non se plantexaría, xa que se podería estimar a distribución marxinal da serie sen ningún tipo de hipótese previa agás condicións de regularidade desa función de densidade marxinal, como continuidade ou diferenciabilidade da mesma. É por iso que nesta sección tentaremos estimar a distribución que segue cada unha das series temporais que describimos anteriormente dun xeito non paramétrico. Para iso, empregaremos os selectores do parámetro ventá estudados no Capítulo 4: h_{PI}, h_{MCV} e h_{BOOT}; realizando un estudo comparativo entre eles.

Para seleccionar o parámetro de suavizado óptimo, partiremos dun conxunto de catro ventás equiespaciadas entre os valores 0,01 e 10, e procederemos de xeito análogo a como se describíu no Capítulo 5 para a selección do parámetro ventá óptimo.

Comezaremos co conxunto de datos `lynx`. Como xa dixemos, traballaremos coa transformación logaritmo neperiano dos mesmos, considerando a serie temporal log(`lynx`). Proporcionaremos a continuación o gráfico coa estimación da densidade con cada selector de ventá (Figura 6.9), os parámetros ventá resultantes por cada método (Cadro 6.1), e os tempos de CPU necesarios en cada execución (Cadro 6.2). Ademais, representarase a curva da densidade que se obtivo na Sección 6.1 para o modelo paramétrico (6.1).

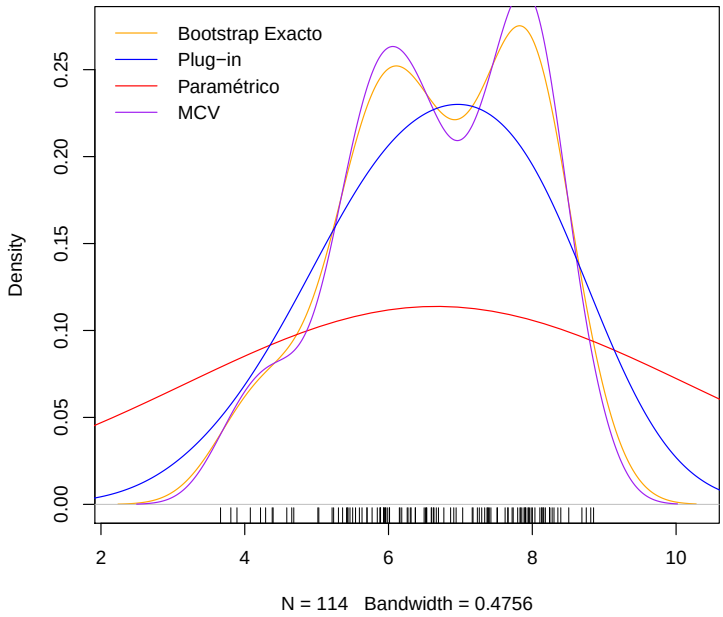


Figura 6.9: Estimación da función de densidade para os datos `lynx` empregando distintos selectores do parámetro ventá.

h_{PI}	h_{MCV}	h_{BOOT}
1,01897	0,38897	0,47559

Cadro 6.1: Parámetros ventá para os datos `lynx` seleccionados polos métodos plug-in, validación cruzada modificada e bootstrap (empregando a expresión exacta para o $MISE^*(h)$).

CPU_{PI}	CPU_{MCV}	CPU_{BOOT}
30,71	57,47	28,99

Cadro 6.2: Tempos de execución da CPU (en segundos) para obter os parámetros ventá seleccionados polos métodos estudados no Capítulo 4 para os datos `lynx`.

A partir do Cadro 6.1 e da Figura 6.9 vemos que tanto a ventá h_{MCV} como a ventá h_{BOOT} proporcionan unha boa estimación da función de densidade, xa que incluso son capaces de captar as dúas modas. En cambio, a ventá h_{PI} proporciona de novo un estimador sobresuavizado (é dicir, h_{PI} é un valor alto). En canto á curva da densidade obtida a partir do modelo paramétrico proposto (6.1), aínda que a curva da densidade se axusta ós datos, non é a curva máis axeitada, xa que é demasiado suave, máis incluso que a obtida polo selector de ventá h_{PI}. Por outra banda, atendendo ó Cadro 6.2, concluímos que o tempo requerido para obter o selector de ventá h_{MCV} é moito maior que os demais. Ademais, o selector de ventá que menos tempo require a súa execución é h_{BOOT}.

Estudaremos agora o conxunto de datos `sunspot.year`, dos cales eliminaremos os datos que son 0, para posteriormente transformar coa función logaritmo neperiano, considerando a serie temporal log(`sunspot.year`). De xeito análogo a como procedimos co conxunto de datos `lynx`, reuniremos nun gráfico as estimacións da densidade cos selectores de ventá $h_{PI}, h_{MCV}, h_{BOOT}$ (Figura 6.10). De seguido, presentaremos os selectores do parámetro de ventá h_{PI}, h_{MCV} e h_{BOOT} no Cadro 6.3, e, finalmente, analizaremos os tempos de CPU requeridos en cada caso, recollidos no Cadro 6.4.

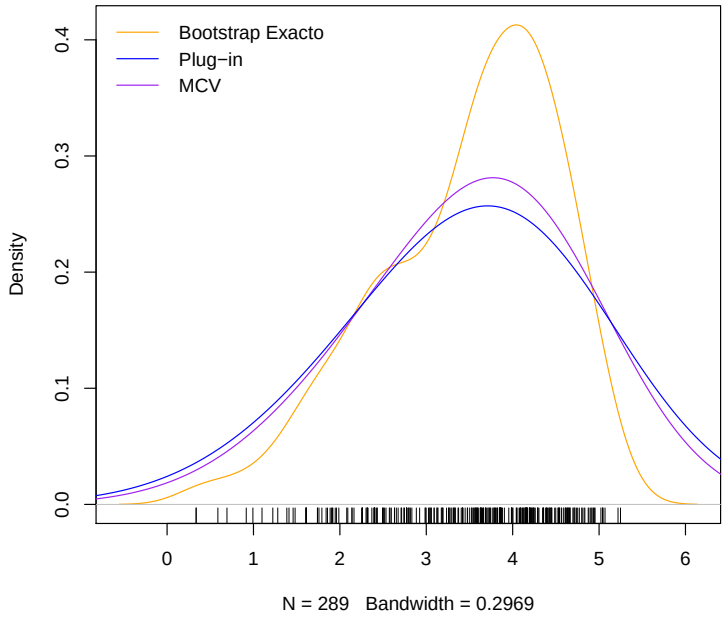


Figura 6.10: Estimación da función de densidade para os datos `sunspot.year` empregando distintos selectores do parámetro ventá.

h_{PI}	h_{MCV}	h_{BOOT}
1,14731	0,97095	0,29693

Cadro 6.3: Parámetros ventá para os datos `sunspot.year` seleccionados polos métodos plug-in, validación cruzada modificada e bootstrap (empregando a expresión exacta para o $MISE^*(h)$).

CPU_{PI}	CPU_{MCV}	CPU_{BOOT}
213,54	366,22	212,7

Cadro 6.4: Tempos de execución da CPU (en segundos) para obter os parámetros ventá seleccionados polos métodos estudados no Capítulo 4 para os datos `sunspot.year`.

Neste caso, atendendo ó Cadro 6.3 e á Figura 6.10, observamos que tanto a ventá h_{PI} como a ventá h_{MCV} devolven un estimador sobresuavizado, que non é capaz de reflexar a densidade dos datos con total fidelidade. Sen embargo, o selector h_{BOOT} é o que produce para este conxunto de datos un estimador da densidade máis fidedigno. Por outra banda, á vista do Cadro 6.4, temos de novo que o selector h_{MCV} require moito máis tempo de CPU para proporcionar un resultado; mentres que h_{BOOT} e h_{PI} precisan aproximadamente o mesmo tempo de execución. Se comparamos estes resultados cos do Cadro 6.2, vemos que en xeral os tempos requeridos polos datos `sunspot.year` son moito maiores. Isto débese a que a serie de tempo log(`sunspot.year`) conta con máis observacións que a serie log(`lynx`).

Capítulo 7

Conclusións

Unha das funcións que caracterizan a distribución dunha variable aleatoria é a función de densidade. Logo, é de vital importancia contar con ferramentas para poder coñecer dita función. Unha das formas de proceder é empregando un enfoque paramétrico, isto é, tratando de axustar unha distribución que estea determinada por un número finito de parámetros aos datos empregados. Sen embargo, un dos principais problemas da estatística paramétrica é que, aínda que un modelo paramétrico pareza unha boa elección para determinar a distribución da poboación a estudar, é posible que existisen outros modelos distintos, tamén paramétricos, que determinasen correctamente a distribución dos datos. Polo tanto, o principal problema deste tipo de procedementos será, unha vez atopado un modelo paramétrico que axuste a distribución dos datos, que dito modelo proposto sexa o mellor que se pode atopar. Por outra banda, en moitas ocasións nin sequera está claro qué modelo paramétrico é o axeitado.

Neste momento, teñen cabida as técnicas non paramétricas. Por un lado, estas poderían ser útiles para estimar de xeito non paramétrico a densidade co fin de describir o comportamento dunha determinada variable aleatoria. Por outra banda, a partir dun estimador non paramétrico da densidade, poderíase contrastar se un determinado modelo paramétrico se axusta de xeito satisfactorio ós datos. Neste traballo empregouse un enfoque non paramétrico para a estimación da densidade.

O estimador non paramétrico clásico para a función de densidade é o histograma. Non obstante, aínda que é unha ferramenta útil na análise exploratoria dos datos para facernos unha idea de como pode ser a distribución dos mesmos, este estimador depende do ancho de banda, h, e da selección do punto de orixe, x_0, co que é construido, ademais de que é un estimador non continuo da función de densidade. Consecuentemente, xorde outro estimador non paramétrico da densidade, coñecido como o estimador naive ou histograma móbil. Este, se ben segue a ser discontinuo e depende do ancho de banda, h, xa non depende do punto inicial, x_0. Finalmente, o estimador tipo núcleo da densidade, tamén coñecido como estimador de Parzen-Rosenblatt, foi introducido a comezos dos anos 60 do século XX. Este estimador, no cal nos centramos neste traballo, depende unicamente do parámetro de ventá, h, xa que o núcleo seleccionado non inflúe na apariencia do estimador obtido.

Unha das preguntas que poden xurdir a raíz do emprego do estimador tipo núcleo de Parzen-Rosenblatt é como afecta o parámetro ventá, h, á estimación da densidade. En efecto, realizando un estudo máis pormenorizado do sesgo e da varianza de dito estimador, vemos que o sesgo diminúe cando h é pequeno, polo que valores de h pequenos proporcionarán estimadores centrados. Sen embargo, isto conlevaría a un incremento da varianza. Por outro lado, se h é grande, conseguiríamos reducir a varianza, pero derivaría nun aumento do sesgo. Polo tanto, un dos problemas centrais da estatística non paramétrica será seleccionar un valor de h que proporcione un estimador que atope un balance entre o sesgo e a varianza. Entón, é preciso o uso dun criterio de erro que evalúe o comportamento do estimador da densidade, é dicir, que relacione a densidade real con dito estimador. Deste xeito, a partir

do criterio de erro poderase extraer un valor óptimo para o parámetro de ventá, h. Concretamente, o foco de estudo deste traballo foi a elección deste selector de ventá, h, nun contexto de dependencia dos datos, empregando como criterio de erro a minimizar o erro cuadrático medio integrado, ou $MISE(h)$.

Nembargantes, ademais de revisar as propostas de Hall *et al.* (1995) e Cox e Kim (1997) (métodos plug-in e validación cruzada modificada, respectivamente) para a selección do parámetro de suavizado, abordouse este problema co uso dunha das técnicas máis empregadas de remostraxe: o bootstrap. Estas técnicas cobran gran importancia cando o tamaño mostral non é suficientemente grande, xa que con elas é posible xerar unha nova mostra do tamaño mostral que se desexe, e que teña unha distribución similar á da poboación orixinal. Así, unha vez realizada unha revisión bibliográfica do bootstrap por bloques, bootstrap estacionario e o método de submostraxe; describíuse unha suavización do bootstrap estacionario de Politis e Romano (1994a), para datos dependentes, e presentouse unha expresión exacta para a versión bootstrap do $MISE(h)$, o $MISE^*(h)$, de xeito que se poida obter directamente o parámetro de suavizado, h, sen máis que minimizar dita expresión. En efecto, a implementación da expresión exacta para o $MISE^*(h)$ supón un enfoque diferente na obtención do parámetro de suavizado, h, baixo a hipótese de dependencia dos datos empregados, evitando realizar unha aproximación de dito criterio de erro por Monte Carlo, e polo tanto, evitando tamén un erro de aproximación.

Finalmente, comprobouse a eficacia deste novo procedemento, e comparouse co comportamento dos parámetros ventá obtidos polos métodos plug-in e validación cruzada modificada; así como coa metodoloxía bootstrap empregando a aproximación por Monte Carlo. Realizouse para iso un estudo de simulación con diferentes modelos, descritos no Capítulo 5, e aplicáronse os distintos métodos a dous conxuntos de datos reais no Capítulo 6.

Por unha banda, corroborouse empiricamente o bo comportamento do parámetro de suavizado obtido minimizando a expresión exacta para o $MISE^*(h)$, ó que denotamos por h_{BOOT}, xa que en todos os casos que analizamos no estudo de simulación, o valor proporcionado por este último era cercano ó valor obtido tras minimizar o $MISE(h)$. Ademais, o selector h_{BOOT} é o que menos tempo de computación require para obter un resultado. Por outra banda, o selector de ventá resultante tras empregar o método de validación cruzada modificada, h_{MCV}, aínda que ofrece valores bastante cercanos ó selector h_{MISE}, tamén é certo que en ocasións produce estimadores infrasuavizados, e por tanto, demasiado rugosos (consecuencia dun parámetro h_{MCV} baixo). Asimesmo, h_{MCV} é o parámetro de suavizado que maior tempo de computación require para obter un resultado. Por último, o parámetro ventá obtido polo método plug-in de Hall *et al.* (1995), h_{PI}, produce estimadores sobresuavizados, é dicir, os valores obtidos para h_{PI} son demasiado grandes. En conclusión, obtivemos un bo comportamento empírico do parámetro de suavizado h_{BOOT}, tanto a nivel do grao de suavización do estimador da función de densidade que proporciona, como para os tempos de computación necesarios, sendo este o que mellores resultados ofrece.

Bibliografía

[1] Billingsley, P. (1968). *Probability and Measure*. Jonh Wiley & Sons.

[2] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71, 353-360.

[3] Cao, R. (1993). Bootstrapping the mean integrated squared error. J. Mult. Anal. 45, 137-160.

[4] Cao, R. (1999). An overview of bootstrap methods for estimating and predicting in time series. Test, 8, 95-116.

[5] Cao, R., Quintela-del-Río, A. and Vilar-Fernández, J.M. (1993). Bandwidth Selection in Nonparametric Density Estimation Under Dependence. A simulation study. Com. Statist., 8, 313-332.

[6] Cox, D. and Kim, TY.(1997). A study on bandwidth selection in density estimation under dependence. J. Mult. Anal., 62, 190-203.

[7] Doukhan, P., Lang,G., Surgallis, D. and Teyssière, G. (2010). *Dependence in Probability and Statistics*. Springer.

[8] Efron, B. y Tibishirani, R. (1986). *An Introduction to the Bootstrap*. Chapman and Hall.

[9] Hall, P., Lahiri, N. S. and Truong, YT.(1995). On bandwidth choice for density estimation with dependent data. Ann. Statist., 23, 6, 2241-2263.

[10] Hwang, E. and Shin, DW. (2012). Stationary bootstrap for kernel density estimators under psi-weak dependence. Com. Statist. and Data Analysis, 56, 1581-1593.

[11] Kreiss, JP. and Paparoditis, E. (2011). Bootstrap methods for dependent data: A review. J. Korean Statist. Soc., 40, 357-378.

[12] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. Ann. Statist., 17, 1217-1241.

[13] Liu, R.Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In Exploring the limits of bootstrap (R. LePage and L Billard, Eds.), pp. 225-248. New York: Wiley.

[14] Politis, D.N. and Romano, J.R. (1994a). The stationary bootstrap. J. Amer. Statist. Assoc., 89, 1303-1313.

[15] Politis, D.N. and Romano, J.R. (1994b). Large sample confidence regions based on subsamples under minimal assumptions. Ann. Statist., 22, 2031-2050.

[16] Radulovic, D. (1996). The bootstrap of the mean for strong mixing sequences under minimal conditions. Statistics & Probability Letters, 28, 65-72.

[17] Sheather, S. J. y Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B, 53, 683-690.