USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Estimación tipo núcleo da función de densidade

Álvaro Arrojo Vázquez

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Estimación tipo núcleo da función de densidade

Álvaro Arrojo Vázquez

Setembro, 2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estadística e Investigación Operativa
Título: Estimación tipo núcleo da función de densidade
Breve descrición do contido
Dada unha variable aleatoria continua, a función de densidade permite coñecer como se distribúe dita variable dado que a probabilidade de que a variable aleatoria caia nunha rexión específica do espazo de posibilidades estará dada pola integral da densidade desta variable entre un e outro límite de dita rexión. Por este motivo, a estimación da función de densidade é un tema de gran interese no ámbito da Estatística. Ao longo deste traballo, prestarase especial atención á estimación tipo núcleo da función de densidade. Neste caso, resultará fundamental a selección do parámetro de suavizado de cara a obter unha boa estimación da función de densidade. Presentaranse e compararanse diferentes estratexias dispoñibles na literatura para levar a cabo a selección do parámetro de suavizado. A título orientativo, o traballo podería organizarse nas seguintes seccións: ▪ A función de densidade. Estimación. ▪ Estimación tipo núcleo da función de densidade. ▪ Selección do parámetro de suavizado.

- Estudo comparativo dos diversos selectores do parámetro de suavizado.

Presentaranse diferentes estimacións da función de densidade aplicadas tanto a conxuntos de datos tanto reais como simulados. Para elo empregarase o software libre R (https://www.r-project.org/).

Recomendacións

Outras observacións

Índice xeral

Resumo

A estimación da función de densidade é un tema de gran importancia no campo da Estatística xa que calquera variable aleatoria continua queda completamente caracterizada grazas á súa función de densidade. Dada a importancia da estimación da función de densidade, este tema foi abordado dende diferentes puntos de vista. Presentouse dende unha primeira aproximación grazas a unha representación gráfica (coñecida como histograma) ata métodos non paramétricos máis complexos como a estimación tipo núcleo.

No presente traballo introdúcese a estimación tipo núcleo da función de densidade así como diferentes criterios de erro asociados ao mencionado estimador. Ademais, abordarase o problema da selección do parámetro de suavizado e mostraranse diferentes propostas de selectores dispoñibles na literatura. Para poder comparar os diferentes selectores do parámetro de suavizado propostos deseñouse un completo estudo de simulación que permitirá extraer conclusións sobre as súas propiedades.

Por outra banda, propoñerase unha estimación da función de densidade asociada a unha mostra de datos reais para ilustrar a utilidade do estimador tipo núcleo na práctica. Finalmente, presentaranse as principais conclusións deste traballo. Resultan palabras clave do mesmo as seguintes:

- estimación non paramétrica
- función de densidade
- ancho de ventá.

Abstract

The estimation of density function is a real important topic in the field of Statistics due to the fact that every continuous variable is completely defined by its density function. Given the density function importance, this subject was approached by different points of view. It was presented from a first graphic approach (known as histogram) to non parametric complex methods as kernel density estimation.

On the present project the kernel density estimation and some different error criteria related with the given estimator are introduced. It is also address the smooth parameter selection and it is shown different selector proposals that are present in the literature. To be able to compare the different selectors of the proposed smoothing parameter, a complete simulation study was designed that will allow conclusions about their properties.

On the other hand, an estimation of the density function associated with a sample of real data will be proposed in order to illustrate the usefulness of the kernel density estimator in practice. Finally, the main conclusions of this project will be presented. The keywords of this project are:

- nonparametric estimation
- density function
- bandwidth.

Resumen

La estimación de la función de densidad es un tema de gran importancia en el campo de la Estadística ya que cualquiera variable aleatoria continua queda completamente caracterizada gracias a su función de densidad. Dada la importancia de la estimación de la función de densidad, este tema fue abordado desde diferentes puntos de vista. Se presentó desde una primera aproximación gracias a una representación gráfica (conocida como histograma) hasta métodos no paramétricos más complejos como la estimación tipo núcleo.

El el presente trabajo se introduce la estimación tipo núcleo de la función de densidad así

como diferentes criterios de error asociados al mencionado estimador. Además, se aborda el problema de la selección del parámetro de suavizado y se muestran diferentes propuestas de selectores disponibles en la literatura. Para poder comparar los diferentes selectores del parámetro de suavizado propuestos se diseñó un completo estudio de simulación que permitirá extraer conclusiones sobre sus propiedades.

Por otra banda, se propondrá una estimación de la función de densidad asociada a una muestra de datos reales para ilustrar la utilidad del estimador tipo núcleo en la práctica. Finalmente, se presentarán las principales conclusiones de este trabajo. Resultan palabras clave del mismo las siguientes:

- estimación no paramétrica
- función de densidad
- ancho de ventana.

Capítulo 1

Introdución

Unha **variable aleatoria** X é unha función definida sobre un espazo de probabilidade que asigna un valor numérico ao resultado dun experimento aleatorio. Formalmente, un espazo de probabilidade é unha terna $(\Omega, \mathcal{A}, \mathcal{P})$, onde Ω é o espazo mostral, é dicir, os posibles resultados do experimento aleatorio, $\mathcal{A}$ é unha $\sigma-$álxebra sobre Ω e $\mathcal{P}$ é unha función de probabilidade que asigna unha probabilidade a cada suceso.

Dicimos que unha **variable aleatoria** X é **discreta** se toma un número finito ou infinito numerable de valores cunhas certas probabilidades asociadas a cada valor. O resultado tras o lanzamento dun dado é un exemplo de variable aleatoria discreta.

Pola contra, unha **variable aleatoria** X **continua** é aquela que toma un número infinito de valores nun intervalo, nunha unión de intervalos ou en toda a recta real. A estatura dunha certa poboación é un exemplo de variable aleatoria continua. Ao longo do presente traballo consideraremos exclusivamente variables aleatorias continuas, aínda que non se mencione explicitamente.

Toda variable aleatoria está caracterizada pola **función de distribución**, que denotaremos por F, e vén dada por

$$\begin{aligned} F : \Omega &\to [0,1] \\ x &\to F(x) = \mathbb{P}(X \leq x). \end{aligned} \tag{1.1}$$

É dicir, representa a probabilidade acumulada ata un certo valor $x \in \mathbb{R}$. A derivada da función de distribución é a **función de densidade** (que habitualmente denotaremos por

f), isto é

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Dita función só pode definirse para variables aleatorias continuas.

Para calquera variable aleatoria continua, a función de densidade satisfai as seguintes propiedades:

1. A área entre o percorrido da función f e o eixo $y = 0$ ten valor 1, é dicir:

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

2. Só toma valores positivos, é dicir: $f(x) \geq 0$ para todo $x \in \mathbb{R}$.

3. A probabilidade de que a variable X tome valores no intervalo $[a, b]$ é igual á área baixo a curva da función de densidade en dito intervalo. É dicir:

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a).$$

No caso de variables aleatorias continuas a función de densidade terá unha gran importancia xa que determina como se distribúe unha variable aleatoria. Como consecuencia, é de gran interese saber como estimar a función de densidade asociada a unha mostra de datos, dado que con ela podemos, por exemplo, estimar probabilidades ou medidas características asociadas a esa variable como

$$\widehat{\mathbb{P}(X \leq a)} = \int_{-\infty}^{a} \widehat{f}(x)dx$$

ou

$$\widehat{\mathbb{E}[X]} = \int_{-\infty}^{\infty} x\widehat{f}(x)dx$$

onde $\widehat{f}$ denota a estimación da función de densidade.

Concretamente, consideraremos $\{x_1, ..., x_n\}$ unha mostra aleatoria simple[1] dunha variable aleatoria continua X para a cal queremos estimar a súa función de densidade f. Se os nosos datos non seguen unha distribución coñecida (ou non podemos asumir que a seguen), non coñeceremos unha fórmula para calcular f e trataremos de dar unha aproximación da mesma.

[1] Unha mostra dise aleatoria simple se é identicamente distribuída e con x_i,x_j independentes para todo i, j tal que $i \neq j$.

O primeiro intento de aproximación da función de densidade foi o **histograma**, que consiste nun gráfico con barras no que a área de cada barra é proporcional á frecuencia dos valores no intervalo que conforma a base da barra. Porén, esta aproximación é moi sensible ao ancho dos intervalos do histograma. Ademais, o histograma presenta outros problemas ou limitacións: intenta aproximar unha función continua por a través dunha función escalonada, polo que deixamos de ter unha variable continua e pasamos a ter unha variable discretizada, e ademais, o histograma non se pode estender ao caso multivariante.

Ilustraremos a construción do histograma mediante un exemplo axudándonos dunha mostra aleatoria de 100 datos dunha distribución normal[2] de media $\mu = 0$ e varianza $\sigma^2 = 1$, que se coñece habitualmente como normal estándar.

Na Figura 1.1 representamos diferentes histogramas xunto coa función de densidade teórica asociada aos datos simulados. No histograma (b) da Figura 1.1, R non ten ningunha indicación a cerca do número de divisións a facer, digamos que de algún xeito escolle o número óptimo de caixas conforme á distribución dos datos[3]. Nos dous restantes histogramas o número de divisións está modificado. Notemos que na parte (a), a gráfica da función de densidade, que é a liña sólida que acompaña ao histograma, non é satisfactoriamente aproximada por este. De igual modo, no gráfico (c) parece que a altura das caixas varía moito máis que a altura da gráfica da función densidade.

Dadas as limitacións do histograma, ao longo deste traballo imos introducir un método para poder estimar a función de densidade dado un conxunto de datos sobre o cal non temos ningunha información sobre a súa distribución.

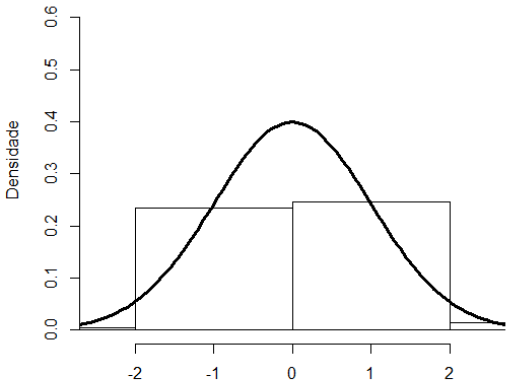
Este método é coñecido como estimación tipo núcleo, e aparece a mediados do século XX. A continuación trátase a estimación tipo núcleo univariante, dando a súa definición, abordando o problema de elección do ancho de banda e analizando varios métodos para a súa elección[4].

---

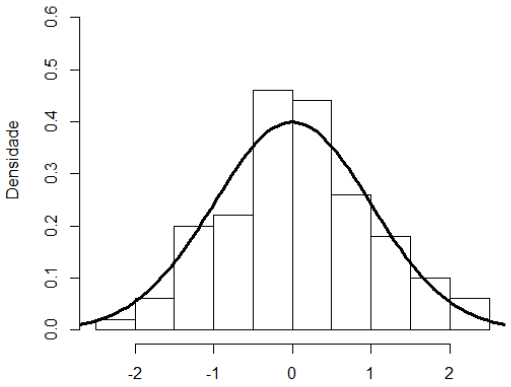
[2]Dentro das variables continuas, unha das distribucións máis importante é a distribución normal cuxa función de densidade é $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$. Como podemos observar, f depende dos parámetros μ, que é a media, e σ, que é a desviación típica.

[3]Neste caso, R emprega o método de Sturges para seleccionar un número adecuado de intervalos de clase que ven dado por $1 + \log_2(n) = 1 + \log_{10}(n)/\log_{10}(2)$ sendo n o tamaño de mostra. Para máis información ver [6].

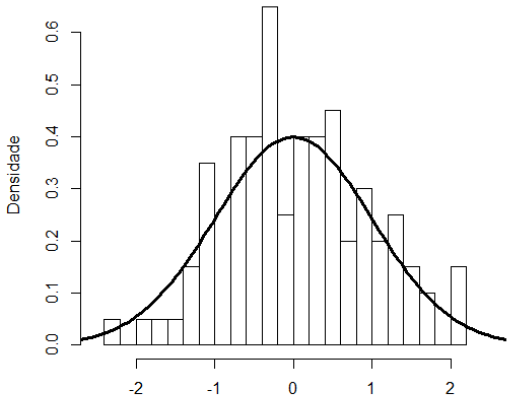
[4]Isto é o que se coñece como Estadística Non Paramétrica posto que non se realiza ningunha suposición sobre a distribución dos datos que se pretenden analizar.



(a) Histograma con dous intervalos



(b) Histograma con dez intervalos



(c) Histograma con vinte e tres intervalos

Figura 1.1: Representación dos histogramas con distintos intervalos de clase xunto coa función de densidade teórica asociada, a normal estándar.

Capítulo 2

Estimación tipo núcleo da función de densidade

A estimación non paramétrica da función de densidade resulta especialmente efectiva cando os modelos paramétricos son inadecuados. O método máis habitual en estimación non paramétrica da función de densidade é o estimador tipo núcleo, que presentamos a continuación.

Para a realización deste capítulo seguiremos o capítulo 2 do libro [7] e impoñeremos as seguintes condicións sobre a variable de interese X e sobre a súa función de densidade f, así como sobre o parámetro h e a función kernel K.

- Existe $f^{(2)}$, que ademais é continua, de cadrado integrable e monótona nos intervalos $(-\infty, M)$ e (M, ∞) para algún $M > 0$.

- A secuencia dos posibles valores positivos de h, que denotaremos por h_n, tende a cero, pero cunha orde menor do que o fai a sucesión $\frac{1}{n}$. É dicir, $\lim_{n\to\infty} h_n = 0$ e $\lim_{n\to\infty} n \cdot h_n = 0$.

- O núcleo K é unha función de densidade limitada, simétrica en torno á orixe e con momento de orde catro finito.

2.1. Estimador tipo núcleo

Ao longo deste capítulo considérase a X unha variable aleatoria continua que ten función de densidade f. Dada $\{x_1, ..., x_n\}$ unha mostra aleatoria simple da variable X imos estimar a función de densidade f mediante o **estimador tipo núcleo**, que se define da seguinte forma:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{2.1}$$

onde K denomínase función **núcleo** (ou *kernel*) que é unha función simétrica en torno ao cero e satisfai

$$\int_{\mathbb{R}} K(x)dx = 1.$$

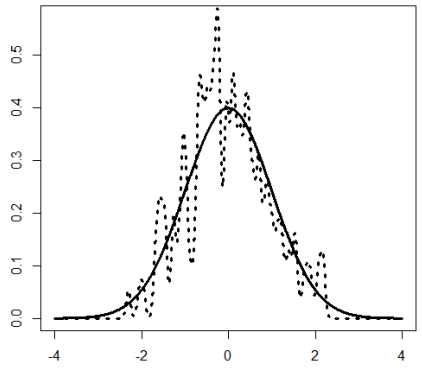
O parámetro $h > 0$ chámase **ventá ou ancho de banda**.

A ecuación (2.1) pódese abreviar introducindo a notación $K_h(u) = h^{-1}K(u/h)$, e obtense como resultado

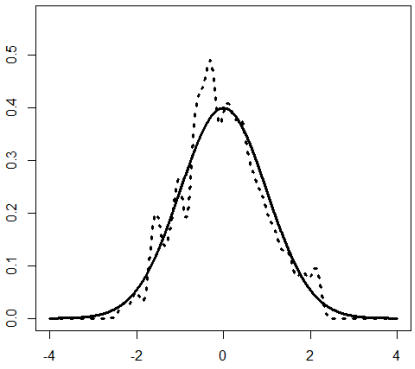
$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) \tag{2.2}$$

A idea deste tipo de estimación baséase en definir unha función núcleo K e centrar dita función nun entorno de lonxitude $2h$ en torno a cada punto x_i da mostra. Nesta situación, se sumamos todas as funcións núcleo obteremos a estimación tipo núcleo da función densidade da nosa mostra. O $1/nh$ na ecuación (2.1) ten a misión de que a integral do estimador teña valor 1, propiedade fundamental dunha función de densidade. Deste xeito obtemos estimacións puntuais da función de densidade. Realizando este proceso para diferentes puntos $x \in \mathbb{R}$ obtemos a estimación tipo núcleo da función de densidade no intervalo desexado e co parámetro ventá h e función núcleo K elixidos.

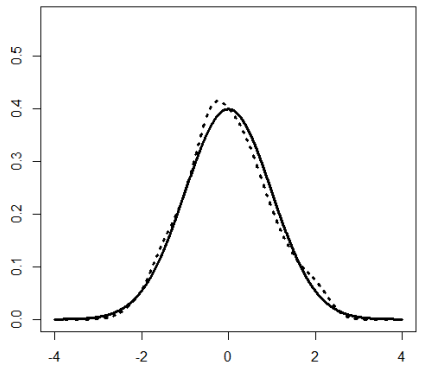
Veremos máis adiante que, na ecuación (2.1), a elección de K non é moi importante. En cambio, unha boa elección do valor de h será fundamental para acadar unha boa aproximación da función de densidade. Isto ilústrase mediante as Figuras 2.1 e 2.2, onde se representan estimacións tipo núcleo da densidade para diferentes valores do parámetro ventá h. Na Figura 2.1 representamos a estimación da función de densidade para unha mostra de tamaño 150 dunha variable con distribución normal estándar, mentres que na Figura 2.2 se representa unha mostra dunha distribución χ-cadrado con tres graos de liberdade. En ambos casos, a estimación tipo núcleo foi xerada cunha función núcleo K normal estándar. En ambas figuras a liña sólida representa a función de densidade teórica



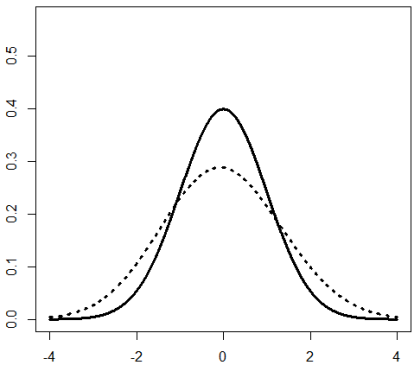
(a) $h = 0.05$



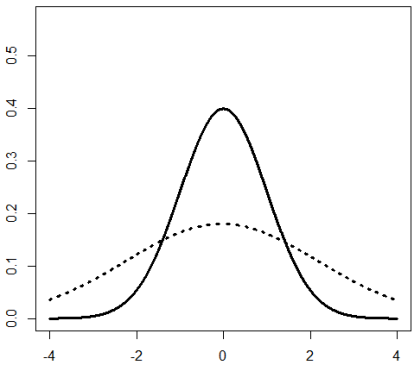
(b) $h = 0.1$



(c) $h = 0.3$

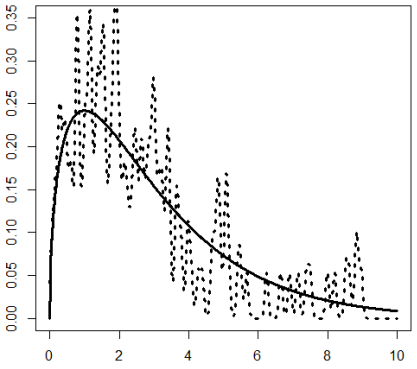


(d) $h = 1$

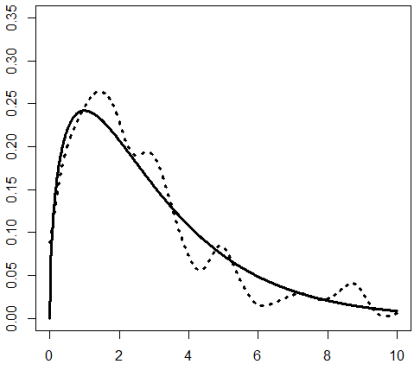


(e) $h = 2$

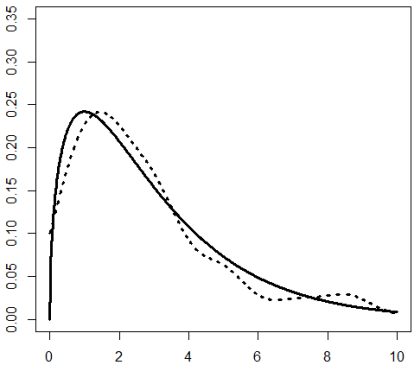
Figura 2.1: Estimación tipo núcleo da función de densidade baseada nunha mostra de 150 datos que seguen unha distribución normal estándar. A liña sólida representa a función densidade teórica mentres que a liña punteada representa a estimación tipo núcleo obtida para os diferentes valores do parámetro h.



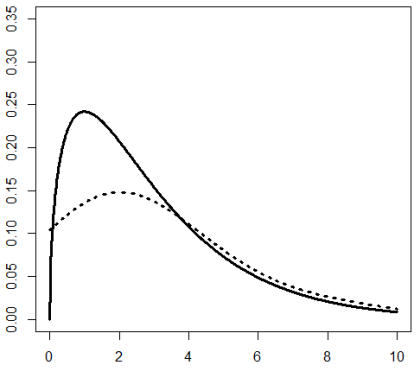
(a) $h = 0.05$



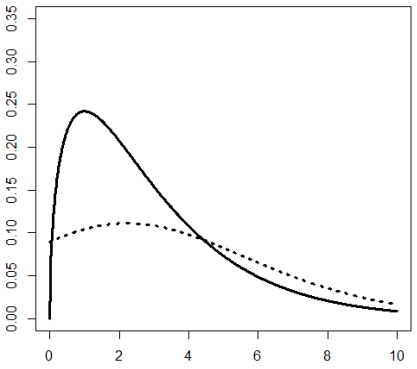
(b) $h = 0.3$



(c) $h = 0.6$



(d) $h = 2$



(e) $h = 3$

Figura 2.2: Estimación tipo núcleo da función de densidade baseada nunha mostra de 150 datos que seguen unha distribución χ-cadrado con tres grados de liberdade. A liña sólida representa a función densidade teórica mentres que a liña punteada representa a estimación tipo núcleo obtida para os diferentes valores do parámetro h.

e a liña punteada representa a estimación tipo núcleo.

Á vista das Figuras 2.1 e 2.2 observamos que unha boa ventá h depende da distribución dos datos, é dicir, non hai unha ventá óptima para todas as distribucións. Neste caso, mentres para a mostra que provén dunha distribución normal o valor $h = 0.3$ parece xerar unha boa aproximación, para a χ-cadrado ese mesmo valor semella demasiado pequeno.

No Capítulo 3 abordaremos a elección da ventá h óptima, pero primeiro débense introducir os diferentes criterios de erro que se utilizarán na procura da mellor h.

2.2. Propiedades do estimador tipo núcleo

Para saber canto de boa é a estimación da función densidade empregando o estimador tipo núcleo, basearémonos en varios criterios de erro, tanto puntuais como en toda a recta real. En concreto, o criterio de erro puntual que se utilizará será o erro cadrático medio que definimos a continuación, pero debemos recordar previamente a definición de esperanza.

Definición 2.1. Dada unha variable aleatoria continua X con función de densidade f, defínese a súa **esperanza** mediante a seguinte expresión:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

Definición 2.2. Sexa $\widehat{\theta}$ un estimador dun parámetro descoñecido θ. O **erro cadrático medio** (en adiante ECM) dun estimador $\widehat{\theta}$ con respecto ao parámetro descoñecido θ defínese como

$$\mathrm{ECM}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta} - \theta)^2.$$

Podemos descompoñer o ECM en termos de varianza[1] e sesgo[2] do seguinte xeito:

$$\begin{aligned}
\mathrm{ECM}[\widehat{\theta}] &= \mathbb{E}[\widehat{\theta} - \theta]^2 = \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right) + \left(\mathbb{E}(\widehat{\theta}) - \theta\right)\right]^2 = \\
&= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right)^2 + \left(\mathbb{E}(\widehat{\theta}) - \theta\right)^2 + 2\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right)\left(\mathbb{E}(\widehat{\theta}) - \theta\right)\right] = \\
&= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}(\widehat{\theta}) - \theta\right)^2\right] + 2\mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right)\left(\mathbb{E}(\widehat{\theta}) - \theta\right)\right] =
\end{aligned}$$

[1] Defínese como varianza dunha variable aleatoria á esperanza do cadrado da desviación de dita variable respecto á súa media. É dicir, $Var(X) = \mathbb{E}[(X - \mu)^2]$

[2] Definimos como sesgo de un estimador á diferencia entre a súa esperanza e o valor que estima. É dicir $Sesgo(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta$.

$$= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right)^2\right] + \mathbb{E}\left[\mathbb{E}(\widehat{\theta}) - \theta\right]^2 + 2\left(\mathbb{E}(\widehat{\theta}) - \theta\right)\mathbb{E}\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right) =$$
$$= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right)^2\right] + \mathbb{E}\left[\mathbb{E}(\widehat{\theta}) - \theta\right]^2 = \mathrm{Var}\left(\widehat{\theta}\right) + \mathrm{Sesgo}^2\left(\widehat{\theta}\right). \tag{2.3}$$

No caso que tratamos temos unha estimación da función de densidade $\widehat{f_h}(x)$. Coa descomposición (2.3) podemos calcular o ECM coas expresións da varianza e do sesgo. É dicir, que se obtemos a varianza e esperanza de $\widehat{f_h}(x)$ teremos o seu ECM.

Partindo da ecuación (2.2) e tendo en conta que $\{x_1, ..., x_n\}$ é unha mostra aleatoria simple podemos dicir que

$$\mathbb{E}[\widehat{f_h}(x)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[K_h(x - x_i)] =$$
$$= \frac{1}{n}\cdot n \cdot \mathbb{E}[K_h(x - X)] = \mathbb{E}[K_h(x - X)] =$$
$$= \int_{-\infty}^{+\infty} K_h(x - y)f(y)dy. \tag{2.4}$$

Unha vez coñecemos a expresión da esperanza, para calcular o sesgo introducimos a seguinte expresión:

$$(f * g)(x) = \int f(x - y)g(y)dy,$$

onde $*$ denota o operador de convolución, xa que nos permite escribir

$$\mathrm{Sesgo}(\widehat{f_h}(x)) = \mathbb{E}[\widehat{f_h}(x)] - f(x) = (K_h * f)(x) - f(x).$$

Para calcular a expresión da varianza séguese un proceso semellante:

$$\mathrm{Var}(\widehat{f_h}(x)) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)\right) = \frac{n}{n^2}\mathrm{Var}\left(K_h(x - X)\right)$$
$$= \frac{1}{n}\left(\mathbb{E}[(K_h(x - X))^2] - \mathbb{E}^2[K_h(x - X)]\right)$$
$$= \frac{1}{n}\left(\int (K_h(x - y))^2 f(y)dy - \left(\int K_h(x - y)f(y)dy\right)^2\right)$$
$$= \frac{1}{n}\left[(K_h^2 * f)(x) - (K_h * f)^2(x)\right]. \tag{2.5}$$

Finalmente, obtidas as expresións do sesgo e da varianza, temos a expresión do erro cadrático medio da estimación tipo núcleo da función de densidade, que sería:

$$\begin{aligned}\mathrm{ECM}(\widehat{f_h}(x)) &= [\mathrm{Sesgo}(\widehat{f_h}(x))]^2 + \mathrm{Var}(\widehat{f_h}(x)) = \\ &= \big[(K_h * f)(x) - f(x)\big]^2 + \frac{1}{n}\big[(K_h^2 * f)(x) - (K_h * f)^2(x)\big] \qquad (2.6)\end{aligned}$$

Por certo lado, dado que o ECM está fixado nun punto x, poderiamos pensar nun criterio de erro global que medira a distancia entre as funcións $\widehat{f_h}(\cdot)$ e f ao longo de toda a recta real. Nesta liña atópase o **erro cadrático integrado**, coñecido polas súas siglas en inglés ISE (*Integrated Squared Error*) e defínese como:

$$\mathrm{ISE}(\widehat{f_h}) = \int_{-\infty}^{+\infty} \big[\widehat{f_h}(x) - f(x)\big]^2 dx.$$

O problema do ISE é que depende da verdadeira (e descoñecida) densidade f, do estimador utilizado e do tamaño da mostra. Incluso con estas tres cantidades fixas, o ISE é unha función dunha realización en n puntos. É por iso que resultará máis apropiado utilizar como criterio de erro o seu valor esperado. Deste xeito definimos o **erro cadrático medio integrado**, coñecido polas súas siglas en inglés, MISE (*Mean Integrated Squared Error*), que se define como:

$$\mathrm{MISE}[\widehat{f_h}] = \mathbb{E}[\mathrm{ISE}(\widehat{f_h})] = \mathbb{E}\left[\int_{-\infty}^{+\infty} \big[\widehat{f_h}(x) - f(x)\big]^2 dx\right].$$

Se permutamos a esperanza coa integral, obtemos unha expresión do MISE en función do ECM. En concreto:

$$\mathrm{MISE}[\widehat{f_h}] = \int_{-\infty}^{+\infty} \mathbb{E}[\widehat{f_h}(x) - f(x)]^2 dx = \int_{-\infty}^{+\infty} \mathrm{ECM}[\widehat{f_h}(x)] dx. \qquad (2.7)$$

Introducindo agora a seguinte notación para dúas sucesións de números reais a_n e b_n,

$$a_n = o(b_n),\ n \to \infty \Leftrightarrow \lim_{n\to\infty} |\frac{a_n}{b_n}| = 0$$

e facendo uso do desenrolo en serie de Taylor obteremos expresións máis sinxelas para o ECM e o MISE que dependerán dunha forma máis clara do parámetro h.

Deste xeito, partindo de (2.4) e utilizando o desenrolo de Taylor de $f(x - hz)$ sobre x chegamos á seguinte expresión do sesgo:

$$\mathbb{E}[\widehat{f}(x,h)] - f(x) = \frac{1}{2}h^2\mu_2(K)f^{(2)}(x) + o(h^2) \qquad (2.8)$$

sendo $\mu_2(g) = \int_{-\infty}^{\infty} z^2 g(z)dz$.

Para a varianza, partindo de (2.5)

$$\begin{aligned}\mathrm{Var}(\widehat{f}_h(x)) &= \frac{1}{nh}\int K(z)^2 f(x-hz)dz - \frac{1}{n}[\mathbb{E}\widehat{f}(x,h)]^2 \\ &= \frac{1}{nh}\int K(z)^2 dz f(x) + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}R(K)f(x) + o\left(\frac{1}{nh}\right) \end{aligned} \tag{2.9}$$

con $R(g) = \int_{-\infty}^{\infty} g(z)^2 dz$. Xuntando agora a ecuación (2.9) co cadrado de (2.8) obtemos

$$ECM(\widehat{f}_h(x)) = \frac{1}{nh}R(K)f(x) + \frac{1}{4}h^4\mu_2(K)^2 f^{(2)}(x)^2 + o\left(\frac{1}{nh} + h^4\right)$$

e integrando esta expresión obtemos

$$MISE[\widehat{f}_h] = AMISE[\widehat{f}_h] + o\left(\frac{1}{nh} + h^4\right)$$

Na práctica será de interese coñecer o comportamento límite do MISE. Este límite chámase AMISE, que son as siglas en inglés de *Asymptotic Mean Integrated Squared Error*, e vén dado por:

$$\mathrm{AMISE}[\widehat{f}_h] = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2 R(f^{(2)}) \tag{2.10}$$

onde $g^{(i)}$ representa a derivada de orde i dada unha certa función g.

Dada a fórmula do AMISE, veremos a continuación que, na ecuación (2.1), a elección de K non é moi importante. Como dixemos antes, a función núcleo verifica $\int_{\mathbb{R}} K(x)dx = 1$ e é simétrica en torno ao cero. Naturalmente, existen moitas funcións que verifican estas dúas propiedades, polo que é natural preguntarse que funcións núcleo son mellores. Para ver isto, partimos da fórmula do AMISE dada na expresión (2.10) e seguindo [7] consideramos o núcleo $K_\delta(\cdot) = K(\cdot/\delta)/\delta$ con $\delta_0 = (R(K)/\mu_2(K)^2)^{1/5}$. Nesta situación:

$$\begin{aligned}R(K_{\delta_0}) &= \int K_{\delta_0}(z)^2 dz = \int \frac{1}{\delta_0^2}K(z/\delta_0)^2 dz = \\ &= \frac{1}{\delta_0^2}\int \delta_0 K(w)^2 dw = \frac{1}{\delta_0}\int K(w)^2 dw = \frac{1}{\delta_0}R(K)\end{aligned}$$

e

$$\mu_2(K_{\delta_0}) = \int z^2 K_{\delta_0}(z)dz = \int z^2 \frac{1}{\delta_0}K(z/\delta_0)dz =$$

$$= \int (\delta_0 w)^2 \frac{1}{\delta_0} K(w) \delta_0 dw = \delta_0^2 \int w^2 K(w) dw =$$
$$= \delta_0^2 \mu_2(k)$$

onde consideramos o cambio de variable $w = z/\delta_0$. Con isto obtemos unha expresión de AMISE que separa a dependencia de h e de K:

$$\begin{aligned}
\text{AMISE}[\widehat{f}_h] &= \frac{1}{nh} R(K_{\delta_0}) + \frac{1}{4} h^4 \mu_2(K_{\delta_0})^2 R(f^{(2)}) = \\
&= \frac{1}{nh} \frac{R(K)}{\delta_0} + \frac{1}{4} h^4 (\mu_2(K)\delta_0^2)^2 R(f^{(2)}) = \\
&= \frac{1}{nh} \left(\frac{\mu_2(K)^2}{R(K)} \right)^{1/5} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 \left(\frac{R(K)}{\mu_2(K)^2} \right)^{4/5} R(f^{(2)}) = \\
&= C(K) \left[\frac{1}{nh} + \frac{1}{4} h^4 R(f^{(2)}) \right]
\end{aligned}$$

sendo $C(K) = \left[R(K)^4 \mu_2(K)^2\right]^{1/5}$.

Dada a expresión anterior resulta máis sinxelo abordar o problema de determinar a función núcleo óptima, xa que só necesitaremos escoller o K que minimice $C(K)$ coas restricións:

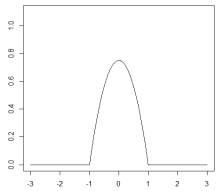
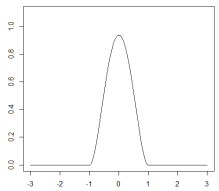
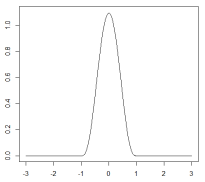
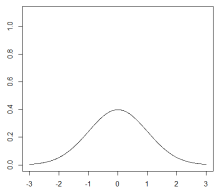
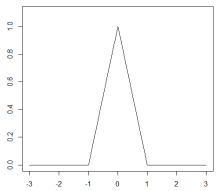
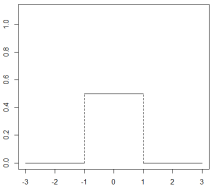
- $\int_{\mathbb{R}} K(x)dx = 1$,
- $\int_{\mathbb{R}} xK(x)dx = 0$,
- $\int_{\mathbb{R}} x^2 K(x)dx = a^2 < \infty$

con $K(x) \geq 0$ para todo $x \in \mathbb{R}$. Nesta situación obtense como núcleo óptimo o núcleo de Epanechnikov, que ten a seguinte expresión:

$$K_{epa}(x) = \frac{3}{4}[1 - x^2] I_{|x|<1}$$

onde $I_{|x|<1}$ é a función que toma valor 1 se $x \in (-1, 1)$ e valor 0 noutro caso.

Comparemos agora o núcleo óptimo que acabamos de obter con outros núcleos que tamén son utilizados na práctica. Para iso faremos a operación $[C(K_{epa})/C(K)]^{5/4}$, que recibe o nome de eficiencia de K con respecto a K_{epa}. Os valores de dita operación para as principais funcións núcleo están representados na terceira columna da Táboa 2.1. Ademais, na cuarta columna da Táboa 2.1 dáse unha representación gráfica das principais funcións núcleo que se poderían considerar.

Núcleo	Expresión matemática	Eficiencia	Representación gráfica
Epanechnikov	$\frac{3}{4}(1-x^2)I_{\|x\|<1}$	1.000	
Biweight	$\frac{15}{16}(1-x^2)^2 I_{\|x\|<1}$	0.994	
Triweight	$\frac{35}{32}(1-x^2)^3 I_{\|x\|<1}$	0.987	
Normal	$(2\pi)^{-1/2}e^{-x^2/2}$	0.951	
Triangular	$(1-\|x\|)I_{\|x\|<1}$	0.986	
Uniforme	$\left(\frac{1}{2}\right) I_{\|x\|<1}$	0.930	

Táboa 2.1: Eficiencia dos núcleos máis utilizados con respecto ao núcleo de Epanechnikov.

Na Táboa 2.1 vese que non son moi grandes as diferenzas entre a eficiencia do núcleo de Epanechnikov e os demais núcleos, xa que tódolos valores están próximos ao valor 1. É por iso que a elección do núcleo non será moi importante para a estimación tipo núcleo da función de densidade.

Capítulo 3

Selector do parámetro ventá

Como anticipamos no Capítulo 2 e comprobamos graficamente nas Figuras 2.1 e 2.2, escoller un bo parámetro ventá h resulta de vital importancia para obter unha boa estimación tipo núcleo da función de densidade.

Nesta liña irá dirixido este capítulo, onde veremos tres métodos de selección e optimización do parámetro h que estarán fundados nos diferentes criterios de erro vistos no Capítulo 2. Unha vez máis, X será unha variable aleatoria continua, con función de densidade asociada f, e traballaremos cunha mostra aleatoria simple $\{x_1, ..., x_n\}$ procedente da variable X.

Dadas as fórmulas do ECM (ver ecuación (2.6)) e do MISE (ver ecuación (2.7)) obtidas anteriormente, é difícil ver a influencia da ventá h na estimación tipo núcleo. Para o desenrolo deste capítulo seguiremos o Capítulo 3 do libro [7], e impoñeremos as mesmas condicións que se establecían no Capítulo 2 sobre a variable de interese X e sobre a súa función de densidade f, así como sobre o parámetro h e a función kernel K. Recordamos a continuación ditas condicións:

- Existe $f^{(2)}$, que ademais é continua, de cadrado integrable e monótona nos intervalos $(-\infty, M)$ e (M, ∞) para algún $M > 0$.
- A secuencia dos posibles valores positivos de h, que denotaremos por h_n, tende a cero, pero cunha orde menor do que o fai a sucesión $\frac{1}{n}$. É dicir, $\lim_{n\to\infty} h_n = 0$ e $\lim_{n\to\infty} n \cdot h_n = 0$.
- O núcleo K é unha función de densidade limitada, simétrica en torno á orixe e con

momento de orde catro finito.

Derivando a ecuación (2.10)

$$\text{AMISE}[\widehat{f}_h] = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f^{(2)})$$

con respecto ao parámetro h obtemos:

$$\frac{\partial AMISE[\widehat{f}_h]}{\partial h} = \frac{-nR(K)}{(nh)^2} + h^3\mu_2(K)^2R(f^{(2)})$$

e igualando esa expresión a cero, chegamos a que a ventá óptima para o AMISE é

$$h_{AMISE} = \left(\frac{R(K)}{\mu_2(K)^2R(f^{(2)})n}\right)^{1/5} \tag{3.1}$$

onde $R(g) = \int_{-\infty}^{+\infty} g(z)^2dz$ e $\mu_2(g) = \int_{-\infty}^{+\infty} z^2g(z)dz$.

Notemos que na ecuación (3.1) todo é coñecido coa excepción de $R(f^{(2)})$, xa que non coñecemos a expresión de $f^{(2)}$. Polo tanto, será fundamental coñecer un bo estimador de $R(f^{(2)})$ para poder obter un bo parámetro h. Nesta situación temos varias formas de afrontar o problema. A continuación veremos tres posibilidades.

3.1. Regra do polgar

Este método de selección do parámetro ventá foi introducido por Silverman en [5] e ten por obxectivo propoñer un estimador rápido e sinxelo do parámetro ventá h. Para isto a **regra do polgar** baséase na suposición de que a variable X de estudo segue unha distribución normal con media e varianza descoñecidas, e consiste en estimar a ventá h_{AMISE} supoñendo que a variable de interese segue unha distribución normal. Neste contexto, se f se corresponde coa función de densidade asociada a unha distribución normal con media 0 e varianza σ^2 pasamos a ter unha expresión para $R(f^{(2)})$, que antes era descoñecido. Neste caso,

$$f(x) = \frac{1}{\sigma\sqrt{2}\pi}e^{-\frac{x^2}{2\sigma^2}}$$

e derivando dúas veces obtemos

$$f^{(2)}(x) = \frac{1}{\sigma^3\sqrt{2\pi}}\left(\left(\frac{x}{\sigma}\right)^2 - 1\right)e^{-\frac{x^2}{2\sigma^2}}.$$

Recordando agora $R(g) = \int_{-\infty}^{+\infty} g(x)^2 dx$ teremos que:

$$R(f^{(2)}) = \int_{\mathbb{R}} f^{(2)}(x)^2 dx = \int_{\mathbb{R}} \frac{1}{\sigma^6 2\pi} \left(\left(\frac{x}{\sigma} \right)^2 - 1 \right)^2 e^{-\frac{x^2}{\sigma^2}} dx = \frac{3}{8\sigma^5 \sqrt{\pi}}.$$

Con isto temos unha expresión do termo descoñecido $R(f^{(2)})$ en función de σ. Substituíndo en (3.1) obtemos,

$$h_{\text{AMISE}} = \left[\frac{8\sqrt{\pi} R(K)}{3\mu_2(K)^2 n} \right]^{1/5} \sigma.$$

Con esta expresión só teremos pendente aproximar o valor de σ para obter unha estimación do parámetro h. Aproximaremos σ por $\widehat{\sigma}$, que normalmente será a cuasi desviación típica mostral, ou versións máis robustas como o rango intercuantílico estandarizado. As expresión para a cuasi desviación típica mostral e para o rango intercuantílico estandarizado veñen dadas respectivamente por:

- $\widehat{\sigma}_s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $\widehat{\sigma}_{RIQ} = \frac{rango \quad intercuantlico}{\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)} = \frac{Q_3 - Q_1}{\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)}$

onde Φ^{-1} representa a inversa da función de distribución normal estándar e Q_1 e Q_3 denota os cuantís de orde 0.25 e 0.75, respectivamente, da mostra $\{x_1, ..., x_n\}$. Nótese que grazas a $\widehat{\sigma}_{RIQ}$ obteremos estimacións máis robustas da desviación típica que se empregamos $\widehat{\sigma}_s$, posto que o emprego dos cuantís suaviza o efecto de posibles datos atípicos.

Este método facilita unha primeira e rápida aproximación do parámetro h, e proporciona resultados realmente coherentes e satisfactorios cando a distribución dos datos se aproxima a unha distribución normal.

3.2. Selector de validación cruzada de mínimos cadrados

Este método, que está baseado nas ideas que plasman Rudemo en [3] e Bowman en [1], ten por obxectivo atopar un parámetro h óptimo que minimice o MISE. Partindo da

ecuación (2.7)

$$\text{MISE}[\widehat{f}_h] = \int_{-\infty}^{+\infty} \mathbb{E}[\widehat{f}_h(x) - f(x)]^2 dx = \int_{-\infty}^{+\infty} \text{ECM}[\widehat{f}_h(x)]dx,$$

resulta inmediata a obtención da expresión

$$\begin{aligned}\text{MISE}\left[\widehat{f}_h\right] &= \int \mathbb{E}[\widehat{f}_h(x) - f(x)]^2 dx \\ &= \mathbb{E}\left[\int \widehat{f}_h(x)^2 dx\right] - 2\mathbb{E}\left[\int \widehat{f}_h(x) f(x) dx\right] + \int f(x)^2 dx. \end{aligned} \tag{3.2}$$

Dado que tratamos de atopar un h óptimo e o último termo da fórmula non depende do parámetro h, minimizar (3.2) será equivalente a minimizar:

$$\arg\min_h \text{MISE}\left[\widehat{f}_h\right] = \arg\min_n \mathbb{E}\left[\int \widehat{f}_h(x)^2 dx - 2\int \widehat{f}_h(x) f(x) dx\right], \tag{3.3}$$

e ademais podemos identificar o segundo termo da parte dereita da ecuación como a esperanza do estimador $\widehat{f}_h$, que se pode estimar empregando a media mostral. Nesta situación o noso problema está reducido a estimar a función de densidade en cada punto x_i, e logo calcular a súa media, pero non podemos empregar ese mesmo punto x_i, pois estariamos empregando o mesmo dato para estimar e para avaliar a estimación obtida. Introdúcese neste contexto o concepto de **validación cruzada**. Este método consiste en empregar toda a mostra $\{x_1, ..., x_n\}$ menos o dato x_i para estimar a función de densidade e logo avaliala no punto x_i.

Esta idea lévanos á seguinte ecuación, que é un estimador insesgado da expresión (3.3).

$$\text{VC}(h) = \int \widehat{f}_h(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_h^{-i}(x_i) \tag{3.4}$$

onde $\widehat{f}_h^{-i}$ representa un estimador tipo "*leave-one-out*" que se define como

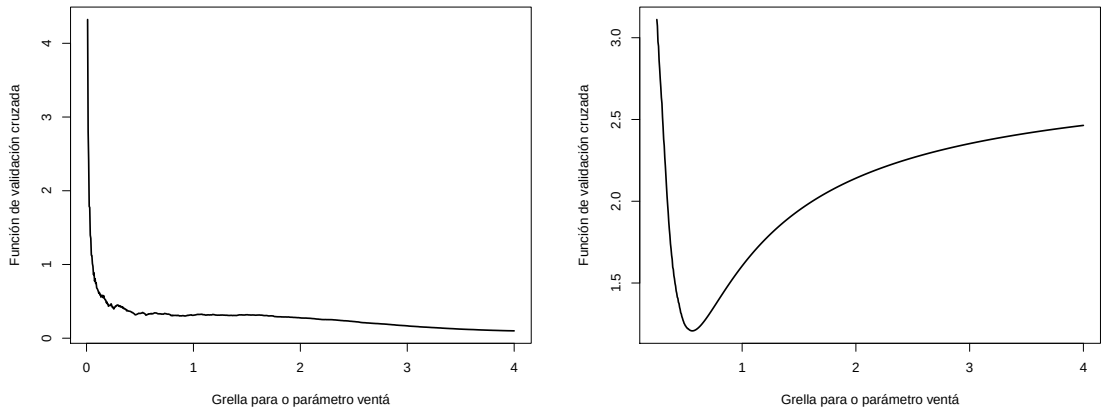
$$\widehat{f}_h^{-i}(x) = \frac{1}{n-1}\sum_{j\neq i}^{n} K_h(x_i - x_j).$$

Polo tanto, o selector $\widehat{h}_{VC}$ proposto polo método de validación cruzada será o valor de h que minimice a función VC definida en (3.4). É dicir,

$$\widehat{h}_{VC} = \arg\min_h \text{VC}(h).$$

Cabe destacar que nalgunhas ocasións a función de validación cruzada ten máis dun mínimo local. Neste caso pódese dar un fenómeno de confusión, polo que é preciso ter

precaución ao atopar $\widehat{h}_{VC}$ na práctica. Este fenómeno de confusión dáse cando temos o mínimo absoluto da función de validación cruzada fora da grella de ventás considerada na cal avaliamos dita función. O seguinte exemplo reflexa este caso. Podemos ver na Figura 3.1, dúas situacións. Na primeira, Figura 3.1 (a), vemos que a función de validación cruzada non ten un mínimo absoluto, e o mínimo local que posúe está influenciado polo intervalo que collamos para determinar a h. Deste xeito, se o intervalo de ventás onde queremos avaliar a función de validación cruzada é $(0, x)$, o mínimo para dito intervalo será x, pero non temos un mínimo global para $h > 0$, como se pretende no método de validación cruzada. En cambio, para a Figura 3.1 (b), vemos que a función de validación cruzada sí que ten un mínimo absoluto no intervalo de ventás considerado.



(a) Función de validación cruzada fronte a parámetro ventá dunha distribución normal estándar

(b) Función de validación cruzada fronte a parámetro ventá dunha distribución normal con media 0 e desviación típica 0.1

Figura 3.1: Representación da función de validación cruzada de dúas distribucións fronte ao parámetro ventá.

3.3. Selector plug-in

A regra do polgar asume a normalidade da función de densidade f para obter o selector do parámetro h. No método **plug-in** asumiremos de novo normalidade, pero esta vez nunha etapa máis avanzada. Isto producirá unha menor influenza da normalidade na selección do parámetro ventá.

Partindo da fórmula (3.1) e como mencionabamos anteriormente, tódolos termos son coñecidos coa excepción de $R(f^{(2)})$. Recordemos que por definición tiñamos $R(g) = \int_{-\infty}^{+\infty} g(x)^2 dx$, e aplicando a fórmula de integración por partes obtemos:

$$R(f^{(2)}) = \int f^{(2)}(x)^2 dx = (-1)^2 \int f^{(4)}(x) f(x) dx = \mathbb{E}[f^{(4)}(x)].$$

onde $f^{(r)}$ representa a $r-$ésima derivada da función de densidade f.

Se introducimos a notación

$$\psi_r = \int f^{(r)}(x) f(x) dx = \mathbb{E}[f^{(r)}(x)]$$

obtemos que $R(f^{(2)}) = \mathbb{E}[f^{(4)}(x)] = \psi_4$, e substituíndo na ecuación (3.1) o termo $R(f^{(2)})$ por ψ_4 chegamos á expresión

$$h_{AMISE} = \left(\frac{R(K)}{\mu_2(K)^2 \psi_4 n} \right)^{1/5}. \tag{3.5}$$

Mais de igual modo que $R(f^{(2)})$, o termo ψ_4 é descoñecido, polo que deberemos de novo buscar unha estimación do mesmo. Dita estimación vén dada por:

$$\widehat{\psi}_4(g) = \frac{1}{n} \sum_{i=1}^{n} \widehat{f}_g^{\,(4)}(x_i),$$

onde g será un novo parámetro ventá, que en xeral será distinta de h. Recordando a expresión do estimador tipo núcleo, é dicir, $\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - x_i)$, e substituíndo na fórmula anterior obtemos

$$\widehat{\psi}_4(g) = \frac{1}{n^2} \sum_{i=1}^{n} L_g^{(4)}(x - x_i).$$

onde L é unha función núcleo que pode ser distinta da función núcleo K. A proposta do estimador $\widehat{\psi_4}$ pode verse en [7], (páxina 67).

Deste xeito, substituíndo na ecuación (3.5) o termo descoñecido ψ_4 polo seu correspondente estimador $\widehat{\psi}_4(g)$, obtemos a aproximación de h polo método plug-in, que denotaremos por $\widehat{h}_{\mathrm{PI}}$:

$$\widehat{h}_{\mathrm{PI}} = \left(\frac{R(K)}{\mu_2(K)^2 \widehat{\psi}_4 n} \right)^{1/5}. \tag{3.6}$$

Mais notemos que a función $\widehat{h}_{\mathrm{PI}}$ depende da elección do parámetro g, que é de novo descoñecido. Para realizar unha aproximación de dito parámetro debemos ter que conta

o erro cadrático medio asintótico, que denotaremos por AMSE debido ás súas siglas en inglés, do estimador $\widehat{\psi}_4$. Pode verse en [7] (ver páxina 71) que a ventá óptima para $\widehat{\psi}_4(g)$ seguindo este criterio de erro é:

$$g_{\widehat{\psi}_4(g)} = \left[\frac{2K^{(4)}(0)}{-\mu_2(K)\psi_6 n} \right]^{1/7}. \tag{3.7}$$

Pero con este procedemento de selección do parámetro vemos que g ten o mesmo problema que para definir $\widehat{h}_{\mathrm{PI}}$, xa que depende de ψ_6 o cal é descoñecido. Se tratamos de estimar ψ_6, a ventá óptima para dita aproximación dependerá de ψ_8 do seguinte xeito:

$$g_{\widehat{\psi}_6(g)} = \left[\frac{2K^{(6)}(0)}{-\mu_2(K)\psi_8 n} \right]^{1/9}. \tag{3.8}$$

É dicir, a ventá óptima para a estimación de ψ_r depende de ψ_{r+2}. Para solucionar este problema acostúmase estimar ψ_r, para un certo r, mediante a regra do polgar.

Nesta situación podemos preguntarnos cantas estimacións de funcións ψ_r debemos realizar antes de asumir normalidade para estimar a función ψ_{r+2}. Sobre este problema non existe un número exacto que diga cantas iteracións se deben facer, pero consideracións teóricas determinan que deben facerse, polo menos, dúas iteracións. É dicir, no caso de facer dúas iteracións deberiamos asumir a normalidade de ψ_8 ou o que é o mesmo, asumiremos que a función de densidade f segue uns distribución normal para estimar a súa derivada de orde 8.

Debemos ter en conta que dada f unha función de densidade dunha variable normal con varianza σ^2, en [7] pode verse que:

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}. \tag{3.9}$$

A continuación preséntase un exemplo do proceso para calcular o selector proposto pola regra plug-in realizando dúas iteracións. Dito selector foi proposto por Sheather e Jones e pode consultarse en [4] e divídese en varios pasos:

Paso 1: Dado que imos realizar dúas iteracións, o primeiro será asumir a normalidade de ψ_8. Usando a expresión (3.9), a estimación de ψ_8 ven dada por

$$\widehat{\psi}_8 = \frac{105}{32\pi^{1/2}\widehat{\sigma}^9}$$

sendo $\widehat{\sigma}$ a estimación de σ.

Paso 2: Tendo en conta agora a ecuación (3.8) e substituíndo nela o valor descoñecido de ψ_8 pola súa estimación $\widehat{\psi}_8$ obtida no paso anterior, podemos calcular o valor de g_1 que ven dado por

$$g_1 = \left[\frac{-2K^{(6)}(0)}{\mu_2(K)\widehat{\psi}_8 n}\right]^{1/9}.$$

e estimamos ψ_6 mediante $\widehat{\psi}_6(g_1)$.

Paso 3: De novo procedemos como no paso anterior, unha vez obtido o valor de $\widehat{\psi}_6(g_1)$ e tendo en conta a ecuación (3.7), substituímos na mesma o valor de ψ_6 pola súa estimación $\widehat{\psi_6}$. Desta maneira estimamos ψ_4 mediante $\widehat{\psi}_4(g_2)$ con

$$g_2 = \left[\frac{-2K^{(4)}(0)}{\mu_2(K)\widehat{\psi}_6(g_1) n}\right]^{1/7}.$$

Paso 4: Finalmente, substituíndo na ecuación (3.6) o valor de $\widehat{\psi}_4$ obtido no paso anterior, obtemos o sector do parámetro ventá h proposto polo método plug-in:

$$\widehat{h}_{\mathrm{PI}} = \left[\frac{R(K)}{\mu_2(K)^2\widehat{\psi}_4(g_2) n}\right]^{1/5}.$$

Capítulo 4

Estudo de simulación

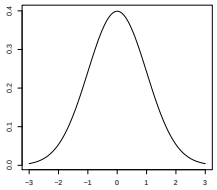
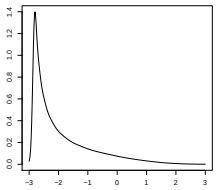
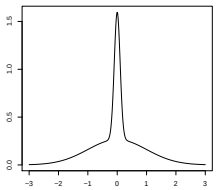
Neste capítulo compáranse, mediante un estudo de simulación, os diferentes métodos de selección do parámetro ventá h que se presentaron no Capítulo 3. Para isto programamos os métodos de polgar e de validación cruzada, aínda que os resultados que aparecen ao longo deste capítulo referentes á validación cruzada non foron obtidos con dita programación propia, senón mediante a función *h.ucv* que se atopa no paquete *kedd* de R. Así mesmo, empregouse a función *dpik* da libraría *KernSmooth* para obter os resultados do método plug-in. Pódese atopar máis información do paquete *kedd* na páxina https://cran.r-project.org/web/packages/kedd/kedd.pdf e do paquete *KernSmooth* na seguinte ligazón https://cran.r-project.org/web/packages/KernSmooth/KernSmooth.pdf.

Para iso empregarase o método de Monte Carlo, que se basea en xerar moitas mostras e operar con elas co obxectivo de ver como se comporta cada un dos métodos para as distintas mostras xeradas. En concreto, os pasos a seguir son os seguintes:

1. Xerar unha mostra aleatoria simple da variable de interese.
2. Estimar para a mostra anterior o parámetro ventá conforme a cada un dos diferentes métodos presentados no Capítulo 3.
3. Calcular o ISE para cada método de selección do parámetro ventá.
4. Repetir os pasos 1-3 anteriores un número M grande de veces.

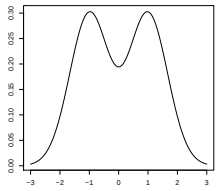
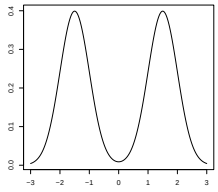
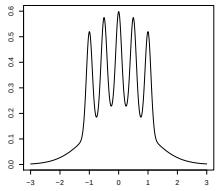
Deste xeito, podemos concluír que método de selección de ventá se axusta mellor á distribución de interese.

En concreto, para este capítulo, xeramos con  1000 mostras aleatorias simples de varias distribucións diferentes e con tamaños de mostra $n = 100$, $n = 250$ e $n = 500$, co obxectivo de ver cal é a tendencia a medida que aumentamos o número de datos. Ademais, utilizaranse os modelos que aparecen nas Táboas 4.1 e 4.2, do conxunto de funcións de densidade empregadas en [2] (ver páxina 717). Nestas táboas introdúcese unha nova notación pola cal se referencia, por exemplo, o modelo 3 das distribucións de [2] como M3. Nótese que $N(\mu, \sigma^2)$ representa unha variable normal con media μ e varianza σ^2.

Modelo	Expresión Matemática	Gráfica
M1	N(0,1)	
M3	$\sum_{i=0}^{7} \frac{1}{8} N\left(3\left(\left(\frac{2}{3}\right)^{i} - 1\right), \left(\frac{2}{3}\right)^{2i}\right)$	
M4	$\frac{2}{3}N(0,1) + \frac{1}{3}N\left(0, \left(\frac{1}{10}\right)^{2}\right)$	

Táboa 4.1: Modelos considerados en [2] xunto coa súa expresión en función de distribucións normais así como a representacións das correspondentes funcións de densidade.

Estas funcións de densidade que eliximos teñen por obxectivo ver como se comportan os diferentes selectores de ventá en diferentes escenarios. Deste xeito eliximos formas suaves, como por exemplo a función de densidade normal, que é o modelo M1, e formas máis rugosas como por exemplo a "garra", que se corresponde co modelo M10.

Modelo	Expresión Matemática	Gráfica
M6	$\frac{1}{2}N\left(-1,\left(\frac{2}{3}\right)^2\right)+\frac{1}{2}N\left(1,\left(\frac{2}{3}\right)^2\right)$	
M7	$\frac{1}{2}N\left(-\frac{3}{2},\left(\frac{1}{2}\right)^2\right)+\frac{1}{2}N\left(\frac{3}{2},\left(\frac{1}{2}\right)^2\right)$	
M10	$\frac{1}{2}N(0,1)+\sum_{i=0}^{4}\frac{1}{10}N\left(\frac{i}{2}-1,\left(\frac{1}{10}\right)^2\right)$	

Táboa 4.2: Modelos considerados en [2] xunto coa súa expresión en función de distribucións normais así como a representacións das correspondentes funcións de densidade.

Con estas mostras realizamos o procedemento descrito anteriormente e, facendo a media dos ISE obtidos para cada un dos métodos de selección do parámetro ventá, chegamos ao MISE do estimador tipo núcleo conforme a ese método de selección. Deste xeito podemos comparar os erros obtidos ao empregar cada un dos métodos de selección de h, e á súa vez concluír que método é mellor estimar o parámetro ventá para cada unha das distribucións consideradas. Ademais, ao ir incrementando o tamaño da mostra podemos ver o efecto que ten dito incremento sobre o valor da estimación do parámetro h e sobre os erros ISE e MISE.

Preséntase a continuación, na Táboa 4.3, os datos obtidos tras realizar o proceso explicado anteriormente en  partindo da semente 123. En dita táboa indícase o MISE obtido para cada unha das distribucións de datos utilizadas, con cada tamaño de mostra, e para cada método de selección do parámetro h. Nótese que o código empregado para obter

a Táboa 4.3 pode consultarse no Anexo deste documento.

Distribución	Tamaño	Polgar	VC	Plug-in	AMISE
M1	n=100	0.3342	0.3377	0.3313	0.3275
	n=250	0.3115	0.3167	0.3125	0.3117
	n=500	0.3012	0.3058	0.3033	0.3032
M3	n=100	0.1250	0.6040	0.2655	0.7538
	n=250	0.1794	0.6511	0.4176	0.6849
	n=500	0.2571	0.6427	0.5476	0.6490
M4	n=100	0.9355	0.9243	0.9231	0.9088
	n=250	0.8808	0.7831	0.8023	0.7725
	n=500	0.8483	0.7278	0.7411	0.7211
M6	n=100	0.2698	0.2837	0.2777	0.2825
	n=250	0.2656	0.2673	0.2641	0.2635
	n=500	0.2590	0.2564	0.2549	0.2539
M7	n=100	0.3674	0.3696	0.3611	0.3620
	n=250	0.3625	0.3353	0.3341	0.3300
	n=500	0.3499	0.3173	0.3179	0.3145
M10	n=100	0.4401	0.4509	0.4347	0.6061
	n=250	0.4016	0.4376	0.4046	0.4944
	n=500	0.3906	0.4107	0.4247	0.4490

Táboa 4.3: Error cadrático medio integrado (MISE) obtido tras o estudo de simulación para os modelos M1, M3, M4, M6, M7 e M10 considerados por [2] para cada un dos méritos considerados regra do polgar (baixo o título `Polgar`). Na última columna aparece o MISE da ventá AMISE, que debería ser o MISE máis pequeno.

Como podemos ver na Táboa 4.3, na maioría de modelos os MISE fanse máis pequenos a medida que aumentamos o tamaño de mostra. Isto é algo esperable que ocorre en todos os casos con excepción do modelo M3. Este caso é algo especial xa que se produce o que se chama efecto fronteira. Este fenómeno ocorre cando temos gran cantidade de datos preto da fronteira do intervalo no cal se define a función de densidade. Podemos ver na Táboa 4.1 que o modelo M3 concentra a maior cantidade de datos xusto onde comeza o intervalo de definición da función de densidade, o que fomenta este problema do efecto fronteira. Dada a complexidade do modelo non conseguimos mellorar os erros obtidos aumentando o tamaño da mostra, e para unha satisfactoria estimación da función de densidade necesitaríamos

utilizar técnicas específicas. Por outro lado vemos que tódolos resultados dos MISE se sitúan en torno ao resultado da última columna, que fai referencia ao MISE da ventá h_{AMISE}. Isto é algo bo xa que a ventá h_{AMISE} é, asintóticamente, a mellor ventá posible.

Por outro lado, podemos dicir que non resulta claro o feito de que un tipo de selector de parámetro ventá sexa mellor que os outros para tódolos casos. Por exemplo, para o modelo M1 o menor MISE obtense co método do polgar, mentres que para o modelo M7 o MISE que se obtén co método do polgar é o maior de todos, mentres que o menor obtense co método do plug-in. Isto indica que deberemos ter en contra a mostra para elixir o método polo cal debemos seleccionar a ventá.

Presentamos a continuación unhas gráficas en forma de boxplot dos parámetros ventá h obtidos por cada un dos diferentes métodos. Isto pode dar unha idea da dispersión que teñen as estimacións das ventás obtidas por cada método, así como se presentan moitos datos atípicos. Ademais, a liña horizontal punteada que aparece acompañando á gráfica fai referencia á ventá h_{AMISE}.

Para a Figura 4.1, os datos foron xerados seguindo unha distribución normal estándar. Este feito produce que a regra do polgar produza os resultados que vemos na Figura 4.1, na que a regra do polgar presenta pouca dispersión para todos os tamaños de mostra, incluso menos que o método do plug-in. Por outro lado vemos que a validación cruzada presenta moita dispersión para as ventás. Isto é algo esperable xa que a función de distribución normal é un modelo moi suave, situación na que a validación cruzada non proporciona bos resultados. Finalmente mirando cara a ventá h_{AMISE}, vemos que tende a cero a medida que o tamaño de mostra aumenta.

O modelo M4 é en certo sentido complicado, xa que parte dunha normal e presenta un crecemento e seguidamente un decrecemento moi grandes. O feito de que sexa un modelo complexo pode verse, na Figura 4.2, mirando a cantidade de datos atípicos que presenta para tamaño de mostra $n = 100$. Nesta situación, e como se pode ver na Figura 4.2, o método do polgar non proporciona resultados satisfactorios. Isto débese a que o modelo M4 dista moito do modelo M1. En canto á validación cruzada, a rugosidade do modelo fai que este método proporcione bos resultados, sen ter moita variabilidade nas ventás. Finalmente dicir que o método do plug-in converxe a h_{AMISE} e presenta moi pouca dispersión, polo que para tamaños de mostra grandes este método será o mellor que poidamos usar.

Na Figura 4.3 podemos ver que, ao igual que para o modelo M4, o método do polgar

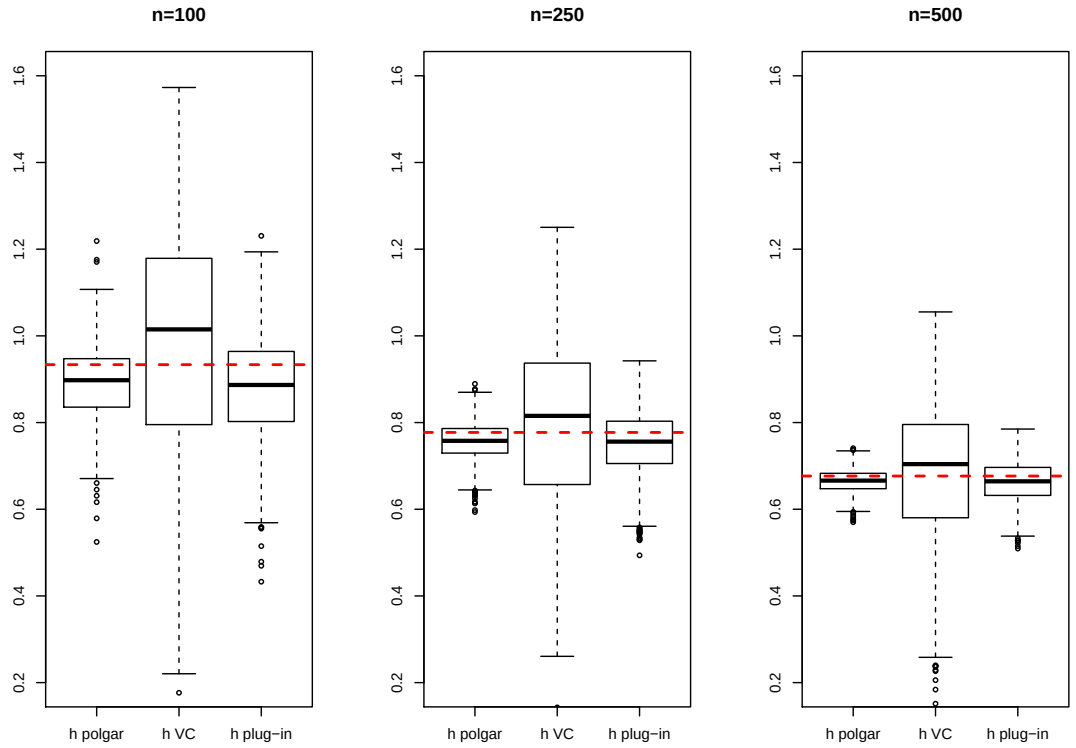


Figura 4.1: Diagramas de caixa dos diferentes selectores considerados para o modelo M1 de [2] asociados a mostras de tamaño n. Nótese que a liña vermella representa a ventá h_{AMISE}.

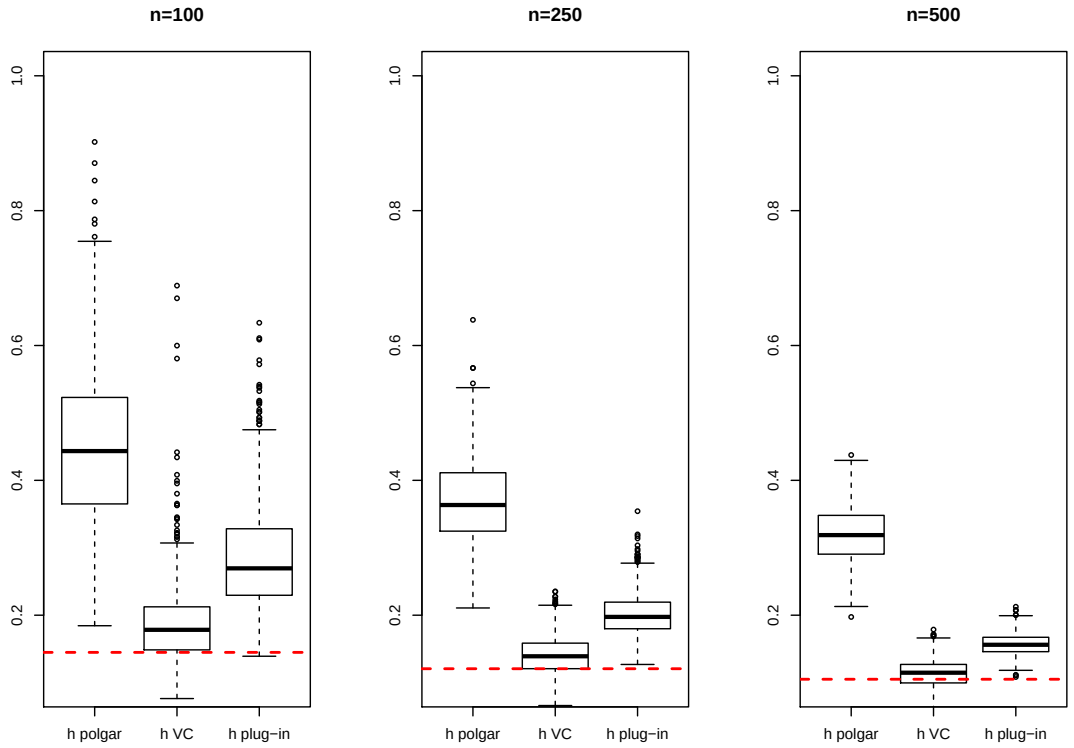


Figura 4.2: Diagramas de caixa dos diferentes selectores considerados para o modelo M4 de [2] asociados a mostras de tamaño n. Nótese que a liña vermella representa a ventá h_{AMISE}.

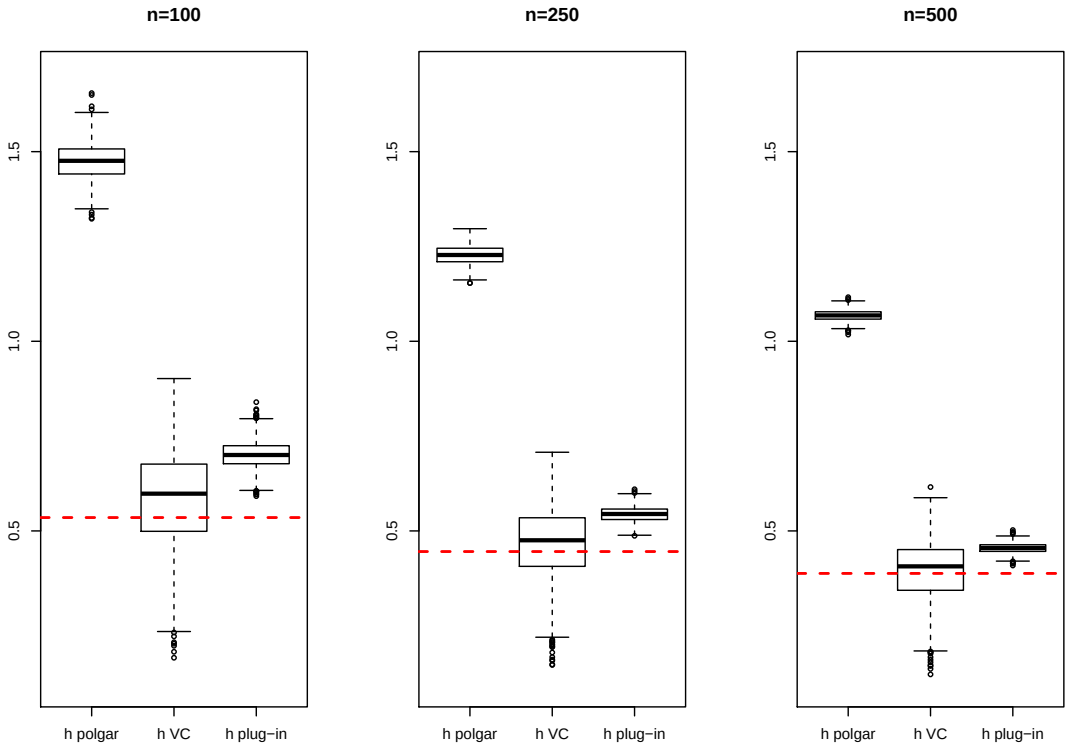


Figura 4.3: Diagramas de caixa dos diferentes selectores considerados para o modelo M7 de [2] asociados a mostras de tamaño n. Nótese que a liña vermella representa a ventá h_{AMISE}.

non proporciona bos resultados. De novo isto sucede debido a que o modelo M7 dista moito dunha distribución normal. Cando isto sucede, a regra do polgar non proporciona resultados satisfactorios aínda que se aumente o tamaño da mostra. En canto ao método de validación cruzada vemos que presenta bastante dispersión, pero vaise corrixindo a medida que aumentamos o tamaño da mostra. Finalmente en canto ao método plug-in, a pesar de que presenta un sesgo coa ventá h_{AMISE}, converxe á ventá AMISE a medida que o tamaño de mostra n crece.

Capítulo 5

Aplicación a datos reais

Ao longo deste capítulo, aplicaremos a metodoloxía desenrolada neste traballo a unha base de datos reais. Con este obxectivo, empregaremos a base de datos `faithful` de R, que contén información sobre o famoso géyser Old Faithful do parque nacional de Yellowstone situados nos Estados Unidos (ver Figura 5.1). Pode consultarse máis información sobre esta base de datos en http://127.0.0.1:16629/library/datasets/html/faithful.html. En concreto, nesta base de datos recóllese o tempo transcorrido entre erupcións (baixo o nome `eruptions`) e a duración destas mesmas (baixo o nome `waiting`).

Figura 5.1: Imaxe do géyser Old Faithful en plena erupción no ano 2013.

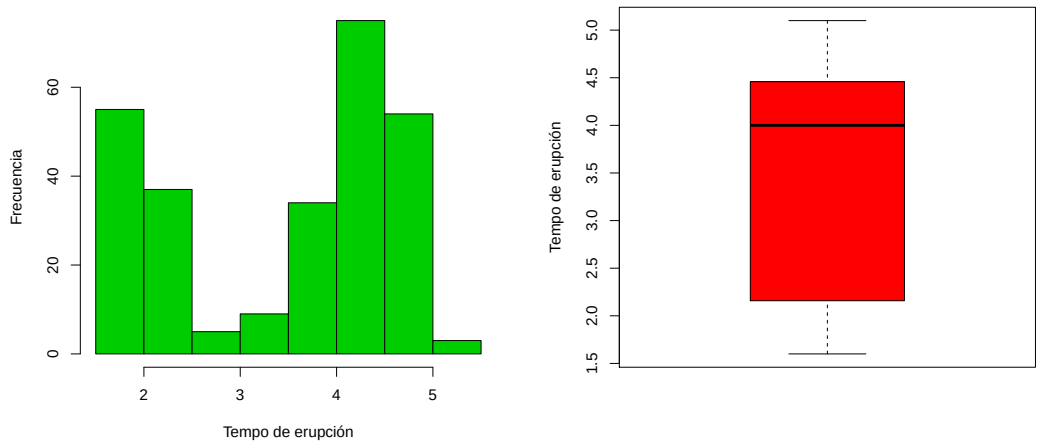
Tras cargar dita base de datos en R, o noso obxectivo será estimar a función de densidade do tempo transcorrido entre erupcións. Para iso podemos, en primeiro lugar, facer un análise descritivo da variable de estudo, que será a variable `eruptions`, a cal recolle a duración de cada erupción do geyser observado. Os resultados obtidos aparecen resumidos na Táboa 5.1. Dado que a variable de estudo está dada en minutos, este análise descritivo da variable dinos, por exemplo, que o tempo mínimo de erupción é de 1.60 minutos, e que o tempo medio de erupción é de 3.48. Ademais sabemos que o 75 % das erupcións teñen un tempo de duración superior a 2.16 minutos.

Medidas Características	Valor (en minutos)
Media	3.48
Desviación Típica	1.14
Mínimo	1.60
Máximo	5.10
Rango	3.50
Cuantil 0.25	2.16
Mediana (Cuantil 0.5)	4.00
Cuantil 0.75	4.45
Rango Intercuantílico	2.29
Moda	1.86

Táboa 5.1: Principais medidas características asociadas á variable `eruptions` asociada á base de datos `faithful` en R.

De cara a continuar co resumo da variable `eruptions` podemos realizar representacións gráficas como o histograma (ver parte (a) da Figura 5.2) que nos dará unha idea da forma da función de densidade que queremos estimar. Ademais, gran parte das medidas características recollidas na Táboa 5.1 podemos resumilas grazas a un diagrama de caixa (ver parte (b) da Figura 5.2) que tamén nos informa sobra a presenza de datos atípicos. Á vista de ditas representacións podemos dicir que a variable `eruptions` segue unha distribución bimodal, onde unha das modas sitúase entre os valores 1 e 2 e a outra entre os valores 4 e 5. Ademais o diagrama de caixas mostra que a mediana está notablemente máis cerca do cuantil 3 que do cuantil 2, o que indica que gran parte dos datos están nun intervalo relativamente pequeno, comprendido entre os 4 e os 4.45 minutos.

Como comentamos anteriormente, o noso obxectivo é presentar unha estimación da

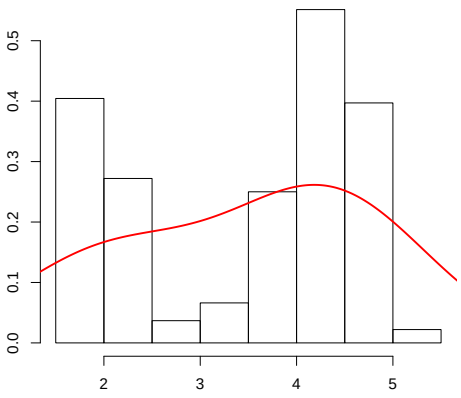


(a) Histograma (b) Diagrama de caixas

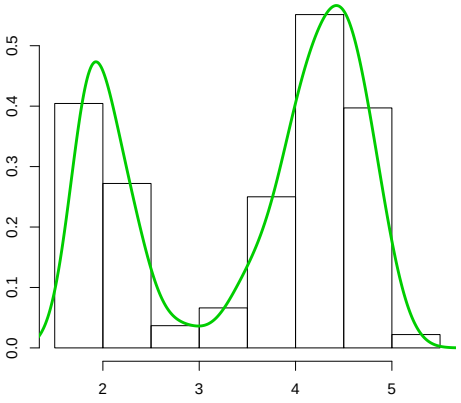
Figura 5.2: Representación mediante histograma e diagrama de caixas da variable `eruptions` correspondente á base de datos `faithful` de R.

función de densidade da variable `eruptions` e para iso empregaremos o diferentes métodos presentados ao longo do Capítulo 3. Na Figura 5.3 representamos o histograma da variable `eruptions` xunto coas estimacións da función de densidade empregando tanto a regra do polgar, como o método de validación cruzada ou a regra plug-in.

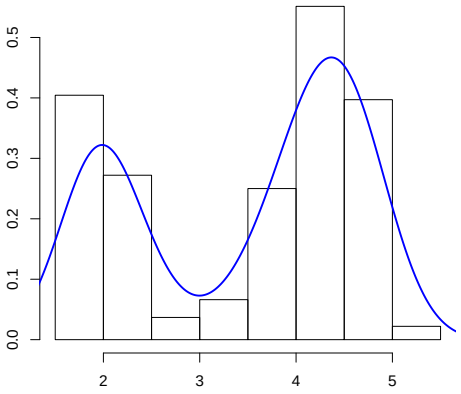
A Figura 5.3 móstranos que a estimación da función de densidade que ofrece o método do polgar non se aproxima de maneira axeitada ao histograma dos nosos datos. Isto é razoable, pois este método de selección de h baséase na suposición de que a variable de estudo, neste caso `eruptions`, segue unha distribución normal con media e varianza descoñecidas. En concreto, para os datos da variable `eruptions` dita suposición é completamente errónea, xa que a variable de estudo ten a menor densidade de datos xusto onde a distribución normal conta con maior densidade de datos. Polo tanto, o feito de supoñer distribución normal é o que produce que a estimación tipo núcleo obtida grazas á regra do polgar non mostre un bo axuste aos datos. Para ver de maneira máis formal que os datos da variable *eruptions* non seguen unha distribución normal podemos presentar un gráfico qq-plot (ver Figura 5). Recordemos que nun gráficos qq-plot representarmos os cuantís dunha variable normal fronte aos cuantís da variable de estudo, e así, no caso de que a variable seguise unha distribución normal, a gran maioría dos puntos deberían situarse entre as dúas liñas punteadas que se presentan na Figura 5.



(a) Regra do polgar



(b) Validación cruzada



(c) Regra Plug-in

Figura 5.3: Representación, mediante histograma, da variable `eruptions` correspondente á base de datos `faithful` de R, acompañada coa estimación tipo núcleo da función densidade con distintos parámetros ventá h. A primeira imaxe correspóndese coa elección de ventá empregando a regra do polgar, para a segunda imaxe obtívose h mediante o método de validación cruzada e na terceira imaxe utilizouse a regra plug-in.

Por outra banda, tamén poderiamos aplicar un test de normalidade, como por exemplo o test de Shapiro-Wilk. Recordemos que o test de Shapiro-Wilk ten como hipótese nula que a mostra, neste caso `eruptions`, provén dunha poboación normal e como hipótese

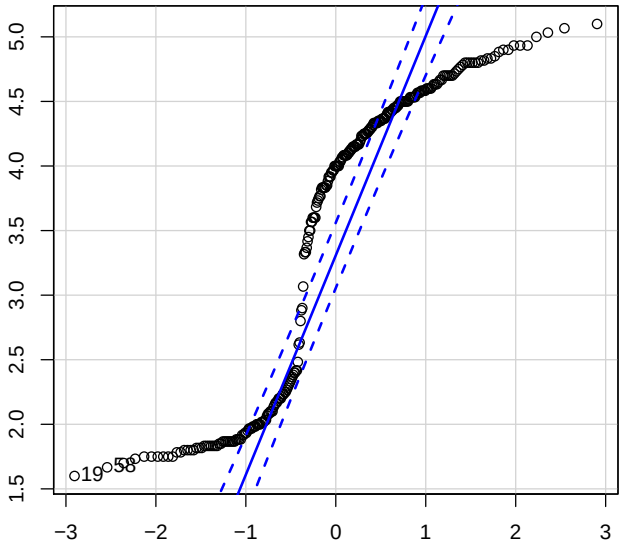


Figura 5.4: Gráfico qq-plot da variable `eruptions` correspondente á base de datos `faithful` dispoñible en R .

alternativa que a mostra non provén dunha distribución normal. Se aplicamos o test de Shapiro-Wilk empregando o software R obtense o seguinte resultado:

```
> shapiro.test(eruptions)

        Shapiro-Wilk normality test

data:  eruptions
W = 0.84592, p-value = 9.036e-16
```

Deste xeito, a obtención dun p-valor de $9.036e^{-16}$ indica o rechazo da hipótese nula en favor da hipótese alternativa. É dicir, temos evidencias estadísticamente significativas de que a variable `eruptions` non segue una distribución normal. Por este motivo non é satisfactoria a estimación tipo núcleo da función de densidade utilizando a ventá h obtida polo método do polgar.

En relación ás outras dúas estimacións da función de densidade, ambas producen resultados moi similares, mais parece que a mellor de elas é a que fai uso do parámetro h

obtido polo método de validación cruzada. Sen embargo, a estimación tipo núcleo que se proporciona utilizando a ventá obtida co método do plug-in poderíamos dicir que se axusta relativamente ben á función distribución real dos datos. Notemos tamén que o histograma só nos proporciona unha idea orientativa da forma da función de densidade, e pode variar en función do tamaño que elixamos para o ancho dos intervalos de clase, polo que non é sinxelo dicir cal das dúas estimacións da función de densidade se axusta mellor á función de densidade real dos datos. Concluímos entón dicindo que as estimacións tipo núcleo da función de densidade obtidas co parámetro ventá dado polo método do plug-in e co parámetro ventá dado polo de validación cruzada producen resultados moi semellantes.

Capítulo 6

Conclusións

Calquera variable aleatoria continua queda completamente caracterizada grazas á súa función de densidade. A estimación da función de densidade é un problema dunha gran importancia xa que, no caso de coñecer a función de densidade asociada a unha mostra, teríamos toda a información posible da mesma. Debido á importancia da función de densidade, neste traballo preséntanse a **estimación tipo núcleo** da mesma. Dada $\{x_1, ..., x_n\}$ unha mostra aleatoria simple da variable X o **estimador tipo núcleo** defínese da seguinte forma:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

onde K denomínase función **núcleo** ou *kernel* e $h > 0$ denota a **ventá ou ancho de banda**. Tamén se estudan as propiedades teóricas deste estimador en termos de erro cadrático medio (ECM), erro cadrático integrado (ISE), erro cadrático medio integrado (MISE) e erro cadrático medio integrado asintótico (AMISE), co obxectivo de determinar canto de boa é dita estimación. Este estudo do estimador tipo núcleo e dos erros mencionados introduce de maneira natural o problema da selección da función núcleo K que minimiza o erro cadrático medio integrado asintótico, que como se mostra no Capítulo 2 é o núcleo de Epanechnikov. Así mesmo, tamén se introduce o problema de selección do parámetro ventá, ao que se lle adica un capítulo completo, posto que o seu efecto sobre o estimador é moito máis importante que o da función núcleo.

A bondade do estimador tipo núcleo está claramente influenciada polo **parámetro de suavizado**, tamén coñecido como ancho de banda ou ventá. No Capítulo 3 preséntase a elección do parámetro ventá h e estúdanse tres métodos de selección de dito parámetro ventá. Selección que, como vimos, terá máis importancia que a elección do núcleo K á

hora de obter unha boa estimación da función de densidade. Os métodos de selección do parámetro ventá que se presentan podemos dividilos en función do obxectivo que perseguen. Por un lado, preséntase o método de validación cruzada que intenta minimizar o erro cadrático medio integrado. Por outra banda, detállanse a regra do polgar e a regra plug-in que intentan estimar a ventá que minimiza o erro cadrático medio integrado asintótico que recordemos vén dada por:

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

onde $R(g) = \int_{-\infty}^{+\infty} g(z)^2 dz$, $\mu_2(g) = \int_{-\infty}^{+\infty} z^2 g(z) dz$ e $f^{(2)}$ denota a segunda derivada da función de densidade que queremos estimar. Nótese que a regra do polgar baséase en asumir que a variable de interese sigue unha distribución normal para poder estimar $R(f^{(2)})$ mentres que o método do plug-in baséase en estimar dita cantidade empregando estimadores non paramétricos. Nótese que a regra plug-in tamén vai asumir a normalidade da variable de interese pero nunha etapa posterior, co que consigue reducir notablemente o efecto de dita suposición.

De cara a comparar os diferentes selectores do ancho de banda, no Capítulo 4 deséñase un estudo de simulación que permitirá establecer unha comparativa xusta entre os diferentes selectores en termos do erro cadrático medio integrado. Considerаranse diferentes modelos de distribucións presentados en [2] e con diferentes formas. Deste xeito, poderanse extraer conclusións sobre a utilidade de cara selector en función da forma da densidade que se pretende estimar. Compararemos os diferentes selectores tanto a través do erro cadrático medio integrado en canto á calidade da estimación, así como en termos de estimadores das ventás óptimas dende un punto de vista teórico.

Para rematar, no Capítulo 5 aplícase o estimador estudado nos capítulos anteriores a unha base de datos reais. En concreto emprégase a base de datos `faithful` de  , na cal se recollen datos da duración da erupción do geyser Old Faithfull. Con estes datos, dado que non coñecemos ningunha información sobre a función de densidade que seguen, trataremos de estimar a función de densidade de ditos datos empregando os diferentes métodos de selección de ventá vistos no Capítulo 3. Grazas a esta base de datos reais, ilustramos os motivos polos que a regra do polgar non proporciona unha boa estimación da función de densidade para certos conxuntos de datos, como por exemplo, este caso. É dicir, se os datos considerados non se asemellan para nada a unha distribución normal, a regra do polgar non proporcionará unha boa estimación da función de densidade. Ademais, mostraranse as ferramentas necesarias para contrastar a hipótese de normalidade na práctica a través de ferramentas como o test de normalidade de Shapiro-Wilk ou representacións gráficas como

o qq-plot.

Ao final do traballo, baixo o título de Anexo  , aparece o código utilizado para a programación dos métodos, para a realización do estudo de simulación e para a representación dalgunhas figuras utilizadas ao longo do traballo.

Anexo: Código de 

6.1. Código correspondente á Figura 1.1

```
set.seed(123)
z<-rnorm(100)

#Histograma 1
hist(z,freq=F,main="",xlab="",breaks=2,ylim=c(0,0.65),ylab="Densidade",
    xlim=c(-2.5,2.5))
curve(dnorm(x,0,1),xlim=c(-3,3),main="",ylab="Densidade",add=T,lwd=3)

#Histograma 2
hist(z,freq=F,main="",xlab="",ylim=c(0,0.65),ylab="Densidade",
     xlim=c(-2.5,2.5))
curve(dnorm(x,0,1),xlim=c(-3,3),main="",add=T,lwd=3)

#Histograma 3
hist(z,freq=F,main="",xlab="",breaks=30,ylim=c(0,0.65),ylab="Densidade",
     xlim=c(-2.5,2.5))
curve(dnorm(x,0,1),xlim=c(-3,3),main="",ylab="Densidade",add=T,lwd=3)
```

6.2. Funcións programadas

Nesta sección aparecen as funcións que foron programadas para a realización do traballo. Cabe destacar que a programación da función de validación cruzada realizouse co obxectivo de axudar a comprender dito método, mais non se utilizou para a obtención dos datos. Para este fin utilizouse unha función xa programada de R por unha cuestión de eficiencia .

6.2.1. Polgar

```
hpolgar<-function(datos){
zetaK<-function(x){(3/4)*(1-x^2)*x^2}
KEpacadrado<-function(x){((3/4)*(1-x^2))^2}
Rk<-integrate(KEpacadrado,lower=-1,upper=1)$value
mudous<-integrate(zetaK,lower=-1,upper=1)$value
mudouscadrado<-mudous^2

p1<-(IQR(datos)/(qnorm(0.75)-qnorm(0.25)))
p2<-sd(datos)

p3<-min(p1,p2)

if(p3==0){max(p1,p2)*((8*pi^(1/2)*Rk)/(3*mudouscadrado*length(datos)))^(1/5)}
else{p3*((8*pi^(1/2)*Rk)/(3*mudouscadrado*length(datos)))^(1/5)}
}
```

6.2.2. Validación cruzada

```
simpson<-function(fxs,a,b){
np<-length(fxs)
h=(b-a)/(np-1)

int<-3*(fxs[1]+fxs[np])/8+7*(fxs[2]+fxs[np-1])/6+
+23*(fxs[3]+fxs[np-2])/24+sum(fxs[4:(np-3)])
```

```
return(int*h)
}


f<-function(xn,x,h){
epa<-function(x){(3/4)*(1-x)^2*(abs(x)<=1)}
n=length(xn)
(n*h)^(-1)*sum(epa((x-xn)/h))
}

hvc<-function(datos,hgrella){
n<-length(datos)
fcvxenerico<-matrix(0,ncol=length(hgrella),nrow=length(datos))

for(j in 1:length(hgrella)){
for (i in 1:length(datos)){
datoscv<-datos[-i]
x<-datos[i]
fcvxenerico[i,j]<-f(datoscv,x,hgrella[j])
}
}

fcv.vector<-colSums(fcvxenerico)

xseq=seq(min(datos),max(datos),by=0.01)
fx.hatx<-matrix(0,nrow=length(xseq),ncol=length(hgrella))
for(j in 1:length(hgrella)){
for(i in 1:length(xseq)){
fx.hatx[i,j]<-f(datos,xseq[i],hgrella[j])
}
}

vcxenerico<-numeric(length(hgrella))
for(i in 1:length(hgrella)){
vcxenerico[i]<-simpson(fx.hatx[,i]^2,min(datos),max(datos))-2*fcv.vector[i]/n
}
```

```
return(hgrella[which.min(vcxenerico)])
}
```

6.2.3. Regra de Simpson

```
simpson<-function(fxs,a,b){

np<-length(fxs)
h=(b-a)/(np-1)

int<-3*(fxs[1]+fxs[np])/8+7*(fxs[2]+fxs[np-1])/6+23*(fxs[3]+fxs[np-2])/24+
       sum(fxs[4:(np-3)])
return(int*h)
}
```

6.3. Estudo de simulación

Dado que o estudo de simulación é moi similar para todos os modelos, escribimos aquí o código só para o modelo 4.

```
set.seed(123)
library(KernSmooth)
library(Deriv)
library(kedd)
library(kernelboot)
library(nor1mix)


source("funcions_auxiliares.R")

Calculo.MISE.m4<-function(n,M){
set.seed(123)
```

```
###Ventá AMISE
f4<-function(x){(1/2)*1/((1/2)*sqrt(2*pi))*exp(-(x+3/2)^2/(2*1/4))+
+(1/2)*1/((1/2)*sqrt(2*pi))*exp(-(x-3/2)^2/(2*1/4))}

f4_1<-Deriv(f4,"x")
f4_2<-Deriv(f4_1,"x")


zetaK<-function(x){(3/4)*(1-x^2)*x^2}
KEpacadrado<-function(x){((3/4)*(1-x^2))^2}

Rk<-integrate(KEpacadrado,lower=-1,upper=1)$value

mudous<-integrate(zetaK,lower=-1,upper=1)$value
mudouscadrado<-mudous^2

f4_2cadrado<-function(x){f4_2(x)^2}
Rf2<-integrate(f4_2cadrado,lower=-Inf,upper=+Inf)$value


polgar=vc=vc1=plug_in<-numeric(M)
ise_polgar=ise_vc=ise_vc1=ise_plug_in=ise_amise<-numeric(M)

h_amise<-(Rk/(mudouscadrado*n*Rf2))^(1/5)
h_amise

pb=txtProgressBar(style=3)
kind=1
u=proc.time()

for (i in 1:M){
setTxtProgressBar(pb,kind/M)

xn<-rnorMix(n,MW.nm4)
polgar[i]<-hpolgar(xn)
```

```
hos=sd(xn)*(243*Rk/(35*mudouscadrado*n))^(1/5)
hgrella=seq(0.1*hos,2*hos,by=0.01)
vc1[i]<-h.ucv(xn, deriv.order = 0, kernel = "epanechnikov")$h
vc[i]<-hvc(xn,hgrella)
plug_in[i]<-dpik(xn,kernel="epanech")


### CÁLCULO DO ISE
xseq<-seq(min(xn),max(xn),0.01)
f_teorica=dnorMix(xseq,MW.nm4)

### ESTIMACIONS DA DENSIDADE CON DIFERENTES VENTÁS
f_hat_polgar<-numeric(length(xseq))
f_hat_vc=f_hat_vc1<-numeric(length(xseq))
f_hat_plug_in<-numeric(length(xseq))
f_hat_AMISE<-numeric(length(xseq))
for (j in 1:length(xseq)){
f_hat_polgar[j]<-f(xn,xseq[j],polgar[i])
f_hat_vc[j]<-f(xn,xseq[j],vc[i])
f_hat_vc1[j]<-f(xn,xseq[j],vc1[i])
f_hat_plug_in[j]<-f(xn,xseq[j],plug_in[i])
f_hat_AMISE[j]<-f(xn,xseq[j],h_amise)
}

### ISE PARA AS DIFERENTES VENTÁS
ise_polgar[i]<-simpson((f_hat_polgar-f_teorica)^2,min(xn),max(xn))
ise_vc[i]<-simpson((f_hat_vc-f_teorica)^2,min(xn),max(xn))
ise_vc1[i]<-simpson((f_hat_vc1-f_teorica)^2,min(xn),max(xn))
ise_plug_in[i]<-simpson((f_hat_plug_in-f_teorica)^2,min(xn),max(xn))
ise_amise[i]<-simpson((f_hat_AMISE-f_teorica)^2,min(xn),max(xn))

kind=kind+1
}
```

```
Calculo.MISE.m4(100,1000)$MISE
Calculo.MISE.m4(250,1000)$MISE
Calculo.MISE.m4(500,1000)$MISE
```

Bibliografía

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

[2] Marron, J. S. e Wand, M. P. (1992). *Exact mean integrated squared error.* Institute of Mathematical Statistics, 712-736.

[3] Rudemo, M. (1982). *Empirical choice of histograms and kernel density estimators.* Scandinavian Journal of Statistics 9, 65-78.

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

[5] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[6] Sturges H. A. (1926). *The choice of a class interval.* Journal of the American Statistical Association, 21, 65–66.

[7] Wand, M. P. e Jones, M. C. (1995). *Kernel Smoothing.* Chapman & Hall.