

Selección del parámetro de suavizado en estimación noparamétrica de la densidad y de la regresión

María Leyenda Rodríguez
Silvia Suárez Crespo

15 de abril de 2010

Índice general

Índice general	3
1. Introducción	5
2. Estimación noparamétrica de la densidad	7
2.1. Objetivos	7
2.2. Marco teórico	8
2.3. Simulación	13
2.3.1. Algoritmos	13
2.3.2. Proceso de simulación	16
2.4. Resultados y conclusiones	21
3. Estimación noparamétrica de la regresión	33
3.1. Objetivos	33
3.2. Análisis de las distribuciones univariantes de las variables en estudio	34
3.3. Estimador lineal local con ventana de validación cruzada	37
3.3.1. Marco teórico	38
4. Anexos	41
4.1. Código necesario para el Capítulo 2	41
4.2. Código necesario para el Capítulo 3	41

Capítulo 1

Introducción

Las técnicas de estimación no paramétrica surgen, a grandes rasgos, como un complemento o una alternativa a las técnicas que se utilizan en el contexto paramétrico. La estimación tipo núcleo es un caso particular de las técnicas no paramétricas, y será la que se utilice en el presente estudio. Los primeros trabajos que existen sobre las técnicas tipo núcleo aparecen a finales de la década de los 50 y principios de los 60, de la mano de [?] y [?], respectivamente. Debido a la sencillez e interpretabilidad de su expresión, los estimadores tipo núcleo se han convertido en una importante herramienta para el análisis exploratorio de datos, así como para el desarrollo de técnicas de inferencia estadística basadas en dichos estimadores.

Dentro del contexto de la Estadística no paramétrica, existen dos campos de estudio claramente diferenciados: la estimación no paramétrica de la densidad y la estimación no paramétrica de la regresión. Sobre estos dos campos versarán los dos siguientes capítulos, realizando en cada uno de ellos un estudio centrado en una de las problemáticas más importantes que surgen al utilizar técnicas no paramétricas: la selección del parámetro de suavizado.

Cada capítulo presentará una estructura similar, en la se incluirán el planteamiento de los objetivos, el análisis de las variables que se utilizarán (sólo en el Capítulo 3), la descripción de los conceptos teóricos necesarios para el estudio y los correspondientes algoritmos, el estudio de simulación a realizar (sólo en el Capítulo 2) y finalmente la presentación de los resultados y el análisis de los mismos. Como complemento se adjuntará en un cuarto capítulo la implementación de los códigos utilizados para la realización de los estudios de simulación.

El objetivo final del presente documento será comprobar el funcionamiento de los métodos de selección del parámetro de suavizado, tanto para el estimador tipo núcleo de la densidad como para el estimador local lineal de la regresión. Para ello, se utilizarán todas las herramientas teóricas necesarias y se hará uso del paquete estadístico **R** (y de sus bases de datos cuando sea preciso) para realizar los estudios de simulación. Dicho paquete proporciona una gran variedad de técnicas gráficas y estadísticas y está disponible como software libre (ver [?]).

Capítulo 2

Estimación noparamétrica de la densidad

2.1. Objetivos

El objetivo de este capítulo será comprobar el funcionamiento de la selección del parámetro de suavizado en la estimación núcleo de la densidad univariante utilizando el método de validación cruzada. Se realizará para ello un pequeño estudio de simulación, utilizando para el mismo un núcleo gaussiano. La primera tarea a realizar será la implementación de la ventana de validación cruzada, \hat{h}_{CV} utilizando el paquete R. Como *competidor* del método analizado se considerará la ventana normal, \hat{h}_{NS} , basada en estimar la ventana AMISE suponiendo que la distribución de los datos es normal. Como criterio de error se empleará el error cuadrático integrado dado por la expresión :

$$\text{ISE}(\hat{f}_{h,K}) = \int (\hat{f}_{h,K}(x) - f(x))^2 dx. \quad (2.1)$$

Este criterio de error permitirá decidir cuál de las dos ventanas seleccionadas comete menos error. Como modelos de prueba se considerarán las 15 densidades descritas en [?]. Estas densidades se han convertido en un estándar para validar cualquier método de estimación de la densidad y están programadas en R (en la librería `normix`).

2.2. Marco teórico

En esta sección introduciremos los conceptos teóricos necesarios para la realización del estudio, comenzando con la definición del estimador tipo núcleo de la densidad y finalizando con la obtención de los criterios de selección de ventana que utilizaremos en este estudio. Se tratará además el desarrollo teórico oportuno para la obtención de una ventana óptima teórica (h_{AMISE}), la imposibilidad de su utilización en la práctica y la justificación de los criterios de selección para la obtención de *buenas* aproximaciones de la ventana óptima teórica.

Supóngase que se parte de una muestra aleatoria simple (m.a.s) X_1, \dots, X_n de una variable aleatoria (v.a) X absolutamente continua con distribución $F(\cdot)$ (de ahora en adelante, $X \sim F$) y función de densidad $f(\cdot)$. La expresión del estimador núcleo de la densidad univariante $f(\cdot)$ viene dada por:

$$\hat{f}_{h,K}(x) = (nh)^{-1} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (2.2)$$

donde $K(\cdot)$ es una función de densidad (y por tanto verifica $\int K(x)dx = 1$ y $K(x) \geq 0$) denominada *núcleo* (en inglés, *kernel*) y $h > 0$ es el parámetro o ventana de suavizado. La técnica que utiliza este estimador es situar sobre cada una de las observaciones una función que otorga el mayor peso a la propia observación y va disminuyendo el peso que aportan las restantes observaciones según estas se van alejando de la observación en cuestión. En definitiva, el estimador (2.2) se construye como suma de funciones (*campanas*), donde la función núcleo $K(\cdot)$ determina la forma de las campanas situadas sobre cada observación y h el ancho de las mismas (a las observaciones que se encuentran fuera del rango establecido por h les viene asociado un peso igual a cero). Una formulación equivalente a (2.2) es la siguiente:

$$\hat{f}_{h,K}(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i), \quad (2.3)$$

donde $K_h(u) = h^{-1}K(u/h)$ se denomina núcleo reescalado.

Algunas propiedades elementales de las estimaciones tipo núcleo se derivan directamente de su definición. Supuesto que el núcleo es una función de densidad, se sigue de (2.2) que la estimación $\hat{f}_{h,K}(\cdot)$ será también una función de densidad. Además, esta estimación heredará todas las propiedades de continuidad y diferenciabilidad del núcleo $K(\cdot)$.

Existen distintos tipos de núcleos que se utilizan en la práctica (Gaussiano, Uniforme,

Epanechnikov,...), todos ellos funciones de densidad simétricas (es decir, verifican además $\int uK(u) du = 0$). Se puede ver que todos ellos son *eficientes* (ver [?]), obteniéndose por tanto que la decisión de escoger un núcleo u otro puede basarse en criterios más generales como la eficiencia computacional. En este trabajo se utilizará un núcleo Gaussiano, cuya expresión viene dada por:

$$K(x) = \frac{1}{(2\pi)^{1/2}} \exp\left(\frac{-x^2}{2}\right). \quad (2.4)$$

La selección de la ventana de suavizado es, sin embargo, de crucial importancia en la estimación de la densidad. La elección de un valor muy pequeño puede producir una estimación de la densidad *infrasuavizada*, es decir, una aproximación basta en el sentido de que la estimación muestra la variabilidad asociada a cada observación individual en lugar de la estructura general subyacente. En el otro extremo, si la ventana que se utiliza es muy grande, se obtendrá previsiblemente una estimación *sobresuavizada* en la que se oscurece la estructura de la densidad subyacente al realizarse el proceso de suavizado en una región muy grande.

El cálculo de las ventanas óptimas (teóricas) se basa en la minimización de las expresiones de distintos criterios de error, como pueden ser los criterios MSE (*Mean Square Error*) y MISE (*Mean Integrated Square Error*). El primero de ellos resulta ser un criterio local, es decir, se evalúa $\hat{f}_{h,K}(x)$ como estimador de $f(x)$, mientras que el segundo se centra en analizar el comportamiento de $\hat{f}_{h,K}(\cdot)$ como estimador de $f(\cdot)$ (criterio global). En la práctica, para el cálculo de las ventanas óptimas se minimizan las expresiones asintóticas de las anteriores (AMSE y AMISE, respectivamente). De forma esquemática, dado que no es el objetivo de este trabajo el estudio teórico del estimador núcleo de la densidad univariante, definiremos brevemente dichos criterios de error y las ventanas óptimas asociadas. Los razonamientos teóricos pertinentes pueden verse en [?].

De forma general, el criterio MSE para $\hat{f}_{h,K}(x)$ como estimador de $f(x)$ se escribe como:

$$\begin{aligned} MSE(\hat{f}_{h,K}(x)) &= \mathbb{E}(\hat{f}_{h,K}(x) - f(x))^2 = \text{Sesgo}_{\hat{f}_h(x)}^2(f(x)) + \text{Var}(\hat{f}_h(x)) \\ &= (\mathbb{E}\hat{f}_h(x) - f(x))^2 + \text{Var}(\hat{f}_h(x)). \end{aligned}$$

Bajo condiciones de regularidad sobre $f(\cdot)$ y suponiendo que $\mu_2(K) = \int u^2 K(u) du < \infty$ se pueden calcular las expresiones de sesgo y varianza, obteniéndose la expresión:

$$MSE(\hat{f}_{h,K}(x)) = \frac{1}{nh} R(K) f(x) + \frac{1}{4} h^4 \mu_2^2(K) (f''(x))^2 + o((nh)^{-1} + h^4),$$

donde $R(K) = \int K^2(u) du$. Así, la expresión asintótica AMSE en un punto x viene dada

por:

$$AMSE(\hat{f}_{h,K}(x)) = \frac{1}{nh}R(K)f(x) + \frac{1}{4}h^4\mu_2^2(K)(f''(x))^2. \quad (2.5)$$

Integrando el MSE en x se tiene:

$$MISE(\hat{f}_{h,K}) = \frac{1}{nh}R(K) + \frac{1}{4}h^2\mu_2^2(K)R(f''(x)) + o((nh)^{-1} + h^4) \quad (2.6)$$

y la versión asintótica

$$AMISE(\hat{f}_{h,K}) = \frac{1}{nh}R(K) + \frac{1}{4}h^2\mu_2^2(K)R(f''(x)). \quad (2.7)$$

Se tiene que las ventanas óptimas asociadas a la minimización de las expresiones (2.5) y (2.7) son, respectivamente,

$$h_{AMSE(x)} = \left(\frac{9f(x)}{2n(f''(x))^2} \right)^{1/5} \quad (2.8)$$

y

$$h_{AMISE} = \left(\frac{9}{2nR(f'')} \right)^{1/5}. \quad (2.9)$$

Sendas ventanas resultan inaplicables en la práctica, pues dependen de cantidades a priori desconocidas (sin ir más lejos, dependen de la propia función de densidad $f(\cdot)$). Además, la expresión 2.8, aún cuando ignorásemos su dependencia respecto a la función de densidad, depende del punto x en el que se evalúa dicha función, por lo que no es constante a lo largo del dominio de estimación y resulta por ende poco práctica. Por todos estos motivos es necesaria la obtención de ventanas aproximadas que presenten un *buen comportamiento* a la hora de realizar la estimación núcleo de la función de densidad.

Los métodos de selección de ventana son aquéllos que utilizan la información proporcionada por la m.a.s X_1, \dots, X_n para la construcción de una ventana aproximada \hat{h} . Estos métodos pueden dividirse en dos clases bien diferenciadas. Por una parte se encuentran los denominados *selectores de ventana simples*, cuyo objetivo es obtener estimaciones núcleo de la densidad de una forma rápida. El segundo tipo de selectores podría calificarse como un conjunto de *selectores de ventana automáticos*, pues están basados en argumentos matemáticos y requieren mayor esfuerzo computacional a cambio de proporcionar mejores estimaciones. En este trabajo estudiaremos el selector de ventana de escala Normal y el selector de ventana por validación cruzada cuadrática, siendo el primero uno de los selectores simples y el segundo uno de los selectores automáticos. Dentro de los selectores automáticos se encuentran otros selectores muy conocidos y muy aplicados en la práctica, como es, por ejemplo, el método *plug-in* o más conocido como método de Sheather and Jones (ver [?]).

Una aproximación muy sencilla y natural es utilizar una familia de distribuciones estándar para asignar un valor al término $R(f'')$ en la expresión (2.9). Entonces, suponiendo que $f(\cdot)$ es una función de densidad Normal con varianza σ^2 puede verse que

$$h_{\text{AMISE}} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2n} \right]^{\frac{1}{5}} \sigma \quad (2.10)$$

(ver [?]). La *ventana Normal* se obtiene por tanto reemplazando en (2.10) el parámetro σ por una estimación $\hat{\sigma}$:

$$\hat{h}_{\text{NS}} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2n} \right]^{\frac{1}{5}} \hat{\sigma}. \quad (2.11)$$

Además, en nuestro caso particular en el que estamos utilizando un núcleo Gaussiano, la ventana resultante que resulta de realizar los cálculos pertinentes en (2.11) es ([?]):

$$\hat{h}_{\text{NS}} = (4\pi)^{-1/10} \frac{3}{8} \pi^{-1/2} \hat{\sigma} n^{-1/5} = 1,06 \hat{\sigma} n^{-1/5}. \quad (2.12)$$

Como estimaciones de σ pueden pensarse varias opciones, entre ellas la más común que es la cuasidesviación típica muestral S . Sin embargo, cabe esperar que utilizando una medida de variabilidad más robusta se obtengan mejores resultados. En [?] se propone como opción deseable tomar como $\hat{\sigma}$:

$$A = \min \left(S, \frac{RIC}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)} \right), \quad (2.13)$$

denotando por $RIC = Q_3 - Q_1$ al rango intercuartílico con Q_1 y Q_2 denotando, respectivamente, el primer y el tercer cuantil de la función de densidad Normal con desviación típica σ , y Φ^{-1} la función cuantil de la $N(0, 1)$. Es decir, propone utilizar el mínimo entre la cuasidesviación típica muestral y el rango intercuartílico estandarizado. Utilizando por tanto la estimación propuesta en (2.13) y substituyéndola en (2.12) obtenemos la ventana Normal, \hat{h}_{ns} ,

$$\hat{h}_{\text{NS}} = 1,06 \min \left(S, \frac{RIC}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)} \right) n^{-1/5}. \quad (2.14)$$

La *validación cruzada cuadrática* (LSCV, *Least Squares Cross-Validation*) (ver [?] e [?]) se motiva en la reexpresión de (2.6) en la forma:

$$MISE(\hat{f}_{h,K}) = \mathbb{E} \int \hat{f}_{h,K}^2(x) dx - 2 \mathbb{E} \int \hat{f}_{h,K}(x) f(x) dx + \int f^2(x) dx. \quad (2.15)$$

Dado que el término $\int f^2(x) dx$ no depende de h , la minimización de (2.15) es equivalente a la minimización de:

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

Denotemos:

$$LSCV(h) = \int \hat{f}_h^2(x) dx - 2 \int \hat{f}_h(x) f(x) dx \quad (2.16)$$

El primer término de (2.16) es fácil de calcular a partir de la muestra:

$$\begin{aligned} \int \hat{f}_h^2(x) dx &= \int \frac{1}{n^2 h^2} \sum_{i=1}^n \sum_{j=1}^n K\left(\frac{x - X_i}{h}\right) K\left(\frac{x - X_j}{h}\right) dx \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \int K(u) K\left(u - \frac{X_i - X_j}{h}\right) du \\ &= \frac{1}{n^2 h} \sum_{i,j} K \star K\left(\frac{X_i - X_j}{h}\right), \end{aligned} \quad (2.17)$$

onde \star denota a operación convolución.¹

El segundo término de (2.16) puede escribirse como:

$$2 \int \hat{f}_h(x) f(x) dx = 2 \mathbb{E} \left(\hat{f}_{h,K}(X) \right),$$

donde X es una v.a con densidad $f(\cdot)$. Por tanto, este término no se puede calcular directamente de la muestra (pues utilizaríamos dicha muestra tanto para la construcción del estimador como para su evaluación, por lo que introduciríamos sesgo), problema que se solventaría tomando una m.a.s Y_1, \dots, Y_n de X independiente de la muestra X_1, \dots, X_n y construyendo el estimador (insesgado):

$$\frac{1}{n} \sum_{i=1}^n \hat{f}_{h,K}(Y_i).$$

La obtención de dicha muestra alternativa no parece un proceso viable para la realización de la estimación. En su lugar, se optará por reutilizar la m.a.s inicial X_1, \dots, X_n para construir otra muestra que sea independiente de la misma. El procedimiento será construir para cada X_i , $i = 1, \dots, n$ el estimador utilizando toda la muestra excepto esa observación:

$$\hat{f}_{h,K}^{-i}(X_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right) = \frac{2}{(n-1)h} \sum_{j > i} K\left(\frac{X_i - X_j}{h}\right).$$

¹Si f y g son dos funciones reales de variable real, se define la convolución $f \star g$ como otra función real de variable real dada por:

$$(f \star g)(x) = \int f(x-y)g(y)dy.$$

Se tendría entonces un estimador insesgado del segundo término de (2.16) sería:

$$\frac{2}{n} \sum_{i=1}^n \hat{f}_{h,K}^{-i}(X_i) = \frac{4}{n(n-1)h} \sum_{i=1}^n \sum_{j>i}^n K\left(\frac{X_i - X_j}{h}\right). \quad (2.18)$$

Finalmente, dadas las expresiones (2.17) y (2.18), se tiene que:

$$\widehat{LSCV}(h) \equiv LSCV(h) = \frac{1}{n^2 h} \sum_{i,j} K \star K\left(\frac{X_i - X_j}{h}\right) - \frac{4}{n(n-1)h} \sum_{i=1}^n \sum_{j>i}^n K\left(\frac{X_i - X_j}{h}\right) \quad (2.19)$$

es un estimador insesgado de (2.16). El método de validación cruzada selecciona como parámetro de suavizado (*ventana de validación cruzada*) el valor h_{CV} que minimiza la expresión (2.19). Puede ocurrir en algunos casos que la expresión (2.19) posea más de un mínimo local [?].

2.3. Simulación

Como bien se especificaba en la Sección 1.1, el fin de este estudio será comprobar el funcionamiento del criterio de selección de ventana por validación cruzada. Lo que se realizará realmente será observar el comportamiento que tiene tanto la ventana de validación cruzada como la ventana Normal, para luego compararlas (utilizando el criterio de error (2.1)) con el objetivo de concluir qué tipo de procedimiento funciona mejor en según qué casos. Para ello será necesario implementar un código en el que se obtengan los valores de interés para el estudio (se utilizará el paquete **R**). En la primera parte de esta sección se explicarán los algoritmos que se emplearán en dicha implementación. Por otra parte, para obtener conclusiones fiables será necesario realizar un estudio de simulación, que consistirá en la repetición del algoritmo un número (elevado) de veces de forma que se puedan extraer conclusiones generales sobre los resultados (medias, porcentajes,...). En la segunda parte de esta sección se describirá el proceso de simulación realizado, incluyendo además algunos aspectos técnicos a tener en cuenta a la hora de llevar a cabo dicho proceso.

2.3.1. Algoritmos

El esquema del algoritmo principal en el que se basará la implementación será el siguiente:

Algoritmo 1. Algoritmo principal: Dada una m.a.s M

1. Cálculo de h_{CV} (mediante la minimización de la expresión (2.19) utilizando la m.a.s M).
2. Cálculo de h_{NS} (mediante la implementación de la fórmula (2.14) utilizandos la m.a.s M).
3. Cálculo de $ISE(\hat{f}_{h_{CV}}) = \int (\hat{f}_{h_{CV}}(x) - f(x))^2 dx$, donde $\hat{f}_{h_{CV}}(\cdot)$ es el estimador núcleo univariante de la función de densidad $f(\cdot)$ construído utilizando un núcleo Gaussiano y la ventana h_{CV} (notemos que hemos suprimido el subíndice K en la expresión del estimador núcleo para una notación más sencilla).
4. Cálculo de $ISE(\hat{f}_{h_{NS}}) = \int (\hat{f}_{h_{NS}}(x) - f(x))^2 dx$, donde $\hat{f}_{h_{NS}}(\cdot)$ es el estimador núcleo univariante de la función de densidad $f(\cdot)$ construído utilizando un núcleo Gaussiano y la ventana h_{NS} .

A su vez, en el paso 1. del algoritmo 1, será necesaria la minimización de la expresión (2.19) en un determinado intervalo (adecuado). Debido al problema que mencionábamos en la Sección 2.2 de que pudiesen existir varios mínimos locales, podríamos pensar en al menos dos posibilidades:

Algoritmo 2. Algoritmo de minimización de (2.19) - (1): Dado un intervalo I ,

1. Dividir el intervalo I en m subintervalos I_1, \dots, I_m más pequeños.
2. Efectuar la minimización respecto de h de la expresión (2.19) en cada uno de los subintervalos elaborados en el paso 1. y escoger el mínimo.

Algoritmo 3. Algoritmo de minimización de (2.19) - (2): Dado un intervalo I ,

1. Construír una rejilla “suficientemente fina” que cubra el intervalo I (es decir, una secuencia de ventanas h_1, \dots, h_k) y evaluar la expresión (2.19) en cada uno de las ventanas.
2. Escoger el mínimo de los valores obtenenidos en en apartado 1.

En la siguiente subsección comentaremos las ventajas y desventajas computacionales que tiene cada uno de estos algoritmos.

Por otra parte, los pasos 3. y 4. del Algoritmo principal consisten, a grandes rasgos, en el cálculo de integrales de una resta de funciones en un intervalo al que podemos denotar por $J = [a, b]$ (más tarde hablaremos de cómo escoger tal intervalo). Optaremos en este caso por la integración numérica de dichas integrales (ver [?]); en concreto utilizaremos la Regla del Trapecio Compuesta, que consiste en dividir el intervalo $[a, b]$ en s subintervalos, dentro de los cuales se realiza la aproximación utilizando la Regla del Trapecio Simple, y sumando finalmente todos los resultados. De forma genérica, denotamos el largo de los subintervalos por $h = \frac{b-a}{s}$, $s \geq 1$, y los extremos de los subintervalos por $x_j = a + jh$, $0 \leq j \leq s$. Se tiene entonces que, para cualquier integral $I(g) = \int_a^b g(x) dx$,

$$I(g) = \sum_{j=1}^s \int_{x_{j-1}}^{x_j} g(x) dx.$$

Entonces, aplicando la Regla del Trapecio Simple en cada uno de los subintervalos, una buena aproximación ² de $I(g)$ viene dada por:

$$I(g) \sim \frac{h}{2} (g(x_0) + g(x_s)) + h \sum_{j=1}^{s-1} g(x_j). \quad (2.20)$$

Por tanto, los pasos a seguir para realizar el paso 3. del Algoritmo principal (en el caso del paso 4 sería análogo substituyendo la ventana h_{CV} por h_{NS}) serían:

Algoritmo 4. Algoritmo para la construcción de $ISE(\hat{f}_{h_{CV}})$: Dado un intervalo $J = [a, b]$

1. Dividir el intervalo J en s subintervalos, $J_1 = [a = x_0, x_1]$, $J_2 = [x_1, x_2]$, ... $J_s = [x_{s-1}, x_s = b]$.
2. Evaluar las funciones $\hat{f}_{h_{CV}}(\cdot)$ y $f(\cdot)$ en los extremos de los intervalos obtenidos en el paso 1.
3. Calcular los valores $\hat{f}_{h_{CV}}(z) - f(z)$ para $z = x_0, x_1, \dots, x_s$.
4. Realizar la aproximación propuesta en (2.20) con los valores obtenidos en el paso 3.

²La aproximación será buena si los intervalos son lo suficientemente pequeños, pues en tal caso el integrando será prácticamente una función lineal en cada subintervalo y la Regla del Trapecio Simple proporciona buenos resultados.

2.3.2. Proceso de simulación

Para realizar el estudio de simulación se generarán $B = 500$ muestras de tamaño $n = 100$ de cada una de las 15 densidades que aparecen en el documento de [?] (cuyas expresiones pueden verse en la Tabla 1 de dicho documento). Al igual que aparece en el artículo, nos referiremos a las densidades por #1, #2, y así sucesivamente hasta la densidad #15. Para cada muestra se estimará la ventana h usando los dos métodos analizados se evaluará el error cometido. Así, para cada densidad, se tendrán dos series de B números que se corresponden con las ventanas (\hat{h}_{CV} y \hat{h}_{NS}) obtenidas para cada simulación, así como otras dos series de B números que representan el error cometido por cada uno de los dos métodos. Notemos que en la implementación de la simulación se fijará semilla 0 (`set.seed(0)`), de forma que los resultados que aportamos en este documento se obtendrán automáticamente utilizando el código que anexamos en el Capítulo 4, sección 1.

Como explicábamos en la subsección anterior, es necesario optimizar de alguna forma la minimización de la expresión (2.16) de forma que se tome realmente el mínimo absoluto y no un mínimo local. El primer paso del algoritmo 2 será de utilidad para obtener gráficas de las expresiones 2.16 para un rango de ventanas, de forma podamos observar en cuáles de las densidades se obtienen gráficas más complicadas (en las que aparezcan mínimos locales). Tomando una rejilla de extremos 0 y 4 y paso 0.01, las gráficas que se obtienen en las cinco primeras simulaciones para las quince densidades consideradas se presentan en las Figuras 2.1, 2.2 y 2.3. Como se puede observar en dichas gráficas, en prácticamente todas las curvas se observa único mínimo, a excepción de las densidades #2 y #10, en las que se pueden observar mínimos locales bastante próximos. Se procederá pues a aplicar los algoritmos 2 y 3 utilizando, por ejemplo, la densidad #2, para ver si no existen diferencias entre los resultados que se obtienen en ambos casos, en cuyo caso se seleccionará el método que sea computacionalmente más eficiente. Utilizaremos para el algoritmo 1 la misma rejilla que hemos utilizado para las representaciones gráficas realizadas en las Figuras 2.1, 2.2 y 2.3. Para el algoritmo 2 consideraremos, en primer lugar, 8 subintervalos dentro del intervalo $I = [0, 4]$ (que por tanto tendrán longitud 0.5), y en segundo lugar 4 intervalos (de longitud 1). Las ventanas que se obtienen al minimizar la expresión (2.16) para las 30 primeras simulaciones de la densidad #2 utilizando los algoritmos propuestos pueden observarse en el Cuadro 2.1.

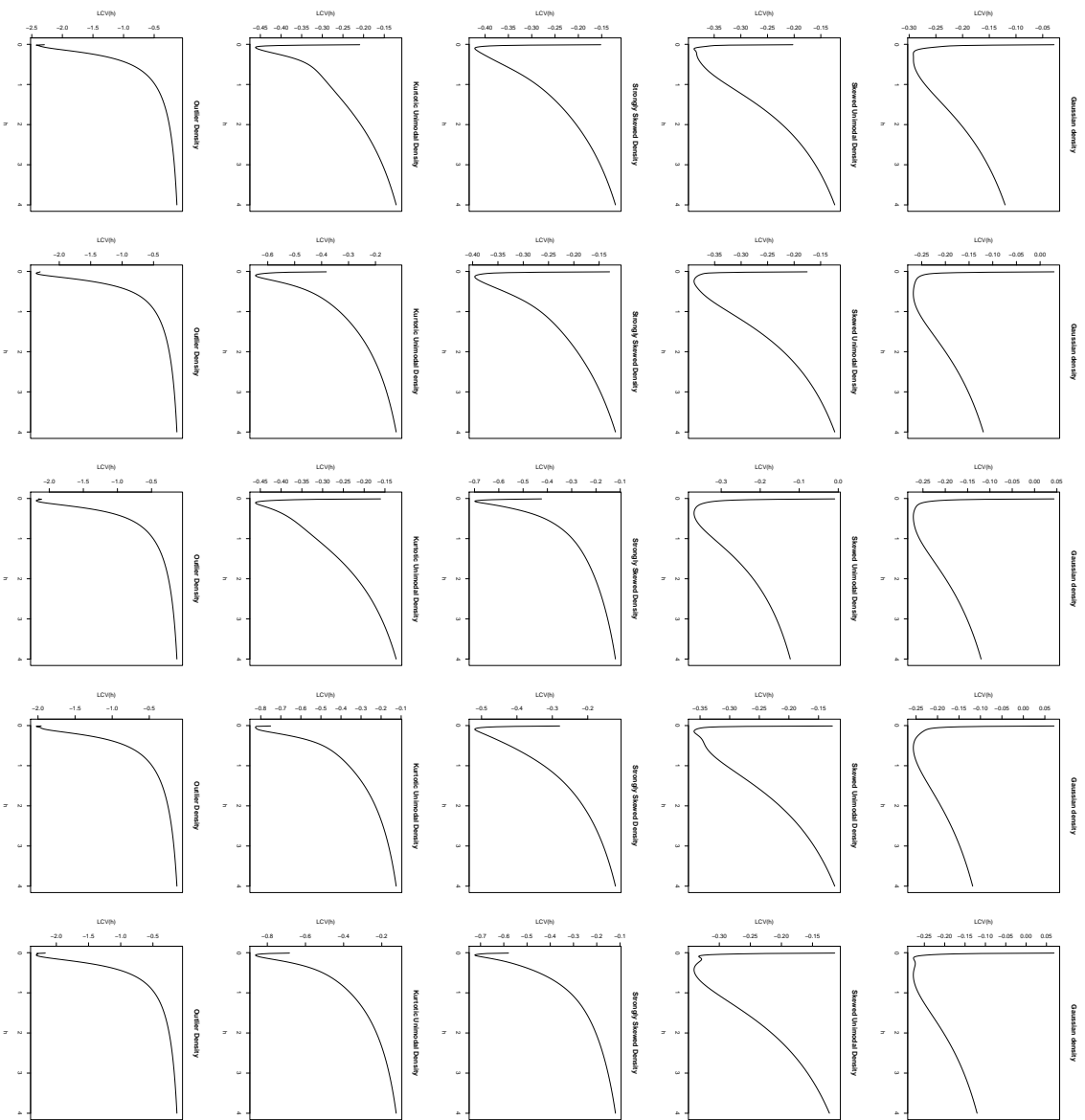


Figura 2.1: De izquierda a derecha y de arriba a abajo, representaciones de la curva (2.16) para las cinco primeras simulaciones de las densidades #1, #2, #3, #4 y #5.

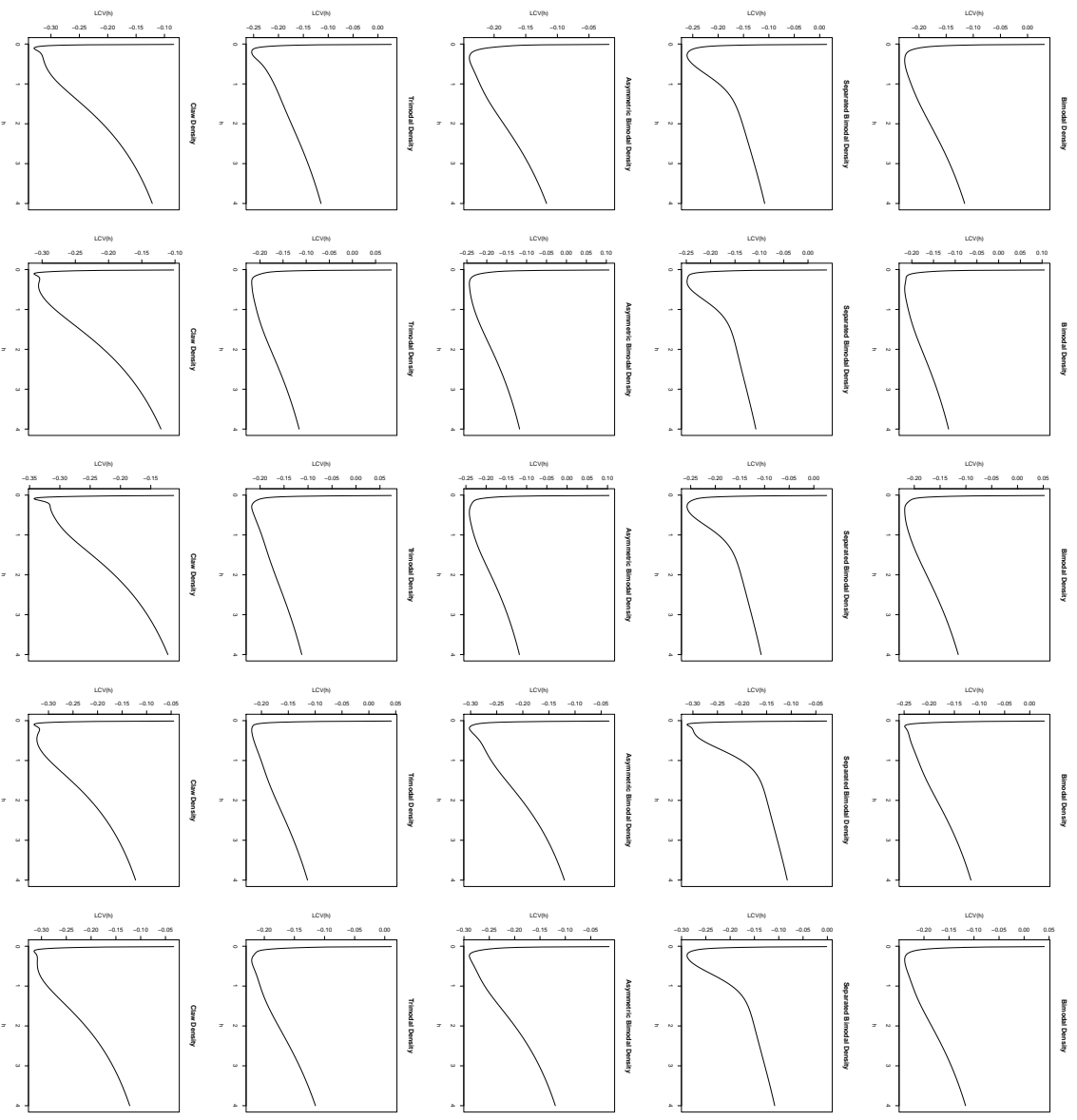


Figura 2.2: De izquierda a derecha y de arriba a abajo, representaciones de la curva (2.16) para las cinco primeras simulaciones de las densidades #6, #7, #8, #9 y #10.

2.3. SIMULACIÓN

19

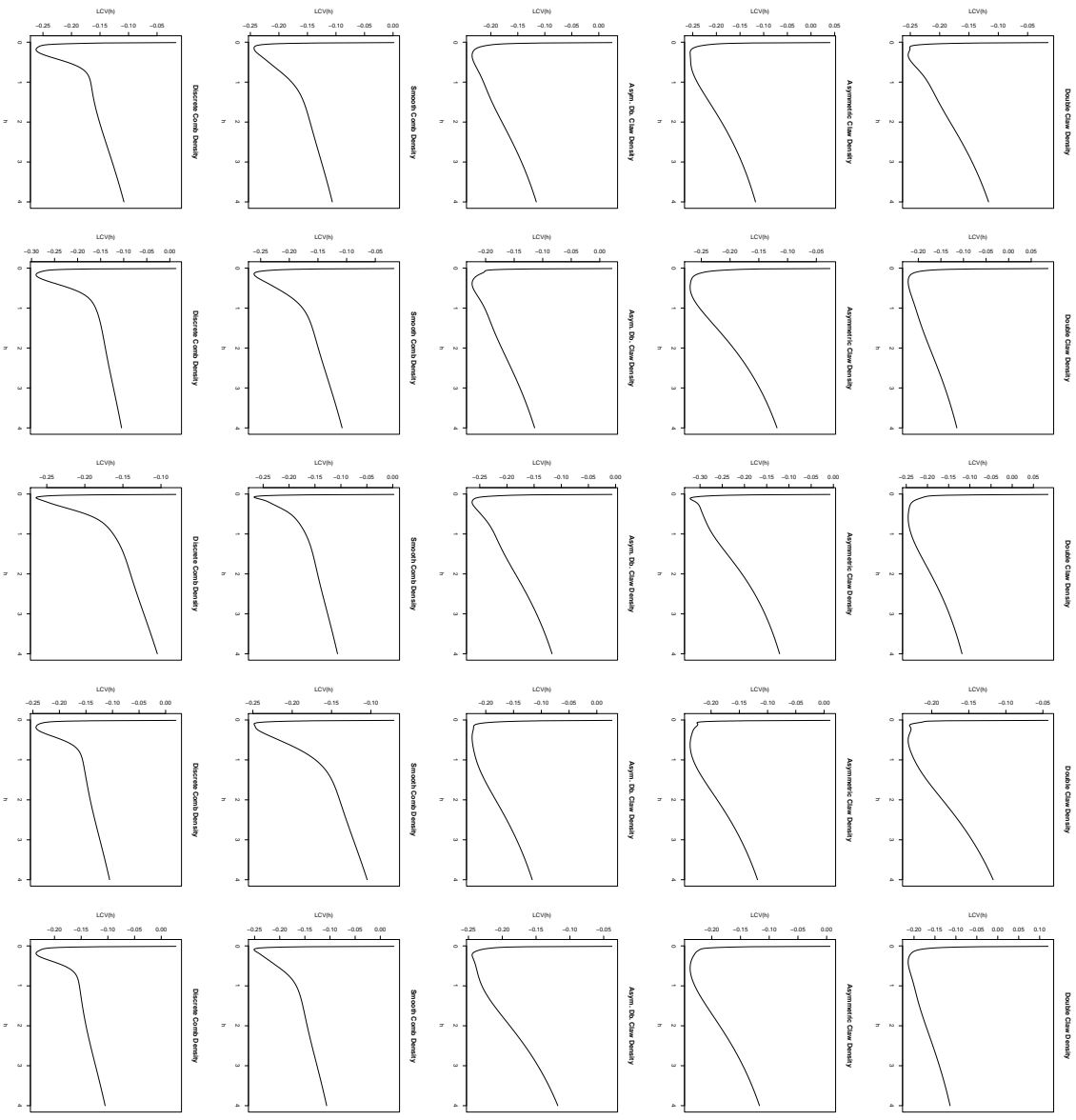


Figura 2.3: De izquierda a derecha y de arriba a abajo, representaciones de la curva (2.16) para las cinco primeras simulaciones de las densidades #11, #12, #13, #14 y #15.

Simulación	Algoritmo 1	Algoritmo 2 (1)	Algoritmo 2 (2)
1	0.35	0.3511	0.3511
2	0.30	0.2983	0.2983
3	0.39	0.3920	0.3920
4	0.38	0.3839	0.3839
5	0.30	0.2989	0.2989
6	0.29	0.2942	0.2942
7	0.33	0.3251	0.3251
8	0.30	0.2989	0.2989
9	0.31	0.3145	0.3145
10	0.36	0.3567	0.3567
11	0.35	0.3506	0.3506
12	0.44	0.4361	0.4361
13	0.33	0.3254	0.3254
14	0.29	0.2927	0.2927
15	0.44	0.4430	0.4430
16	0.32	0.3175	0.3175
17	0.31	0.3131	0.3131
18	0.28	0.2788	0.2788
19	0.37	0.3665	0.3665
20	0.38	0.3798	0.3798
21	0.37	0.3653	0.3653
22	0.24	0.2439	0.2439
23	0.41	0.4134	0.4134
24	0.34	0.3407	0.3407
25	0.30	0.3029	0.3029
26	0.27	0.2737	0.2737
27	0.23	0.2292	0.2292
28	0.27	0.2716	0.2716
29	0.31	0.3056	0.3056
30	0.21	0.2145	0.2145

Cuadro 2.1: Valores de las ventanas de validación cruzada que resultan de aplicar el algoritmo 1 (utilizando una rejilla de extremos 0 y 1 y paso 0.05) y el algoritmo 2 ((1) utilizando 8 subintervalos; (2) utilizando 4 subintervalos) para la densidad #2.

Dados los resultados obtenidos en el Cuadro 2.1, concluimos que en los tres casos se obtienen los mismos valores de ventana (excepto algún cambio en el cuarto decimal), si bien en el primero de los casos tan sólo se consigue una aproximación de dos decimales. Debemos tener en cuenta que el algoritmo 2 es computacionalmente más eficiente que el algoritmo 1 (y más aún cuantos menos subintervalos tenga en cuenta). Aún así, dado que tan sólo hemos considerado 30 simulaciones de una densidad, seremos precavidos y escogeremos para realizar el proceso de simulación completa el algoritmo 2 con 8 subintervalos.

Las restantes partes del proceso de simulación consistirán en la implementación directa de los algoritmos definidos en la Sección 2.2. Utilizaremos como intervalo de integración $J = [-4, 4]$ en el algoritmo 4, dado que este intervalo es válido sea cual sea la densidad $f(\cdot)$ (ningún valor de los generados en la simulación de las densidades se encuentra fuera de este intervalo). El código correspondiente se anexa en el Capítulo 4, Sección 1.

2.4. Resultados y conclusiones

En esta sección se comentarán los resultados obtenidos con la simulación realizada, de manera que se sea posible extraer conclusiones generales sobre cuándo funciona mejor un método u otro (selección por validación cruzada o selección por escala normal) y por qué. Para ello se utilizarán las siguientes herramientas:

- **Representación gráfica de las estimaciones núcleo $\hat{f}_{h,K}(\cdot)$ de la función de densidad $f(\cdot)$ construídas con las ventanas \hat{h}_{CV} y \hat{h}_{NS} .** Se realizarán para las cinco primeras simulaciones de cada densidad, cuyas gráficas pueden observarse en las Figuras 2.4, 2.5, 2.6 y 2.7. En ellas, aparece con línea continua la curva de la función de densidad teórica, con línea discontinua la curva de la estimación núcleo realizada con \hat{h}_{CV} y con línea punteada la estimación núcleo realizada con \hat{h}_{NS} . Basándonos en dichas gráficas analizamos que la estimación núcleo con ventana \hat{h}_{NS} parece ajustarse relativamente bien a las densidades unimodales no excesivamente asimétricas (densidades #1, #2 y #5), es decir, a las densidades próximas a una Normal. Para las densidades fuertemente asimétricas, como es el caso de la densidad #3, el hecho de utilizar \hat{h}_{NS} se traduce en una aproximación sobresuavizada. Para las cinco primeras densidades, las diferencias entre las aproximaciones obtenidas utilizando una y otra ventana no son muy grandes. Sin embargo, al introducirse el fenómeno de la multimodalidad en las densidades, la ventana Normal sobresuaviza de forma drástica, eliminándose casi por completo el efecto multimodal en la estimación, mientras que

la ventana de validación cruzada intenta aproximarse a las modas. El ejemplo más claro se observa para las densidades #14 y #15, en donde la ventana Normal no es capaz de captar en absoluto la mayoría de las modas que aparecen a la derecha de las representaciones. Se debe notar también que el intento de la ventana de validación cruzada por acercarse a las modas se traduce en estimaciones que presentan ruido en las colas (por ejemplo, en las aproximaciones de las densidades #3, #4, #5 y #10).

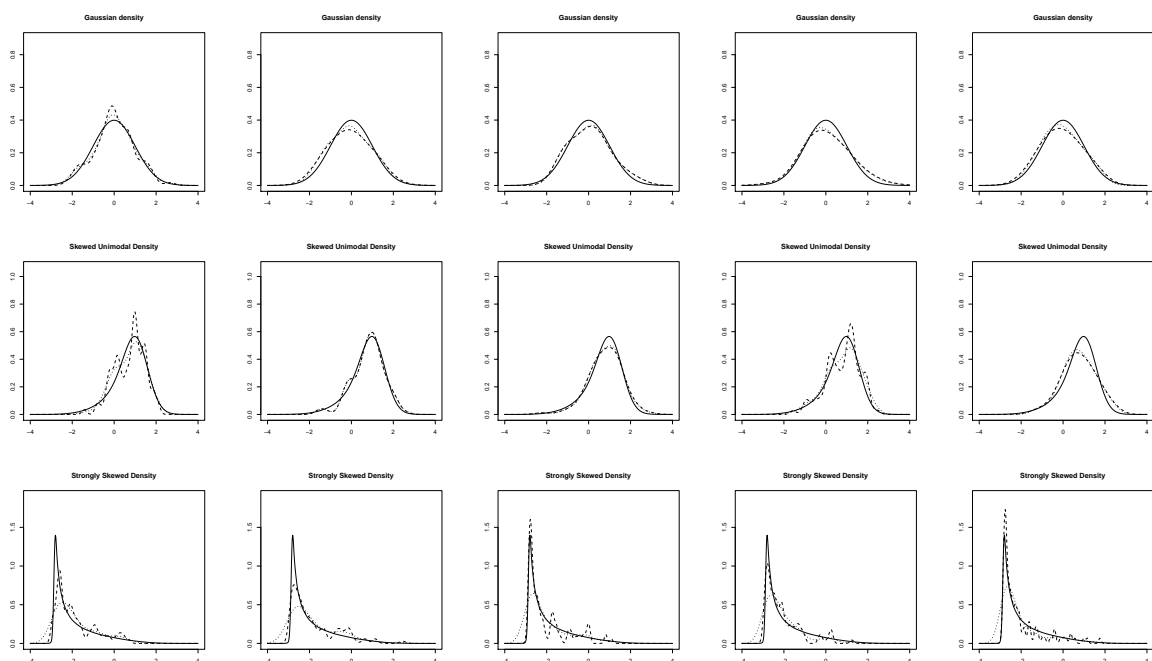


Figura 2.4: Representación de la curva de densidad teórica (línea sólida) acompañada de las estimaciones núcleo con ventanas \hat{h}_{CV} (línea discontinua) y \hat{h}_{NS} (línea punteada) para las densidades #1, #2 y #3.

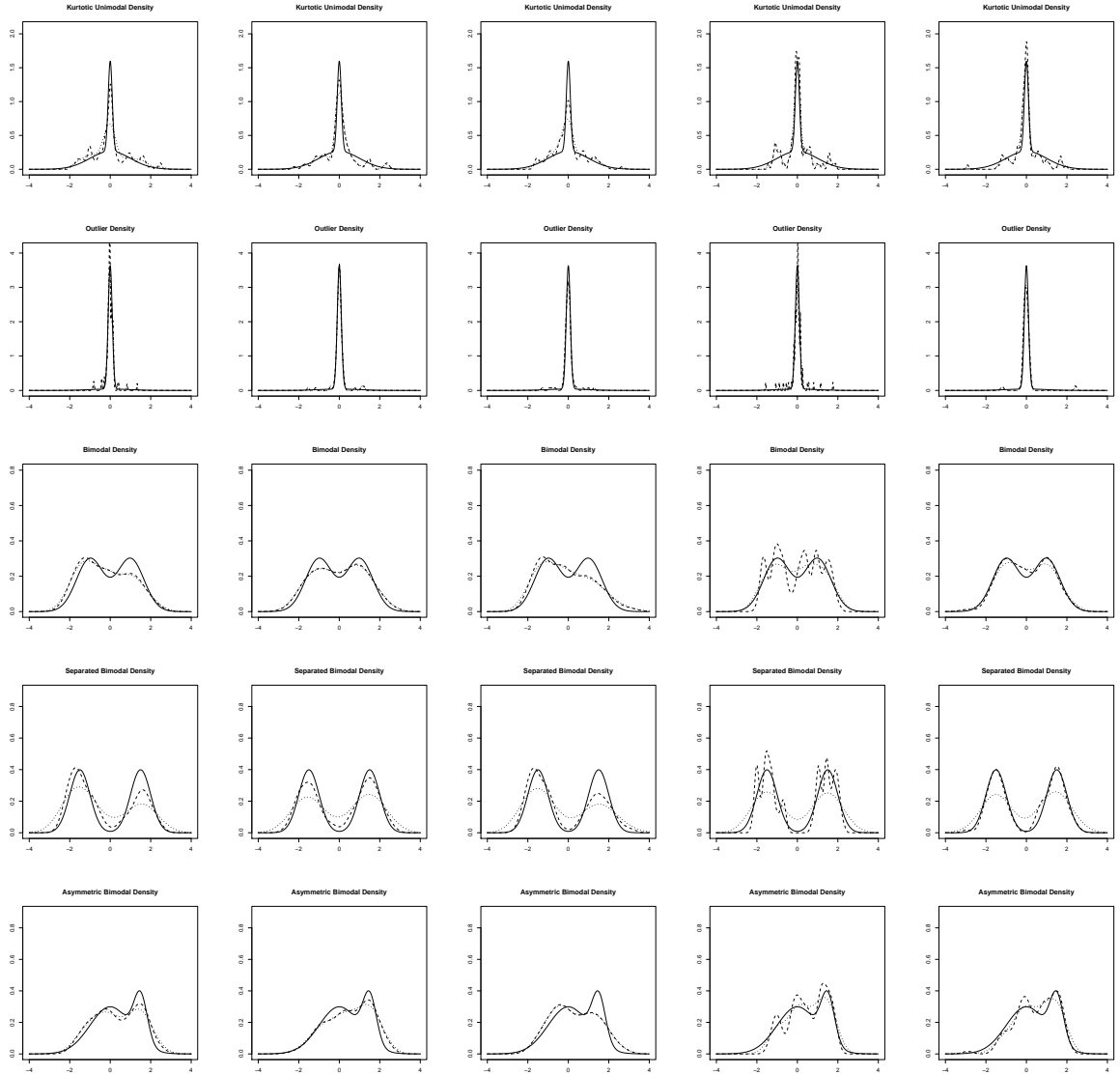


Figura 2.5: Representación de la curva de densidad teórica (línea sólida) acompañada de las estimaciones núcleo con ventanas \hat{h}_{CV} (línea discontinua) y \hat{h}_{NS} (línea punteada) para las densidades #4, #5, #7, #7 y #8.

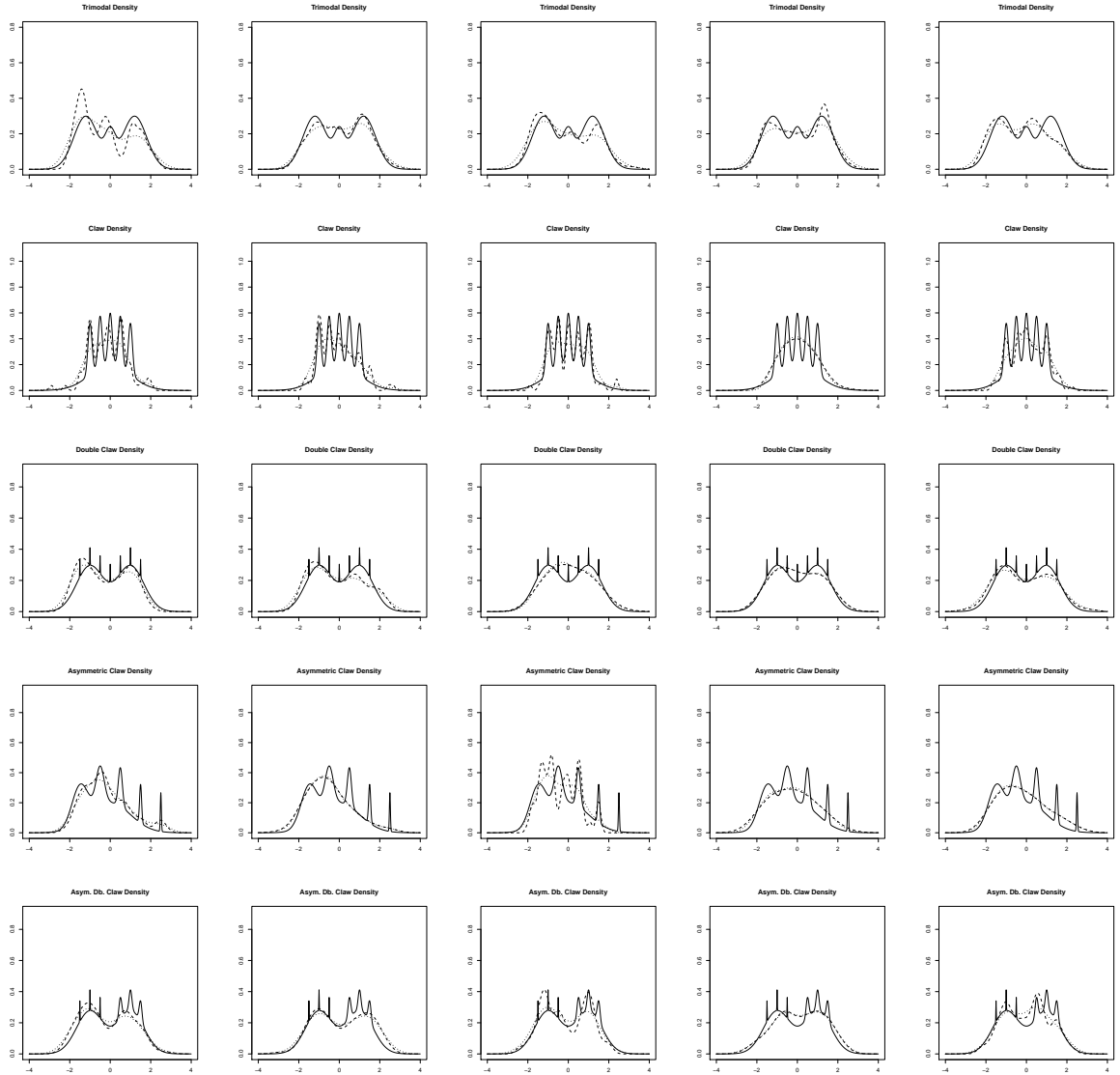


Figura 2.6: Representación de la curva de densidad teórica (línea sólida) acompañada de las estimaciones núcleo con ventanas \hat{h}_{CV} (línea discontinua) y \hat{h}_{NS} (línea punteada) para las densidades #9, #10, #11, #12 y #13.

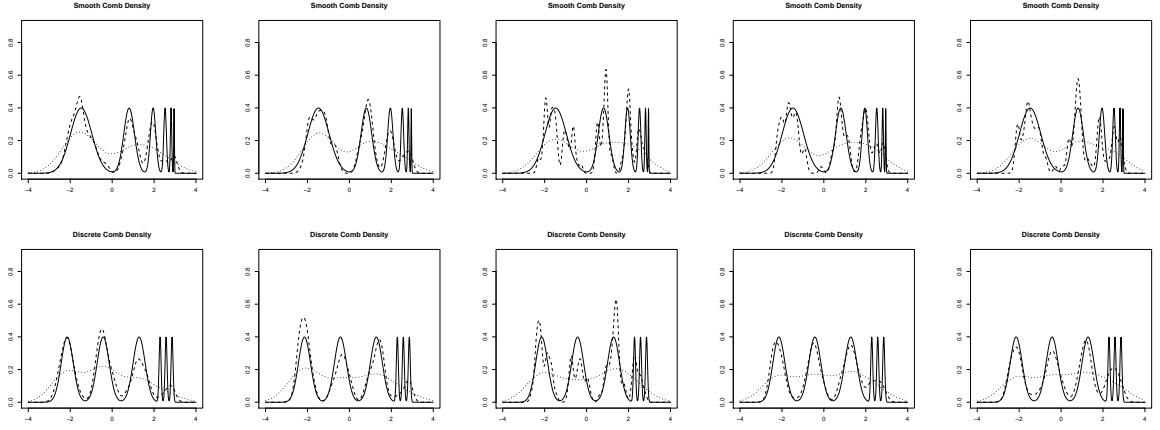


Figura 2.7: Representación de la curva de densidad teórica (línea sólida) acompañada de las estimaciones núcleo con ventanas \hat{h}_{CV} (línea discontinua) y \hat{h}_{NS} (línea punteada) para las densidades #14 y #15.

- **Resumen de estadísticos descriptivos de las simulaciones de las ventanas \hat{h}_{CV} y \hat{h}_{NS} y de los valores de las expresiones $ISE(\hat{f}_{h_{CV}})$ y $ISE(\hat{f}_{h_{NS}})$ asociadas, para cada densidad.** Dichos valores se resumen en los Cuadros 2.2 y 2.3 (se denota por Dens. la densidad, por Mín. el mínimo, por Med. la mediana, por Máx. el máximo y por Sd la desviación típica).
- **Gráficos boxplot comparativos entre las dos ventanas en estudio (\hat{h}_{CV} y \hat{h}_{NS}) y de los errores cuadráticos integrados asociados a cada una de ellas ($ISE(\hat{f}_{h_{CV}})$ y $ISE(\hat{f}_{h_{NS}})$), para cada densidad, contruídos a partir de las simulaciones.** Las gráficas se encuentran en las Figuras 2.8 y 2.9, respectivamente.
- **Resumen del porcentaje de veces que la ventana normal \hat{h}_{NS} ofrece mejores resultados en las simulaciones que la ventana de validación cruzada \hat{h}_{CV} (es decir, el porcentaje de veces asociado a que $ISE(\hat{f}_{h_{NS}})$ es menor que $ISE(\hat{f}_{h_{CV}})$), para cada densidad.** Dichos porcentajes se resumen en el Cuadro 2.4.

	Dens.	Mín.	Media	Med.	Máx.	Sd
h_{CV}	#1	0.0404	0.4394	0.4696	0.6681	0.1308
	#2	0.0241	0.3123	0.3271	0.5160	0.0826
	#3	0.0138	0.0903	0.0888	0.1965	0.0310
	#4	0.0160	0.0834	0.0825	0.2242	0.0255
	#5	0.0071	0.0464	0.0485	0.0729	0.0122
	#6	0.0712	0.4122	0.4149	0.7393	0.1360
	#7	0.0662	0.2625	0.2738	0.3872	0.0607
	#8	0.0528	0.3418	0.3313	0.6521	0.1274
	#9	0.0584	0.3717	0.3681	0.7674	0.1302
	#10	0.0299	0.1970	0.1098	0.5446	0.1453
	#11	0.0840	0.4029	0.3976	0.7780	0.1336
	#12	0.0475	0.2888	0.2144	0.7201	0.1706
	#13	0.0734	0.3827	0.3833	0.6958	0.1310
	#14	0.0319	0.1458	0.1455	0.2758	0.0486
	#15	0.0307	0.1523	0.1663	0.2486	0.0474
h_{NS}	#1	0.2707	0.4049	0.4100	0.5022	0.0393
	#2	0.2156	0.3055	0.3045	0.4109	0.0353
	#3	0.2087	0.3805	0.3797	0.5413	0.0609
	#4	0.0831	0.2016	0.2001	0.3590	0.0513
	#5	0.0286	0.0469	0.0465	0.0644	0.0055
	#6	0.4281	0.5067	0.5059	0.6018	0.0267
	#7	0.6072	0.6672	0.6672	0.7412	0.0214
	#8	0.3598	0.4621	0.4627	0.5432	0.0282
	#9	0.4538	0.5375	0.5372	0.6196	0.0263
	#10	0.2676	0.3591	0.3583	0.4378	0.0275
	#11	0.4296	0.5049	0.5060	0.5892	0.0261
	#12	0.3547	0.4658	0.4662	0.5530	0.0322
	#13	0.4143	0.5011	0.5015	0.5971	0.0254
	#14	0.6199	0.6918	0.6911	0.7697	0.0283
	#15	0.6067	0.7110	0.7126	0.8032	0.0345

Cuadro 2.2: Estadísticos descriptivos de \hat{h}_{CV} y \hat{h}_{NS} obtenidos tras la realización del proceso de simulación.

	Dens.	Mín.	Media	Med.	Máx.	Sd
ISE($\hat{f}_{h_{CV}}$)	#1	0.0014	0.0696	0.0573	0.2888	0.0493
	#2	0.0035	0.0693	0.0542	0.3832	0.0550
	#3	0.0269	0.0967	0.0771	0.4619	0.0605
	#4	0.0214	0.0895	0.0702	0.8180	0.0693
	#5	0.0063	0.0667	0.0567	0.3427	0.0458
	#6	0.0083	0.0771	0.0653	0.2723	0.0486
	#7	0.0064	0.0683	0.0607	0.2297	0.0374
	#8	0.0077	0.0851	0.0735	0.3158	0.0516
	#9	0.0090	0.0830	0.0718	0.3141	0.0506
	#10	0.0187	0.1974	0.0989	0.6998	0.1712
	#11	0.0361	0.1172	0.1014	0.4039	0.0604
	#12	0.0297	0.1570	0.1059	0.5471	0.1054
	#13	0.0439	0.1276	0.1047	0.3945	0.0673
	#14	0.0465	0.1217	0.1099	0.3801	0.0495
	#15	0.0510	0.1191	0.1137	0.2827	0.0377
ISE($\hat{f}_{h_{NS}}$)	#1	0.0032	0.0524	0.0449	0.2699	0.0372
	#2	0.0040	0.0575	0.0454	0.2600	0.0438
	#3	0.3529	1.1133	1.0904	2.1936	0.3626
	#4	0.0472	0.4505	0.3992	1.4928	0.2818
	#5	0.0078	0.0578	0.0461	0.2743	0.0412
	#6	0.0144	0.0849	0.0814	0.2261	0.0373
	#7	0.4253	0.6192	0.6161	0.9169	0.0834
	#8	0.0297	0.1084	0.1022	0.2909	0.0406
	#9	0.0391	0.1180	0.1129	0.2688	0.0387
	#10	0.2933	0.3799	0.3697	0.5613	0.0483
	#11	0.0484	0.1375	0.1321	0.3726	0.0433
	#12	0.1863	0.2594	0.2518	0.4770	0.0413
	#13	0.0716	0.1601	0.1545	0.3948	0.0453
	#14	1.0385	1.2583	1.2525	1.6337	0.1017
	#15	1.3216	1.6354	1.6320	1.9609	0.1077

Cuadro 2.3: Estadísticos descriptivos de $ISE(\hat{f}_{h_{CV}})$ y $ISE(\hat{f}_{h_{NS}})$ obtenidos tras la realización del proceso de simulación.

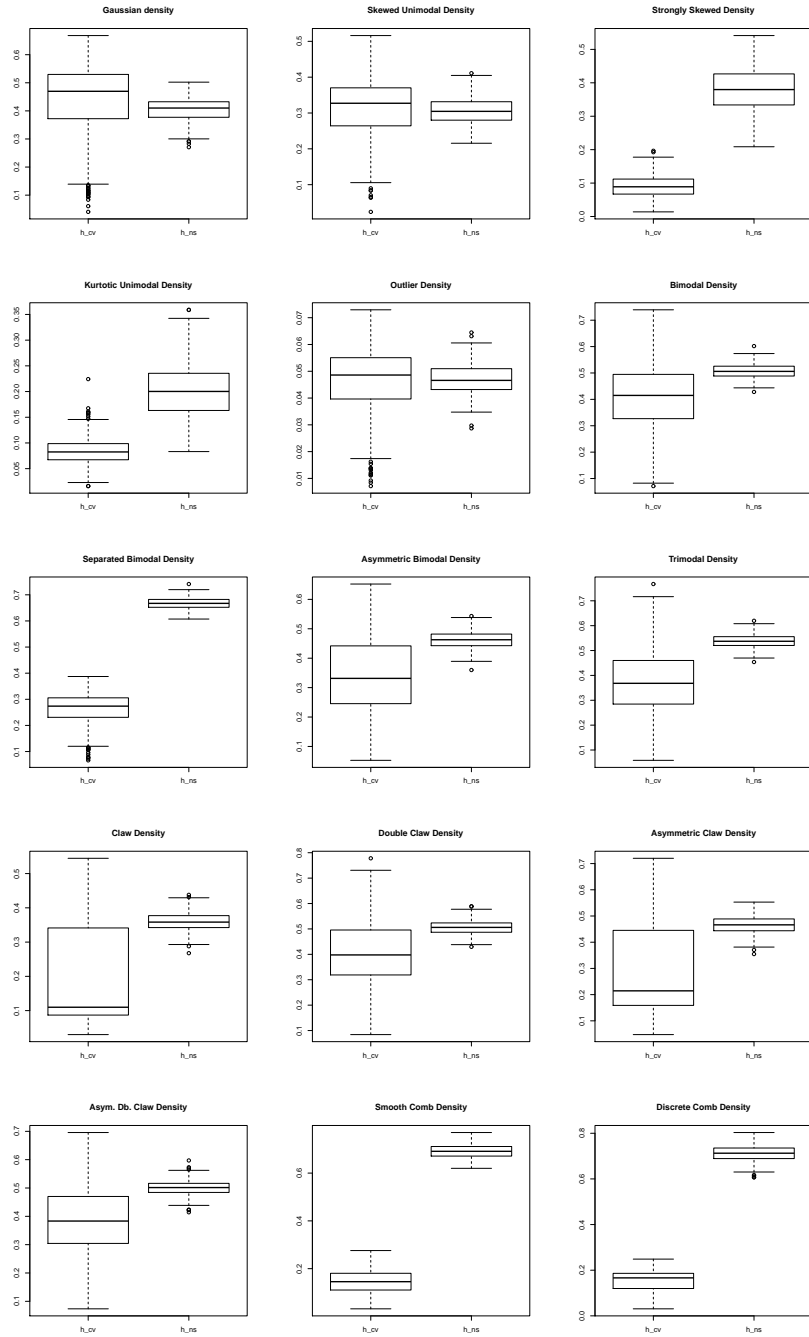


Figura 2.8: De izquierda a derecha, y de arriba a abajo, gráficos boxplot comparativos asociados a las simulaciones de las ventanas \hat{h}_{CV} y \hat{h}_{NS} , de la densidad #1 hasta la #15.

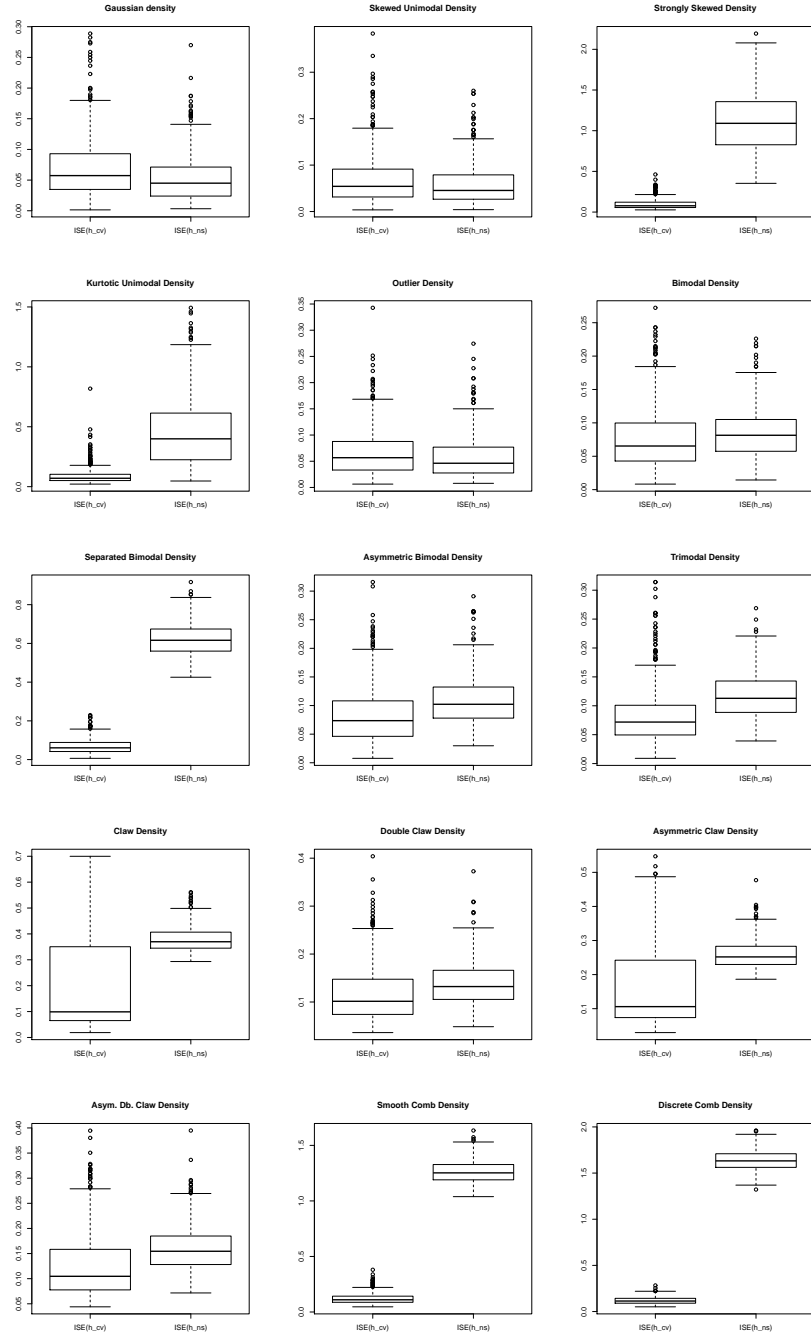


Figura 2.9: De izquierda a derecha, y de arriba a abajo, gráficos boxplot comparativos asociados a las simulaciones de las ventanas $ISE(\hat{f}_{h_{cv}})$ y $ISE(\hat{f}_{h_{ns}})$, de la densidad #1 hasta la #15.

Densidad	#1	#2	#3	#4	#5	#6	#7	#8
Porcentaje	74.6 %	71.6 %	0.0 %	1.4 %	68.4 %	29.2 %	0.0 %	23.0 %
Densidad	#9	#10	#11	#12	#13	#14	#15	
Porcentaje	13.6 %	23.0 %	24.6 %	22.6 %	19.0 %	0.0 %	0.0 %	

Cuadro 2.4: Porcentaje de simulaciones en las que $ISE(\hat{f}_{h_{NS}})$ es menor que $ISE(\hat{f}_{h_{CV}})$, para cada una de las densidades.

Atendiendo a los resultados que se presentan en el Cuadro 2.2, se observan diferencias entre las desviaciones típicas asociadas a las ventanas normales y las asociadas a las ventanas de validación cruzada: la variabilidad de las primeras es pequeña y prácticamente constante a lo largo de las densidades (la gran mayoría de las desviaciones típicas rondan el 0.03), mientras que el rango de variabilidad de las segundas no es independiente de las densidades (las desviaciones típicas varían desde el valor 0.0474 alcanzado para la densidad #15 hasta el valor 0.1706 obtenido para la densidad #12) y es en general mayor que el alcanzado para las ventanas normales. Si nos fijamos en los valores medios de ambas ventanas se tiene que en general, las ventanas normales son más grandes que las ventanas de validación cruzada (excepto para las densidades #1 y #2). En el Cuadro 2.3 se puede observar como, en media, el error cuadrático integrado asociado a la ventana de validación cruzada en general es más pequeño que el asociado a la ventana normal (excepto para las densidades #1, #2 y #5).

Los gráficos que aparecen en las Figuras 2.8 y 2.9 generalizan los resultados que acabamos de comentar en el párrafo anterior. Cabe destacar el caso de la densidad Normal: aún cuando la ventana Normal proporciona buenas aproximaciones de una forma fácil y sencilla, el proceso de selección de ventana por el método de validación cruzada no acaba de definirse por un intervalo de ventanas concreto, obteniéndose así una gran cantidad de *outliers* en el gráfico que aparece en la Figura 2.8 para esta densidad (llega a escoger en algún caso un valor de ventana menor que 0.5, infrasuavizando la estimación núcleo de la densidad de forma extrema). Llama también la atención los gráficos de la Figura 2.9 asociados a las densidades #14 y #15, en las que se ve claramente que el error cuadrático integrado asociado a la ventana de validación cruzada es mucho menor que el asociado a la ventana Normal.

Los porcentajes que aparecen en el Cuadro 2.4 no hacen sino ratificar lo analizado hasta ahora: la ventana Normal funciona mejor para las densidades #1, #2 y #4, mientras los resultados son mejores para el resto de densidades utilizando la ventana de validación cruzada. Se observa también como para las densidades #3, #7, #14 y #15 no existe

ninguna simulación en la que el error cuadrático integrado asociado a la ventana Normal haya sido menor que el asociado a la ventana de validación cruzada.

Una vez analizadas todas las herramientas de las que disponemos, las conclusiones que podemos extraer sobre el funcionamiento del parámetro de suavizado utilizando el método de validación cruzada en comparación con el método de escala Normal son las siguientes:

1. Los selectores de ventana de escala Normal proporcionan una primera aproximación de forma fácil y rápida, y cabe esperar resultados razonables cuando la muestra de la que se dispone está próxima a la de una Normal. Para este tipo de densidades funciona mejor la ventana Normal, pues la ventana de validación cruzada tiende a infrasuavizar en algunos casos.
2. Para densidades unimodales fuertemente asimétricas, la ventana Normal tiende a sobresuavizar. En estos casos se prefiere la ventana de validación cruzada.
3. Para densidades unimodales leptocúrticas (exceptuando los casos extremos como la densidad #5), se prefiere la ventana Normal, pues la ventana de validación cruzada en general se aproxima más a la moda pero a base de incluir mucho ruido en las colas.
4. Para densidades multimodales no se debe utilizar la ventana Normal, pues tiende a sobresuavizar en exceso y a enmascarar la existencia de modas. En este caso conviene utilizar la ventana de validación cruzada.

Capítulo 3

Estimación noparamétrica de la regresión

3.1. Objetivos

El objetivo de este capítulo será comprobar el funcionamiento de la selección del parámetro de suavizado en la estimación núcleo de la regresión, realizando en este caso el análisis sobre un conjunto de datos reales. El conjunto de datos a emplear será el *data frame* denominado *airquality*, que contiene diversas medidas de la calidad del aire en Nueva York entre mayo y septiembre de 1973. El análisis se centrará en las medidas de temperatura y niveles de ozono. Se desea analizar la relación existe entre la temperatura y la concentración de ozono en la ciudad de Nueva York, siendo la variable independiente la temperatura. Para hacer el análisis de regresión se considerará el método local lineal, abordando los siguientes problemas:

1. Analizar la distribución univariante de las dos variables involucradas en el estudio. ¿Qué se puede decir sobre el clima en Nueva York (al menos en la década de los 70)?
2. Escribir una función en R que, dada una muestra $\{(x_i, y_i) : i = 1, \dots, n\}$ y una ventana h , devuelva el valor del estimador local lineal en un vector cualquiera $t = (t_1, \dots, t_s)$ (en particular devuelve el valor del estimador en el vector $x = (x_1, \dots, x_n)$).
3. Escribir una función en R que permita calcular la función de validación cruzada para el estimador local lineal. ¿Qué ventana selecciona el criterio de validación cruzada para los datos analizados?

4. Dibujar los datos analizados y el ajuste realizado por el método local lineal, usando como parámetro de suavizado la ventana de validación cruzada. ¿Resulta satisfactorio el ajuste realizado?
5. En regresión también existe una alternativa *plug-in* a la selección del parámetro de suavizado, implementada dentro de la librería `KernSmooth`. La función `dpill` permite calcular la ventana *plug-in* para la regresión. Utilizar dicha función para seleccionar un parámetro de suavizado alternativo al calculado por validación cruzada. ¿Cuál de las dos alternativas parece más razonable? ¿Por qué?
6. Utilizar el estimador de Nadaraya-Watson, utilizando la ventana seleccionada por el método de validación cruzada para este estimador, para analizar los datos. ¿Dónde se observan mayores diferencias con el método local lineal?
7. A la vista de los resultados anteriores, ¿se podría considerar un modelo sencillo para la descripción de los datos? Utilizar la regresión lineal simple para estimar dicho modelo. Comparar los resultados de ese ajuste lineal con los obtenidos con las técnicas no paramétricas.

3.2. Análisis de las distribuciones univariantes de las variables en estudio

Como comentábamos en la Sección 1 de este capítulo, las variables con las que trabajaremos serán la temperatura (medida en grados Fahrenheit) y la concentración de ozono (medida en partes por billón -*ppb*-) en la ciudad de Nueva York en el período que abarca del 1 de mayor al 30 de septiembre de 1973, guardadas en el *data frame* denominado `airquality`. Debemos tener en cuenta que en el ozono existen datos faltantes, por lo que se realizará el análisis sobre los restantes datos (eliminando en la temperatura también los índices correspondientes a dichos datos faltantes).

En primer lugar se realizará un estudio descriptivo de las variables. En el Cuadro 3.1 se resumen los estadísticos descriptivos, denotando por Var. la variable, Mín. el mínimo, Med. la media, Máx. el máximo y Sd la desviación típica. A dicho cuadro le acompañan los digramas de cajas de ambas variables que aparecen en la Figura ??, en la que se puede observar la existencia de datos atípicos en la concentración de ozono

3.2. ANÁLISIS DE LAS DISTRIBUCIONES UNIVARIANTES DE LAS VARIABLES EN ESTUDIO3

Var.	Mín.	Media	Med.	Máx.	Sd
Ozono	1.00	42.13	31.50	168.00	32.98788
Temperatura	57.00	77.87	79.00	97.00	9.485486

Cuadro 3.1: Estadísticos descriptivos asociados a las variables ozono y temperatura.

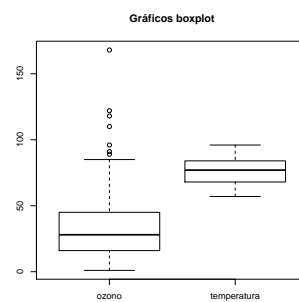


Figura 3.1: Diagrama de cajas del ozono y de la temperatura.

Se investigará ahora la distribución de las variables. En la Figura 3.2 se presentan los histogramas de cada una de las variables, y la estimación de la densidad utilizando el estimador núcleo con ventana de validación cruzada. En el ozono obtenemos una estimación de una densidad fuertemente asimétrica, por lo que cabe esperar que la distribución del mismo no sea Normal. Sin embargo, para la temperatura, si bien se aprecia un pequeño grado de asimetría, la gráfica se adecúa bastante a la de una variable con distribución Normal.

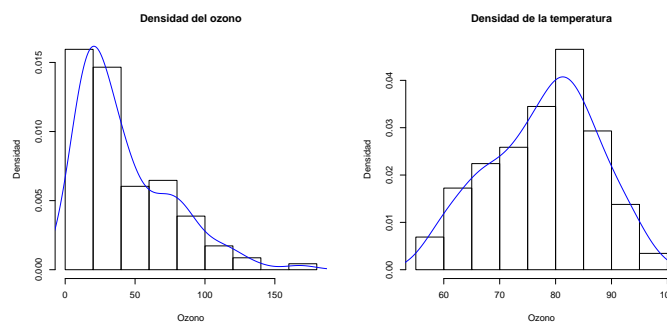


Figura 3.2: De izquierda a derecha, histogramas del ozono y de la temperatura, respectivamente, acompañados de las correspondientes estimaciones núcleo realizadas con la ventana de validación cruzada.

Se realiza a continuación el test de normalidad de Shapiro-Wilks para contrastar la normalidad de las variables. Para el ozono se obtiene un p -valor de $2.790\text{e-}08$, por lo que se rechaza la hipótesis de normalidad; para la temperatura, el p -valor obtenido es 0.0719 , por lo que se acepta normalidad con un nivel de confianza de 0.95 .

Poderíamos plantearnos qué forma tiene la serie de tiempo de la temperatura asociada a este período (se realiza sobre la variable sin eliminar los índices correspondientes a los datos faltantes del ozono, ver Figura 3.3). Como es de esperar, las temperaturas son mayores en los meses centrales (julio y agosto) y más bajas en los restantes meses. Podría pensarse en realizar el estudio análogo para la concentración de ozono, pero la base de datos posee datos faltantes en esta variable, por lo que no tendría sentido.

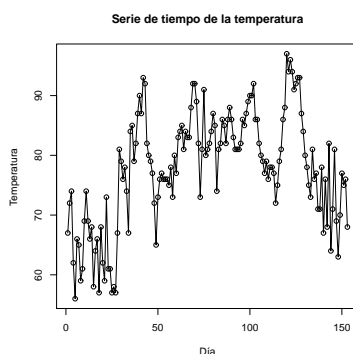


Figura 3.3: Serie de tiempo asociada a la variable temperatura, en el período en el que registran las observaciones.

En la primera gráfica de la Figura 3.4 se presenta el gráfico de dispersión de la concentración de ozono con respecto a la temperatura. Como puede observarse, parece que existe relación entre la temperatura y la concentración de ozono, de forma que la concentración de ozono se mantiene prácticamente constante durante un rango de temperaturas, para luego aumentar de forma lineal a partir de un cierto valor de la temperatura. Podría pensarse en un ajuste mediante un modelo de regresión lineal simple, pero como puede observarse en la segunda gráfica de la Figura 3.4, el ajuste no es bueno (además, el estadístico R^2 ajustado es 0.4832). Así, si realizamos los diagramas de cajas de la concentración de ozono sobre dos rangos de temperaturas (menores o iguales a 75 y mayores que 75), se observa como los rangos entre los que varía la concentración de ozono son muy distintos (ver Figura 3.5).

3.3. ESTIMADOR LINEAL LOCAL CON VENTANA DE VALIDACIÓN CRUZADA 37

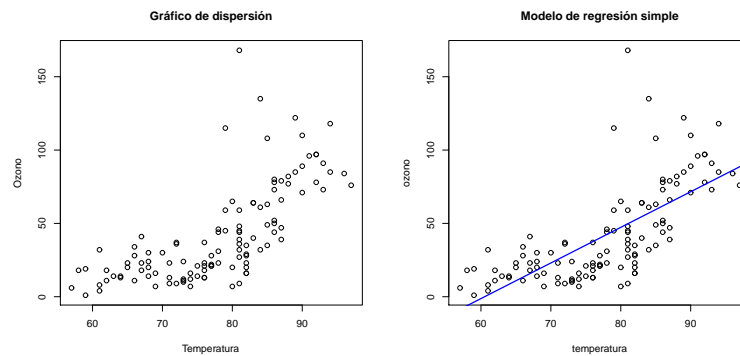


Figura 3.4: De izquierda a derecha, gráfico de dispersión de la concentración de ozono con respecto a la temperatura y ajuste obtenido mediante el método de regresión lineal simple.

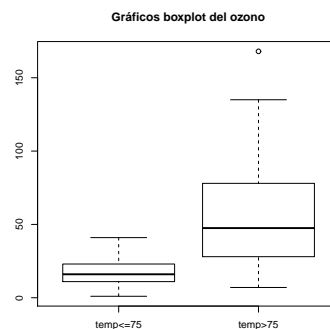


Figura 3.5: Diagramas de cajas de la concentración de ozono sobre dos rangos de temperaturas (menores o iguales a 75 y mayores que 75).

Concluimos tras este estudio la existencia de relación entre ambas variables, no explicándose la variable correspondiente a la concentración de ozono con un modelo lineal simple sobre la temperatura. Será necesario por tanto utilizar otros métodos, entre ellos la regresión no paramétrica, que será el objeto de estudio de las siguientes secciones.

3.3. Estimador lineal local con ventana de validación cruzada

No sabemos que escribir

3.3.1. Marco teórico

Considérese una muestra de una v.a bidimensional (X, Y) , es decir, un conjunto de pares $(X_1, Y_1), \dots, (X_n, Y_n)$ independientes e idénticamente distribuidos a (X, Y) y $\mathbb{X} = (X_1, \dots, X_n)$. Entonces, si la variable dependiente Y depende de la variable independiente X , se tiene que

$$Y = m(X),$$

donde $m(\dots)$ es una función, a priori desconocida. En términos generales, la regresión suele formalizarse como la media condicional de la variable respuesta en función del valor que tome la variable explicativa. La función de regresión es, por tanto, de la forma

$$m(X) = \mathbb{E}(Y \mid X = x), \text{ para cualquier valor } x \text{ de } X. \quad (3.1)$$

En consecuencia, la variable respuesta se puede descomponer en función del resultado de X más un error de media 0:

$$Y = m(X) + \varepsilon$$

donde ε se conoce como error y verifica $\mathbb{E}(\varepsilon \mid X = x_i) = 0$, $i = 1, \dots, n$.

El objetivo sea estimar la función de regresión que viene expresada en (3.1). De ahora en adelante, se denotará por $f(\cdot)$ a la función de densidad de la variable X . De forma análoga a como se describió en el capítulo anterior, la función $K(\cdot)$ será usualmente una función de densidad simétrica denominada núcleo que será la que asigne los pesos a los promedios locales (se denotará por $K_h(\cdot)$ al núcleo reescalado).

La idea que subyace a la construcción del estimador lineal local es, para un x dado, modelar $m(\cdot)$ de forma lineal en un entorno de x cuyo tamaño está determinado por una ventana h [?]. Así, para obtener la curva de regresión en un punto x dado, se debe aplicar la técnica de la regresión lineal en una franja de datos del entorno de x , es decir,

$$Y_i = a(x) + b(x)X_i + \varepsilon_i, \text{ para } X_i \in x \pm h,$$

donde $a(\cdot)$ y $b(\cdot)$ dependen de x y h es la ventana (tamaño del entorno de x). Asignando el peso $K_h(x - X_i)$ para cada punto (X_i, Y_i) , con el objetivo de darle menos importancia a las contribuciones de los datos que se encuentren lejos de x , se tiene la siguiente función objetivo del problema de mínimos cuadrados ponderados:

$$\sum_{i=1}^n (Y_i - a(x) - b(x)X_i)^2 K_h(x - X_i). \quad (3.2)$$

3.3. ESTIMADOR LINEAL LOCAL CON VENTANA DE VALIDACIÓN CRUZADA 39

Denotando por $W_{hi} = \frac{K_h(x-X_i)}{\sum_{j=1}^n K_h(x-X_j)}$, (3.2) se puede escribir de la siguiente forma:

$$\sum_{i=1}^n (Y_i - a(x) - b(x)X_i)^2 W_{hi}. \quad (3.3)$$

Resolviendo (3.3), la solución obtenida es

$$(\hat{a}, \hat{b}) = \arg \min_{(a(x), b(x))} \sum_{i=1}^n (Y_i - a(x) - b(x)X_i)^2 W_{hi}.$$

Entonces, estimador lineal local en el punto x es:

$$\hat{m}_{LL}(x) = \hat{a}(x) + \hat{b}(x)x.$$

El estimador lineal local es un estimador lineal, es decir, se puede escribir de la forma

$$\hat{m}(x) = \sum_{j=1}^h l_j(x) Y_j$$

donde $\sum_{j=1}^n l_j(x) = 1$.

Teorema 3.3.1. *El estimador local lineal se puede escribir en la forma $\hat{m}_{LL}(x) = \sum_{j=1}^h l_j(x) Y_j$ donde $l_j(x) = \frac{b_j(x)}{\sum_{k=1}^n b_k(x)}$, con*

$$b_j(x) = K\left(\frac{X_j - x}{h}\right) (S_{n,2}(x) - (X_j - x)S_{n,1}(x))$$

y

$$S_{n,r}(x) = \sum_{j=1}^n K\left(\frac{X_j - x}{h}\right) (X_j - x)^r, \quad r = 1, 2.$$

Una posibilidad para elegir el parámetro de suavizado es utilizar el método de validación cruzada adaptado al contexto de la regresión. Esta técnica, como ya se describió en el Capítulo 2, consiste en, para cada i , con $i = 1, \dots, n$, tomar los datos $\{(X_j, Y_j), j \neq i\}$ para construir una estimación de la regresión $\hat{m}_{LL,-i}(\cdot)$ y realizar después la validación del modelo examinando el error de predicción $Y_i - \hat{m}_{LL,-i}(X_i)$. Así, la función de validación cruzada se define como

$$CV(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}_{LL,-i}(X_i))^2$$

donde $\hat{m}_{LL,-i}(\cdot)$ denota al estimador lineal local construido a partir de la muestra original después de eliminar el par (X_i, Y_i) .

En concreto, para cualquier estimador lineal $\hat{m}(\cdot) = \sum_{j=1}^n l_j(x)Y_j$ se define

$$\hat{m}_{-i}(x) = \sum_{j=1}^n l_{j,-i}(x)Y_j,$$

donde

$$l_{j,-i}(x) = \begin{cases} 0 & \text{si } j = i \\ \frac{l_j(x)}{\sum_{k \neq i} l_k(x)} & \text{si } j \neq i \end{cases}$$

Capítulo 4

Anexos

4.1. Código necesario para el Capítulo 2

4.2. Código necesario para el Capítulo 3

Bibliografía

- A. W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.
- J. Fan and J. S. Marron. Best possible constant for bandwidth selection. *The Annals of Statistics*, 20(4):2057–2070, 1992.
- P. Hall and J. S. Marron. Local minima in cross-validation functions. *J. Roy. Statist. Soc. Ser. B*, 53(1):245–252, 1991.
- J. S. Marron and M. P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992.
- E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076, 1962.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
- M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27:832–837, 1956.
- M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, 9(2):65–78, 1982.
- D. W. Scott. *Multivariate density estimation*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1992. Theory, practice, and visualization, A Wiley-Interscience Publication.
- S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 53(3):683–690, 1991.

- B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.
- J. Stoer and R. Bulirsch. *Introduction to numerical analysis*. Springer-Verlag, New York, 1980. Translated from the German by R. Bartels, W. Gautschi and C. Witzgall.
- M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1995.