Trabajo Fin de Máster

Estimación de la densidad con núcleo variable

Autora: MªIsabel Borrajo García
Tutor: Alberto Rodríguez Casal

Máster en Técnicas Estadísticas
Universidad de Santiago de Compostela
Enero 2014

Trabajo Fin de Máster

Estimación de la densidad con núcleo variable

Autora: MªIsabel Borrajo García
Tutor: Alberto Rodríguez Casal

El presente documento recoge el Trabajo Fin de Máster para el Máster en Técnicas Estadísticas realizado por Dña. Mª Isabel Borrajo García con el título "Estimación de la densidad con núcleo variable".

Ha sido realizado bajo la dirección de D. Alberto Rodríguez Casal que lo considera terminado y da su conformidad para la presentación y defensa del mismo.

Santiago de Compostela, a 10 de enero de 2014.

Fdo. Alberto Rodríguez Casal

Índice general

Resumen

La estimación de la densidad ha sido un tema estudiado en gran profundidad y mediante diferentes perspectivas a lo largo de los años.

Desde la introducción del histograma hasta llegar al estimador tipo núcleo con los diferentes métodos de selección del parámetro ventana, se han hecho numerosos avances. En el presente trabajo se aborda la estimación no paramétrica de la densidad desde un punto de vista que difiere del habitual y plantea por consiguiente, nuevos retos y desafíos.

El estudio que hemos realizado, propone eliminar la influencia de la ventana para centrarse exclusivamente en la función núcleo, lo que se consigue a través de un estimador similar al tipo núcleo pero sin ventana. Una vez definido el estimador, el objetivo se centra en buscar el núcleo que minimice el error cometido.

En las páginas de esta memoria se detallan varios procedimientos de estimación de la función núcleo y se realiza un estudio de simulación para compararlos. Además, los selectores de ventana más habituales para el estimador tipo núcleo también se incluyen en dicha comparación.

Para finalizar, se ilustra la aplicación práctica de las técnicas presentadas, abordando la estimación de la densidad de un conjunto de datos reales. Dichos datos son de gran interés socio-económico puesto que están relacionados con el volumen de renta de las familias gallegas, y deja patente la utilidad de las técnicas descritas en disciplinas alejadas de la estadística metodológica.

Capítulo 1

Introducción

La estadística es una ciencia transversal que se basa en el estudio de variables aleatorias y sus propiedades. Una de las formas de caracterizar una variable aleatoria es mediante su distribución, para la que la representación matemática más habitual es la función de distribución o la función de densidad.

Este trabajo se centra en la estimación de la función de densidad, cuya existencia está garantizada bajo el único supuesto de que la variable aleatoria que se esté estudiando sea absolutamente continua. La estimación de la densidad posee una gran variedad de aplicaciones en diversos ámbitos: a nivel exploratorio permite obtener información acerca de la estructura de un conjunto de datos; en teoría de la probabilidad se puede definir una medida empírica sobre la σ-álgebra de Borel con mejores propiedades que la medida empírica habitual[1]...

Una de las aplicaciones más útiles de la estimación de la densidad es la simulación. Es habitual, en diferentes disciplinas, obtener una muestra de la población de interés, pero no disponer de recursos o medios suficientes para poder recabar más datos. Es ahí donde entran en juego la estimación de la densidad y las técnicas de remuestreo, ya que aplicándolas, se pueden generar tantos datos como se quieran con distribución similar a la de la población inicial, sin más que tener una muestra representativa. Los nuevos datos reflejarán tanto mejor la estructura de la población cuanta más información dispongamos de la misma. Además, al contrario de lo que ocurre con la distribución empírica, al estimar la densidad se pueden simular fácilmente datos de una distribución absolutamente continua.

En la estimación de la densidad, como en la inferencia en general, existen dos posibles vías de estudio. Por una parte está la estimación paramétrica, en la que *a priori* se asume una determinada distribución de la variable y se emplean los datos en la estimación de los correspondientes parámetros. Por la otra, la estimación no paramétrica, que no asume ninguna hipótesis inicial y utiliza únicamente la información proporcionada por la muestra.

Tanto la estadística paramétrica como la no paramétrica poseen numerosos simpatizantes y detractores, pues ambas metodologías de trabajo tienen ventajas e inconvenientes que han sido ampliamente estudiadas a lo largo de los años.

La suposición inicial de que la población de la que proceden los datos sigue un modelo paramétrico puede limitar mucho el ajuste del modelo. En caso de ser correcta dicha suposición,

[1] La medida empírica habitual es aquella que asigna peso $1/n$ a cada uno de los datos, y a partir de los mismos, por propiedades de aditividad se puede definir la medida de cualquier conjunto.

el ajuste será muy bueno, pero si el modelo paramétrico es incorrecto, las conclusiones podrían ser totalmente erróneas. Por ello es deseable considerar técnicas no paramétricas que olviden cualquier hipótesis previa y trabajen únicamente con la información que proporcionan los datos; teniendo siempre presente la aleatoriedad intrínseca a los mismos.

Los principios de la estimación no paramétrica de la densidad datan de finales del s.XIX, cuando Karl Pearson introdujo el histograma, que no es más que la representación de las frecuencias por clases. El histograma es un estimador discontinuo, que además depende de la elección de un punto inicial y de un parámetro ventana, con gran influencia por parte de ambos en el resultado final. Para solventar el problema de la dependencia del punto inicial, hay que esperar hasta mediados del s.XX, cuando se desarrolló el denominado histograma móvil o estimador naive, que sigue siendo discontinuo y dependiendo de la ventana. Posteriormente, Parzen (1962) y Rosenblatt (1956), propusieron el estimador tipo núcleo, que sí es continuo y que por lo tanto, en la mayor parte de las ocasiones, se ajusta mejor a la realidad de los modelos estudiados, aunque también depende en gran medida de la elección de un parámetro ventana.

En la literatura estadística ha sido ampliamente estudiado el papel fundamental del parámetro ventana en el estimador tipo núcleo. Dicho parámetro es el que controla el grado de suavización del estimador, y una mala elección del mismo puede derivar en un estimador tanto infra como sobresuavizado. Debido a esto, la segunda mitad del s.XX fue muy prolífica en cuanto a métodos de selección de ventana, entre los que destacan el propuesto por Silverman (1986), el método de Sheather y Jones (1991) y el de Bowman (1984).

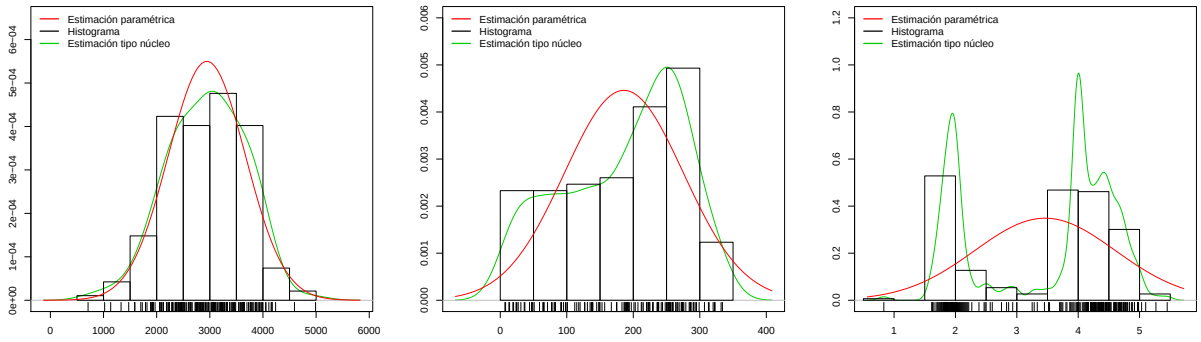
Una manera sencilla para hacerse una idea de las características de los diferentes estimadores de la densidad, es mediante la ilustración de los mismos a través de un ejemplo concreto con un conjunto de datos reales. En la Figura 1.1 se presentan los estimadores ya mencionados para tres conjuntos de datos a los que se tiene acceso desde el software R Core Team (2012). La base de datos *Birthwt* (librería MASS), que recoge diversas variables sobre 189 recién nacidos a lo largo del año 1986 en el Baystate Medical Center (Springfield, Massachusetts); el conjunto *Airquality* que contiene medidas diarias de variables de calidad de aire en la ciudad de Nueva York recogidas entre mayo y septiembre de 1973; y finalmente el conocido conjunto *Geyser* (librería MASS), muy empleado en la ilustración de la estimación de la densidad, y que recoge una versión de los datos de erupción de los geyser de Old Faithful en el parque nacional de Yellowstone entre el 1 y el 15 de agosto de 1985.

En el conjunto de datos *Birthwt* se ha escogido para esta ilustración la variable `bwt`, que es el peso al nacer, en gramos, de los bebés incluidos en el estudio. En *Airquality* la variable `Solar.R` que contiene los datos de radiación solar, medida en langleys, en Central Park entre las 8 y las 12 de la mañana. Y en *Geyser*, se estudia la variable `duration` que recoge, como su nombre indica, la duración de las erupciones en minutos.

Para cada una de las variables mencionadas, se representa una estimación paramétrica de la densidad, en este caso presentamos el modelo normal con media y varianza estimadas por máxima verosimilitud. Además se han obtenido el histograma y el estimador tipo núcleo con núcleo gaussiano y ventana escogida por la regla de Sheather y Jones (1991)

Analizando la Figura 1.1, la hipótesis de normalidad parece asumible para el primer conjunto de datos, pero no para el segundo (presentan asimetría) ni para el último (presentan una bimodalidad muy marcada). Nótese que no se apreciarían estas características si no dispusiésemos de

un método no paramétrico.



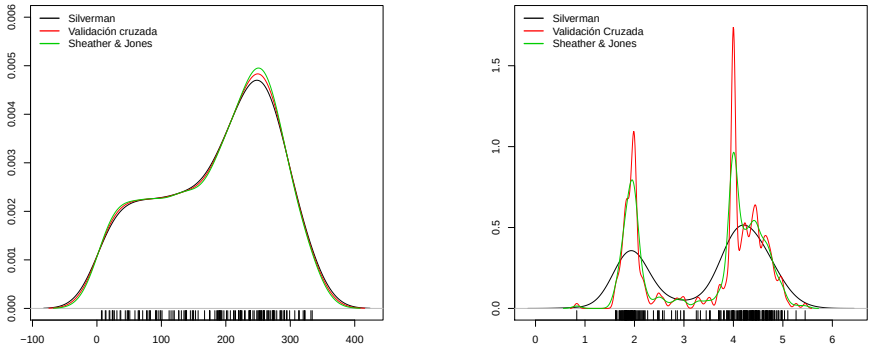
(a) Birthwt (b) Airquality (c) Geyser

Figura 1.1: Representación de estimaciones paramétrica y no paramétricas de la densidad para los datos de (a)*Birthwt*, (b)*Airquality* y (c)*Geyser*.

En lo relativo a los histogramas, se podría decir que, a pesar del problema de la discontinuidad, captan la estructura básica de los datos, aunque si bien es cierto que modificando el tamaño de la ventana, se podrían obtener estimaciones que difiriesen mucho de la dada. Por último, el estimador tipo núcleo parece proporcionar, tal y como era esperado, la estimación más ajustada a la información de los datos.

Observando los gráficos (b) y (c) de la Figura 1.1, queda patente el hecho que se comentaba anteriormente a propósito de la estadística paramétrica, y la posibilidad de cometer graves errores en los procedimientos inferenciales asociados si nuestros datos no cumplen la correspondiente hipótesis.

Es importante mencionar que, para los distintos procedimientos de selección del parámetro ventana, las estimaciones podrían ser muy diferentes. Se presenta en la Figura 1.2 la estimación de la densidad de los datos de *Airquality* y *Geyser* con el estimador tipo núcleo para los tres selectores citados previamente en esta Introducción.



(a) Airquality (b) Geyser

Figura 1.2: Representación del estimador tipo núcleo con los tres selectores de ventana más empleados para los datos de (a)*Airquality* y (b)*Geyser*.

En la Figura 1.2, se puede apreciar como en el primer conjunto de datos la discrepancia entre las estimaciones es mínima, mientras que en el segundo se ven claramente esas diferencias. Son especialmente marcadas con el selector de Silverman (1986), que es claro que tiende a sobresuavizar, a pesar de que sí llega a captar la bimodalidad presente en los datos.

Una aportación posterior a la estimación no paramétrica de la densidad, se presenta en Watson y Leadbetter (1963), en donde se propone una generalización del estimador tipo núcleo. La diferencia esencial con éste es que se ha suprimido el parámetro ventana, dejando que toda la variabilidad del estimador recaiga sobre la función núcleo.

El interés principal de este estimador radica en, evitar muchas de las condiciones de regularidad y suavidad impuestas sobre la densidad teórica en el proceso de estimación para el estimador tipo núcleo. Esto permite una aplicación más general y precisa, pues dado que la densidad teórica es desconocida, la mayoría de esas condiciones no son comprobables en el ámbito práctico, y simplemente se asumen como ciertas bajo la convicción de que el modelo que sigue la población de interés no sea demasiado complejo.

Aunque la introducción de este estimador, al que nos referiremos como estimador con núcleo variable, no es reciente, únicamente en Bernacchia y Pigolotti (2011b) se ha retomado esta línea de trabajo proponiendo un procedimiento eficiente para la estimación de la función núcleo óptimo[2]. Nuestra idea es determinar un nuevo método de estimación del núcleo óptimo que mejore, o al menos sea comparable con el existente, y que sea también competitivo con los selectores de ventana habituales para el estimador tipo núcleo. De este modo, se obtendrían resultados razonables sin necesidad de asumir condiciones de regularidad sobre el modelo teórico.

La presente memoria del trabajo fin de máster se organiza de la siguiente forma. En el Capítulo 2 se hace referencia a la definición formal del estimador tipo núcleo de la densidad y se presenta el estimador con núcleo variable de Watson y Leadbetter (1963). Se detallan los criterios de error que se emplearán a lo largo del trabajo y se explica como obtener un estimador óptimo en términos de uno de ellos. Posteriormente se elabora una breve comparativa entre el estimador óptimo de Watson y Leadbetter (1963) y el estimador tipo núcleo para dos casos particulares.

En el Capítulo 3, se expone todo el proceso de estimación necesario para llevar a la práctica los conceptos y desarrollos teóricos que constituyen el Capítulo 2. Se definen y detallan los distintos selectores que se han desarrollado para el estimador de Watson y Leadbetter (1963). Se trata de cuatro procedimientos distintos, algunos inéditos y otros ya existentes, inspirados en mayor o menor medida en los conocidos selectores de la ventana para el estimador tipo núcleo, a los que también se hace referencia.

El Capítulo 4 consta de una primera parte en la que se explica la teoría necesaria para la implementación del estimador junto con los selectores definidos en el Capítulo 3. En la segunda parte se realiza un completo estudio de simulación que permite comparar los selectores del Capítulo 3 entre sí y con los ya conocidos para el parámetro ventana del estimador tipo núcleo.

En el Capítulo 5 se presenta la aplicación de la técnicas expuestas a lo largo de la memoria sobre un conjunto de datos reales. De este modo se ilustra el proceso necesario para implementar en un contexto práctico todas las ideas planteadas, y se extraen las conclusiones correspondien-

[2]Se explica el concepto de núcleo óptimo y se detalla su cálculo en el Capítulo 2 del presente trabajo

tes a dicho estudio.

Al final del trabajo pueden consultarse los apéndices. En ellos se especifican los modelos de Marron y Wand (1992) que se emplean tanto en la ilustración de ejemplos como en el estudio de simulación. Además se presentan con detalle, una serie de cálculos necesarios en el desarrollo de la metodología estudiada, pero que no se especifican completamente en la parte principal de la memoria.

Capítulo 2

Estimación de la densidad

2.1. El estimador tipo núcleo y criterios de error

La estimación no paramétrica de la densidad ha sido uno de los temas más estudiados en inferencia estadística durante la segunda mitad del s.XX. Dada una muestra aleatoria simple (m.a.s.) $X_1, \cdots, X_n$ de la variable de interés X con densidad f, el método más habitual en estimación no paramétrica de la densidad es el denominado estimador tipo núcleo, propuesto por Parzen (1962) y Rosenblatt (1956) y cuya expresión es la siguiente:

$$\hat{f}_{nh}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

siendo $h > 0$ la ventana (o parámetro de suavizado) y K una función núcleo, esto es, una función real, no negativa, integrable con $\int K = 1$, unimodal y simétrica respecto al origen.

El estimador (2.1) tiene dos elementos desconocidos. Por una parte está el parámetro ventana, cuya importancia en la estimación de la densidad es conocida y hemos visto algún ejemplo en la Introducción del presente trabajo. Y por la otra está la función núcleo, K, que habitualmente es fijada antes de iniciar el estudio por el investigador, pero sobre la que también se ha desarrollado metodología.

Para comparar los diversos procedimientos de estimación es necesario disponer de mecanismos que midan la bondad de ajuste de los mismos, comúnmente denominados criterios de error. Existen diversas posibilidades para medir este ajuste: empleando el valor absoluto de la diferencia, la medida del espacio L_2[1], ponderaciones de esta última... En este trabajo se empleará, en general como criterio de error, el Error Cuadrático Integrado (ISE; *Integrated Squared Error*) que se define para un estimador de la densidad $\hat{f} \in L_2$ como:

$$ISE(\hat{f}) = \int (\hat{f}(x) - f(x))^2 dx.$$

Bajo la condición de que $f \in L_2$, la principal característica del ISE es que es un criterio de error global, es decir, que no depende del punto en el que se evalúa el estimador. Sin embargo, este criterio sí depende de la muestra de datos, y por tanto se está introduciendo una cierta variabilidad intrínseca a la propia muestra y no al estimador. Por ello se define el Error Cuadrático

[1]Los espacios L_p con $p \in [1, \infty)$ son espacios vectoriales normados que se definen sobre un espacio de medida (Ω, Σ, ν) como el espacio de todas las funciones medibles g que cumplen $\int_\Omega |g|^p d\nu < \infty$.

Medio Integrado (MISE; *Mean Integrated Squared Error*) que suprime la aleatoriedad procedente de cada muestra individual promediando los resultados obtenidos para varias:

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

Por consiguiente, el MISE es un criterio de error global que no depende de la muestra empleada.

En general, para el estimador (2.1) no se puede obtener una expresión exacta del MISE. Haciendo desarrollos de Taylor y bajo ciertas condiciones de regularidad sobre f (ver, por ejemplo, Teorema 2.31, Chacón Durán, 2004), se obtiene una aproximación asintótica cuya expresión, supuesto que h depende de n y cumple que $h \to 0$ y $nh \to \infty$, viene dada por:

$$MISE(\hat{f}_{nh}) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)R(f'') + o((nh)^{-1} + h^4) \equiv AMISE(\hat{f}_{nh}) + o((nh)^{-1} + h^4), \tag{2.2}$$

donde R es una aplicación que asigna a cualquier función de L_2 la integral de su cuadrado, esto es, si $g \in L_2$ entonces $R(g) = \int g^2(x)dx$; μ_2 es otra aplicación definida como $\mu_2(h) = \int x^2 h(x)dx$, siempre que este valor sea finito, y AMISE denota el denominado *Asympthotic Mean Integrated Squared Error* que es la parte del MISE que conocemos de manera exacta.

El AMISE es el que se emplea como criterio de error a optimizar en la obtención de la expresión de la ventana óptima. Para ello, se minimiza en h la expresión del AMISE y se obtiene:

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

Si se sustituye este valor en la expresión (2.2), se puede ver que la tasa de convergencia del error óptimo para el estimador es $O(n^{-4/5})$, ver Wand y Jones (1995, Cap.2).

Hasta ahora, se ha presentado brevemente la teoría de optimización relativa al parámetro ventana, que es la que más frecuentemente se emplea en el estimador tipo núcleo. Pero como se ha comentado, también existe teoría de optimización sobre la función núcleo, ya que entre todas las posibles funciones que cumplen las condiciones exigidas al núcleo K, sería deseable determinar cual es "la mejor".

Esta teoría se desarrolla con detalle en Wand y Jones (1995, Cap.2), y se basa en poder expresar el error como producto de dos factores, de manera que uno de ellos dependa únicamente de h y el otro de K. Para esto se considera la función núcleo reescalada $K_\delta(\cdot) = K(\cdot/\delta)/\delta$, pues de no hacerlo así, la expresión AMISE dada en (2.2) no se puede minimizar fácilmente en K, ya que el efecto de la ventana y el núcleo van emparejados.

Tomando $\delta_0 = \left(R(K)/\mu_2(K)^2\right)^{1/5}$, el AMISE puede factorizarse en dos términos que separan la dependencia de h y K:

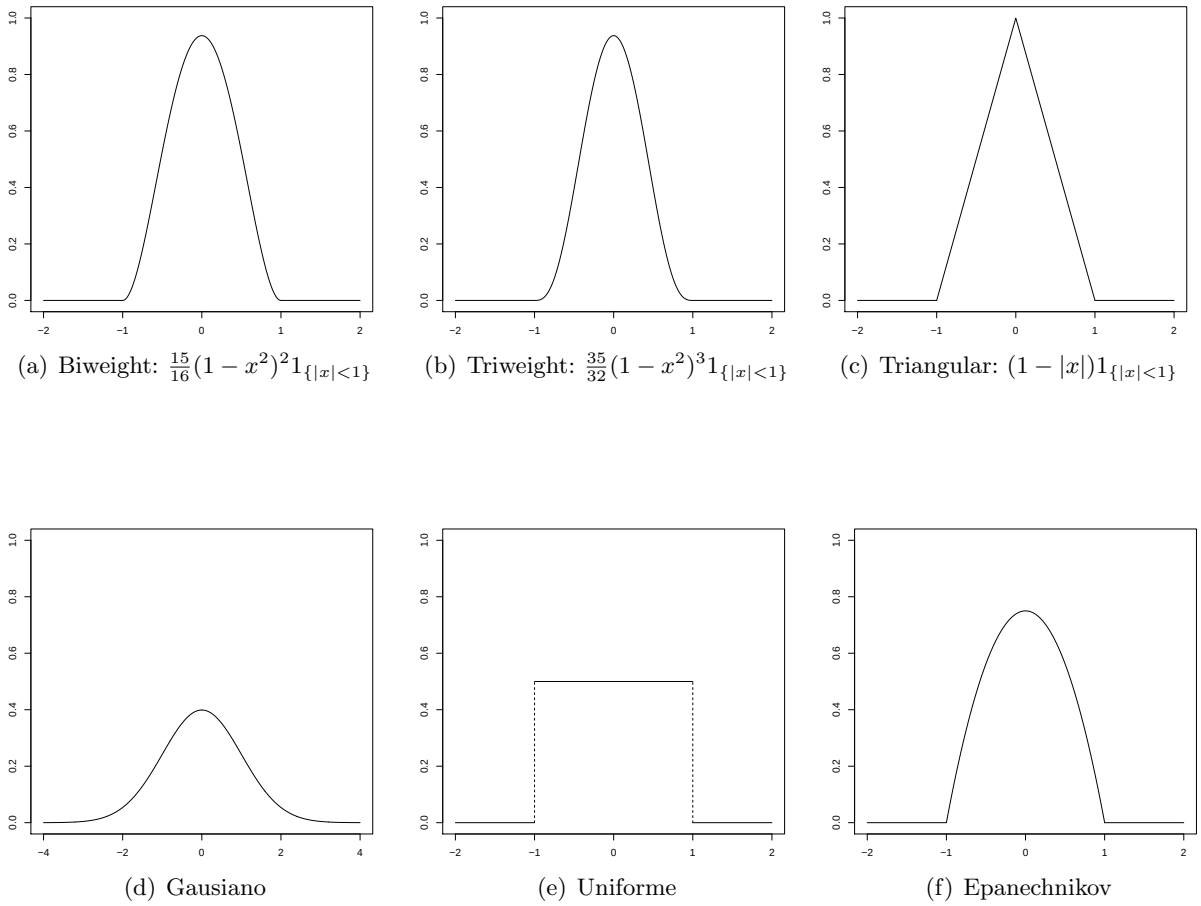
$$AMISE(\hat{f}_{nh}) = C(K_{\delta_0})\left[\frac{1}{nh} + \frac{1}{4}h^4 R(f'')\right], \tag{2.3}$$

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

Para obtener el núcleo óptimo, se debe minimizar en K la expresión (2.3), esto es, basta determinar la función K que minimice $C(K_{\delta_0})$.

Finalmente, tal y como indican en Wand y Jones (1995), se obtiene como núcleo óptimo el de Epanechnikov, cuya expresión es $K_{epa}(x) = \frac{3}{4}(1-x^2)I_{\{|x|<1\}}$.

La teoría relativa a la selección del núcleo óptimo no ha tenido demasiados seguidores. Esto se debe a que las diferencias, en términos de error, para los estimadores resultantes con uno u otro núcleo son casi inexistentes. Algunas de las funciones más comúnmente empleadas pueden verse en la Figura 2.1 que se muestra a continuación.



(a) Biweight: $\frac{15}{16}(1-x^2)^2 1_{\{|x|<1\}}$ (b) Triweight: $\frac{35}{32}(1-x^2)^3 1_{\{|x|<1\}}$ (c) Triangular: $(1-|x|)1_{\{|x|<1\}}$

(d) Gausiano (e) Uniforme (f) Epanechnikov

Figura 2.1: Representación de las funciones núcleo habitualmente más empleadas en la estimación tipo núcleo de la densidad.

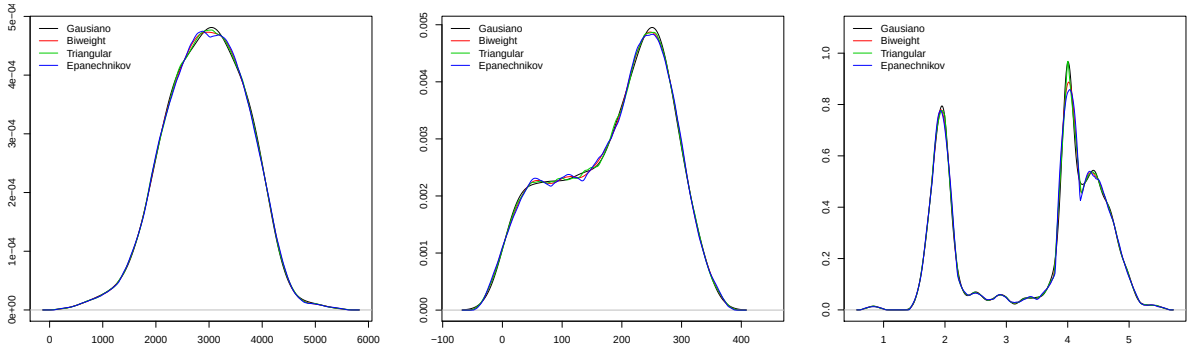
La Tabla 2.1 aparece recogida en Wand y Jones (1995, Cap.2), y se exponen en ella los valores $(C(K_{epa})/C(K))^{5/4}$. Este cociente se suele denominar eficiencia de K relativa a K_{epa}, pues representa, para una densidad f dada, la proporción del tamaño muestral necesario para obtener el mismo AMISE empleando K_{epa} que usando K. Así, por ejemplo, si se tiene una eficiencia de 0.95, esto quiere decir que con el núcleo K_{epa} se alcanzará el mismo AMISE empleando únicamente el 95 % de los datos que para K.

Núcleo	Eficiencia(K_{epa}/K)
Epanechnikov	1.000
Biweight	0.994
Triweight	0.987
Gausiano	0.951
Tringular	0.986
Uniforme	0.930

Tabla 2.1: Eficiencias de varios núcleos empleados habitualmente relativos al óptimo, K_{epa}.

En efecto, los cocientes del error cometido para cada núcleo con respecto al óptimo son muy próximos a 1, lo que justifica, en cierto modo, la escasa relevancia de esta teoría. Aunque como veremos más adelante en este mismo capítulo, si "suprimimos" el parámetro ventana y dejamos que el núcleo varíe de forma, entonces este último adquiere mucha más relevancia y su estudio constituye una línea de trabajo bastante interesante.

Esta escasa diferencia entre los núcleos que se ha puesto de manifiesto a nivel teórico, también se ve reflejada en la práctica con la similitud entre las estimaciones asociadas a cada uno de ellos. Para ilustrar este hecho, se presentan en la Figura 2.2 las estimaciones con cuatro de los núcleos de la Figura 2.1 y ventana escogida por la regla de Sheather y Jones (1991), sobre las variables ya empleadas en la Introducción de las bases de datos *Birthwt*, *Airquality* y *Geyser*.



(a) Birthwt (b) Airquality (c) Geyser

Figura 2.2: Representación del estimador tipo núcleo de la densidad con varios núcleos y ventana escogida por Sheather y Jones (1991) para los datos de (a)*Birthwt*, (b)*Airquality* y (c)*Geyser*.

2.2. El estimador con núcleo variable

Se va a presentar de manera formal el estimador de la densidad con núcleo variable, que como ya se avanzaba en la Introducción de la memoria, se propone en Watson y Leadbetter (1963). Su expresión es:

$$\hat{f}_{nK}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(x - X_i\right), \tag{2.4}$$

donde la única condición que se exige sobre la función núcleo es que $K \in L_2$.

En Watson y Leadbetter (1963) se puede ver que la tasa de convergencia del MISE óptimo

del estimador (2.4) para modelos cuya función característica[2] decrece exponencialmente[3], es $O(\log(n)/n)$, lo que supone una mejoría respecto al estimador tipo núcleo. Entendiendo como MISE óptimo el error mínimo del "mejor" estimador de la familia dada en (2.4), cuyo cálculo se detalla posteriormente en este mismo capítulo.

Además de la velocidad de convergencia, otro punto a favor de este estimador es que se puede calcular, como veremos más adelante, una expresión del MISE óptimo explícita y no asintótica; hecho que no se verifica para el estimador tipo núcleo en el se emplea el AMISE. Es necesario notar que, en general, éste es un estimador más complejo que (2.1), ya que se pasa de un problema de optimización en $\mathbb{R}^+$ a uno de optimización funcional sobre el espacio de funciones L_2.

Por otra parte, también es un estimador más general que el estimador tipo núcleo definido en (2.1), pues podemos obtener este último sin más que tomar como núcleo $K = K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. Esto prueba que la teoría desarrollada para (2.4) es más amplia e incluye al estimador (2.1). En particular, el MISE óptimo será siempre menor o igual para (2.4) que para (2.1).

Es fundamental mostrar el procedimiento completo ideado por Watson y Leadbetter (1963) que permite obtener la expresión óptima en términos de MISE para el estimador (2.4). Esta expresión óptima se construye bajo la única condición de que las funciones K, $f \in L_2$. Hay que tener en cuenta que a lo largo del desarrollo teórico se empleará de manera recurrente tanto la igualdad de Parseval[4] de paso al dominio de frecuencias como la transformada de Fourier[5].

Para utilizar el paso al dominio de frecuencias es necesario conocer la función característica del estimador:

$$\varphi_{\hat{f}_{nK}}(t) = \int e^{itx}\hat{f}_{nK}(x)dx = \int e^{itx}\frac{1}{n}\sum_{l=1}^{n}K(x-X_l)dx = \frac{1}{n}\sum_{l=1}^{n}\int e^{itx}K(x-X_l)dx$$
$$\overset{(*)}{=} \frac{1}{n}\sum_{l=1}^{n}\int e^{it(z+X_l)}K(z)dz = \frac{1}{n}\sum_{l=1}^{n}e^{itX_l}\int e^{itz}K(z)dz = \varphi_n(t)\varphi_K(t), \qquad (2.5)$$

$(*)$ cambio de variable $z = x - X_l$
donde $\varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n}e^{itX_j}$ es la función característica empírica.

[2]Dada una función $f \in L_2$, se define su función característica, φ_f, como

$$\varphi_f(t) = \int e^{itx}f(x)dx.$$

[3]La función característica de una función $f \in L_2$ se dice que posee decrecimiento exponencial de coeficiente p si

$$|\varphi_f(t)| \leq Ae^{-p|t|} \qquad \text{para alguna constante } A > 0 \text{ y todo } t,$$

y además,

$$\lim_{t\to\infty}\int_0^1 \left(1 + e^{2pt^r}|\varphi_f(tx)|^2\right)^{-1}dx = 0.$$

[4]Dada $g \in L_2$, entonces $\varphi_g \in L_2$ y se cumple

$$\int |g(x)|^2dx = \frac{1}{2\pi}\int |\varphi(t)|^2dt.$$

[5]La definición y aplicación de la transformada de Fourier se detalla en la primera sección del Capítulo 4.

Una vez determinada esta expresión, se puede proceder al cálculo del MISE del estimador:

$$MISE(\hat{f}_{nK}) = \mathbb{E}\left[\int (\hat{f}_{nK}(x) - f(x))^2 dx\right] = \mathbb{E}\left[||\hat{f}_{nK} - f||_2^2\right] \overset{(*)}{=} \mathbb{E}\left[\frac{1}{2\pi}||\varphi_{\hat{f}_{nK}} - \varphi_f||_2^2\right]$$
$$= \frac{1}{2\pi}\mathbb{E}\left[\int |\varphi_{\hat{f}_{nK}}(t) - \varphi_f(t)|^2 dt\right] \overset{(**)}{=} \frac{1}{2\pi}\mathbb{E}\left[\int |\varphi_K(t)\varphi_n(t) - \varphi_f(t)|^2 dt\right], \quad (2.6)$$

$(*)$ igualdad de Parseval
$(**)$ ecuación (2.5)
donde $||\cdot||_2$ denota la norma del espacio funcional L_2.

La ecuación (2.6) se puede reescribir, siguiendo las indicaciones que aparecen en Watson y Leadbetter (1963), como:

$$MISE(\hat{f}_{nK}) = \frac{1}{2\pi}\mathbb{E}\left[\int |\varphi_K(t)/n \sum_{r=1}^{n} e^{iX_r t} - \varphi_f(t)|^2 dt\right]$$
$$= \frac{1}{2\pi}\int \left[(1/n)|\varphi_K(t)|^2\left(1 - |\varphi_f(t)|^2\right) + |\varphi_f(t)|^2(|1 - \varphi_K(t)|^2)\right] dt$$
$$= \frac{1}{2\pi}\left[\int \left(\frac{1}{n} + \frac{n-1}{n}|\varphi_f(t)|^2\right)\left(\left|\varphi_K(t) - \frac{|\varphi_f(t)|^2}{(1/n) + [(n-1)/n]|\varphi_K(t)|^2}\right|^2\right) dt\right.$$
$$\left. + \int \frac{|\varphi_f(t)|^2(1 - |\varphi_f(t)|^2)}{1 + (n-1)|\varphi_f(t)|^2} dt\right]. \quad (2.7)$$

Esta expresión del MISE depende únicamente de φ_f y φ_K. Si se minimiza en φ_K, se obtiene fácilmente una expresión para la función característica del núcleo óptimo, K^*, atendiendo a que únicamente el segundo factor del primer sumando depende del elemento sobre el que se va a optimizar:

$$\varphi_{K^*}(t) = \frac{n|\varphi_f(t)|^2}{1 + (n-1)|\varphi_f(t)|^2}. \quad (2.8)$$

Aplicando la transformada inversa de Fourier, se puede determinar la expresión para el núcleo óptimo

$$K^*(x) = \frac{1}{2\pi}\int e^{-itx}\varphi_{K^*}(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{n|\varphi_f(t)|^2}{1 + (n-1)|\varphi_f(t)|^2}dt,$$

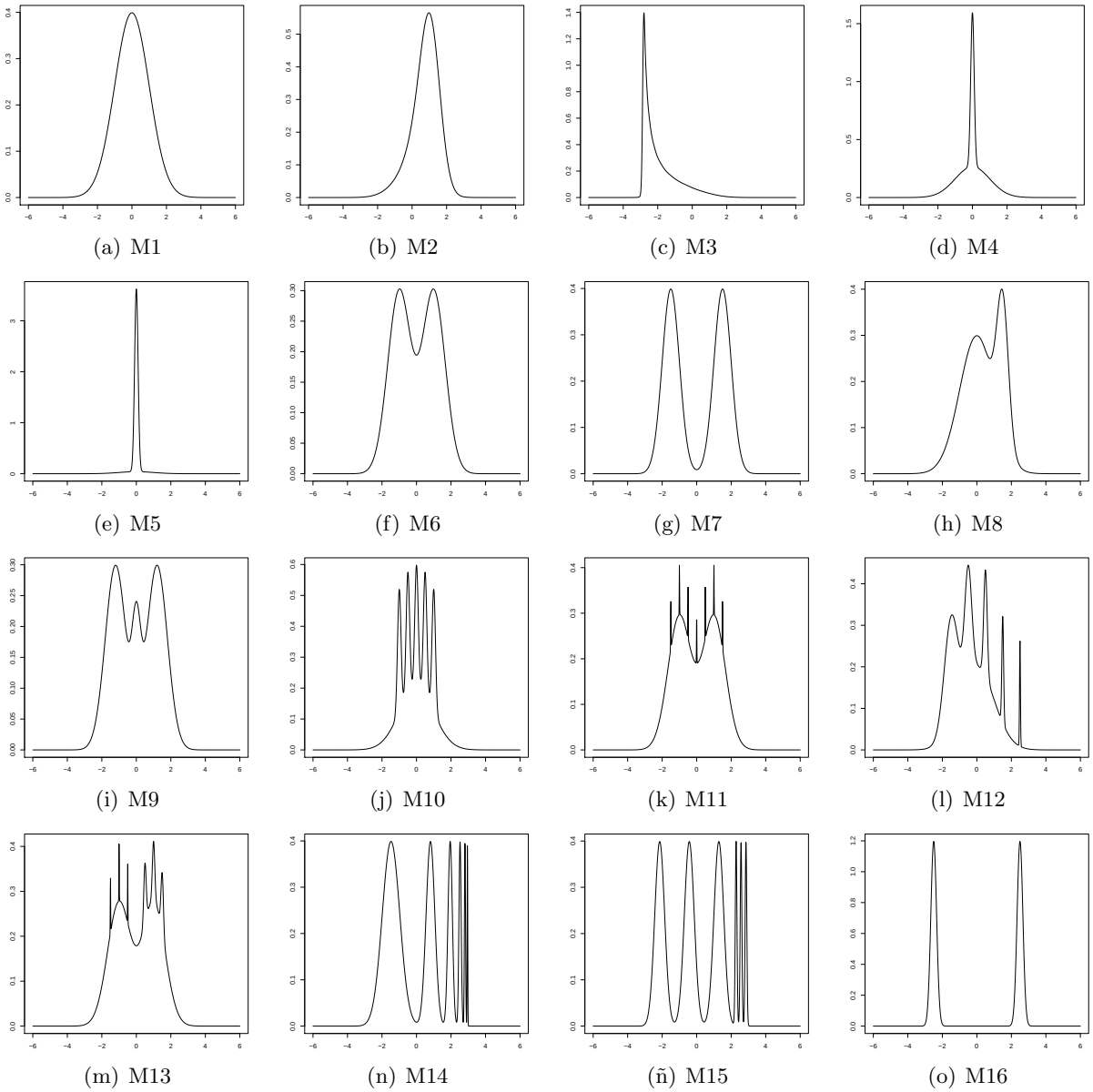
que al sustituirla en (2.4), da lugar al estimador que se denomina óptimo y que se denotará por $\hat{f}_{nK^*}$. El MISE asociado, y por consiguiente el MISE óptimo es:

$$MISE^* = MISE(\hat{f}_{nK^*}) = \frac{1}{2\pi}\int \frac{|\varphi_f(t)|^2\left(1 - |\varphi_f(t)|^2\right)}{1 + (n-1)|\varphi_f(t)|^2}. \quad (2.9)$$

La expresión (2.9) proporciona una manera exacta y no asintótica de obtener el mejor estimador en términos de MISE para la familia de estimadores dada por (2.4). Recordemos además que la única restricción impuesta sobre la función núcleo es que pertenezca al espacio L_2.

El problema que se plantea ahora es que el núcleo óptimo, K^*, depende de la función característica de la densidad que es desconocida; pues recuérdese que la función de densidad constituye el objetivo inicial de la estimación. Esto se solventará en el Capítulo 3 del presente trabajo en el que se detallan varios procedimientos de estimación para K^*. Sí se verán más adelante, en este mismo capítulo, representaciones del núcleo óptimo teórico así como del estimador asociado.

A lo largo de esta memoria, se empleará en la ilustración de los procedimientos la familia de densidades presentadas en Marron y Wand (1992). Todas ellas son mixturas de normales, lo que como se verá posteriormente, posibilitará la realización de algunos cálculos. Este conjunto es especialmente útil, pues cubre un amplio abanico en cuanto a tipos de densidades con diversas características (simetría, modas, oscilaciones...) y permite por tanto el análisis de prácticamente cualquier procedimiento que involucre a la estimación de la densidad. Las fórmulas explícitas para cada modelo se han añadido en el Apéndice A de esta memoria y en la Figura 2.3 se incluye su representación gráfica.



(a) M1 (b) M2 (c) M3 (d) M4

(e) M5 (f) M6 (g) M7 (h) M8

(i) M9 (j) M10 (k) M11 (l) M12

(m) M13 (n) M14 (ñ) M15 (o) M16

Figura 2.3: Representación de las funciones de densidad teóricas para los modelos de Marron y Wand (1992).

Se ha aplicado el desarrollo detallado anteriormente al modelo M1, es decir, se ha calculado el núcleo óptimo K^* (que en este caso sí se puede obtener puesto que se trata de un escenario controlado en el que se conoce la densidad teórica), y se ha sustituido en la expresión (2.4) para

obtener el estimador óptimo. De esta manera podemos visualizar cómo sería el mejor estimador posible para un tamaño muestral dado.

En la Figura 2.4 podemos ver la representación del núcleo óptimo para dicho modelo, junto con el estimador óptimo, f_{nK^*}. Para poder hacer esta representación se ha programado un código en el software R, en el que se emplea la transformada rápida de Fourier para reducir el coste computacional[6].

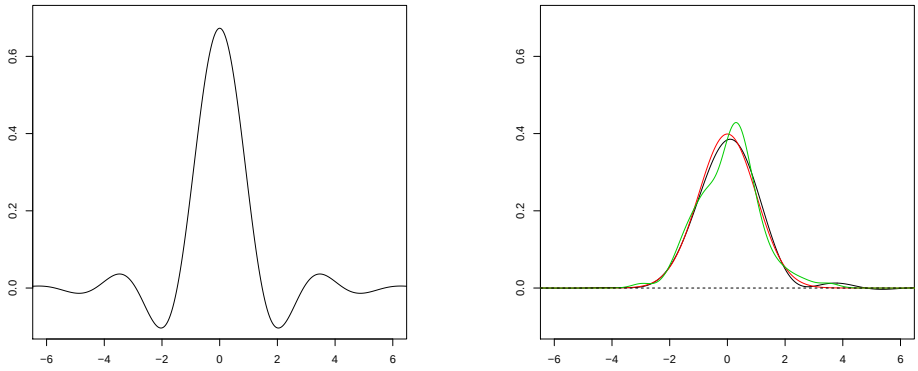


Figura 2.4: A la izquierda se representa el núcleo óptimo para M1 con tamaño muestral $n = 100$; a la derecha aparecen el estimador óptimo (línea negra) para una muestra dada de ese modelo con dicho tamaño, el estimador tipo núcleo con ventana escogida por Sheather y Jones (1991) (línea verde) y la densidad teórica (línea roja).

En este modelo, que se corresponde con una densidad gaussiana, el resultado obtenido para el estimador (2.4) no difiere mucho de lo que se obtiene para el estimador (2.1) con núcleo gaussiano y ventana escogida por Sheather y Jones (1991). Sin embargo, en la Figura 2.5 podemos ver las mismas gráficas para el modelo M10, en las que sí existen diferencias muy notables.

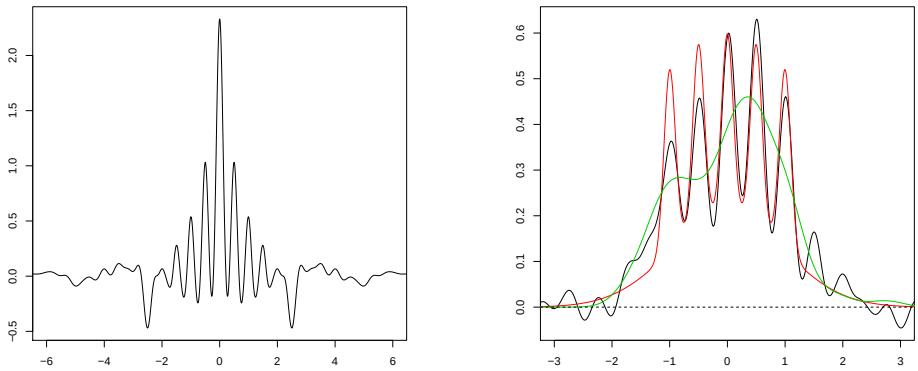


Figura 2.5: A la izquierda se ve la representación el núcleo óptimo para M10 con tamaño muestral $n = 100$, a la de la derecha se presenta el estimador óptimo (línea negra) para una muestra dada de ese modelo con dicho tamaño, el estimador tipo núcleo con ventana escogida por Sheather y Jones (1991) (línea verde) y la densidad teórica (línea roja). (Nótese que ambas gráficas no están en la misma escala).

En efecto, observando la Figura 2.5, se aprecia el núcleo óptimo difiere mucho de cualquier función núcleo que se pueda emplear en el estimador (2.1), pues no es unimodal, ni no nega-

[6]Esta técnica se detalla en la primera sección del Capítulo 4.

tiva y tampoco integra necesariamente la unidad. La única característica que comparte con las funciones núcleo habituales es la simetría respecto del origen, que se debe a que su función característica dada en (2.8) toma únicamente valores reales. Recuérdese que el proceso de optimización, se hace sobre el espacio L_2, no sobre un subconjunto propio del mismo sometido a restricciones.

Una desventaja clara que presenta el estimador (2.4) es que puede tomar valores negativos, lo que implica que no es una densidad y que no será posible emplearlo en un proceso de generación de datos. Este problema se resuelve fácilmente utilizando la teoría propuesta en Glad, Hjort, y Ushakov (2003) y que se detalla brevemente en el Capítulo 4. Esta corrección consiste básicamente en anular las partes negativas del estimador y reescalar las positivas para que la función resultante integre uno. Además, el estimador que se obtiene es siempre mejor que el anterior en términos de ISE. En la Figura 2.6 podemos ver una gráfica del estimador de la Figura 2.4 y Figura 2.5 una vez aplicada la corrección:

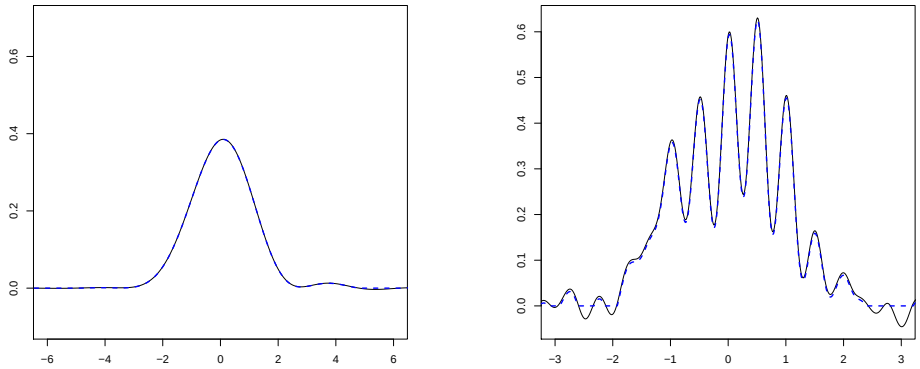


Figura 2.6: Representación del estimador óptimo para las muestras de tamaño $n = 100$ de los modelos M1 y M10 (izquierda y derecha respectivamente), empleadas en la Figura 2.4 y Figura 2.5 (línea negra continua), junto con el estimador aplicada la corrección (línea azul discontinua).

2.3. Eficiencia relativa para muestras finitas

Se han expuesto diversas ventajas de (2.4) sobre los estimadores habituales (expresión exacta del MISE, tasa de converencia, ...). Un nuevo punto de interés, sobre todo a nivel práctico, es ver qué ocurre para un valor de n fijo, es decir, analizar cuál de los dos estimadores, el tipo núcleo o el de núcleo variable, proporciona un menor MISE para un tamaño muestral dado.

La conclusión a este respecto es inmediata, pues por construcción, el estimador (2.4) sigue siendo mejor, y proporciona siempre valores de MISE menores que el estimador (2.1), ya que recuérdese que el tipo núcleo puede verse como un caso particular de éste, tal y como se explica al comienzo de este mismo capítulo.

Una vez constatado el mejor comportamiento del estimador con núcleo variable, sería interesante poder cuantificarlo de algún modo. Para ello se va a realizar una comparación entre dicho estimador y dos casos particulares del estimador tipo núcleo.

Se consideran los modelos de Marron y Wand (1992), que pueden ser expresados como $f = \sum_{j=1}^{p} \omega_j f_{\mu_j \sigma_j}$, donde ω_j son los pesos tales que $\sum_{j=1}^{p} \omega_j = 1$ y $\omega_j > 0$, y $f_{\mu_j \sigma_j}$ la función de

densidad de una normal de media μ_j y desviación típica σ_j. Se considera también el estimador (2.1) con núcleo gaussiano, $\hat{f}_{nhG}$, y con núcleo el Sinc[7], $\hat{f}_{nhS}$. Esta segunda función núcleo se denomina habitualmente en la literatura supernúcleo (*high order kernel*), y suele ser empleado para reducir el sesgo del estimador. Además mejoran el orden de convergencia con respecto a los núcleos convencionales alcanzando incluso la tasa del estimador con núcleo variable para densidades suficientemente suaves, como se ha demostrado en Davis (1977).

A diferencia del caso general, para las mixturas de normales se puede calcular una expresión exacta del MISE para el estimador (2.1). La expresión relativa al núcleo gaussiano se ha calculado basándose en las indicaciones dadas en Wand y Jones (1995, Cap. 2) y cuya expresión puede consultarse también en Marron y Wand (1992); mientras que las del Sinc, se basan en los cálculos de Davis (1977). Las expresiones resultantes (en al Apéndice B puede consultarse el desarrollo en detalle de las mismas), son:

$$\begin{aligned}\mathrm{MISE}(\hat{f}_{nhG}) = \frac{1}{2\sqrt{\pi}nh} + \left(1-\frac{1}{n}\right)\sum_{j=1}^{p}\sum_{l=1}^{p}\omega_j\omega_l f_{0,(2h^2+\sigma_j^2+\sigma_l^2)^{1/2}}(\mu_j-\mu_l) \\ - 2\sum_{j=1}^{p}\sum_{l=1}^{p}\omega_j\omega_l f_{0,(h^2+\sigma_j^2+\sigma_l^2)^{1/2}}(\mu_j-\mu_l) + \sum_{j=1}^{p}\sum_{l=1}^{p}\omega_j\omega_l f_{0,(\sigma_j+\sigma_l)^{1/2}}(\mu_j-\mu_l),\end{aligned} \tag{2.10}$$

$$\mathrm{MISE}(\hat{f}_{nhS}) = \frac{1}{nh\pi} - \frac{n+1}{n\pi}\int_0^{1/n}\rho_f(t)dt + \frac{1}{\pi}\int_0^{\infty}\rho_f(t)dt, \tag{2.11}$$

donde se denota $\rho_f(t) = |\varphi_f(t)|^2$.

El único parámetro desconocido en (2.10) y (2.11) es la ventana h, por tanto, podemos minimizar dichas expresiones y obtener el valor óptimo. Al sustituirlo en el MISE correspondiente, proporcionaría el menor error que se podría cometer en la estimación con $\hat{f}_{nhG}$ y $\hat{f}_{nhS}$, y que denotamos por $\mathrm{MISE}^*(G)$ y $\mathrm{MISE}^*(S)$, respectivamente. Así, se pueden comparar los valores de estos MISE óptimos con el del estimador con núcleo variable dado en (2.9).

Una manera sencilla para efectuar esa comparación, es mediante los cocientes de dichos errores óptimos, esto es, $\mathrm{MISE}^*(G)/\mathrm{MISE}^*$ y $\mathrm{MISE}^*(S)/\mathrm{MISE}^*$. Se denominarán cocientes de eficiencia, ya que miden cómo de bien lo podrían llegar a hacer, a nivel teórico, los estimadores $\hat{f}_{nhG}$ y $\hat{f}_{nhS}$ con respecto a $\hat{f}_{nK^*}$. Recuérdese que éste último es el objetivo a alcanzar, pues es el óptimo de una familia de estimadores que incluye a los otros dos.

En la Tabla 2.2 se presentan los citados cocientes de eficiencia para tres tamaños muestrales diferentes, $n = 100, 400$ y 1600. En primer lugar se observa que todos los valores son mayores que 1, que era lo que justamente se esperaba por el razonamiento en torno a la generalidad del estimador (2.4). El caso del modelo M1 es de especial intéres, pues a pesar de corresponderse con la densidad gaussiana, tanto el estimador (2.4) como el estimador $\hat{f}_{nhS}$, consiguen una disminución bastante notable del MISE con respecto al de $\hat{f}_{nhG}$.

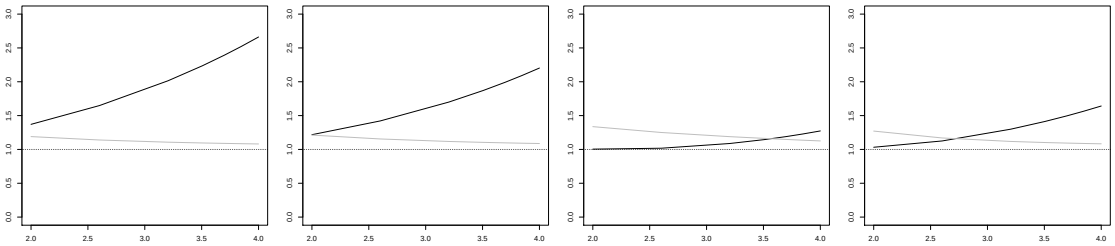
[7]La función Sinc se define para todo punto $x \in \mathbb{R}$ como $S(x) = \frac{\sin(x)}{\pi x}$

	n = 100		**n = 400**		**n = 1600**	
	Gaus/WL	**Sinc/WL**	**Gaus/WL**	**Sinc/WL**	**Gaus/WL**	**Sinc/WL**
M1	1.3706	1.1896	1.6485	1.1392	2.0145	1.1070
M2	1.2177	1.2119	1.4217	1.1545	1.7001	1.1177
M3	1.0043	1.3367	1.0178	1.2503	1.0876	1.1887
M4	1.0321	1.2711	1.1253	1.1693	1.2990	1.1180
M5	1.2925	1.1822	1.5437	1.1337	1.8783	1.1030
M6	1.2153	1.4291	1.3690	1.2119	1.6304	1.1007
M7	1.3792	1.4321	1.5375	1.2162	1.8024	1.1006
M8	1.0579	1.4442	1.0853	1.2378	1.2214	1.1468
M9	1.1680	1.3465	1.1351	1.4289	1.1557	1.2662
M10	1.8650	2.1676	1.6899	1.4838	1.6607	1.3237
M11	1.1797	1.3491	1.2499	1.1518	1.3837	1.2129
M12	1.4093	1.7414	1.3334	1.5636	1.2728	1.5900
M13	1.1727	1.2784	1.3585	1.3241	1.3380	1.5685
M14	1.0673	1.2972	1.0552	1.3146	1.0467	1.4070
M15	1.3769	1.4442	1.4025	1.7733	1.3248	1.3376
M16	1.3816	1.3598	1.5356	1.2627	1.7851	1.1746

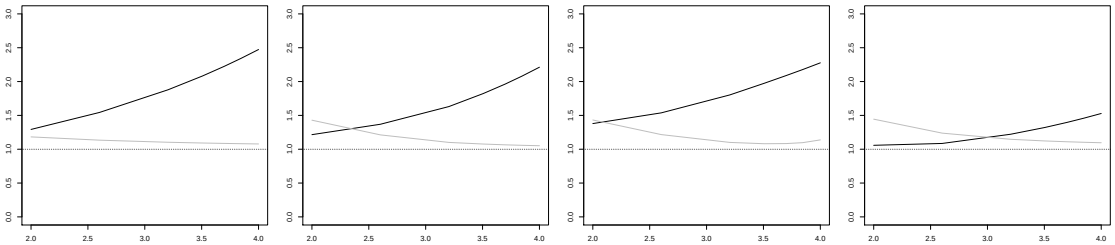
Tabla 2.2: Cocientes de eficiencia de los MISE mínimos para $\hat{f}_{nhG}$ y $\hat{f}_{nhS}$ entre el MISE óptimo (2.9).

Al incrementar el tamaño muestral, el MISE del estimador $\hat{f}_{nhS}$, se aproxima más al valor óptimo dado por (2.9) que $\hat{f}_{nhG}$. Esto concuerda con la teoría relativa a órdenes de convergencia introducida previamente en este mismo capítulo. Recuérdese que mientras el estimador (2.1) con núcleo gaussiano y supuestas ciertas condiciones de suavidad sobre f, tiene una tasa de convergencia de $O(n^{-4/5})$, el estimador (2.1), empleando como núcleo el Sinc, y para modelos con funciones características con decrecimiento exponencial, iguala el orden de convergencia, $O(log(n)/n)$, del estimador (2.4), tal y como se prueba en Davis (1977).

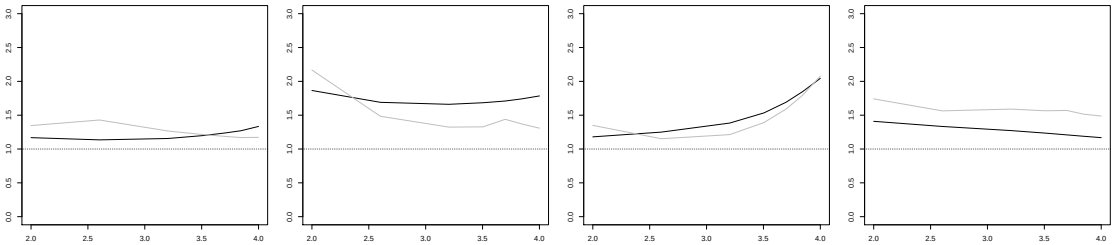
Una forma más visual de enfrentarse a estos resultados es mediante una representación gráfica de los cocientes de eficiencia en función del tamaño muestral (en escala logarítmica). Lo ideal sería que estas curvas fuesen monótonas crecientes, esto evidenciaría que el MISE óptimo para el estimador (2.4) fuese mejor a medida que se incrementa el tamaño muestral con respecto al de los otros dos estimadores con los que comparamos. Este hecho cabría esperarlo para el de núcleo gaussiano ya que la tasa de convergencia es mayor que la del estimador de Watson y Leadbetter (1963), pero no para el Sinc, ya que como se ha dicho, para esta familia de modelos, alcanza la misma tasa que el estimador con núcleo variable.



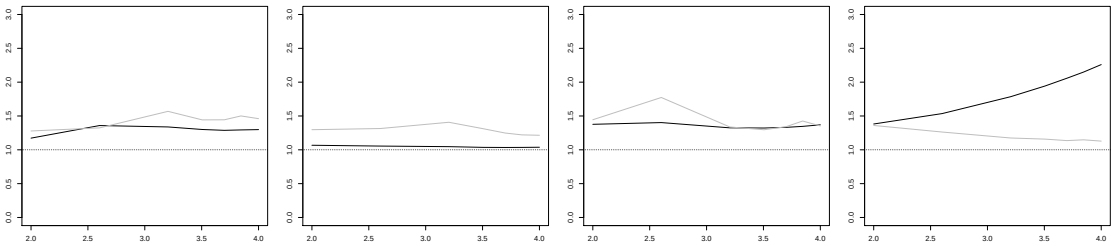
(a) M1 (b) M2 (c) M3 (d) M4



(e) M5 (f) M6 (g) M7 (h) M8



(i) M9 (j) M10 (k) M11 (l) M12



(m) M13 (n) M14 (ñ) M15 (o) M16

Figura 2.7: Gráfico de cociente de eficiencias del estimador tipo núcleo con núcleo gaussiano (línea negra) y con núcleo el Sinc (línea gris) entre el MISE mínimo para el estimador de Watson y Leadbetter (1963).

En la Figura 2.7 no se cumple, en general, esa condición de monotonía de la que se hablaba anteriormente, exceptuando algunos casos con el gaussiano. Para modelos sencillos, el cociente correspondiente al Sinc tiende a 1 con bastante rapidez, es decir, que convergería hacia el óptimo obtenido en (2.9). En los modelos complejos esa convergencia, aunque teóricamente también se da, es más lenta.

En los últimos modelos, salvo en el M16, se observa que la diferencia entre el gaussiano y el Sinc no es tan evidente y de hecho, las gráficas se llegan a cruzar incluso en varias ocasiones.

Además, también es especialmente interesante lo que ocurre en los modelos M12, M13 y M14, ya que lo que se esperaría es que el gaussiano se aleje de 1, y no que, como ocurre en estos casos, parece que tenga una asíntota. Incluso en algunos de ellos, funciona mejor que el Sinc. Nótese que dichos modelos no son especialmente suaves ni sencillos.

Con estos resultados termina la presentación de los estimadores incluidos en el trabajo, de las propiedades teóricas más destacables de cada uno de ellos, así como de una breve comparación de su eficiencia para muestras finitas. En el Capítulo 3, nos centramos en los procedimientos de estimación del núcleo óptimo para el estimador de Watson y Leadbetter (1963), que permitirán llevar a la práctica la metodología desarrollada.

Capítulo 3

Selectores del núcleo

3.1. Introducción

Hasta este momento, se ha desarrollado la teoría necesaria para abordar el problema de estimación de la densidad, objetivo fundamental del presente trabajo. Se ha presentado el estimador a emplear, y se ha obtenido una expresión para el óptimo teórico en un contexto general en el que los datos procedan de una variable aleatoria absolutamente continua con densidad f.

En este capítulo se va a tratar el aspecto empírico, es decir, el paso de los conceptos teóricos a métodos que sean aplicables en un contexto práctico. En el Capítulo 2 se ha visto como el estimador (2.4) depende de la función núcleo, para la que se ha obtenido la expresión óptima. El contenido del presente capítulo consiste fundamentalmente en la presentación de varios procedimientos para la estimación de esa función de manera apropiada.

Como ya hemos mencionado, el principal problema en el estimador (2.1) es la selección del parámetro ventana o parámetro de suavizado. A lo largo de las últimas décadas del s.XX se han ideado numerosos métodos que permiten estimar, con mejor o peor resultado, dicho valor. Los procedimientos más habituales son los propuestos por Silverman (1986), Sheather y Jones (1991) y Bowman (1984), que se supondrán conocidos en lo que sigue, y en ellos se basan las nuevas propuestas para la selección del núcleo en el estimador (2.4).

3.2. Proceso de estimación

En el Capítulo 2, se había obtenido la expresión (2.8), que se corresponde con la función característica del núcleo óptimo para el estimador con núcleo variable. La principal dificultad existente, es que depende de la función característica de la densidad teórica, que es desconocida.

Recuérdese que para aligerar la notación, se empleará $\rho_g(t) = |\varphi_g(t)|^2$ para cualquier función $g \in L_2$.

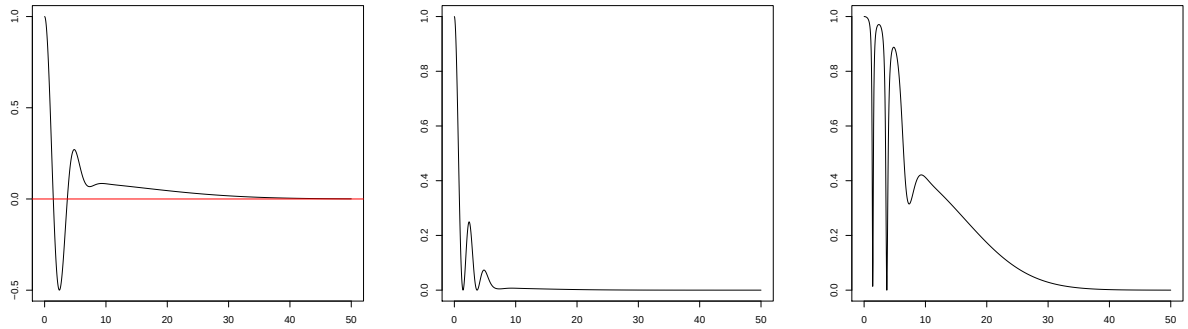
El primer intento de estimación es claramente intuitivo y consiste en sustituir ρ_f por su estimador empírico, esto es,

$$\rho_n(t) = |\varphi_n(t)|^2 \text{ donde recuérdese que } \varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j},$$

y así obtener una estimación de la característica del núcleo óptimo determinada en (2.8),

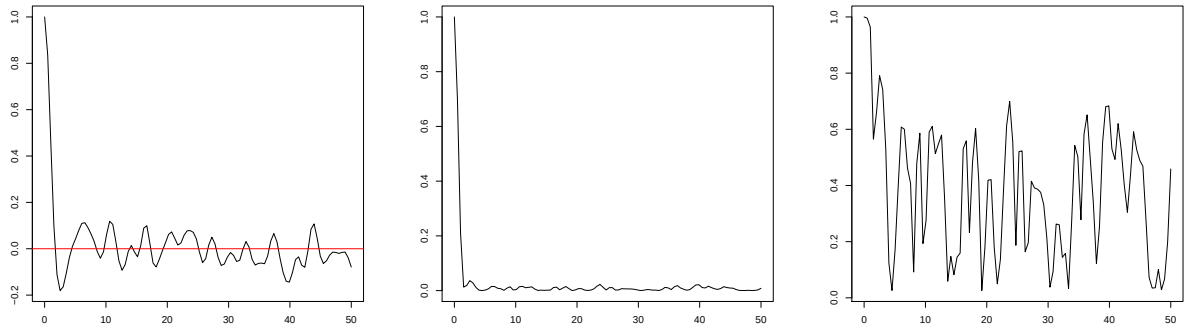
$$\hat{\varphi}_{K^*}(t) = \frac{n\rho_n(t)}{1 + (n-1)\rho_n(t)}. \tag{3.1}$$

Es interesante visualizar la representación gráfica de los valores teóricos y las estimaciones de la función característica, φ_f, su módulo al cuadrado, ρ_f y la característica del núcleo óptimo para un modelo dado. En este caso se ha escogido el modelo M9 de las densidades de Marron y Wand (1992), cuya representación de la densidad puede verse en la Figura 2.3.



(a) $\varphi_f(t)$ (b) $\rho_f(t)$ (c) φ_{K^*}

Figura 3.1: Funciones teóricas correspondientes al modelo M9.



(a) $\varphi_n(t)$ (b) $\rho_n(t)$ (c) $\hat{\varphi}_{K^*}$

Figura 3.2: Funciones estimadas para una muestra de tamaño $n = 100$ del modelo M9.

Al comparar la Figura 3.1 y Figura 3.2, se observa que esta estimación de la función característica para el núcleo óptimo no es demasiado buena, pues presenta muchas más oscilaciones que la función teórica. Además, este desajuste entre estimación y función a estimar por medio de oscilaciones, se aprecia también en la característica de la densidad y en el cuadrado del módulo, aunque en menor medida en esta última.

A pesar de que este es un ejemplo para un modelo concreto, el problema de las oscilaciones se mantiene para todos los modelos de Marron y Wand (1992), que como es sabido, abarcan una amplia selección. Se trata por tanto de un problema generalizado, por lo que se desecha la función característica empírica como estimador y se debe intentar determinar un nuevo procedimiento que lo mejore.

Este hecho modifica el objetivo inicial del capítulo de conseguir una buena estimación de la función núcleo, pues como se ha visto, dicho problema pasa por obtener una buena estimación de la función característica de la densidad teórica y sustituirla en (2.8).

A continuación se presentan cada uno de los procedimientos desarrollados, y cuyo comportamiento se comparará en el Capítulo 4 mediante un exhaustivo estudio de simulación.

3.3. Regla del pulgar

Inspirada en la idea detallada en Silverman (1986) para el estimador tipo núcleo, se trata de un procedimiento sencillo, de fácil comprensión e implementación. En primer lugar retomemos las ideas de Silverman, que basó su teoría en la fórmula de la ventana AMISE óptima, cuya expresión recuérdese, viene dada por:

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

La propuesta de Silverman consiste en sustituir $R(f'')$ por su valor bajo el supuesto de que f siga una distribución normal:

$$h_{NS} = \left(\frac{8\sqrt{\pi} R(K)}{3\mu_2(K)^2 n} \right)^{1/5} \hat{\sigma},$$

siendo $\hat{\sigma}$ un estimador apropiado de la desviación típica de los datos.

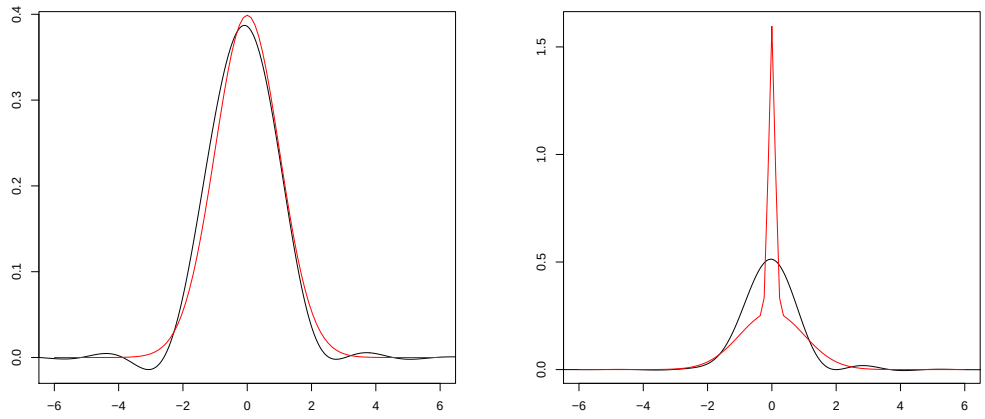
El procedimiento que planteamos en esta sección es análogo, y consiste en sustituir en la expresión (2.8) la función φ_f por la función característica de una densidad normal de media cero y desviación típica σ. En este caso se ha fijado a cero el parámetro de localización μ, ya que se puede ver fácilmente que no influye en los cálculos. Así, un estimador de φ_{K^*} sería

$$\hat{\varphi}_{K^*} = \frac{n e^{-t^2\hat{\sigma}^2}}{1 + (n-1)e^{-t^2\hat{\sigma}^2}}, \tag{3.2}$$

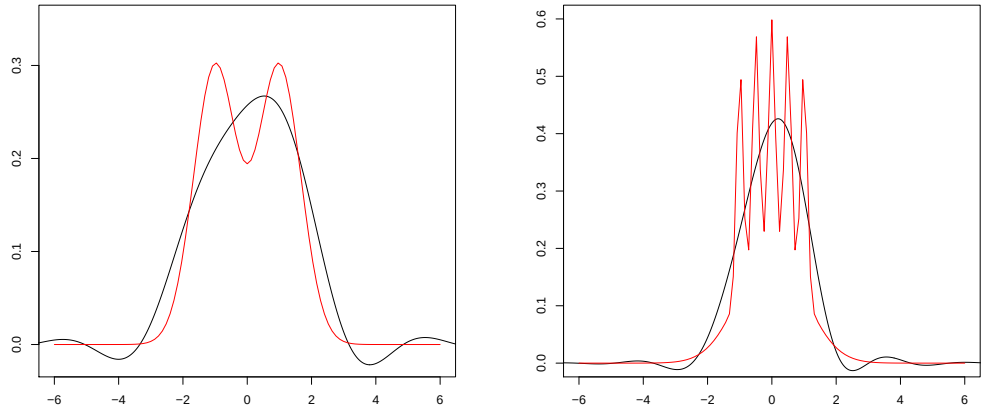
siendo $\hat{\sigma}^2$ un estimador apropiado de la varianza, como por ejemplo la varianza muestral.

Aplicando la transformada de Fourier inversa se obtiene la expresión para la función núcleo y sustituyéndola en (2.4) se tiene el estimador aplicando la regla del pulgar en el proceso de selección del núcleo.

Pueden verse en la Figura 3.3 algunos ejemplos del estimador obtenido con este selector para muestras de tamaño $n = 100$ procedentes de varios de los modelos de Marron y Wand (1992), junto con las correspondientes densidades teóricas. Se han escogido cuatro modelos que abarcan muchas de las posibles características de las densidades: la suavidad del modelo M1, la bimodalidad del M6 y los efectos especialmente marcados de los modelos M4 y M10.



(a) M1 (b) M4



(c) M6 (d) M10

Figura 3.3: Representación del estimador (2.4) (línea negra) con la regla del pulgar para muestras de tamaño $n = 100$ de los modelos M1, M4, M6 y M10 de Marron y Wand (1992), junto con las correspondientes densidades teóricas (línea roja). (Nótese que las gráficas no están todas en la misma escala).

Nuestra regla del pulgar tiende a sobresuavizar en exceso la estimación, por lo que no es capaz de hacer un buen ajuste de densidades que presentan multimodalidades o efectos extremos, de hecho, en este ejemplo sólo tendríamos un ajuste razonable para la normal estándar (modelo M1).

Esta apreciación visual se puede corroborar calculando el ISE para cada una de las muestras empleadas. En el caso de la normal es aproximadamente 0.002, mientras que para la garra está en torno a 0.05, lo que supone un valor 25 veces mayor. Es importante decir que el cálculo de la integral del ISE se hace empleando la regla de integración numérica de Simpson sobre una partición de $M = 2^{14}$ nodos en el intervalo $[-100, 100]$, lo que garantiza una partición lo suficientemente fina para que el error numérico cometido sea irrelevante.

3.4. Regla plug-in

La regla del pulgar que venimos de presentar, se basa en asumir la normalidad de la densidad teórica en una etapa inicial de la estimación, lo que podría ser la razón de una sobresuavización no deseada. Es por ello que se plantea esta nueva idea, que también asume normalidad pero en

una fase más avanzada de la estimación, esperando una menor influencia de dicha hipótesis en el resultado.

Al igual que ocurre en la regla del pulgar, este método consiste en sustituir en (2.8) la función ρ_f por un estimador apropiado. En este caso, la propuesta se basa en que para determinar la función característica de f, debemos conocer la función asociada, por lo tanto, se sustituirá la densidad teórica por un estimador de la forma (2.4), esto es

$$\hat{\rho}_{nL}(t) = |\varphi_{\hat{f}_{nL}}(t)|^2,$$

donde L será una función núcleo por determinar.

El problema es ahora conocer la función característica asociada a un estimador de ese tipo con función núcleo L. Esto ya lo hacíamos en la expresion (2.5), obteniendo

$$\varphi_{\hat{f}_{nL}}(t) = \varphi_n(t)\varphi_L(t).$$

Por tanto,

$$\hat{\rho}_{nL}(t) = |\varphi_n(t)|^2|\varphi_L(t)|^2 \equiv \rho_n(t)\rho_L(t). \tag{3.3}$$

Dada una muestra, el único valor desconocido de esta expresión es la función núcleo L, así que se propone minimizar en L el MISE de $\hat{\rho}_{nL}$:

$$\min_L MISE(\hat{\rho}_{nL}) = \min_L \mathbb{E}\left[\int (\hat{\rho}_{nL}(t) - \rho_f(t))^2 dt\right].$$

Para obtener el MISE, se calculará el sesgo y la varianza del estimador. En este proceso se necesitan una serie de conceptos y resultados enmarcados en la teoría de U-estadísticos que se introducen a continuación, y se particularizan para este caso concreto. Una visión más amplia de dicha teoría puede consultarse en Lee (1990).

Recúerdese que el objetivo teórico es la función ρ_f, y queremos estudiar las propiedades de sesgo y varianza de su estimador (3.3). Para ello lo que haremos será construir un estadístico insesgado de ρ_f y determinar la expresión que lo relaciona con el estadístico de interés, $\hat{\rho}_{nL}$.

Para aplicar la teoría de U-estadísticos es necesario que el objetivo sea un parámetro estimable, η, es decir, que exista un estadístico insesgado. Esto puede ser en general difícil de determinar, pero se solventa fácilmente si suponemos que existe una función h medible, conocida y con $\mathbb{E}|h(X_1, \ldots, X_m)| < \infty$ tal que

$$\eta = \mathbb{E}\left[h(X_1, \ldots, X_m)\right],$$

siendo $X_1, \ldots, X_m$ un subconjunto de m valores de la muestra $X_1, \ldots, X_n$.

Nótese que la función h puede asumirse, sin pérdida de generalidad, simétrica, ya que dado un estadístico insesgado de η, el promedio de los valores de dicho estadístico aplicado sobre las permutaciones de las variables sigue siendo insesgado, y además simétrico.

Para definir un estadístico razonablemente eficiente de η hay que tener en cuenta que el parámetro es función de m variables y que se dispone de una muestra de n datos. La construcción pasa por determinar todos los subconjuntos de m datos de la muestra, y realizar un promedio de los valores resultantes de la evaluación del estadístico en dichos subconjuntos (el orden no es

importante dada la simetría de la función h). Sabiendo que existen $\binom{n}{m}$ subconjuntos, se define el U-estadístico con núcleo h y orden m como

$$U_n = \frac{1}{\binom{n}{m}} \sum_P h(X_{i_1}, \ldots, X_{i_m}), \tag{3.4}$$

donde P denota la familia de conjuntos distintos de m elementos que se pueden formar en un conjunto de n elementos.

Se va ahora a particularizar estos conceptos a la función ρ_f. Para ello necesitamos en primer lugar encontrar una expresión adecuada para ρ_f. Usando las propiedades de la función coseno se puede comprobar que:

$$\rho_f(t) = \mathbb{E}\left[\cos(t(X_1 - X_2))\right],$$

así tendríamos que $m = 2$ y la función $h_t(x_1, x_2) = \cos(t(x_1 - x_2))$.

Por consiguiente, y sin más que tener en cuenta la definición dada en (3.4), el U-estadístico sería

$$\begin{aligned}\hat{\rho}_{u_n}(t) &= \frac{1}{\binom{n}{2}} \sum_P h_t(X_{i_1}, X_{i_2}) = \frac{1}{\binom{n}{2}} \sum_{i>j} h_t(X_i, X_j) = \frac{2}{n(n-1)} \sum_{i>j} \cos(t(X_i - X_j)) \\ &= \frac{2}{n(n-1)} \sum_{i \neq j} \frac{\cos(t(X_i - X_j))}{2} = \frac{1}{n(n-1)} \sum_{i \neq j} \cos(t(X_i - X_j)).\end{aligned}$$

Como ya avanzábamos al comienzo del desarrollo de la teoría relativa a U-estadísticos, el estimador muestral ρ_n no coincide con $\hat{\rho}_{u_n}$ por lo que, para obtener la expresión del sesgo del primero de manera sencilla, conviene conocer la relación existente entre ambos.

$$\begin{aligned}\rho_n(t) = |\varphi_n(t)|^2 &= \frac{1}{n^2} \sum_{i,j=1}^{n} \cos(t(X_i - X_j)) = \frac{1}{n^2} \left[\sum_{i=j} \cos(t(X_i - X_j)) + \sum_{i \neq j} \cos(t(X_i - X_j)) \right] \\ &= \frac{1}{n^2} \left[n + \sum_{i \neq j} \cos(t(X_i - X_j)) \right] = \frac{1}{n} + \frac{1}{n^2} n(n-1) \hat{\rho}_{u_n} = \frac{1}{n} + \frac{n-1}{n} \hat{\rho}_{u_n}\end{aligned} \tag{3.5}$$

Se procederá ahora con el cálculo del sesgo y la varianza de $\hat{\rho}_{nL}$.

- Para el sesgo, se necesitan únicamente las propiedades del operador esperanza y la ecuación dada por (3.5)

  $$\mathbb{E}\left[\hat{\rho}_{nL}(t)\right] = \mathbb{E}\left[\rho_n(t)\right] \rho_L(t) = \left(\frac{1}{n} + \frac{n-1}{n} \rho_f(t) \right) \rho_L(t),$$

  de modo que el sesgo es:

  $$B_n(t) = \mathbb{E}\left[\hat{\rho}_{nL}(t)\right] - \rho_f(t) = c_n(t) \rho_L(t) - \rho_f(t), \tag{3.6}$$

  con $c_n(t) = \frac{1}{n} + \frac{n-1}{n} \rho_f(t)$.

- En el caso de la varianza, se emplea, además de las propiedades habituales del operador, la fórmula de la varianza de un U-estadístico con núcleo h y orden m dada por:

  $$\mathrm{Var}(U_n) = \frac{1}{\binom{n}{m}} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2,$$

siendo $\sigma_c^2 = \mathrm{Cov}(h(X_{i_1}, \ldots, X_{i_m}), h(X_{j_1}, \ldots, X_{j_m}))$. Esta expresión se demuestra con detalle en Lee (1990, págs. 10 - 13).

Aplicando dicha fórmula al caso particular de $\hat{\rho}_{nL}$, se tiene:

$$\mathrm{Var}[\hat{\rho}_{nL}(t)] = \rho_L^2(t)\sigma_n^2(t), \tag{3.7}$$

donde, como se desarrolla en Chacón Durán (2010),

$$\begin{aligned}\sigma_n^2(t) &= \left(\frac{n-1}{n}\right)^2 \mathrm{Var}[\hat{\rho}_{u_n}(t)] \\ &= \frac{n-1}{n^3}\left[2(n-2)\{\mathrm{Re}(\varphi_f(2t)\overline{\varphi_f(t)}^2) + \rho_f(t)\} + 1 + \rho_f(2t) - (4n-6)\rho_f^2(t)\right],\end{aligned}$$

siendo $\mathrm{Re}(z)$ la parte real del número complejo z y $\bar{z}$ su conjugado.

El siguiente paso es el de minimizar punto a punto el Error Cuadrático Medio (MSE; *Mean Squared Error*)[1] del estimador, que, empleando las expresiones (3.6) y (3.7), viene dado por:

$$\begin{aligned}\mathbb{E}\left[(\hat{\rho}_{nL}(t) - \rho(t))^2\right] &= B_n^2(t) + \mathrm{Var}[\hat{\rho}_{nL}(t)] \\ &= (c_n(t)\rho_L(t) - \rho_f(t))^2 + \rho_L^2(t)\sigma_n^2(t) \\ &= (c_n^2(t) + \sigma_n^2(t))\rho_L^2(t) - 2c_n(t)\rho_f(t)\rho_L(t) + \rho_f^2(t) \\ &= (c_n^2(t) + \sigma_n^2(t))\left[\rho_L^2(t) - 2\frac{c_n(t)\rho_f(t)}{c_n^2(t) + \sigma_n^2(t)}\rho_L(t) + \frac{\rho_f^2(t)}{c_n^2(t) + \sigma_n^2(t)}\right] \\ &= (c_n^2(t) + \sigma_n^2(t))\left[\left(\rho_L(t) - \frac{c_n(t)\rho_f(t)}{c_n^2(t) + \sigma_n^2(t)}\right)^2 - \left(\frac{c_n(t)\rho_f(t)}{c_n^2(t) + \sigma_n^2(t)}\right)^2 + \right. \\ &\left. + \frac{\rho_f^2(t)}{c_n^2(t) + \sigma_n^2(t)}\right].\end{aligned} \tag{3.8}$$

Al minimizar esta expresión en ρ_L, teniendo en cuenta la no negatividad del primer sumando, y que el segundo y el tercero no dependen de L, se obtiene:

$$\rho_{L^*}(t) = \frac{c_n(t)\rho_f(t)}{c_n^2(t) + \sigma_n^2(t)}. \tag{3.9}$$

Es claro que, una vez obtenido el valor que punto a punto minimiza el MSE del estimador, la función que toma esos valores para cada punto del soporte, minimizará el MISE.

Con esta expresión queda caracterizado el núcleo óptimo, pero no es fácil a partir de ella obtener una expresión explícita para L^*. Sin embargo, al analizar la fórmula (3.3) de la estimación de ρ_f, lo único que se necesita es precisamente ρ_{L^*}.

[1] El error cuadrático medio de un estimador de la densidad $\hat{f}$ se define como

$$MSE(\hat{f}(x)) = \mathbb{E}\left[(\hat{f}(x) - f(x))^2\right]$$

y se trata de un criterio de error que dependerá del punto de evaluación. Además se prueba, sin más que emplear la linealidad del operador esperanza y cálculos algebraicos, que puede expresarse como suma del sesgo al cuadrado y la varianza.

Basándose en el método de selección de ventana para el estimador tipo núcleo propuesto por Sheather y Jones (1991), se estimará ρ_{L^*} bajo la hipótesis de que f sea una densidad normal de media μ y desvación típica σ. Se reemplaza esta expresión junto con un estimador apropiado de σ en (3.3). Se sustituye la estimación resultante en (2.8) para así obtener la función característica del núcleo óptimo y finalmente, aplicando de manera conveniente la transformada de Fourier inversa, se determina un selector plug-in de K^*.

En el caso de que f sea una normal de media 0 y desviación típica σ (tomamos media nula porque el parámetro de localización no influye en los cálculos), se tiene

- $\rho_f(t) = e^{-\sigma^2 t^2}$.

- $c_n(t) = \frac{1}{n} + \frac{n-1}{n} e^{-\sigma^2 t^2}$.

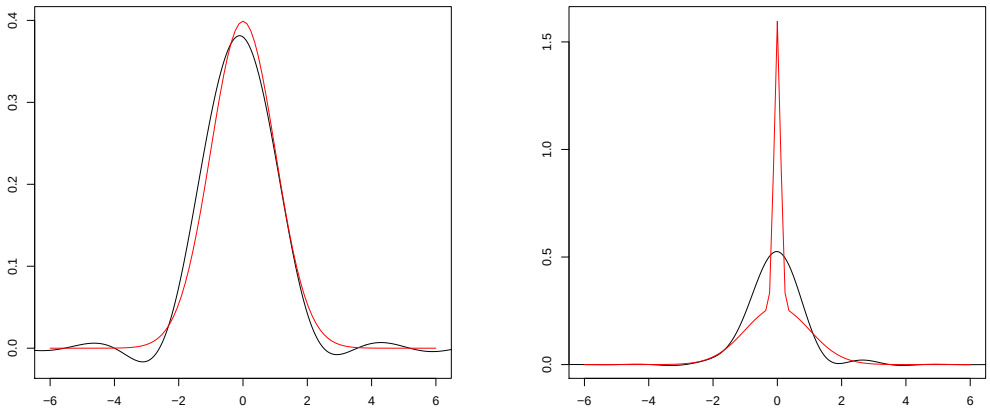
- $\sigma_n^2(t) = \frac{n-1}{n^3} \left[2(n-2) \left(e^{-3\sigma^2 t^2} + e^{-\sigma^2 t^2} \right) + 1 + e^{-4\sigma^2 t^2} - (4n-6) e^{-2\sigma^2 t^2} \right]$.

En la implementación, hemos empleado como estimador de σ^2 la varianza muestral.

En la Figura 3.4 puede verse el estimador (2.4) con el selector plug-in para las mismas muestras que fueron anteriormente empleadas para la regla del pulgar. Y al igual que ocurría con ese procedimiento, tiende a sobresuavizar en exceso las estimaciones. Las funciones obtenidas para cada una de las muestras, son bastante similares entre sí, a pesar de las diferencias existentes entre los modelos que las generan. Por esta misma razón, no es capaz de captar bimodalidades y mucho menos elementos más complejos como las modas del modelo M10 o el pico del modelo M4.

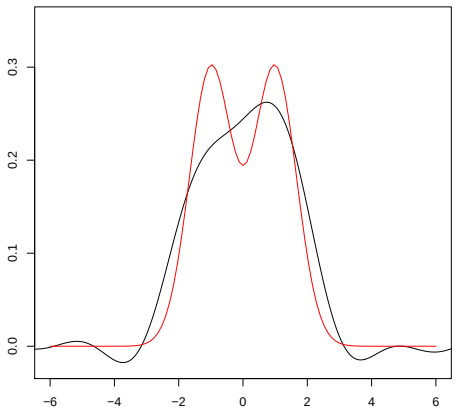
Si se compara las gráficas de la Figura 3.4 con las obtenidas para la regla del pulgar en la Figura 3.3, se ve que las estimaciones obtenidas son muy parecidas. Esto puede ser debido a que, aunque de manera diferente, en ambos casos se emplea la hipótesis de normalidad.

Los ISE obtenidos para la regla plug-in varían entre 0.004 y 0.2, un rango de valores muy parecido al de la regla del pulgar, tal y como era de esperar dada la similitud de las estimaciones. A la vista de los resultados, se podría decir que parece que la hipótesis de normalidad posee más influencia de la deseable y provoca sobresuavizaciones.

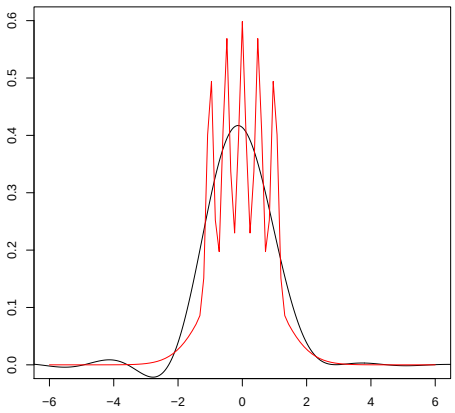


(a) M1

(b) M4



(c) M6



(d) M10

Figura 3.4: Representación del estimador (2.4) (línea negra) con la regla plug-in para muestras de tamaño $n = 100$ de los modelos M1, M4, M6 y M10 de Marron y Wand (1992), junto con las densidades teóricas de cada modelo (línea roja). (Nótese que las gráficas no están en la misma escala).

3.5. El estimador autoconsistente

Este selector, propuesto por Bernacchia y Pigolotti (2011b), se basa en la obtención de una ecuación para la función característica del estimador (2.4) óptimo, a partir de los resultados de Watson y Leadbetter (1963), y su posterior resolución por medio de un método numérico de punto fijo.

Recordemos la expresión correspondiente a la función característica para el estimador (2.4) con núcleo óptimo caracterizado mediante (2.8):

$$\varphi_{\hat{f}^*_{nK}}(t) = \varphi_n(t)\varphi_{K^*}(t) = \varphi_n(t)\frac{n\rho_f(t)}{1+(n-1)\rho_f(t)} = \frac{n\varphi_n(t)}{n-1+|\varphi_f(t)|^{-2}}. \tag{3.10}$$

La propuesta se fundamenta en encontrar un punto fijo para la ecuación (3.10), asumiendo que la función característica del estimador de la densidad y la característica asociada a la densidad teórica deben coincidir. Para ello se necesita un valor inicial φ_0, de manera que sin más que sustituir en el lado derecho de la igualdad, se obtendría una primera aproximación, φ_1. A

continuación se trata de obtener un mejor estimador φ_2 usando el núcleo óptimo para φ_1.

Repitiendo iterativamente este procedimiento, se construye una secuencia de estimadores mejorados, hasta que se alcanza un punto de equilibrio que verifique:

$$\hat{\varphi}_{sc} = \frac{n\varphi_n(t)}{n-1+|\hat{\varphi}_{sc}|^{-2}},$$

cuya solución es calculada en Bernacchia y Pigolotti (2011b):

$$\hat{\varphi}_{sc}(t) = \frac{n\varphi_n(t)}{2(n-1)}\left[1+\sqrt{\left\{1-\frac{4(n-1)}{n^2\rho_n(t)}\right\}}\right] I_A(t). \tag{3.11}$$

A continuación se detalla como se determina ese conjunto A. Para ello se debe profundizar un poco más en la resolución del problema del punto fijo. Se consideran dos situaciones diferentes, por una parte $\rho_n(t) < 4(n-1)/n^2$, en la que la única solución posible será $\hat{\varphi}_{sc}(t) = 0$, y por otra $\rho_n(t) > 4(n-1)/n^2$, en la que el problema de punto fijo planteado anteriormente posee dos soluciones diferentes:

$$\hat{\varphi}^{\pm}_{sc}(t) = \frac{n\varphi_n(t)}{2(n-1)}\left[1\pm\sqrt{\left\{1-\frac{4(n-1)}{n^2\rho_n(t)}\right\}}\right],$$

tales que $\hat{\varphi}^{+}(0) = 1$, mientras que $\hat{\varphi}^{-}$ no lo cumple, lo cual es un problema ya que no verificaría una de las propiedades fundamentales de las funciones características.

En la resolución numérica de ecuaciones es muy importante conseguir soluciones estables, esto es, con derivada menor que 1, puesto que esto garantiza la verificación de un conjunto de "buenas" propiedades. Es claro que la solución nula es siempre estable, en cuanto a las soluciones no triviales, $\hat{\varphi}^{\pm}_{sc}$, se obtiene calculando las derivadas que únicamente $\hat{\varphi}^{+}_{sc}$ lo es, llevando por consiguiente a desechar la otra solución.

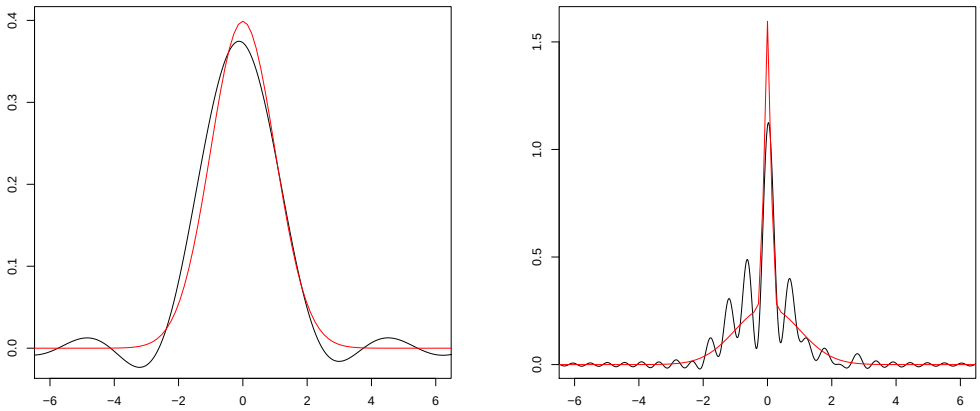
El único elemento que nos resta por conocer de la solución al problema de punto fijo es el conjunto A sobre el que está definina. Para ello, los autores definen el conjunto B:

$$B = \left\{t : \rho_n(t) \geq \frac{4(n-1)}{n^2}\right\}$$

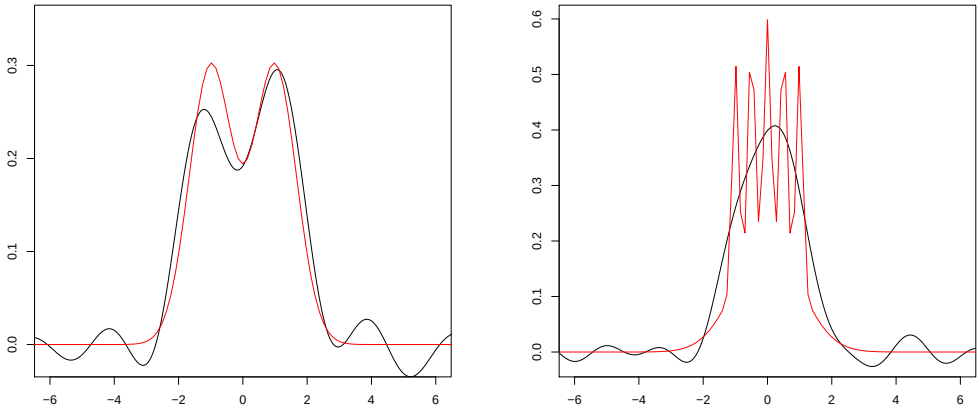
y así, $\hat{\varphi}_{sc}(t) = 0$ cuando $t \notin B$. Sin embargo, que $t \in B$ no garantiza que el proceso iterativo se quede con la solución estable, $\hat{\varphi}^{+}_{sc}$, pues si el valor inicial cumple $|\varphi_0(t)| < |\varphi^{-}_{sc}|$, entonces se llegaría a la solucion nula. Por tanto, sin más que tomar A el subconjunto de elementos de B que cumplen $|\varphi_0(t)| \geq |\varphi^{-}_{sc}|$, se obtendría (3.11). Nótese que el conjunto A no está previamente determinado sino que depende de la condición inicial.

Al igual que hasta ahora, una vez determinada la función característica del estimador, se aplica la transformada inversa de Fourier y se obtiene el correspondiente estimador de la densidad. Se comprueba que integra la unidad y que bajo la hipótesis de que el conjunto A sea acotado, converge casi seguro a la densidad teórica. La demostración de este resultado se puede consultar en el apéndice de Bernacchia y Pigolotti (2011b).

Se presentan en la Figura 3.5 algunos ejemplos del estimador obtenido con este selector para las mismas muestras de los modelos de Marron y Wand (1992) que ya se han empleado para los anteriores selectores, junto con las correspondientes densidades teóricas.



(a) M1 (b) M4



(c) M6 (d) M10

Figura 3.5: Representación del estimador (2.4) (línea negra) con el selector autoconsistente para muestras da tamaño $n = 100$ de los modelos M1, M4, M6 y M10 de Marron y Wand (1992), junto con las correspondientes densidades teóricas (línea roja). (Nótese que las gráficas no están todas en la misma escala).

El selector autoconsistente proporciona estimadores más rugosos, que aunque para modelos suaves como el M1 la estimación es bastante mala, en modelos más complejos hay ganancia con respecto a los selectores previamente analizados. Además, aunque sí capta bimodalidades y levemente picos como el del modelo M4, no es capaz de ajustar bien las diversas modas tan marcadas y próximas del modelo M10. En cuanto a los ISE obtenidos están en los cuatro modelos, entre 0.004 y 0.07, que sigue siendo un rango bastante amplio.

Se ha constatado además, que este selector posee ciertos problemas en la aplicación a datos reales. Pues empleando la función proporcionada por los autores en Bernacchia y Pigolotti (2011a) sobre los datos de *Geyser* presentados en la Introducción, no se consigue obtener una estimación de la densidad de los mismos.

3.6. Validación cruzada

El procedimiento que se presenta a continuación está basado en las ideas de Rudemo (1982) y Bowman (1984). El objetivo inicial del problema pasa por conseguir una buena estimación de φ_f. Recuérdese que al comienzo de este mismo capítulo se proponía el estimador empírico, pero

no se obtenían buenos resultados por las oscilaciones que presenta. Analizando la estructura de las funciones representadas en las gráficas Figura 3.1 y Figura 3.2, se observa que mientras que la función teórica ρ_f es una función suave de integral finita, esto es, $\rho \in L_1$, no ocurre lo mismo con su estimador.

Además, teniendo en cuenta el teorema de Riemann-Lebesgue que puede consultarse en Billingsley (1995, Pag. 345), se sabe que

$$\lim_{t\to\infty} \rho(t) = 0,$$

hecho que tampoco cumple el estimador empírico ρ_n, ya que está oscilando constantemente. Por tanto, se estaría tratando de aproximar una función que va a cero, muchas veces muy rápido, para valores grandes de t, por una que presenta numerosas oscilaciones en las colas, por lo que la discrepancia entre ambas sería mayor que la deseada.

El problema que se plantea entonces es como conseguir que ρ_n no tenga tanto ruido. Para ello existen diversas soluciones. La que se propone en este trabajo consiste en cortar la función en un punto a determinar, es decir, elegir un valor T de abscisa (e implícitamente su opuesto), a partir del cual el estimador empírico de ρ_f se define como nulo (tanto en los valores de abscisa positivos como en los negativos). De este modo se suprimen las oscilaciones en las colas, obteniendo una función con soporte compacto e integral finita. Así, además de cumplir la condición de las colas dada en el teorema de Riemann-Lebesgue, este estimador también formaría parte del espacio L_1.

La nueva propuesta de estimador para ρ_f es:

$$\hat{\rho}_T = \rho_n(t) I_{[-T,T]}(t) = \rho_n(t) I_{\{|t|\leq T\}}.$$

Una vez obtenida la expresión, se sustituye en (2.8), y se tiene un estimador de la función característica del núcleo óptimo:

$$\hat{\varphi}_{K^*}(t) = \frac{n\rho_n(t) I_{|t|\leq T}}{1 + (n-1)\rho_n(t) I_{|t|\leq T}}. \tag{3.12}$$

Se aplica la transformada inversa de Fourier para obtener una expresión de la estimación del núcleo óptimo, y sustituyéndolo en (2.4), se obtendría el estimador asociado, que se denotará por $\hat{f}^*_{nT}$.

Restan por resolver dos cuestiones importantes:

- Demostrar que $\hat{\varphi}_{K^*}$ es integrable para poder aplicar la fórmula de la inversión y obtener una estimación del núcleo óptimo.
- Determinar un método que nos permita obtener un valor de T apropiado con la información que nos proporcionan los datos.

El primer punto tiene fácil solución, ya que la estimación que se obtiene de la función característica es una función acotada definida sobre un soporte compacto, lo que garantiza sin más su integrabilidad.

En cuanto al segundo punto, se ha aplicado el procedimiento estándar de validación cruzada a este caso particular, de manera que para cada muestra, se dispone de un método que nos permite escoger el valor de T más adecuado. Como en este caso se escoge el parámetro para

cada muestra, en lugar de basarse en el MISE, que es un promedio de errores, la elección del parámetro se apoyará en la minimización del ISE.

Dada una m.a.s. $X_1, \ldots, X_n$, es necesario disponer de la expresión del ISE del estimador:

$$\text{ISE}(\hat{f}^*_{nT}) = \int (\hat{f}^*_{nT}(x) - f(x))^2 dx = \int {\hat{f}^*_{nT}}^2(x)dx + \int f^2(x)dx - 2\int \hat{f}^*_{nT}(x)f(x)dx, \quad (3.13)$$

teniendo en cuenta que f es un elemento fijo en T, minimizar la expresión anterior, equivale a:

$$\min_T \left(\int {\hat{f}^*_{nT}}^2(x)dx - 2\int \hat{f}^*_{nT}(x)f(x)dx \right). \quad (3.14)$$

Falta obtener una forma adecuada de la expresión anterior que permita calcularla a partir de un conjunto de datos dado. Para ello se considera el desarrollo de cada sumando por separado:

- $\boldsymbol{\int {\hat{f}^*_{nT}}^2(x)dx}$

  Dada una muestra, el integrando es totalmente conocido y por tanto su integral se puede obtener, con mayor o menor complejidad de cálculo según el caso.

  $$\int {\hat{f}^*_{nT}}^2(x)dx = \frac{1}{2\pi}\int |\varphi_{\hat{f}^*_{nT}}|^2(t)dt = \frac{1}{2\pi}\int |\varphi_n(t)|^2|\hat{\varphi}_{K^*}(t)|^2dt = \frac{1}{2\pi}\int \rho_n(t)|\hat{\varphi}_{K^*}(t)|^2dt,$$

  donde $\hat{\varphi}_{K^*}(t)$, cuya expresión se detalla en (3.12), depende del valor de T.

- $\boldsymbol{\int {\hat{f}^*_{nT}}^2(x)f(x)dx}$

  Este término no es conocido ya que depende de f, pero lo podemos estimar del siguiente modo:

  $$\int {\hat{f}^*_{nT}}^2(x)f(x)dx = \mathbb{E}_f\left[{\hat{f}^*_{nT}}^2(X)\right] \simeq \frac{1}{n}\sum_{i=1}^n \hat{f}^{-i}_{nT}(X_i),$$

  donde $\hat{f}^{-i}_{nT}$ denota el estimador obtenido extrayendo de la muestra el i-ésimo dato.

El hecho de tener que extraer un dato se debe a que se dispone de una única muestra de la variable aleatoria, que se emplea para hacer dos estimaciones, por una parte el estimador de la densidad, y por otra la media de la variable transformada mediante el estimador de la densidad. Por tanto, si se dispusiese de otra muestra $Y_1, \ldots, Y_n$ independiente de la $X_1, \ldots, X_n$, se podría estimar el segundo sumando como

$$\int \hat{f}_{nT}(x)f(x)dx = \mathbb{E}\left[\hat{f}_{nT}(X)\right] \simeq \frac{1}{n}\sum_{i=1}^n \hat{f}_{nT}(Y_i),$$

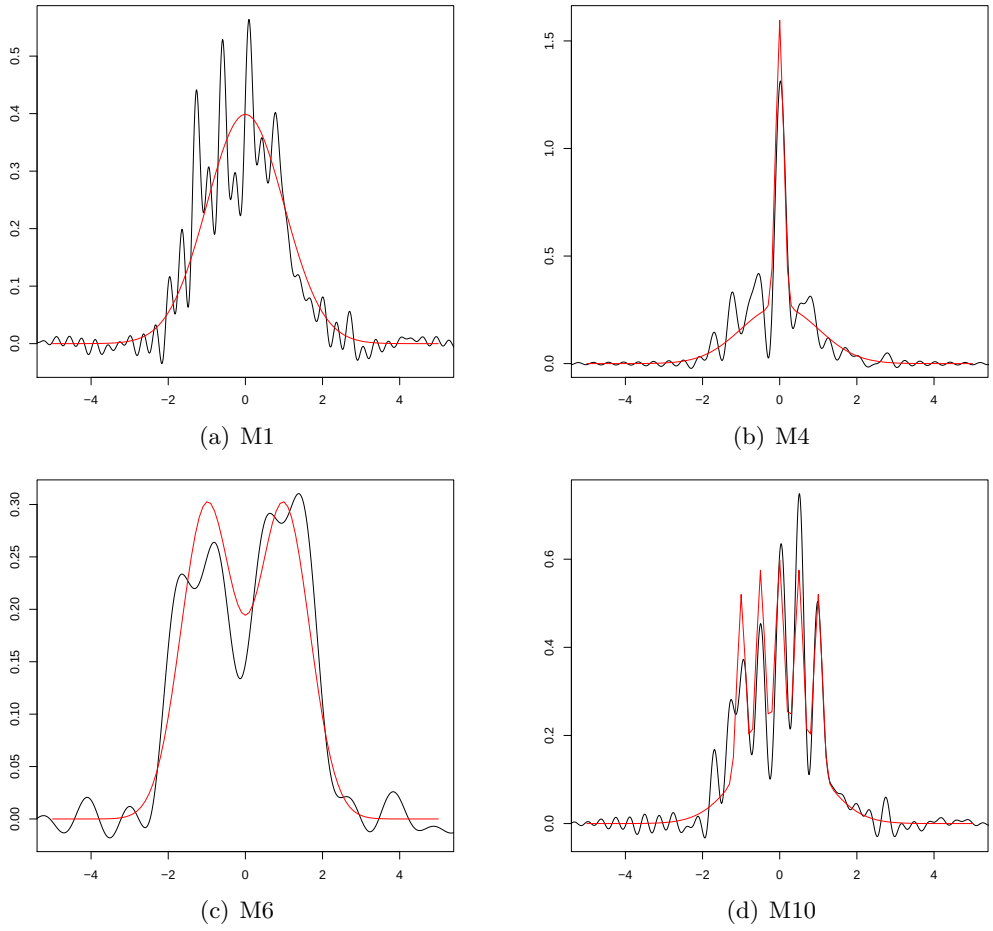
pero como no es así, la solución pasa por, en cada uno de los sumandos de la media muestral de la variable transformada, $f(X)$, extraer ese dato, de manera que no se use la misma información en ambas estimaciones.

En definitiva, T se obtiene minimizando la siguiente expresión, que es una aproximación de (3.14):

$$\min_{T} \left[\frac{1}{2\pi} \int \rho_n(t) \left(\frac{n\rho_n(t) I_{|t|\leq T}}{1 + (n-1)\rho_n(t) I_{|t|\leq T}} \right)^2 dt - 2\frac{1}{n} \sum_{i=1}^{n} \hat{f}_{nT}^{-i}(X_i) \right]. \tag{3.15}$$

En esta expresión, dada una muestra, se pueden determinar todos los elementos involucrados en ella, y por consiguiente se puede obtener el valor de T óptimo.

En la Figura 3.6 se presenta el estimador con el selector de validación cruzada para las mismas muestras que se han empleado en los casos anteriores.



(a) M1 (b) M4

(c) M6 (d) M10

Figura 3.6: Representación del estimador (2.4) con la regla de validación cruzada para muestras de tamaño $n = 100$ de los modelos M1, M4, M6 y M10 de Marron y Wand (1992), junto con las densidades teóricas de cada modelo. (Nótese que las gráficas no están en la misma escala).

Estos estimadores son mucho menos suaves que los obtenidos mediante la regla del pulgar y el método plug-in, y aunque para la densidad normal no funciona bien, vemos como capta por ejemplo las irregularidades del modelo M10 (algo que hasta ahora no ocurría), y las bimodalidades del M6. Obviamente hay que pagar un precio a cambio de obtener un estimador que capte estos efectos extremos, que es el ruido existente en las zonas más suaves. En cuanto a los ISE obtenidos están, en los cuatro modelos, entre 0.01 y 0.05, que son valores más pequeños que los obtenidos para la regla del pulgar, plug-in o incuso el estimador autoconsistente.

En conjunto, al analizar los resultados gráficos obtenidos para los cuatro selectores, pueden clasificarse en dos grupos, por una parte la regla del pulgar y plug-in que claramente obtienen una estimación sobresuavizada; por la otra parte el autoconsistente y validación cruzada, que tienden a infrasuavizar la estimación, especialmente el segundo de ellos. Se ha apreciado que el selector autoconsistente proporciona resultados bastante malos para alguno de los modelos estudiados.

En el Capítulo 4, se podrán analizar los valores numéricos del error, obtenidos mediante un estudio de simulación, para cada uno de los selectores presentados hasta ahora. Además, se hará la comparación con los métodos existentes más comunes en la selección de la ventana, Silverman (1986), Sheather y Jones (1991) y Validación Cruzada (Bowman 1984).

Capítulo 4

Estudio de simulación

Este capítulo está organizado en dos secciones, por un parte la correspondiente a la implementación de las técnicas necesarias para el cálculo y la representación del estimador, el núcleo óptimo o los errores, y por la otra el estudio de simulación en el que se compararán los diferentes selectores.

El estudio de simulación se divide también en dos partes. En la primera de ellas se comparan los resultados obtenidos para los selectores detallados en el Capítulo 3. Una vez hecho este análisis, es natural comparar tanto nuestras propuestas como la de Bernacchia y Pigolotti (2011b) con los métodos habituales de selección de la ventana para el estimador tipo núcleo.

4.1. Técnicas implementadas

Esta sección comprende una explicación en detalle de las técnicas necesarias para implementar todos los conceptos y procedimientos descritos en la memoria.

Para el aspecto computacional se ha empleado únicamente el software estadístico R Core Team (2012). El desarrollo del código ha sido en su gran mayoría personal salvo en lo que concierne a los selectores de la ventana, para lo que se ha empleado las rutinas previamente implementadas a las que hay acceso desde el paquete básico de R, y el código para el selector autoconsistente que se ha obtenido de Bernacchia y Pigolotti (2011a).

En primer lugar se va a detallar el proceso de paso al dominio de frecuencias del estimador (2.4) y la aplicación a este caso particular de la transformada de Fourier, que como ya se ha mencionado, incrementa notablemente la velocidad de cálculo. Posteriormente se comenta con detalle la corrección que se aplica al estimador para que sea una densidad, y que fue ideada por Glad y cols. (2003).

4.1.1. Transformada de Fourier

En general, la transformada discreta de Fourier (DFT, *Discret Fourier Transform*) de un vector $c = (c_1, \ldots, c_M) \in \mathbb{R}^M$ viene dada por $DFT(c) = d$, siendo $d = (d_1, \ldots, d_M) \in \mathbb{R}^M$ con

$$d_j = \sum_{k=1}^{M} c_k e^{2\pi i (l-1)(k-1)/M} \quad l = 1, \ldots, M.$$

Para calcular la DFT de modo rápido, se utiliza el algoritmo conocido como transformada rápida de Fourier (FFT, *Fast Fourier Transform*), que en R Core Team (2012) está implementado en la función `fft` (con argumentos de entrada el vector c e `inverse=TRUE`). Si se cambia este segundo argumento a `FALSE`, lo que hacemos es la trasformada inversa, que no es más que la misma expresión con el exponente cambiado de signo, y permite pasar del dominio de frecuencias al soporte de la función.

Téngase en cuenta que el objetivo es evaluar el estimador en un intervalo acotado $[a, b]$, para lo que se define un rejilla de M puntos equiespaciados $a = r_1 < r_2 < \ldots < r_M = b$. El procedimiento pasa por evaluar $\varphi_{\hat{f}_{nK}}(t) = \varphi_n(t)\varphi_K(t)$ en una rejilla apropiada y aplicar la transformada inversa de Fourier para obtener $\hat{f}_{nK}(x) = \frac{1}{2\pi}\int e^{-itx}\varphi_{\hat{f}_{nK}}(t)dt$. Esto presenta dos grandes dificultades, por una parte se precisa de un método rápido para evaluar φ_n en la rejilla y por otra una manera rápida para la obtención de la transformada inversa.

El primer paso a seguir se basa en la discretización de los datos (*binning*), que consiste en pasar de los puntos $X_1, \ldots, X_n$ con peso $1/n$ a los puntos de la rejilla $r_1, \ldots, r_M$ con pesos $\xi_1, \ldots, \xi_M$. Estos nuevos pesos han de reflejar de algún modo la cantidad de puntos X_i que hay en un entorno de cada uno de los nodos de la rejilla.

Dado que la rejilla está constituida por M puntos equiespaciados, se define el paso de la misma como $\delta = (b-a)/(M-1)$ y así se tendría $r_k = a + (k-1)\delta$ tomando $k = 1, \ldots, M$. Existen múltiples procedimientos para la asignación de pesos, ξ_k. En este caso se ha decidido emplear el conocido como *linear binning*. Esta técnica tiene en cuenta únicamente los datos de la muestra situados en los intervalos contiguos a cada punto de la rejilla, esto es, si consideramos el elemento r_k, se valorarán en cuenta los datos que estén en (r_{k-1}, r_{k+1}). El *linear binning* asigna los pesos mediante un decrecimiento lineal en los citados intervalos; la expresión matemática es:

$$\xi_k = \frac{1}{n}\sum_{i=1}^{n}(1 - |X_i - r_k|/\delta)I_{\{|X_i - r_k| < \delta\}}. \tag{4.1}$$

Al cambiar la muestra $X_1, \ldots, X_n$ con pesos $1/n$ por la rejilla $r_1, \ldots, r_M$ con pesos $\xi_1, \ldots, \xi_M$, lo que se está haciendo es intentar aproximar la medida empírica habitual $\mu_n(A) = \frac{1}{n}\sum_{j=1}^{n} I_{\{X_j \in A\}}$ por la medida discretizada $\nu_M(A) = \sum_{k=1}^{M} \xi_k I_{\{r_k \in A\}}$, de modo que se puede expresar

$$\varphi_n(t) = \sum_{j=1}^{n}\frac{1}{n}e^{itX_j} = \int e^{itx}\mu_n(dx) \approx \int e^{itx}\nu_M(dx) = \sum_{k=1}^{M}\xi_k e^{itr_k} = e^{ita}\sum_{k=1}^{M}\xi_k e^{it(k-1)\delta}.$$

Tomando la rejilla en el dominio de frecuencias, $t_l = 2\pi(l-1)/(\delta M)$ con $l = 1, \ldots, M$ se tiene una expresión en forma de la DFT para la característica empírica:

$$\varphi_n(t_l) \approx e^{it_l a}\sum_{k=1}^{M}\xi_k e^{2\pi i(l-1)(k-1)/M}, \quad l = 1, \ldots, M.$$

Es importante notar que los puntos $\{t_l\}_{l=1}^{M}$ constituyen una rejilla equiespaciada en el intervalo $[0, \frac{2\pi(M-1)^2}{M(b-a)}] \approx [0, \frac{2\pi M}{b-a}]$, que se aproxima a $[0, \infty]$ cuando $M \to \infty$, esto es, cuando la rejilla $\{r_k\}_{k=1}^{M}$ se hace más fina. Por ello es importante mantener un equilibrio entre la longitud del intervalo inicial, $[a, b]$, y el número de puntos que tomamos en él, pues tendrá repercusión directa sobre la rejilla del dominio de frecuencias.

Una vez evaluada la característica empírica, restaría determinar un método rápido para el cálculo de la transformada inversa, para ello se tendrá en cuenta que

$$\xi_k^* = \varphi_{\hat{f}_{nK}}(t_l) = \varphi_n(t_l)\varphi_K(t_l), \quad l = 1, \ldots, M,$$

que se corresponde a evaluar $\varphi_{\hat{f}_{nK}}$ en una rejilla de M puntos equiespaciados con distancia entre ellos $\Delta = t_2 - t_1 = 2\pi/(\delta M)$.

Además, como $t_1 = 0$ y empleando las propiedades generales de la función característica $\varphi_{\hat{f}_{nK}}(-t) = \overline{\varphi_{\hat{f}_{nK}}(t)}$, resulta que el vector $(\overline{\xi_M^*}, \ldots, \overline{\xi_2^*}, \xi_1^*, \ldots, \xi_M^*)$ equivale a evaluar la característica del estimador en los $2M-1$ puntos equiespaciados $-t_M, \ldots, -t_2, 0, t_2, \ldots, t_M$.

Ahora sólo resta calcular la integral correspondiente a la transformada inversa, para lo que se utiliza la aproximación numérica por el método de rectángulos:

$$\begin{aligned}
\hat{f}_{nK}(r_k) &= \frac{1}{2\pi}\int e^{-itr_k}\varphi_{\hat{f}_{nK}}(t)dt \approx \frac{1}{2\pi}\left[\sum_{l=1}^{M}\Delta\xi_l^* e^{-it_l r_k} + \sum_{l=2}^{M}\Delta\overline{\xi_l^*}e^{it_l r_k}\right] \\
&= \frac{1}{2\pi}\left[\sum_{l=1}^{M}\Delta\xi_l^* e^{-it_l r_k} + \sum_{l=1}^{M}\Delta\overline{\xi_l^*}e^{it_l r_k} - \frac{2\pi}{\delta M}\right] = \frac{1}{2\pi}\left[\sum_{l=1}^{M}\Delta 2\mathrm{Re}\left(\xi_l^* e^{-it_l r_k}\right) - \frac{2\pi}{\delta M}\right] \\
&= \frac{2}{\delta M}\mathrm{Re}\left(\sum_{l=1}^{M}\xi_l^* e^{-it_l r_k}\right) - \frac{1}{\delta M} = \frac{2}{\delta M}Re\left(e^{-ita}\sum_{l=1}^{M}\xi_l^* e^{-i(l-1)(k-1)2\pi/M}\right) - \frac{1}{\delta M}.
\end{aligned}$$

Así ya se tiene una expresión con la estructura de la transformada de Fourier discreta y se puede emplear la función `fft` de R Core Team (2012).

Esta metodología se emplea de manera análoga para la representación del núcleo óptimo y para los estimadores obtenidos con cada uno de los selectores descritos.

4.1.2. Corrección del estimador

A lo largo del presente trabajo hemos visto representaciones gráficas del estimador, en la Figura 2.4, Figura 2.5, Figura 3.3, Figura 3.4, Figura 3.5 y Figura 3.6 en las que se observa que el estimador puede tomar valores negativos. Así no sería una densidad y no permitiría, entre otros procedimientos, simular datos a partir de él.

Una solución a este problema se presenta en Glad y cols. (2003). La idea se basa en anular el estimador en aquellos puntos en los que tome valores negativos, pero de este modo se puede variar el valor de la integral, por lo que es necesario reescalar la función para que integre uno. Además se consigue siempre un ISE menor o igual que con el estimador original.

En los casos estudiados, el estimador sin ancho verifica la condición de que su integral en las zonas donde toma valores no negativos es mayor que la unidad, y por tanto la corrección a aplicar que se presenta en Glad y cols. (2003) propone lo siguiente:

$$\tilde{f}(x) = \text{máx}\{0, \hat{f} - \xi\}, \tag{4.2}$$

donde ξ es una constante que se escoge de manera que $\int \tilde{f}(x)dx = 1$.

En el artículo Glad y cols. (2003) se prueba que el nuevo estimador está bien definido, es decir, que esa constante ξ siempre existe y es única. Y demuestran el resultado que pone de manifiesto que esa modificación es siempre al menos tan buena, con respecto al ISE, como el estimador original.

Para poder aplicar esta corrección, se ha creado en R Core Team (2012) una función basada en el algoritmo de dicotomía (también conocido como *caza del león*) sobre ξ. Se toman como condiciones iniciales $\xi_1 = 0$ y $\xi_2 = \text{máx}_x \hat{f}_{nK}(x)$, de manera que se garantiza que $\int (\tilde{f}(x) - \xi_1)dx > 1$ y $\int (\tilde{f}(x) - \xi_2)dx < 1$.

Una vez fijadas las condiciones inciales, se define el nuevo valor de ξ como $\xi_3 = (\xi_1 + \xi_2)/2$, se calcula la integral de $(\tilde{f} - \xi_3)$ y se actualizan los valores de ξ (si esta nueva integral es mayor que la unidad, se tomará $\xi_1 = \xi_3$, mientras que si es menor se hará $\xi_2 = \xi_3$). Este proceso se repite de manera recurrente, y su finalización viene determinada por el denominado criterio de parada. En este caso, este criterio se basa en medir la discrepancia entre la integral del estimador corregido y la unidad, y cuando es lo suficientemente pequeña (valor que determina el usuario) el algoritmo pararía y devolvería la corrección definitiva del estimador.

4.2. Resultados del estudio de simulación

En esta segunda sección se presenta el estudio de simulación realizado, cuyo objetivo, como ya se ha introducido, es el de comparar los distintos selectores presentados en el Capítulo 3 entre sí, y con los selectores más habituales para el estimador tipo núcleo: la regla del pulgar de Silverman (1986), la propuesta de Sheather y Jones (1991) y la validación cruzada desarrollada en Bowman (1984).

Para los selectores del Capítulo 3 se ha elaborado código en R que, dada una muestra, aplica cada uno de los selectores, calcula el estimador de la densidad correspondiente y computa el error cometido con respecto al modelo teórico según el cual se generan los datos. Para el selector propuesto en Bernacchia y Pigolotti (2011b) el código es accesible desde Bernacchia y Pigolotti (2011a). Para el estimador (2.1) se ha utilizado la función `density` que se puede encontrar en el paquete básico de R, empleando por defecto el núcleo gaussiano y, en cada situación, se escoge el selector correspondiente.

Para los dos estimadores se ha empleado la misma rejilla, la que se detalla previamente en este mismo capítulo en el marco de la transformada de Fourier. Es una rejilla equiespaciada en el intervalo $[-100, 100]$ y formada por $M = 2^{14}$ puntos, de manera que el paso de la misma es de aproximadamente 0.0122.

Se utilizan como modelos teóricos el conjunto de densidades de Marron y Wand (1992), cuya representación puede verse en la Figura 2.3. Los tamaños muestrales considerados son $n = 100, 400$ y 1600. Para cada modelo se generarán de modo aleatorio en R Core Team (2012) $B = 500$ muestras partiendo de la semilla `1234567`.

Para cada una de las muestras, se obtiene el estimador evaluado en la rejilla de puntos, se calcula el ISE cometido en cada caso y finalmente se promedian dichos valores en las B muestras para obtener el MISE asociado a cada procedimiento y modelo.

Es interesante recordar que para todos los selectores se emplea el mismo conjunto de muestras, esto es, que el error cometido en cada modelo se calcula sobre los mismos datos, de manera que se suprime la influencia de la variabilidad de la muestra en la comparación.

En este capítulo se introduce nueva notación más compacta para emplear en las tablas y gráficas. Se entenderá por $Pulg$ la regla del pulgar, por PI la de plug-in, por $AutCon$ el selector autoconsistente y por CV la validación cruzada, presentados todos ellos en el Capítulo 3. Además, los ya mencionados selectores habituales para el estimador (2.1), se denotarán como $Pulg_{nc}$ la propuesta de Silverman (1986), por SJ_{nc} la de Sheather y Jones (1991) y CV_{nc} el selector de validación cruzada de Bowman.

Se ha comenzado analizando en comportamiento de los diferentes selectores del Capitulo 3. Se ha hecho dicho estudio en términos de MISE para cada uno de los modelos de Marron y Wand (1992) y se presenta en la Tabla 4.1 el resumen de los resultados:

	Pulg	**PI**	**AutCon**	**CV**
M1	0.4225	0.5141	0.5585	1.0416
M2	0.9180	0.8329	1.0145	1.4753
M3	20.9487	19.3236	6.6426	6.1279
M4	21.6577	18.0780	8.3298	5.9901
M5	63.8234	7.2810	6.5769	5.1776
M6	1.4756	1.6215	0.9595	1.3195
M7	10.5389	12.2834	0.9431	1.4849
M8	1.6633	1.7702	1.4700	1.8588
M9	1.8977	2.1715	1.0392	1.5316
M10	5.4153	5.4265	5.3483	3.8167
M11	1.6067	1.7521	1.1283	1.4324
M12	2.7502	2.8855	2.6689	3.0516
M13	2.1685	2.3580	1.4982	1.9505
M14	11.6825	12.7797	4.7992	4.7392
M15	12.7170	13.2222	8.7175	4.1723
M16	72.2156	75.6293	2.9907	3.5708

Tabla 4.1: MISE's ($\times 100$) para los selectores del Capítulo 3 y tamaño muestral $n = 100$.

El rango de valores de la regla del Pulgar y la de Plug-in son muy superiores a los de los otros dos selectores, salvo en los modelos M1 y M2 que se caracterizan por su enorme suavidad y sencillez. En el modelo M16, la diferencia es extrema, llegando a valores en torno a veinte veces superiores.

A pesar de que se han redondeado los datos a cuatro cifras decimales, la cantidad de valores

presentados en la Tabla 4.1 hacen bastante difícil su interpretación más allá de la comparación más general que se ha hecho en el párrafo anterior.

La solución a este problema pasa por elaborar un gráfico con escala de grises que refleje la clasificación de los selectores empleados. Para cada uno de los modelos, el color más claro indica el valor de MISE más bajo (y por consiguiente lo que sería más deseable) y el tono de gris más oscuro el MISE más elevado. En definitiva, que tanto mejor será el selector globalmente cuanto más clara sea la gama de grises en la que varía sobre el conjunto de los modelos considerados. Nótese que este gráfico es sólo una herramienta cualitativa en la que además no se refleja la magnitud de las diferencias sino únicamente el rango que cada selector ocupa entre los restantes.

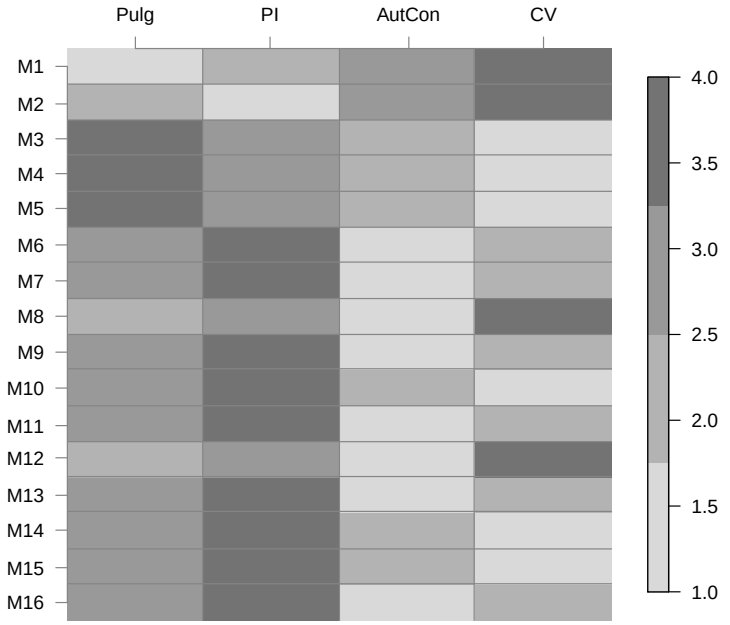


Figura 4.1: Representación mediante una escala de grises de la Tabla 4.1.

En la Figura 4.1 se ve como la regla del pulgar y el selector plug-in aportan resultados muy similares, y además son competitivos únicamente en los modelos M1 y M2, que son los más suaves. Es curioso ver como en el modelo M6 y sucesivos, la regla Plug-in que implementa la hipótesis de normalidad en una segunda fase, proporciona peores resultados que la propia regla del pulgar, que asume, para el cálculo del núcleo, la normalidad de la función de densidad teórica.

Comparando los resultados del autoconsistente y validación cruzada, se puede apreciar un cierto patrón de comportamiento según el cual, validación cruzada funcionaría mejor para modelos menos suaves o con elementos complejos, mientras que el autoconsistente proporciona mejores resultados para modelos menos complejos.

El problema de la infraestimación de las técnicas de validación cruzada es muy común y ha sido tratado en la literatura relativa a los selectores del estimador tipo núcleo. Este hecho,

que en principio supondría un problema, es lo que parece favorecer el buen ajuste para modelos complejos mientras que presenta demasiadas irregularidades en modelos suaves. Se podría decir que parece que la validación cruzada da "demasiada" importancia a la información muestral, conviertiendo lo que pueden ser irregularidades particulares, en características del modelo del que proceden.

De todas maneras, cabe decir que los datos anteriores son para tamaño muestral $n = 100$, por lo que tampoco podemos extraer información concluyente. Lo siguiente que se ha hecho en el estudio ha sido cuadruplicar el tamaño muestral, se espera que los errores disminuyan y que lo hagan además siguiendo aproximadamente las tasas de convergencia expuestas en el Capítulo 2. Veamos en la Tabla 4.2 los resultados para $n = 400$.

	Pulg	**PI**	**AutCon**	**CV**
M1	0.1181	0.1350	0.1649	0.2649
M2	0.3585	0.2431	0.3135	0.4188
M3	19.2520	17.0046	2.5530	2.0964
M4	20.4921	16.0999	2.0541	1.7218
M5	55.3126	1.7242	1.8774	1.4838
M6	1.0692	1.2810	0.2622	0.3713
M7	9.0014	12.0131	0.3333	0.4524
M8	1.3069	1.3547	0.6004	0.5441
M9	1.4126	1.8418	0.4159	0.5285
M10	5.0790	4.9184	4.6389	1.0907
M11	1.2140	1.4214	0.3959	0.4957
M12	2.4824	2.5349	2.3288	1.3096
M13	1.7153	1.9416	0.7130	0.8524
M14	10.9027	12.4751	2.4491	2.2004
M15	12.3286	12.9283	2.3202	1.7361
M16	70.0568	75.1657	0.8594	1.0290

Tabla 4.2: MISE's ($\times 100$) para los selectores del Capítulo 3 y tamaño muestral $n = 400$.

El valor de los MISE's de los modelos ha disminuido claramente, aunque parece que en menor medida en las reglas del pulgar y plug-in, especialmente en los modelos más complejos. Cabe destacar lo bien que parece funcionar la regla de validación cruzada para el modelo M10, pues el valor del error está entre tres y cuatro veces por debajo del de cualquiera de los restantes selectores.

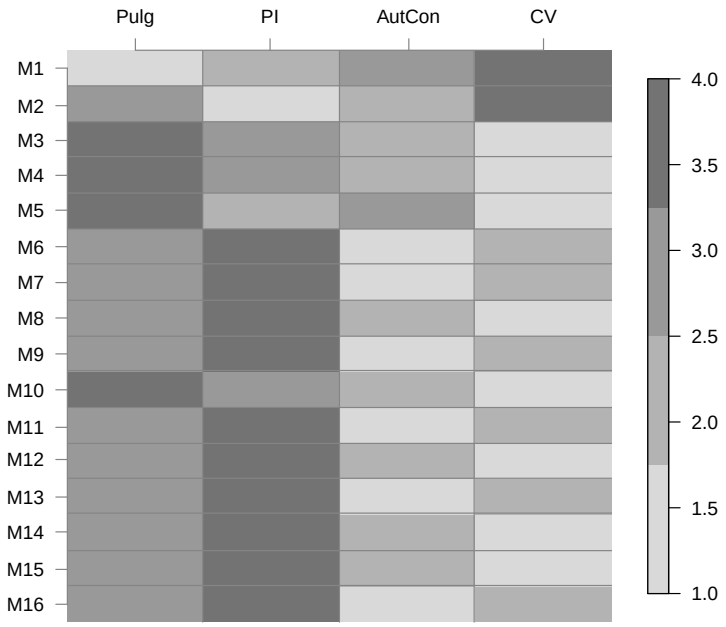


Figura 4.2: Representación mediante una escala de grises de la Tabla 4.2

La Figura 4.2 presenta una clara división de las tonalidades de gris, según la cual, para la gran mayoría de los modelos, el mejor selector es o bien el autoconsistente, o bien validación cruzada. A excepción, por supuesto, de los modelos M1 y M2.

Si se fija solo la atención en el selector autoconsistente y en validación cruzada, vemos que con respecto a los resultados obtenidos para $n = 100$, el primero de ellos gana calidad en los modelos mas suaves y la pierde en los complicados; mientras que al selector de validación cruzada le ocurre exactamente lo contrario. Se podría decir que estos dos procedimientos se complementan muy bien el uno al otro.

Aunque se ha realizado un estudio más completo llegando a analizar los resultados para muestras de tamaño $n = 10000$ para aquellos selectores más competitivos. Sin embargo, los últimos resultados que se presentan en este formato son para $n = 1600$. Esto es debido a que a partir de dicho valor los errores siguen disminuyendo, pero el comportamiento de los selectores es análogo.

Como ya es habitual en este capítulo, se presenta en la Tabla 4.3 los valores de MISE obtenidos, y en la Figura 4.3 su representación gráfica con la correspondiente escala de grises. En dicha figura se aprecia como se mantiene un comportamiento parecido al que ya se explicaba en los resultados para tamaño muestral $n = 400$, aunque obviamente el rango de valores en los que varían los errores es muy inferior. Sí se observa que la dualidad entre validación cruzada y el autoconsistente está más clara, y se aprecia mejor la división entre modelos suaves y complejos.

	Pulg	**PI**	**AutCon**	**CV**
M1	67.3949	74.8761	0.2675	0.3120
M2	12.0410	12.8281	1.6068	0.5782
M3	10.1518	12.3404	1.1663	1.0088
M4	1.3518	1.4053	0.5115	0.4003
M5	2.3764	2.3988	0.7434	0.5189
M6	0.8979	0.9947	0.2206	0.3312
M7	4.9192	4.6376	0.4824	0.3956
M8	1.0255	1.4037	0.1675	0.2278
M9	1.0873	1.0047	0.1522	0.2199
M10	7.2566	11.6857	0.0858	0.1977
M11	0.7566	0.8684	0.0660	0.1803
M12	44.8100	0.4238	0.5035	0.4275
M13	19.3788	14.2228	0.5644	0.4628
M14	17.9535	15.4244	0.7822	0.6206
M15	0.1414	0.0711	0.0876	0.1889
M16	0.0379	0.0416	0.0453	0.1602

Tabla 4.3: MISE's ($\times 100$) para los selectores del Capítulo 3 y tamaño muestral $n = 1600$.

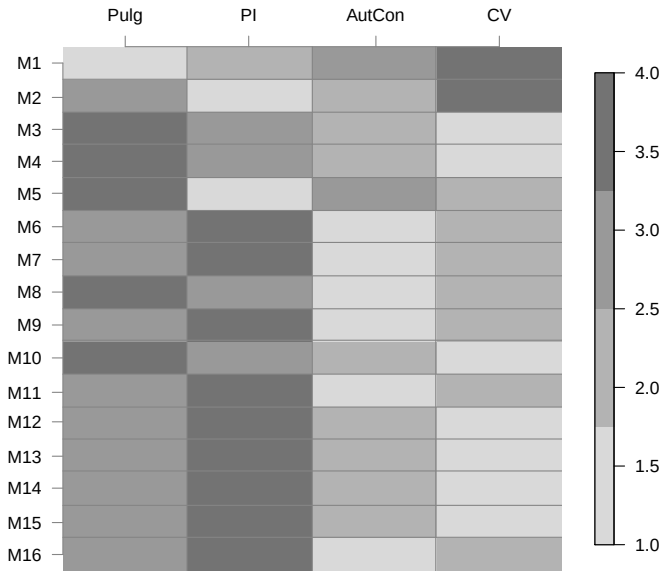


Figura 4.3: Representación mediante una escala de colores de la Tabla 4.3

Una vez realizada la comparación para los selectores ideados para el estimador (2.4), es interesante comparar estas propuestas con las existentes para el estimador tipo núcleo; pues el objetivo de la estimación es el mismo. En la Tabla 4.4, se presentan los resultados de dicho estudio para el tamaño muestral intermedio $n = 400$. Se han incluido en el análisis los selectores citados al comienzo de esta sección: $Pulg$, PI, $AutCon$, CV, $Pulg_{nc}$, SJ_{nc} y CV_{nc}.

	Pulg	**PI**	**AutCon**	**CV**	**$\mathbf{Pulg_{nc}}$**	**$\mathbf{SJ_{nc}}$**	**$\mathbf{CV_{nc}}$**
M1	0.1181	0.1350	0.1649	0.2649	0.2362	0.2363	0.2912
M2	0.3585	0.2431	0.3135	0.4188	0.3460	0.3572	0.4218
M3	19.2520	17.0046	2.5530	2.0964	9.4407	2.4270	1.8255
M4	20.4921	16.0999	2.0541	1.7218	4.6726	1.6222	1.6923
M5	55.3126	1.7242	1.8774	1.4838	2.0964	2.1515	2.5272
M6	1.0692	1.2810	0.2622	0.3713	0.2996	0.2992	0.3466
M7	9.0014	12.0131	0.3333	0.4524	1.5690	0.4243	0.4944
M8	1.3069	1.3547	0.6004	0.5441	0.4642	0.4103	0.4629
M9	1.4126	1.8418	0.4159	0.5285	0.4436	0.3724	0.4130
M10	5.0790	4.9184	4.6389	1.0907	4.4542	3.8454	1.4233
M11	1.2140	1.4214	0.3959	0.4957	0.4363	0.4360	0.4882
M12	2.4824	2.5349	2.3288	1.3096	1.9540	1.5627	1.1763
M13	1.7153	1.9416	0.7130	0.8524	0.7657	0.7343	0.8127
M14	10.9027	12.4751	2.4491	2.2004	5.9761	2.4182	1.7723
M15	12.3286	12.9283	2.3202	1.7361	7.9348	2.3727	1.7222
M16	70.0568	75.1657	0.8594	1.0290	49.4102	1.5056	1.5419

Tabla 4.4: MISE's ($\times 100$) para los selectores del Capítulo 3 y los más habituales para el estimador tipo núcleo con tamaño muestral $n = 400$.

Se ha decidido presentar directamente los resultados para $n = 400$, omitiendo tamaños muestrales inferiores, para obtener resultados fiables y consistentes. Procedamos ahora con el análisis de los valores obtenidos.

La regla del pulgar, la plug-in y la de Silverman presentan valores muy elevados en comparación con los otros selectores, salvo en los modelos M1 y M2. Este hecho podría distorsionar la comparación, así que a partir de ahora vamos a suprimirlos del estudio pues no son competitivos y no aportarían información alguna, permitiendo así un análisis más claro de los restantes.

Al igual que al analizar los selectores del Capítulo 3, es interesante poder ver la gráfica en la gama de grises para que sea más fácil la interpretación. Aunque a diferencia de lo que se ha hecho hasta ahora, para no perder el valor cuantitativo, se realiza un test de igualdad de medias, tomando en cada modelo como datos los errores proporcionados para cada uno de los métodos en las $B = 500$ muestras generadas. De esta manera, podemos apreciar de manera visual tanto

el error del método como su posición en comparación con los restantes sin necesidad de tener que analizar con detalle los valores numéricos.

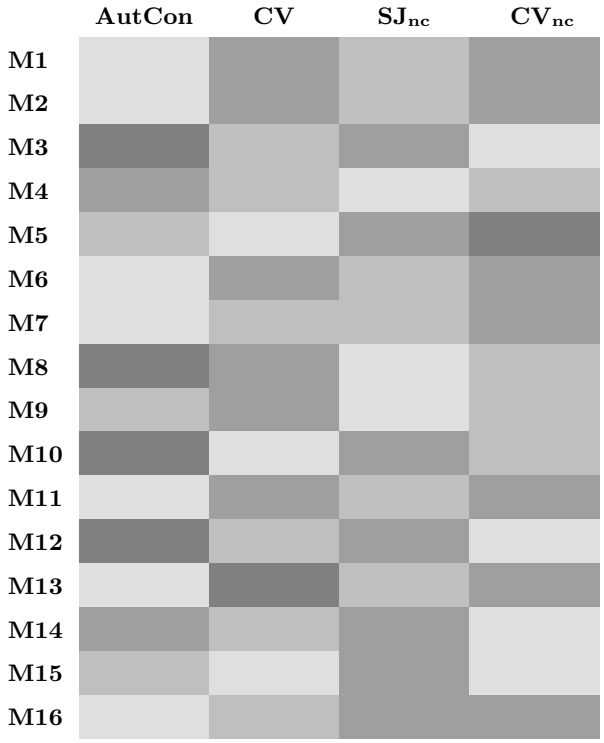


Figura 4.4: Representación en esacla de grises (más claro el menor valor) de los MISE para los selectores considerados sobre los modelos de Marron y Wand (1992).

A la vista de la Figura 4.4 no podemos extraer conclusiones globales sobre los selectores, puesto que en principio no siguen ningún patrón de comportamiento. Lo único que se podría decir es que sorprende que el autoconsistente lo hace muy mal en varios de los modelos considerados, y por lo tanto no parece un selector especialmente prometedor, a pesar de que la tasa de convergencia era teóricamente superior a la de los selectores para el estimador tipo núcleo. Al analizar la Figura 4.4 con un poco más de detalle, se aprecia que los modelos en los que se acepta la hipótesis de igualdad de medias dos a dos para un par de selectores, suele ser bien $AutCon$ y SJ_{nc} o bien CV y CV_{nc}, lo que corrobora el hecho anteriormente citado de la similitud en el comportamiento de esos pares de procedimientos.

Otro aspecto importante en la comparación de los selectores, además de estudiar el valor medio del error, que es lo que básicamente se viene haciendo hasta ahora con el MISE, es estudiar el conjunto de los ISE para cada modelo. De esta manera se puede analizar la variabilidad y la posición relativa de la distribución para cada selector.

A continuación se presenta el gráfico de cajas de los ISE para los selectores $AutCon$, CV, SJ_{nc} y CV_{nc} que son los más competitivos. Se ha decidido suprimir los atípicos en la representación ya que en el análisis que se va a llevar a cabo no tienen relevancia y dificultan la visualización por una cuestión de escala.

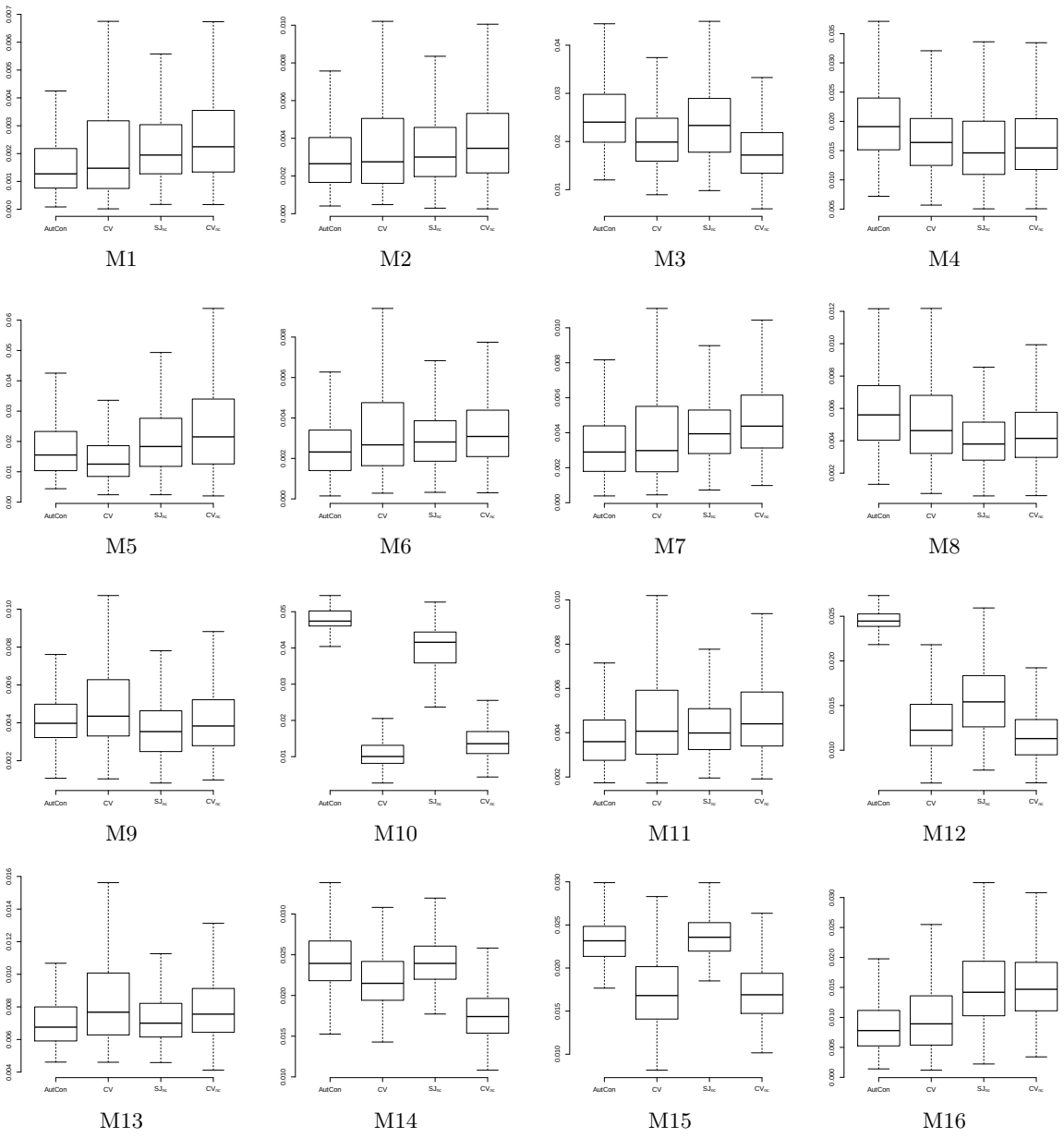


Figura 4.5: Boxplots para los ISE de los modelos de Marron y Wand (1992) con los selectores $AutCon$, CV, SJ_{nc} y CV_{nc} para tamaño muestral $n = 400$.

Se comenzará este análisis de manera individual con cada uno de los selectores, para posteriormente intentar identificar si existe alguna característica común entre ellos.

El selector autoconsistente tiene comportamientos extremos, es decir, si funciona correctamente lo hace muy bien, pero en varios de los modelos proporciona valores muy elevados que se escapan del rango en el que varían los restantes ejemplos. Es importante notar que en general su variabilidad es inferior a la de las técnicas de validación cruzada y muy similar a la regla de Sheather y Jones (1991).

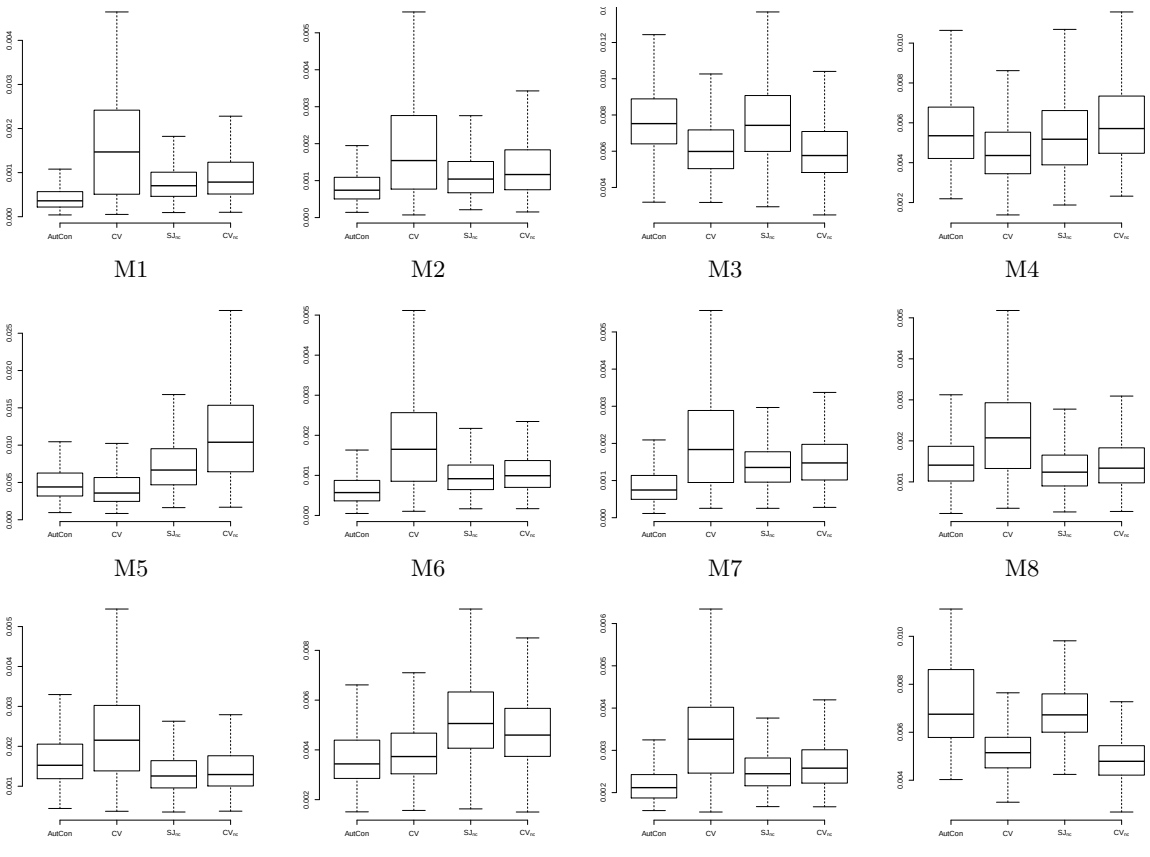
El método CV presenta un comportamiento bastante aceptable. Además, en ninguno de los modelos considerados en el estudio se aprecian valores de este selector que se desvíen de los obtenidos por los otros métodos, aunque como ya se ha dicho, tiene mucha variabilidad.

El selector propuesto por Sheather y Jones (1991), presenta en general un buen comportamiento, aunque existen algunos modelos como el M10 o M12 para los que alcanza valores extremadamente elevados.

Por último, resta estudiar el comportamiento de la regla de Bowman (1984), que en general se asimila mucho al selector CV, aunque cabría destacar su mejor funcionamiento en los modelos M12 y M14, especialmente en este último dista mucho de los otros procedimientos analizados. Quizás, el único punto débil que presenta es el rango de variabilidad, que como se sabe es una característica típica de las técnicas de validación cruzada.

Si hubiese que decantarse por uno de los procedimientos en base a estos resultados, sería sin lugar a dudas el CV, pues aunque posee la desventaja de la variabilidad, tenemos garantizado que no se tendrían valores extremos en ningún caso, y que por tanto, nunca se obtendría un estimador desastrosamente malo.

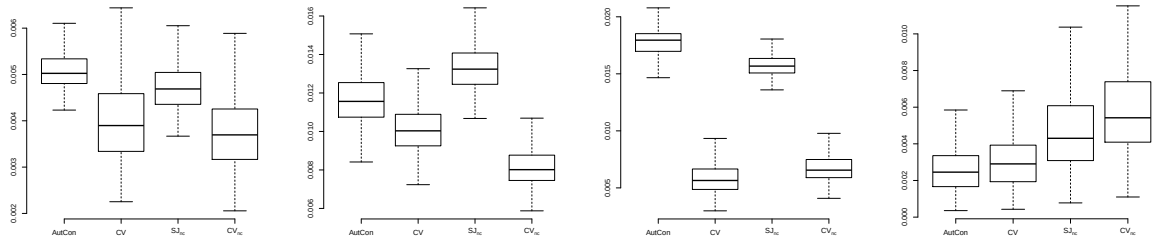
Se presentan también los diagramas de cajas anteriores para tamaño $n = 1600$, pues en algunos modelos hay diferencias bastante marcadas. Por ejemplo, el selector autoconsistente presenta una mejoría muy notable (relativa a los restantes métodos) en los modelos M10 y M12, aunque su comportamiento para los modelos M14 y M15 sigue siendo muy deficiente e incluso parece tener una mayor variabilidad. En los procedimientos que emplean técnicas de validación cruzada, se aprecia un incremento general de la variabilidad, algo que no ocurre con la regla de Sheather y Jones (1991).



M1 M2 M3 M4

M5 M6 M7 M8

M9 M10 M11 M12



M13 M14 M15 M16

Figura 4.6: Boxplots para los ISE de los modelos de Marron y Wand (1992) con los selectores $AutCon$, CV, SJ_{nc} y CV_{nc} para tamaño muestral $n = 1600$.

Estos hechos nos invitan a pensar que los criterios de validación cruzada serían una muy buena opción siempre y cuando pudiesen ser modificados para reducir su variabilidad.

Capítulo 5

Aplicación a datos reales

5.1. Presentación del conjunto de datos

En este capítulo se van a aplicar las técnicas presentadas a lo largo de la memoria a un conjunto de datos recogidos por el Instituto Gallego de Estadística (IGE) en el marco de la Encuesta de Condiciones de Vida de las Familias (ECV). Esta encuesta se lleva a cabo con periodicidad anual desde 1999, aunque en el año 2003 sufrió una leve modificación debida a un cambio en la metodología.

El objetivo de la ECV es obtener información socioeconómica sobre los hogares gallegos, por lo que se recogen multitud de datos demográficos (número de miembros del hogar, edad, sexo, nacionalidad, nivel de estudios ...) de cada uno de los miembros de la unidad familiar así como de tipo económico (ingresos, bienes e inmuebles, gastos ...).

El conjunto de datos seleccionado, que se ha extraído de la web del IGE[1], es la renta por hogar de 9216 familias de Galicia en el año 2011. Es interesante estimar la función de densidad de este conjunto de datos, ya que $\int_a^b f$ representa la proporción de hogares cuya renta está comprendida entre los valores a y b, lo cual constituye una información muy valiosa.

Las variables ingresos, gastos, renta ... han sido objeto de múltiples estudios a lo largo de los años, y se han intentado postular diversos modelos paramétricos que se adecuasen a los datos existentes. Uno de los modelos más empleados es el lognormal. Esto supone que el logaritmo de los datos de la renta sigue una distribución gaussiana de parámetros desconocidos, pero estimables por máxima verosimilitud o algún otro procedimiento válido.

Se va a asumir como cierto dicho modelo en nuestro conjunto de datos. Para realizar el ajuste se han estimado por máxima verosimilitud, la media y la desviación típica de la variable renta por hogar de las familias de Galicia en 2011 que denotaremos por X, obteniendo $\bar{X} = 23030.35$ y $\hat{\sigma}(X) = 15161.34$. Se puede ver a continuación la representación de esta estimación de la densidad.

[1] www.ige.eu

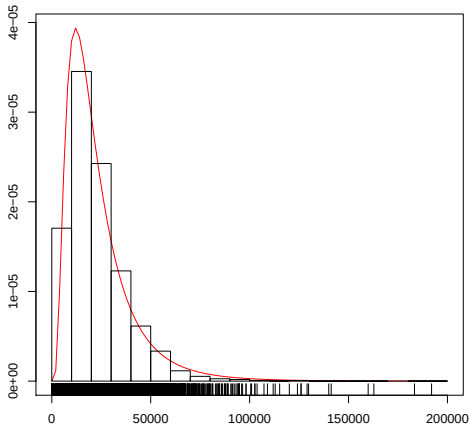


Figura 5.1: Representación de la densidad del modelo lognormal para los datos de renta de las familias gallegas en 2011 estimando los parámetros por máxima verosimilitud, junto con el histograma.

A la vista de los datos, el modelo lognormal no parece una hipótesis disparatada, aunque podrían existir otros modelos distintos del propuesto que aparentemente también proporcionasen un buen ajuste. Por tanto la aplicación de técnicas paramétricas posee, además del problema de la selección del modelo, la comprobación de la idoneidad del mismo.

La inferencia no paramérica permite resolver este tipo de cuestiones, tanto estimar la distribución de los datos sin ninguna hipótesis previa, como determinar si un modelo paramétrico postulado podría ser o no correcto. Es especialmente útil cuando no se dispone de una clara interpretación de los datos que permita proponer *a priori* un modelo coherente.

Para comenzar a trabajar con la muestra es necesario hacer ciertas modificaciones. Por una parte eliminaremos aquellos valores que sean 0, es decir, hogares en los que se supone la renta nula. Por otra parte, y siguiendo las inidicaciones de la literatura económica nos quedaremos sólo con el 97 % de la muestra, suprimiendo ese 3 % de la cola derecha, esto es, los valores de renta más altos. Este corte es relevante, pues de hecho, al suprimir esos datos se obtiene un valor máximo de la muestra notablemente inferior al incial.

Una vez hecho esto y dado que se postulaba un modelo lognormal, se ha aplicado el logaritmo a la variable de interés, de manera que si dicha hipótesis fuese cierta debería de seguir un modelo normal, que estamos más acostumbrados a ver e identificar.

Se comenzará presentando en la Figura 5.2 el ajuste normal de los datos transformados, para lo que se han estimado los parámetros de nuevo por máxima verosimilitud, obteniendo una media muestral de 9.804841 y desviación típica 0.6389567.

En la Figura 5.2 puede verse que a diferencia de lo que ocurre en la visualización del modelo lognormal, parece que se intuye que la hipótesis paramétrica que se ha asumido no es correcta. Si se observa la representación de los datos sobre el eje de abscisas, se aprecia como el máximo de los datos es bastante más pequeño de lo que cabría esperar para una distribución gaussiana, pues no hay datos en la cola derecha y quizás demasiados en la izquierda.

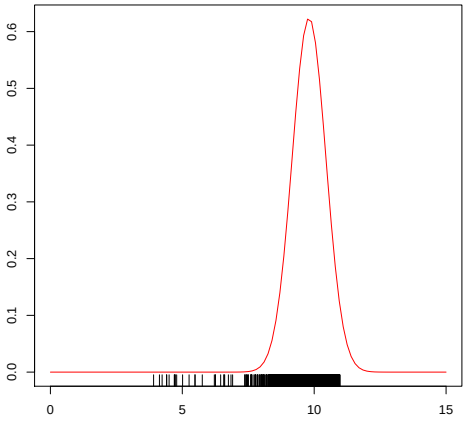
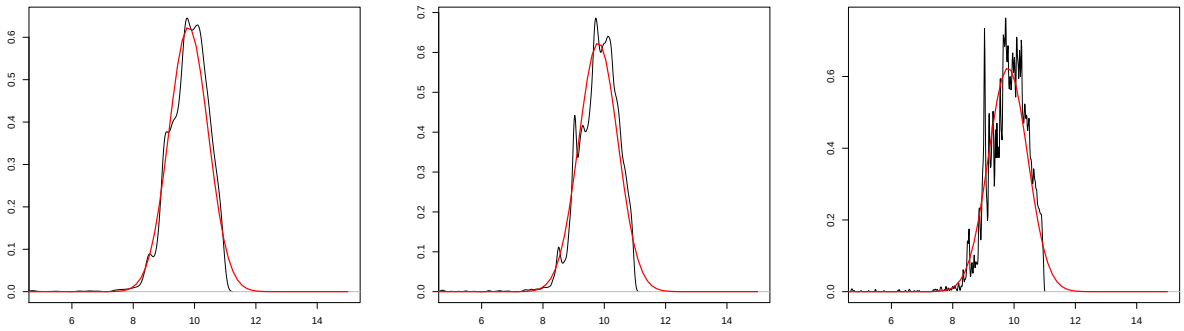


Figura 5.2: Representación de la densidad del modelo lognormal para los datos de renta de las familias gallegas en 2011, estimando los parámetros por máxima verosimilitud.

5.2. Aplicación de las técnicas

Se procederá a continuación con la aplicación de las técnicas no paramétricas descritas en el trabajo. En primer lugar se presentan las estimaciones obtenidas con (2.1) para los tres selectores de ventana con los que se ha trabajado en esta memoria, junto con el resultado de asumir normalidad de la variable transformada mediante logaritmo.



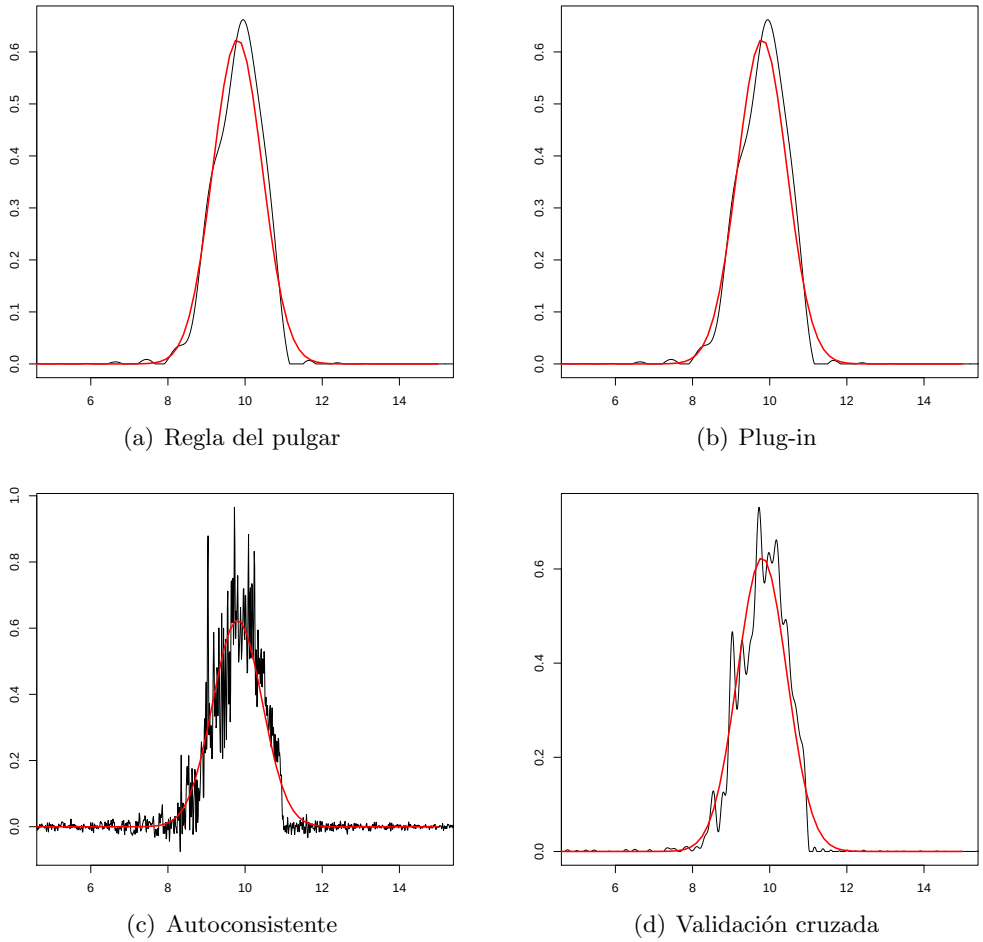
(a) Silverman (1986) (b) Sheather y Jones (1991) (c) Bowman (1984)

Figura 5.3: Representación del estimador tipo núcleo de la densidad con tres selectores diferentes para la variable de renta de los hogares gallegos en 2011 transformada mediante un logaritmo (línea negra), junto con el modelo paramétrico normal (línea roja).

En la Figura 5.3 se presentan las tres estimaciones resultantes. La primera de ellas es la más suave de todas, pero aún así capta cuatro modas que no se corresponderían con un modelo normal, especialmente las dos que se encuentran en torno al punto de abscisa diez. El siguiente selector genera una estimación bastante similar aunque con algún apuntamiento más. Y por último el selector de validación cruzada (Bowman, 1984) proporciona claramente una estimación infrasuavizada con irregularidades no esperadas para la variable de estudio.

Salvo para el selector propuesto en Silverman (1986), el modelo normal no resulta concordante con las estimaciones no paramétricas, por lo que sospecharíamos de dicha hipótesis.

Se incluyen en la Figura 5.4 las estimaciones para estos mismos datos transformados, con los selectores del estimador (2.4) presentados en el Capítulo 3 de esta memoria. Al igual que para el estimador tipo núcleo, también se ha añadido a cada uno de ellos el modelo normal con parámetros estimados por máxima verosimilitud.



(a) Regla del pulgar (b) Plug-in

(c) Autoconsistente (d) Validación cruzada

Figura 5.4: Representación del estimador (2.4) con los cuatro selectores presentados en el Capítulo 3 para la variable de renta de los hogares gallegos en 2011 transformada mediante un logaritmo (línea negra) junto con el modelo normal de parámetros estimados (línea roja).

De manera similar a lo que ocurría en el estimador tipo núcleo hay casos de estimaciones muy suaves, como son la regla del pulgar y la plug-in; y otras con mucho ruido como el autoconsistente. Es curioso que en este caso, la técnica de validación cruzada proporciona un estimador mucho mejor que en el caso del tipo núcleo, de hecho se asemeja a la estimación obtenida empleando Sheather y Jones (1991), aunque identifica alguna moda a mayores y nótese que el grado de rugosidad es menor que con el autoconsistente.

Está claro que el selector autoconsistente y validación cruzada (Bowman, 1984) generan estimaciones que distan mucho de la normal. Si se observan los dos primeros casos, a pesar de la suavidad de la estimación se ve como la cola derecha de la normal es más pesada que la de la estimación. Esto ya lo intuíamos en la Figura 5.2, en la que se aprecia como claramente nuestro conjunto de datos tiene una cota superior más pequeña que lo deseable bajo normalidad.

5.3. Resultados

Dado que se está en un contexto real, no se dispone del modelo teórico que siguen los datos. Por consiguiente no se puede calcular el error cometido en la estimación por cada uno de los métodos de selección del correspondiente parámetro (ya sea la ventana h o el núcleo K).

A la vista de las estimaciones, se puede decir que el modelo que siguen los datos no es excesivamente complejo, pues aunque podría presenta varias modas, estas no son demasiado pronunciadas y van en consonancia con los valores adyacentes. El problema es que esas modas hacen que el modelo no sea lo suficientemente sencillo como para que los procedimientos más simples como Silverman (1986), regla del pulgar o plug-in funcionen correctamente. Pues recordemos que, tal y como se veía en los resultados del estudio de simulación presentados en el Capítulo 4, estos selectores funcionan bien en general para modelos muy suaves.

Parece claro que también destacan los casos de infrasuavizado muy marcado, derivados del selector de validación cruzada (Bowman, 1984) y del autoconsistente. En base a los ya citados resultados recogidos en el Capítulo 4, sí cabría quizás esperar este comportamiento por parte de las técnicas de validación cruzada, pero no del selector autoconsistente, que parecía funcionar bien en modelos de complejidad intermedia.

Quedan el selector de Sheather y Jones (1991) y la propuesta de validación cruzada expuesta en el Capítulo 3. A simple vista estas dos estimaciones son muy parecidas, tal y como se puede apreciar en la Figura 5.5, aunque si bien es cierto que el estimador derivado de validación cruzada con núcleo variable presenta más oscilaciones.

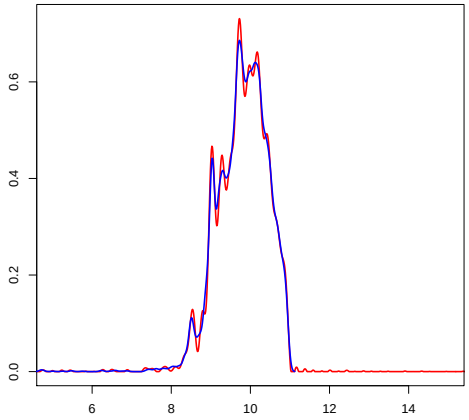


Figura 5.5: Representación conjunta de la estimación de la densidad para los datos de renta de los hogares gallegos en 2011 empleando el selector de Sheather y Jones (1991) (línea azul) y la validación cruzada presentada en el Capítulo 3 (línea roja).

Para decantarse por uno u otro procedimiento en este conjunto de datos, sería necesario realizar alguna prueba a mayores que permita conocer con más detalle las caracterísitcas de la variable en cuestión y así poder decidir cuales de las modas identificadas por el estimador son reales. De todos modos, lo que sí parece claro es que existen motivos para dudar del modelo lognormal para la variable renta de los hogares gallegos en 2011, pues un factor común a todos los selectores, y por tanto se podría asumir como una característica verdadera de la variable, es la bimodalidad presente en los datos, que no se incluye en dicho modelo paramétrico.

Apéndice A

Densidades de Marron y Wand

Los 15 modelos presentados en Marron y Wand (1992) son un conjunto de mixturas de normales que engloban un amplio abanico en lo que a características de las densidades se refiere. Esta familia, se ha convertido con los años, en un elemento fundamental para contrastar cualquier procedimiento nuevo en materia de estimación de la densidad. Se presentan a continuación las expresiones formales para cada uno de ellos. Nótese que el modelo M16 no aparece en el artículo, ya que fue definido posteriormente, pero se ha decidido incluirlo en este apéndice pues aparece en la librería `nor1mix` de R con la que se ha trabajado a lo largo de este proyecto.

Nombre densidad	Expresión($f = \sum_{j=1}^{p} \omega_j N(\mu_j, \sigma_j)$)
M1 - Gaussian	$N(0,1)$
M2 - Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \frac{2}{3}\right) + \frac{3}{5}N\left(\frac{13}{12}, \frac{5}{9}\right)$
M3 - Strongly skewed	$\sum_{l=0}^{7} \frac{1}{8}N\left(3\{(\frac{2}{3})^l - 1\}, (\frac{2}{3})^l\right)$
M4 - Kurtotic unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0, \frac{1}{10})$
M5 - Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N\left(0, \frac{1}{10}\right)$
M6 - Bimodal	$\frac{1}{2}N\left(-1, \frac{2}{3}\right) + \frac{1}{2}N\left(1, \frac{2}{3}\right)$
M7 - Separated bimodal	$\frac{1}{2}N\left(\frac{-3}{2}, \frac{1}{2}\right) + \frac{1}{2}N\left(\frac{3}{2}, \frac{1}{2}\right)$
M8 - Skewed bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N\left(\frac{3}{2}, \frac{1}{3}\right)$
M9 - Trimodal	$\frac{9}{20}N\left(\frac{-6}{5}, \frac{3}{5}\right) + \frac{9}{20}N\left(\frac{6}{5}, \frac{3}{5}\right) + \frac{1}{10}N\left(0, \frac{1}{4}\right)$
M10 - Claw	$\frac{1}{2}N(0,1) + \sum_{l=0}^{4} \frac{1}{10}N\left(\frac{l}{2} - 1, \frac{1}{10}\right)$
M11 - Double claw	$\frac{49}{100}N\left(-1, \frac{2}{3}\right) + \frac{49}{100}N\left(1, \frac{2}{3}\right) + \sum_{l=0}^{6} \frac{1}{350}N\left(\frac{l-3}{2}, \frac{1}{100}\right)$
M12 - Asymmetric claw	$\frac{1}{2}N(0,1) + \sum_{l=-2}^{2} \frac{2^{1-l}}{31}N\left(l + \frac{1}{2}, \frac{2^{-l}}{10}\right)$
M13 - Asymmetric double claw	$\sum_{l=0}^{1} \frac{46}{100}N\left(2l-1, \frac{2}{3}\right) + \sum_{l=1}^{3} \frac{1}{300}N\left(\frac{-l}{2}, \frac{1}{100}\right) + \sum_{l=1}^{3} \frac{7}{300}N\left(\frac{l}{2}, \frac{7}{100}\right)$
M14 - Smooth comb	$\sum_{l=0}^{5} \frac{2^{5-l}}{63}N\left(\frac{65-96(\frac{1}{2})^l}{21}, \frac{\frac{32}{63}}{2^l}\right)$
M15 - Discrete comb	$\sum_{l=0}^{2} \frac{2}{7}N\left(\frac{12l-5}{7}, \frac{2}{7}\right) + \sum_{l=8}^{10} \frac{1}{21}N\left(\frac{2l}{7}, \frac{1}{21}\right)$
M16 - Distant bimodal	$\frac{1}{2}N\left(\frac{-5}{2}, \frac{1}{6}\right) + \frac{1}{2}N\left(\frac{5}{2}, \frac{1}{6}\right)$

Tabla A.1: Expresiones de las densidades de Marron y Wand (1992) de la forma $f = \sum_{j=1}^{p} \omega_j N(\mu_j, \sigma_j)$, siendo $N(\mu_j, \sigma_j)$ una densidad normal de media μ_j y desviación típica σ_j.

Apéndice B

Cálculos detallados del MISE

En el Capítulo 2 de este trabajo, donde se presenta el estimador (2.4) y el (2.1), se emplean unos casos particulares de éste último para realizar la comparación. Dicho análisis se hace en términos de MISE para mixturas de normales, y aunque en esa parte de la memoria aparece únicamente la fórmula resultante, se presentan a continuación todos los cálculos necesarios para la obtención del error exacto y la ventana óptima.

Se comenzará con los cálculos para f_{nhG}, aplicando la fórmula de descomposición del MISE de un estimador como la integral de la suma del sesgo al cuadrado y de la varianza:

$$MISE(\hat{f}_{nhG}) = \int \text{Sesgo}^2\left(\hat{f}_{nhG}(x)\right) dx + \int \text{Var}\left(\hat{f}_{nhG}(x)\right) dx. \tag{B.1}$$

Para la obtención de (B.1), se calcularán por separado cada uno de los sumandos:

$$\begin{aligned}
\text{Sesgo}\left(\hat{f}_{nhG}(x)\right) &= \mathbb{E}\left[\hat{f}_{nhG}(x)\right] - f(x) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} G_h\left(x - X_i\right)\right] - f(x) \\
&= \mathbb{E}\left[G_h\left(x - X\right)\right] - f(x) = \int G_h(x-y)f(y)dy - f(x) \\
&= (G_h * f)(x) - f(x).
\end{aligned}$$

siendo G el núcleo gaussiano y como ya denotábamos en la memoria, G_h esa misma función reescalada.

Entonces,

$$\begin{aligned}
\int \text{Sesgo}^2\left(\hat{f}_{nhG}(x)\right) dx &= \int \left((G_h * f)(x) - f(x)\right)^2 dx \\
&= \int (G_h * f)^2(x)dx + \int f^2(x)dx - 2\int (G_h * f)(x)f(x)dx.
\end{aligned}$$

Por otra parte,

$$\begin{aligned}\operatorname{Var}\left(\hat{f}_{nhG}(x)\right) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} G_h(x-X_i)\right) = \frac{1}{n}\operatorname{Var}\left(G_h(x-X)\right)\\ &= \frac{1}{n}\left[\mathbb{E}[G_h^2(x-X)] - \mathbb{E}^2[G_h(x-X)]\right]\\ &= \frac{1}{n}\left[\int G_h^2(x-y)f(y)dy - \left(\int G_h(x-y)f(y)dy\right)^2\right]\\ &= \frac{1}{n}\left[\left(G_h^2 * f\right)(x) - \left(G_h * f\right)^2(x)\right].\end{aligned}$$

Entonces,

$$\begin{aligned}\int \operatorname{Var}\left(\hat{f}_{nhG}(x)\right)dx &= \frac{1}{n}\int \left(G_h^2 * f\right)(x)dx - \frac{1}{n}\int \left(G_h * f\right)^2(x)dx\\ &= \frac{1}{nh}\int G^2(x)dx - \frac{1}{n}\int \left(G_h * f\right)^2(x)dx.\end{aligned}$$

Teniendo en cuenta los cálculos precedentes junto con (B.1), se tiene que:

$$\begin{aligned}\operatorname{MISE}\left(\hat{f}_{nhG}(x)\right) = \frac{1}{nh}\int G^2(x)dx + \left(1-\frac{1}{n}\right)\int \left(G_h * f\right)^2(x)dx -\\ - 2\int \left(G_h * f\right)(x)f(x)dx + \int f^2(x)dx.\end{aligned} \tag{B.2}$$

Como en el caso que se estaba tratando se asumía que el modelo teórico eran mixturas de normales, entonces recuérdese que se puede expresar $f = \sum_{j=1}^{p}\omega_j f_{\mu_j\sigma_j} \equiv \sum_{j=1}^{p}\omega_j f_j$, con $\sum_{j=1}^{p}\omega_j = 1$, donde ω_j son los pesos y f_j denota la función de densidad de una normal de media μ_j y desviación típica σ_j. A continuación se calculan por separado cada uno de los sumandos de (B.2) teniendo en cuenta la expresión de la densidad y que G denota una normal estándar.

- $\displaystyle\int G^2(x)dx = \int G(x)G(x)dx = \int f_{0,1}(x)f_{0,1}(x) \overset{(*)}{=} f_{0,\sqrt{2}}(0) = \frac{1}{2\sqrt{\pi}}.$

- $$\begin{aligned}\int (G_h * f)(x)dx &= \int G_h(x-y)f(y)dy = \int \frac{1}{h}G\left(\frac{x-y}{h}\right)f(y)dy\\ &= \frac{1}{h}\int \frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-y}{h}\right)^2}\sum_{j=1}^{p}\omega_j f_j(y)dy\\ &= \sum_{j=1}^{p}\omega_j \int f_{0,h}(y-x)f_{0,\sigma_j}(y-\mu_j)dy\\ &\overset{(*)}{=} \sum_{j=1}^{p}\omega_j f_{0,(h^2+\sigma_j^2)^{1/2}}(x-\mu_j).\end{aligned}$$

- $$\int (G_h * f)^2(x)dx = \int (G_h * f)(x)(G_h * f)(x)dx$$
$$= \int \sum_{j=1}^{p} \omega_j f_{0,(h^2+\sigma_j^2)^{1/2}}(x-\mu_j) \sum_{l=1}^{p} \omega_l f_{0,(h^2+\sigma_l^2)^{1/2}}(x-\mu_l)dx$$
$$= \sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l \int f_{0,(h^2+\sigma_j^2)^{1/2}}(x-\mu_j) f_{0,(h^2+\sigma_l^2)^{1/2}}(x-\mu_l)dx$$
$$\overset{(*)}{=} \sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l f_{0,(2h^2+\sigma_j^2+\sigma_l^2)^{1/2}}(\mu_j-\mu_l).$$

- $$\int (G_h * f)(x) f(x)dx = \int \sum_{j=1}^{p} \omega_j f_{0,(h^2+\sigma_j^2)^{1/2}}(x-\mu_j) \sum_{l=1}^{p} \omega_l f_l(x)dx$$
$$= \sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l \int f_{0,(h^2+\sigma_j^2)^{1/2}}(x-\mu_j) f_{0,\sigma_l}(x-\mu_l)dx$$
$$\overset{(*)}{=} \sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l f_{0,(h^2+\sigma_j^2+\sigma_l^2)^{1/2}}(\mu_j-\mu_l).$$

- $$\int f^2(x)dx = \int \left(\sum_{j=1}^{p} \omega_j f_j\right)^2 dx = \int \sum_{j=1}^{p} \omega_j f_j \sum_{l=1}^{p} \omega_l f_l dx$$
$$= \sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l \int f_{0,\sigma_j}(x-\mu_j) f_{0,\sigma_l}(x-\mu_l)dx$$
$$= \sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l f_{0,(\sigma_j+\sigma_l)^{1/2}}(\mu_j-\mu_l).$$

$(*)$ Se aplica la siguiente igualdad algebraica que puede ser consultada en Marron y Wand (1992)[Cap.2]:

$$\int f_{\mu,\sigma}(x) f_{\mu',\sigma'}(x)dx = \int f_{0,\sigma}(x-\mu) f_{0,\sigma'}(x-\mu') = f_{0,(\sigma+\sigma')^{1/2}}(\mu'-\mu) = f_{\mu'-\mu,(\sigma+\sigma')^{1/2}}(0)$$

Sustituyendo estas expresiones en (B.2) se obtiene la expresión exacta del MISE del estimador $\hat{f}_{nhG}$ para mixturas de normales:

$$\begin{aligned}\text{MISE}\left(\hat{f}_{nhG}(x)\right) = \frac{1}{nh}\frac{1}{2\sqrt{\pi}} &+ \left(1-\frac{1}{n}\right)\sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l f_{0,(2h^2+\sigma_j^2+\sigma_l^2)^{1/2}}(\mu_j-\mu_l) \\ &- 2\sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l f_{0,(h^2+\sigma_j^2+\sigma_l^2)^{1/2}}(\mu_j-\mu_l) \\ &+ \sum_{j=1}^{p}\sum_{l=1}^{p} \omega_j\omega_l f_{0,(\sigma_j+\sigma_l)^{1/2}}(\mu_j-\mu_l). \end{aligned} \tag{B.3}$$

Se han realizado los cálculos análogos para el estimador $\hat{f}_{nhS}$, en el que el núcleo es el Sinc, $S(x) = \frac{\sin \pi x}{x}$. Para este caso se emplea el paso al dominio de frecuencias mediante la igualdad

de Parseval y las expresiones de la función característica del estimador y de la densidad teórica.

Como ya se ha visto en esta memoria,

$$\varphi_{\hat{f}_{nhS}}(t) = \varphi_n(t)\varphi_S(t) = \varphi_n(t)I_{[-1/h,1/h]}(t), \tag{B.4}$$

donde recuérdese φ_S denota la función característica del Sinc.

El siguiente paso es obtener la expresión para la función característica de f, teniendo en cuenta que se trata de una mixtura de normales. Se emplean además las siguientes propiedades de las funciones características:

- $\varphi_{N(\mu,\sigma)}(t) = e^{it\mu}e^{-\frac{\sigma^2 t^2}{2}}$.
- $\varphi_{\alpha f}(t) = e^{it}\varphi_f(\alpha t)$ con $\alpha \in \mathbb{R}$.
- $\rho_f(t) = e^{-t^2\beta}$ con $\beta = \sigma_1^2\omega_1^2 + \ldots + \sigma_p^2\omega_p^2$.

Así,

$$\begin{aligned}\varphi_f(t) &= \varphi_{\sum_{j=1}^p \omega_j f_j} = \varphi_{\omega_1 f_1}(t) \cdot \ldots \cdot \varphi_{\omega_p f_p}(t) = e^{itp}\left(\varphi_{f_1}(\omega_1 t) \cdot \ldots \cdot \varphi_{f_p}(\omega_p t)\right) \\ &= e^{itp}e^{it(\mu_1\omega_1+\ldots+\mu_p\omega_p)}e^{-\frac{t^2}{2}(\omega_1^2\sigma_1^2+\ldots\omega_p^2\sigma_p^2)}.\end{aligned} \tag{B.5}$$

El cálculo del MISE es el siguiente:

$$\begin{aligned}\text{MISE}\left(\hat{f}_{nhS}(x)\right) &= \mathbb{E}\left[\text{ISE}\left(\hat{f}_{nhS}(x)\right)\right] = \mathbb{E}\left[\int\left(\hat{f}_{nhS}(x) - f(x)\right)^2 dx\right] = \mathbb{E}\left[||\hat{f}_{nhS} - f||_2^2\right] \\ &\overset{(*)}{=} \frac{1}{2\pi}\mathbb{E}\left[||\varphi_{\hat{f}_{nhS}} - \varphi_f||_2^2\right] = \frac{1}{2\pi}\mathbb{E}\int\left(\varphi_{\hat{f}_{nhS}}(t) - \varphi_f(t)\right)^2 dt \\ &\overset{(**)}{=} \frac{1}{2\pi}\int\left[\frac{1}{n}|I_{[-1/h,1/h]}|^2\left(1 - |\varphi_f(t)|^2\right) + |\varphi_f(t)|^2\left(|1 - I_{[-1/h,1/h]}|^2\right)\right]dt \\ &= \frac{1}{2\pi}\left(\int_{-1/h}^{1/h}\frac{1}{n}\left(1 - \rho_f(t)\right)dt + \int\rho_f(t)\left(1 - I_{[-1/h,1/h]}(t)\right)^2 dt\right) \\ &= \frac{1}{2\pi}\left(\frac{2}{nh} - \frac{2}{n}\int_0^{1/h}\rho_f(t)dt + \int_{-\infty}^{-1/h}\rho_f(t)dt + \int_{1/h}^{\infty}\rho_f(t)dt\right) \\ &= \frac{1}{2\pi}\left(\frac{2}{nh} - \frac{2}{n}\int_0^{1/h}\rho_f(t)dt + 2\int_{1/h}^{\infty}\rho_f(t)dt\right) \\ &= \frac{1}{nh\pi} - \frac{1}{n\pi}\int_0^{1/h}\rho_f(t)dt + \frac{1}{\pi}\int_{1/h}^{\infty}\rho_f(t)dt \\ &= \frac{1}{nh\pi} - \frac{1}{nh}\int_0^{1/h}\rho_f(t)dt\left(-\frac{1}{\pi}\int_0^{1/h}\rho_f(t)dt + \frac{1}{\pi}\int_0^{1/h}\rho_f(t)dt\right) + \frac{1}{\pi}\int_{1/h}^{\infty}\rho_f(t)dt \\ &= \frac{1}{nh\pi} - \frac{n+1}{n}\int_0^{1/h}\rho_f(t)dt + \frac{1}{\pi}\int_0^{\infty}\rho_f(t)dt,\end{aligned} \tag{B.6}$$

$(*)$ igualdad de Parseval.
$(**)$ expresión (2.7).

Así, en (B.3) y (B.6) se han obtenido las expresiones exactas y explícitas del MISE para los estimadores $\hat{f}_{nhG}$ y $\hat{f}_{nhS}$ que se corresponden con las ecuaciones (2.10) y (2.11) y que se emplean en el estudio comparativo del Capítulo 2 de esta memoria.

Bibliografía

Bernacchia, A., y Pigolotti, S. (2011a). *Código del selector autoconsistente.* Descargado de `http://abernacchi.user.jacobs-university.de/software/SCM.R` (Consultada el 14 de diciemre de 2013)

Bernacchia, A., y Pigolotti, S. (2011b). Self-consistent method for density estimation. *Journal of the Royal Statistical Society: Series B*, *73*(3), 407–422.

Billingsley, P. (1995). *Probability and measure.* Wiley.

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

Chacón Durán, J. (2004). *Estimación de densidades: algunos resultados exactos y asintóticos.* Tesis Doctoral no publicada, Universidad de Extremadura.

Chacón Durán, J. (2010). *Aplicación de la teoría de u-estadísticos al cálculo del estimador plug-in.* (Documento sin publicar)

Davis, K. B. (1977). Mean integrated square error properties of density estimates. *Annals of Statistics*, *5*(3), 530–535.

Glad, I. K., Hjort, N. L., y Ushakov, N. G. (2003). Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, *30*(2), 415–427.

Lee, J. (1990). *U-statistics: Theory and practice.* CRC Press.

Marron, J. S., y Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics*, *20*(2), 712–736.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, *33*(3), 1065–1076.

R Core Team. (2012). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de `http://www.R-project.org/` (ISBN 3-900051-07-0)

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 65–78.

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Wand, M. M. P., y Jones, M. C. (1995). *Kernel smoothing*. CRC Press.

Watson, G. S., y Leadbetter, M. R. (1963). On the estimation of the probability density. *Annals of Mathematical Statistics*, *34*, 480–491.