Trabajo Fin de Máster

Estimación tipo núcleo de la densidad para datos agrupados

Mateo Pérez Rodríguez

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación tipo núcleo da densidade para datos agrupados
Título en español: Estimación tipo núcleo de la densidad para datos agrupados
English title: Kernel density estimation for grouped data
Modalidad: A
Autor: Mateo Pérez Rodríguez, Universidade de Santiago de Compostela
Directores: Rosa María Crujeiras Casais, Universidade de Santiago de Compostela; Jose Ameijeiras Alonso, Universidade de Santiago de Compostela.
Breve resumen del trabajo: En este trabajo se considera el problema de estimación de la función de densidad en un contexto de datos agrupados. Para ello, se empleará una modificación del estimador tipo núcleo de Parzen-Rosenblatt, lo cual conducirá a la propuesta de diferentes selectores de ventana, que serán posteriormente evaluados a través de diversos estudios de simulación. Finalmente, se ilustrará el comportamiento de este estimador a través de su aplicación sobre diferentes bases de datos reales.
Recomendaciones: Conocimientos básicos de estimación no paramétrica de la densidad, con especial atención al problema de selección de ventana. Buen manejo de R.
Otras observaciones: Este TFM es una propuesta propiciada por el estudiante Mateo Pérez Rodríguez.

Doña Rosa M. Crujeiras Casais, profesora titular del área de estadística e investigación operativa (Departamento de Estadística, Análisis Matemático y Optimización) de la Universidade de Santiago de Compostela y don Jose Ameijeiras Alonso, profesor ayudante doctor del área de estadística e investigación operativa (Departamento de Estadística, Análisis Matemático y Optimización) de la Universidade de Santiago de Compostela informan que el Trabajo Fin de Máster titulado

Estimación tipo núcleo de la densidad para datos agrupados

fue realizado bajo su dirección por don Mateo Pérez Rodríguez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 31 de enero de 2022

La directora:

CRUJEIRAS CASAIS ROSA MARIA - 52931780V

Firmado digitalmente por CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-52931780V, givenName=ROSA MARIA, sn=CRUJEIRAS CASAIS, cn=CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Fecha: 2022.01.27 21:58:10 +01'00'

Doña Rosa María Crujeiras Casais

El director:

AMEIJEIRAS ALONSO JOSE - 39450236S

Firmado digitalmente por AMEIJEIRAS ALONSO JOSE - 39450236S
Fecha: 2022.01.27 19:47:35 +01'00'

Don Jose Ameijeiras Alonso

El autor:

PEREZ RODRIGUEZ, MATEO (FIRMA)

Firmado digitalmente por PEREZ RODRIGUEZ, MATEO (FIRMA)
Fecha: 2022.01.28 14:11:48 +01'00'

Don Mateo Pérez Rodríguez

Agradecimientos

En primer lugar, mostrar mi más sincero agradecimiento a mis dos tutores de TFM, Rosa M. Crujeiras y Jose Ameijeiras, que han sabido guiarme de la mejor manera posible a lo largo de este proceso, que constituye mi primer acercamiento al mundo de la investigación. Mi especial agradecimiento a ambos, no solo por su indiscutible dedicación, sino también por su cercanía, por todas las buenas palabras que han tenido siempre hacia mí, por valorar (y enseñarme a mí a hacerlo) mi trabajo y por mantenerme motivado hasta en los momentos en los que todo parecía ir en contra. Haber iniciado este pequeño proyecto con vosotros — hace ya unos cuantos meses atrás — fue, sin duda, la mejor decisión que pude tomar.

Agradecer también a la Universidad de Santiago de Compostela y, en particular, al Vicerrectorado de Investigación e Innovación — en colaboración con el Banco Santander — cuya financiación a través de la *Bolsa de iniciación á investigación para alumnos de máster 2020-2021* ha posibilitado el desarrollo de las investigaciones que han permitido elaborar este trabajo. Gracias por promover la investigación entre los jóvenes y por darnos una oportunidad para acceder a este apasionante mundo.

Por último, no quería olvidarme de agradecer también a mi familia, cuyo apoyo incondicional ha sido siempre el motor principal de mi desarrollo académico y personal. También a mis amigos — y en especial a Martín y Aroa — que habéis hecho que estos últimos seis años hayan sido tan especiales y que pueda recordarlos con tanta felicidad. Gracias a todos vosotros por ayudarme a cerrar, de la mejor manera posible, esta etapa de mi vida.

Índice general

Resumen

Resumen en español

En este trabajo se aborda el problema de estimar la función de densidad de una variable aleatoria en aquellos contextos en los que el evento de interés no es observado de forma directa y solo se sabe que ha ocurrido dentro de un intervalo. De esta forma, la estimación se llevará a cabo en base a muestras de datos agrupados, conformadas por intervalos de los que únicamente se conocen sus puntos medios y el número (o, en su defecto, proporción) de observaciones que se han recogido en cada uno de ellos. Este problema se enfocará desde una perspectiva no paramétrica, considerando un estimador tipo núcleo que se obtiene como una generalización del estimador clásico de Parzen-Rosenblatt a este contexto. Asimismo, se abordará el problema de selección de ventana — un problema que ya estaba presente en la construcción del estimador tipo núcleo clásico — a través de la revisión de algunos selectores propuestos en la literatura. También se propondrán otros nuevos que, tras ser implementados en el software estadístico `R`, serán evaluados mediante diversos estudios de simulación. Finalmente, se ilustrará el comportamiento de este estimador a través de su aplicación sobre diferentes bases de datos reales.

English abstract

This work addresses the problem of estimating the density function of a random variable when the event of interest is not directly observed and it is only known to have occurred within a certain interval. The estimation will be based on grouped data samples. These data samples consist of a collection of intervals for which just their midpoints and the number (or, alternatively, the proportion) of observations that have been collected inside are known. This problem will be approached from a nonparametric perspective, considering a kernel-type estimator that is obtained as a generalization of the classic Parzen-Rosenblatt estimator to this context. The bandwidth selection problem, which was already present in the construction of the classical kernel-type estimator, will be also addressed by revising some selectors proposed in the literature. New bandwidth selection procedures will be also proposed. These tools will be implemented in the statistical software R, and the proposals will be evaluated through various simulation studies. Finally, the behavior of this estimator will be illustrated by its application to different real databases.

Capítulo 1

Introducción

Dada una variable aleatoria real y continua X, su función de densidad $f : \mathbb{R} \longrightarrow \mathbb{R}$ se define como una función real de variable real no negativa, que integra uno sobre la recta real y tal que caracteriza la distribución de probabilidad de X de la siguiente forma,

$$\mathbb{P}\left(a \leq X \leq b\right) = \int_a^b f(x)\ \mathrm{d}x,$$

para cualquier $a, b \in \mathbb{R}$ tales que $a < b$. Dado que la función de densidad, desconocida en la práctica, caracteriza la distribución de la variable aleatoria de interés, parece natural pretender estimarla, puesto que ello permitiría analizar el comportamiento de la población en cuestión.

1.1. Estimación paramétrica de la densidad

La vía tradicional de abordar el problema de estimación de la densidad ha sido asumir que esta pertenece a una familia paramétrica de densidades, como puede ser la normal, gamma o exponencial, de manera que el objetivo reside en estimar los parámetros que la definen a partir de una muestra observada. En este contexto destaca la estimación por máxima verosimilitud — cuya definición y propiedades pueden consultarse en el Capítulo 5 de Rossi (2018) — consistente en seleccionar aquel valor del parámetro que asocie una mayor probabilidad a la ocurrencia de la muestra observada; y el método de momentos — empleado por primera vez en Pearson (1894) — que escoge el valor del estimador en base a la idea de que la media muestral se *parezca* a la media poblacional, que es una función del parámetro o parámetros de interés.

En el caso de que la suposición inicial de que la función de densidad siga un determinado modelo paramétrico sea cierta, la estimación realizada será adecuada y gozará además de muy buenas propiedades teóricas. A modo de ejemplo, es bien conocido que, bajo ciertas condiciones de regularidad, el estimador de máxima verosimilitud es un estimador consistente, equivariante, asintóticamente eficiente y con distribución asintóticamente normal, lo cual permite, entre otras cosas, construir intervalos de confianza para el parámetro objetivo de manera sencilla. Sin embargo, en el caso de que esta suposición inicial no sea cierta o, aún siéndolo, si la especificación de la familia paramétrica es incorrecta, entonces la estimación realizada puede conducir a conclusiones totalmente erróneas y contradictorias. Este es el caso de muchas situaciones reales, en las cuales no se conoce ninguna información externa a la muestra o bien existen dudas acerca de su validez. En estas situaciones se requiere de un procedimiento alternativo — y, preferiblemente, algo más flexible — de estimación, como puede ser el enfoque no paramétrico presentado a continuación.

1.2. Estimación no paramétrica de la densidad

Un enfoque alternativo al problema de estimación aquí presentado consiste en no imponer ninguna forma paramétrica a la función de densidad, permitiendo que esta adopte casi cualquier forma posible, exigiendo únicamente ciertas condiciones de regularidad que serán satisfechas en la mayoría de situaciones, como son la continuidad y diferenciabilidad de la misma. Esta es, precisamente, la idea en la que se fundamenta la llamada estimación no paramétrica de la densidad, cuyo origen se remonta a las investigaciones de Fix y Hodges (1951). En este contexto, la estimación de la función de densidad se realiza punto a punto utilizando únicamente la información que proporciona la muestra observada, permitiendo que sean los datos muestrales los que guíen el proceso de inferencia ("*dejar hablar a los datos por sí mismos*"). Es precisamente por este motivo por el cual este enfoque suele requerir de tamaños muestrales más elevados que la estimación paramétrica.

1.2.1. Estimador tipo núcleo de Parzen-Rosenblatt

Entre los métodos no paramétricos enfocados a la estimación de la densidad más conocidos destaca el estimador tipo núcleo, también conocido como estimador de Parzen-Rosenblatt (Parzen (1962) y Rosenblatt (1956)), por ser a quienes generalmente se les atribuye haberlo propuesto — de forma independiente — por primera vez.

Dada una función núcleo K y una ventana $h > 0$, el estimador de Parzen-Rosenblatt construido en base a una muestra aleatoria simple $X_1, \ldots, X_n$ se define como

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

donde $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ denota la función núcleo reescalada por h. De esta manera, este estimador incorpora el concepto de *proximidad*, ya que la estimación de la densidad en un punto se realiza en base a todos los datos muestrales, cada uno con un peso proporcional a su distancia respecto del punto de interés, ponderada por la función K_h. Es importante notar que la construcción del estimador (1.1) requiere de la elección previa de una función núcleo K y de un parámetro de ventana h.

Por un lado, el estimador tipo núcleo (1.1) hereda las propiedades de regularidad de la función núcleo escogida, por lo que, en general, se emplearán núcleos K que sean funciones de densidad continuas (y, preferiblemente, unimodales y simétricas respecto del origen). En este sentido, es conocido que el núcleo más eficiente es el de Epanechnikov (Epanechnikov (1969)). Sin embargo, su uso proporciona un estimador con derivada primera discontinua, haciendo que generalmente sea preferible escoger otro que presente mejores propiedades de regularidad, como es el caso del núcleo gaussiano (de clase $\mathcal{C}^\infty$). De todas formas, la elección de un núcleo óptimo es todavía un problema sin resolver — en la Sección 1.2.4 de Tsybakov (2009) se puede consultar una discusión más detallada de este problema — aunque se ha comprobado en numerosas ocasiones que la influencia de su elección en los resultados es pequeña. Una breve comparativa de la eficiencia de los principales núcleos se puede encontrar en la Sección 2.7 de Wand y Jones (1995).

Ahora bien, no sucede lo mismo con el parámetro ventana h, cuya elección juega un papel fundamental en los métodos de estimación no paramétrica de curvas. En este sentido, valores grandes de h producen estimadores sobresuavizados, con mucho sesgo y poca varianza, que ocultan aspectos importantes de la estructura de probabilidad subyacente. Por el contrario, valores pequeños de h proporcionan estimadores infrasuavizados, con poco sesgo pero mucha varianza, que hacen visibles comportamientos espúreos procedentes exclusivamente de la muestra observada. La elección adecuada de h es un problema clásico que se ha ido abordando mediante la propuesta de un buen número de métodos para su selección. De hecho, su selección óptima constituye también un problema sin resolver, no habiéndose probado todavía que alguno de los selectores propuestos sea el más competitivo en todos los escenarios.

Para evaluar el comportamiento del estimador (1.1) se pueden seguir dos enfoques distintos. Por un lado, es relevante examinar el error de manera local, esto es, el error que se produce al estimar la función de densidad en un punto concreto x, pero también resulta de interés conocer el error global que se comete en la estimación sobre todo el dominio de definición de la función de densidad. En el primer caso, es frecuente considerar como medida de error el Error Cuadrático Medio (MSE por sus siglas en inglés), definido como

$$\mathrm{MSE}(x) = \mathrm{MSE}\left(\hat{f}_n(x)\right) = \mathbb{E}\left[\left(\hat{f}_n(x) - f(x)\right)^2\right] = \mathrm{Sesgo}^2\left(\hat{f}_n(x)\right) + \mathrm{Var}\left(\hat{f}_n(x)\right).$$

Por otro lado, como medida del error global se suele considerar el Error Cuadrático Medio Integrado (MISE, por sus siglas en inglés) definido como la integral del error cuadrático medio sobre el dominio de definición de f,

$$\mathrm{MISE} = \mathrm{MISE}\left(\hat{f}_n\right) = \int \mathrm{MSE}\left(\hat{f}_n(x)\right)\, \mathrm{d}x.$$

Bajo ciertas condiciones de regularidad, en Wand y Jones (1995) se recogen las expresiones analíticas del error cuadrático medio (MSE) y del error cuadrático medio integrado (MISE) de $\hat{f}_n(x)$ y $\hat{f}_n$, respectivamente, dadas por

$$\mathrm{MSE}(x) = \frac{1}{4}\, h^4 \mu_2(K)^2 f''(x)^2 + \frac{1}{nh} R(K) f(x) + o\left((nh)^{-1} + h^4\right), \tag{1.2}$$

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

donde se ha empleado la notación $\mu_2(K) = \int x^2 K(x)\, dx$ y $R(\varphi) = \int \varphi^2(x)\, dx$, siendo φ una función tal que su cuadrado es integrable.

A la vista de la expresión (1.2), cuyo primer sumando se corresponde con el sesgo (al cuadrado) de $\hat{f}_n(x)$, es sencillo comprobar que el estimador de Parzen-Rosenblatt presenta un sesgo negativo en regiones cóncavas (esto es, en regiones donde $f''(x) < 0$, como son las modas de la densidad f) y un sesgo positivo en regiones convexas (esto es, en regiones donde $f''(x) > 0$, como son los valles y colas). De esta forma, cuanta mayor curvatura tenga la densidad teórica, más difícil será de estimar. Por otro lado, la expresión de la varianza — segundo sumando de (1.2) — parece indicar que esta será mayor en aquellos puntos cuya función de densidad asociada sea elevada.

Pues bien, la mayoría de selectores de ventana se aprovechan de estas medidas de error para elegir el parámetro de suavizado h, necesario para la construcción del estimador de Parzen-Rosenblatt. Algunos de ellos parten de la versión asintótica de (1.3) — denotada por AMISE — cuya minimización conduce a la ventana asintóticamente óptima

$$h_{\mathrm{AMISE}} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'') n}\right]^{\frac{1}{5}}, \tag{1.4}$$

y procederían estimando la cantidad $R(f'')$, que es desconocida en la práctica. Este es el fundamento de los selectores clásicos de la regla del pulgar (Silverman (1986)) y del selector plug-in en dos etapas (Sheather y Jones (1991)). Otros selectores, como son los de validación cruzada, parten directamente de aproximar medidas de error. En este sentido, el selector de validación cruzada insesgada (véase Rudemo (1982) y Bowman (1984)) intenta aproximar el MISE, sustituyendo las cantidades desconocidas por estimadores adecuados; mientras que el selector de validación cruzada sesgada (ver Scott y Terrell (1987)) se fundamenta en la aproximación del AMISE. Por último, otros selectores más recientes se basan en la llamada metodología bootstrap — propuesta originalmente por Efron (1979) — pretendiendo aproximar las medidas de error mediante remuestreo. Este tipo de procedimientos requieren de un mayor coste computacional, debido a su estrecha relación con el método Monte Carlo.

1.2.2. Estimación de la densidad en el contexto de datos agrupados

En el ámbito de las ciencias experimentales es frecuente que los datos de interés procedan de mediciones de variables continuas, como pueden ser la temperatura, el tiempo o el peso. Sin embargo, debido a diversos factores de imprecisión — o debido incluso a la propia limitación física de obtener medidas con precisión *infinita* — los verdaderos valores no son observables, existiendo siempre un grado de error o incertidumbre sobre ellos. En definitiva, de una u otra manera todas las variables continuas se miden de forma redondeada, pudiéndose asumir — en un sentido genérico — que todas las medidas se encuentran agrupadas. Parece razonable, pues, que esta incertidumbre sobre la medición pueda tener un impacto sobre la estimación de la función de densidad.

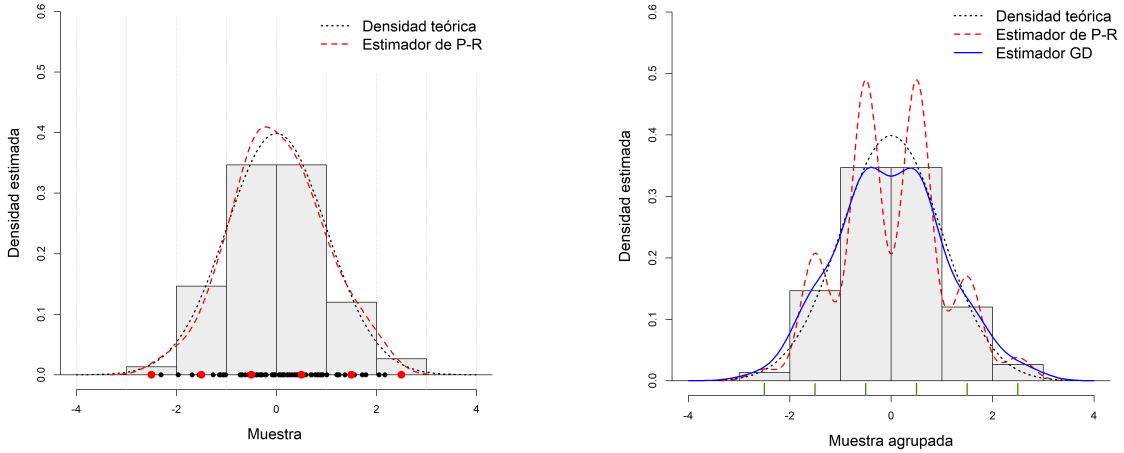
Por otro lado, existen muchas otras situaciones en las que no es realmente necesario obtener medidas con gran precisión, siendo suficiente recoger las observaciones en intervalos de interés. Este tipo de datos (frecuentemente denominados *grouped data* en la literatura anglosajona) también aparecen, por ejemplo, cuando únicamente se sabe que el suceso se ha producido en algún momento dentro de un cierto intervalo. De esta forma, los datos agrupados son comunes en áreas como la ingeniería, economía, ciencias de la salud, etc.

Naturalmente, el problema de estimación de la densidad también se puede plantear en este nuevo contexto de datos agrupados. Una primera aproximación sería ignorar el factor de agrupamiento y emplear el estimador clásico de Parzen-Rosenblatt presentado en la Sección 1.2.1. Sin embargo, en muchas ocasiones — y, sobre todo, cuando el nivel de agrupación es elevado o cuando el tamaño muestral considerado es *pequeño* — este estimador puede conducir a estimaciones que están muy lejos de representar el comportamiento verdadero de la densidad teórica.

A modo de ejemplo, consideremos una muestra de tamaño $n = 75$ procedente de una distribución normal estándar, $X_1, \ldots, X_{75}$, representada con puntos negros sobre la recta $y = 0$ de la Figura 1.1(a). Supongamos ahora la recta real dividida en intervalos de la forma $\{[m, m+1]\}_{m\in\mathbb{Z}}$, todos ellos de longitud unitaria, y recontemos el número de observaciones que se han registrado en cada uno de dichos intervalos, omitiendo aquellos en los que no se ha recogido ninguna observación, $n_1, \ldots, n_k$. A continuación, se considera la muestra $t_1, \ldots, t_k$ conformada por los puntos medios de los intervalos implicados — representada con puntos rojos sobre la recta $y = 0$ de la Figura 1.1(a) — así como sus frecuencias absolutas asociadas $n_1, \ldots, n_k$.

Este proceso de agrupación ilustra muchas situaciones reales, en las que la única información de la que se dispone es, precisamente, la muestra $t_1, \overset{n_1}{\ldots}, t_1, \ldots, t_k, \overset{n_k}{\ldots}, t_k$, siendo la muestra original $X_1, \ldots, X_n$ desconocida. Si ahora se procede a construir el estimador tipo núcleo clásico en base a la muestra de datos agrupados (haciendo uso de la ventana obtenida por la regla del pulgar, véase Silverman (1986)) se obtiene la curva representada con una línea discontinua roja en la Figura 1.1(b). Como vemos, se trata de una estimación excesivamente infrasuavizada, que provoca la formación de hasta seis modas claramente diferenciadas, conduciendo así a una interpretación errónea del comportamiento de la densidad verdadera. Esto se debe a que en torno a cada dato muestral t_j se concentran un total de n_j observaciones, provocando que el estimador tipo núcleo clásico — construido centrando sobre cada observación una función núcleo reescalada por h y sumando las ordenadas resultantes — presente una moda por cada punto medio considerado.

Ante la imposibilidad de obtener una buena estimación de la densidad a través del estimador tipo núcleo clásico — incluso ya en esta situación tan sencilla — parece natural desarrollar alguna modificación del mismo que permita incorporar de alguna manera el factor de agrupamiento en su construcción. Comentar que los primeros estudios relacionados con este tema se recogen en Hall (1982), donde se analiza la influencia de los errores de redondeo sobre la estimación tipo núcleo de la densidad. En este trabajo se considerará el estimador propuesto por Cao et al. (2011) — y posteriormente analizado en Reyes et al. (2016) y Reyes et al. (2017) — que permite trabajar incluso en aquellas situaciones en las que los puntos medios $t_1, \ldots, t_k$ no se encuentran equiespaciados.



(a) Muestra $X_1, \ldots, X_{75}$

(b) Muestra agrupada $t_1, \ldots, t_k$

Figura 1.1: Panel izquierdo: Histograma de la muestra $X_1, \ldots, X_{75}$. Con puntos de colores se representan, sobre la recta $y = 0$, las observaciones en cuestión (puntos negros) y los puntos medios de los intervalos considerados (puntos rojos). Con una línea discontinua de color rojo se representa el estimador de Parzen-Rosenblatt construido en base a la muestra $X_1, \ldots, X_{75}$. Panel derecho: estimador de Parzen-Rosenblatt construido en base a la muestra agrupada $t_1, \ldots, t_k$ (línea discontinua roja) y su respectiva modificación al caso de datos agrupados (línea continua azul). En ambas gráficas se representa, con una curva punteada de color negro, la densidad teórica (densidad normal estándar).

A modo de ejemplo, en la Figura 1.1(b) se representa, con una línea continua azul, el estimador tipo núcleo de Cao et al. (2011) (que hemos denotado por estimador GD, *grouped data*) correspondiente a la muestra agrupada en cuestión. Como vemos, este nuevo estimador se aproxima de manera mucho más satisfactoria a la densidad teórica — representada con una línea negra punteada — y carece de las modas espúreas tan marcadas que se generaban al emplear el estimador de Parzen-Rosenblatt. Comentar que, en este caso, la ventana ha sido escogida a través de un selector análogo al de la regla del pulgar clásica, pero adaptada al caso de datos agrupados, que será introducido posteriormente en el Capítulo 3.

Por último, se presenta a continuación un esquema del trabajo. En el Capítulo 2 se introducirá el estimador no paramétrico de la densidad adaptado al caso de datos agrupados propuesto por Cao et al. (2011), así como sus medidas de error y propiedades básicas. En el Capítulo 3 se abordará el problema de selección del parámetro de suavizado — ya existente en la construcción del estimador de Parzen-Rosenblatt — mediante la revisión de algunos selectores de la literatura (basándonos principalmente en los considerados en Reyes et al. (2017)) y la propuesta de otros nuevos. En el Capítulo 4 se presentarán los resultados relativos a diferentes estudios de simulación — llevados a cabo a través del software estadístico `R` 4.0.3 (R Core Team (2020)) — que permitirán analizar el comportamiento del nuevo estimador en la práctica, así como el de los selectores presentados en el capítulo anterior. A continuación, se presentará en el Capítulo 5 una aplicación de este estimador sobre diferentes bases de datos reales. Para ello, se considerarán en primer lugar los espesores de 485 sellos pertenecientes a la colección de Hidalgo de 1872, elaborados en México y analizadas por primera vez en Wilson (1983). También se abordarán los tiempos de espera medidos entre erupciones sucesivas del géiser Old Faithful (Parque Nacional de Yellowstone, Wyoming), así como los datos relativos a los casos de COVID-19 reportados en España durante diferentes periodos de la pandemia. Finalmente, en los apéndices del trabajo se abordan ciertas cuestiones de corte teórico referidas a los selectores de la regla del pulgar (Apéndice A), plug-in (Apéndice B) y validación cruzada insesgada (Apéndice C) presentados en el Capítulo 3.

Capítulo 2

Estimación tipo núcleo para datos agrupados

La estimación de la densidad de una variable aleatoria, como ya se ha justificado a lo largo de la introducción del trabajo, es un problema clásico que puede ser abordado desde varios enfoques diferentes, entre los cuales se encuentra el no paramétrico. En este contexto surge el estimador tipo núcleo de Parzen (1962) y Rosenblatt (1956), cuya expresión y propiedades se recogen en la Sección 1.2.1. Este estimador suele conducir a resultados satisfactorios en todas aquellas situaciones en las que las observaciones son recogidas de forma *individual*, de manera que en la muestra considerada no haya *demasiados* datos repetidos. En otro caso, ya se ha motivado la necesidad de desarrollar un nuevo estimador que permita trabajar en contextos donde las muestras observadas estén conformadas por datos agrupados en intervalos.

Pues bien, este tema ha recibido cierta atención en la literatura, enfocándose desde diferentes perspectivas. En este sentido, Titterington (1983) aborda dicho problema bajo la suposición de que se dispone de cierta información de la densidad teórica — concretamente, cuando existe un subconjunto de observaciones que no se encuentra afectado por el mecanismo de agrupación — algo que no suele ser frecuente en la realidad. Por otro lado, Wang y Wertelecki (2013) propusieron un estimador no paramétrico de tipo bootstrap, mientras que Blower y Kelsall (2002) abordaron la estimación a través de un esquema iterativo que involucra estimadores tipo núcleo construidos en base a núcleos Gaussianos, cuyo iterante inicial viene dado por el propio histograma construido en base a la muestra agrupada. Finalmente, en Sun (2014) se presenta una generalización de este último estimador que permite trabajar con funciones núcleo asimétricas más generales que la Gaussiana.

Sin embargo, este trabajo se centrará principalmente en las ideas de Reyes et al. (2016) y Reyes et al. (2017), en donde se analiza el estimador propuesto por Cao et al. (2011). Este estimador se fundamenta en las ideas clásicas de Parzen (1962) y Rosenblatt (1956) y, como se mostrará a lo largo de este capítulo, constituye una generalización natural del propuesto por Scott y Sheather (1985), permitiendo trabajar en situaciones más generales, donde las observaciones se encuentran agrupadas en intervalos no necesariamente equiespaciados y en las que únicamente sus proporciones muestrales son conocidas (dando lugar a lo que en literatura inglesa se suele denominar *general grouped data case*).

En lo que sigue, denotaremos por $X_1, X_2, \ldots, X_n$ a una muestra aleatoria simple de la variable de interés X y se considerará un conjunto de k intervalos fijos $\{[y_{j-1}, y_j)\}_{j=1}^{k}$ de longitudes $l_j = y_j - y_{j-1}$ y puntos medios $t_j = (y_{j-1} + y_j)/2$, $j = 1, \ldots, k$. Asumiremos, además, que tanto el número de intervalos como sus longitudes y puntos medios no son aleatorios, pero sí dependientes del tamaño muestral, n, adoptando la notación abreviada $t_j \equiv t_{j,n}$ y $l_j \equiv l_{j,n}$, con $j = 1, \ldots, k$. De esta forma, las muestras observadas estarán conformadas por los puntos medios de los intervalos considerados, repetidos tantas veces como el número de observaciones recogidas en cada uno de ellos, que denotaremos por $n_1, \ldots, n_k$.

2.1. Estimador de la densidad para datos agrupados

Esta sección se iniciará presentando la modificación del estimador tipo núcleo de Parzen-Rosenblatt propuesta por Scott y Sheather (1985), que constituye una primera generalización del estimador clásico al contexto de datos agrupados, bajo la suposición de que todos los intervalos tienen la misma longitud. Esto motivará, de forma natural, la construcción de un estimador alternativo que permita extender la estimación no paramétrica de la densidad a contextos más generales, como es la propuesta en Cao et al. (2011), que constituirá el foco principal de interés a lo largo este trabajo.

2.1.1. Estimador de la densidad de Scott y Sheather (1985)

En Scott y Sheather (1985) se propone una modificación que permite adaptar el estimador de Parzen-Rosenblatt al contexto de datos agrupados en intervalos de longitud constante l (esto es, bajo la suposición $l_1 = \cdots = l_k = l$), en donde el número de observaciones $n_1, n_2, \ldots, n_k$ es conocido.

Sean $t_1, \ldots, t_k$ los puntos medios de los intervalos considerados. Dada una función núcleo L y una ventana $h > 0$, Scott y Sheather (1985) proponen considerar como estimador de la densidad a

$$\hat{f}_n^{g,\mathrm{ss}}(x) = \frac{1}{n} \sum_{j=1}^{k} \frac{n_j}{h} L\left(\frac{x - t_j}{h}\right), \tag{2.1}$$

que no es más que el estimador de Parzen-Rosenblatt aplicado sobre la muestra $t_1, \ldots, t_k$ al cual se le han añadido las frecuencias absolutas n_j como ponderaciones.

Entre otras pruebas, Hall (1982) demostró que, bajo ciertas condiciones de regularidad, la esperanza de (2.1) era idéntica a la del estimador de Parzen-Rosenblatt, pero incrementada por un factor que dependía de la longitud l de los intervalos, como se mostrará en la Sección 2.2.2.

2.1.2. Estimador de la densidad de Cao *et al.* (2011)

Consideremos ahora un contexto de datos agrupados general, en el que las observaciones $X_1, \ldots, X_n$ se encuentran agrupadas en k intervalos de la forma $\{[y_{j-1}, y_j)\}_{j=1}^{k}$ — no necesariamente de la misma longitud — con puntos medios $t_1, \ldots, t_k$ y donde las proporciones muestrales $w = (w_1, \ldots, w_k)$ son conocidas, siendo

$$w_j = \frac{n_j}{n} = F_n(y_j^-) - F_n(y_{j-1}^-), \qquad j = 1, \ldots, k,$$

donde n_j denota el número de observaciones recogidas en el $j-$ésimo intervalo (posiblemente desconocido) y $F_n(y^-)$ denota el límite izquierdo de la función de distribución empírica de X en el punto y.

En este contexto general, en Cao et al. (2011) se propone considerar como estimador de la función de densidad a[1]

$$\hat{f}_w^g(x) = \frac{1}{h} \sum_{j=1}^{k} w_j L\left(\frac{x - t_j}{h}\right) = \sum_{j=1}^{k} w_j L_h(x - t_j), \tag{2.2}$$

donde L es una función núcleo, $h > 0$ el parámetro ventana y donde $L_h(\cdot) = \frac{1}{h} L\left(\frac{\cdot}{h}\right)$ denota a la función núcleo reescalada por h.

[1] En el artículo original de Cao et al. (2011) se reserva la notación $\hat{f}_h$ para el estimador (2.2). Por el contrario, en Reyes et al. (2016) y Reyes et al. (2017) se emplea $\hat{f}_n^g$ y $\hat{f}_h^g$, indistintamente. Sin embargo, en este caso se ha optado por una notación que permite explicitar la dependencia del estimador respecto de los pesos muestrales $w_1, \ldots, w_k$.

De esta forma, la estimación de la densidad en un punto se realiza en función de las distancias a cada dato de la muestra, ahora identificados con los puntos medios de los intervalos, ponderadas por la función núcleo reescalada, L_h. Además, se añaden los pesos w_j, con $j = 1, \ldots, k$, de manera que una distancia tendrá tanto más peso en la estimación cuantas más observaciones se hayan recogido en el intervalo correspondiente.

Ahora bien, aunque es cierto que el estimador (2.2) no requiere de conocer las cantidades $n_1, \ldots, n_k$ para su contrucción — tal y como se comenta en Reyes et al. (2016) — la selección óptima del parámetro de suavizado h requerirá, en general, de conocer el tamaño muestral n. De esta forma, en la práctica, la construcción del estimador de Cao et al. (2011) se restringirá a aquellos contextos en los que las cantidades $n_j = w_j n$, $j = 1, \ldots, k$ sean conocidas, como se mostrará con más detalle en el Capítulo 3.

Observación 2.1. Sustituyendo $w_j = n_j/n$ en (2.2) se llega a la expresión (2.1). De esta manera, el estimador de Cao et al. (2011) puede verse como una extensión natural del estimador de Scott y Sheather (1985) a un contexto general en el cual únicamente las proporciones muestrales w_j son conocidas.

2.2. Resultados asintóticos y medidas de error

Al igual que ya sucedía con el estimador de Parzen-Rosenblatt, resulta interesante disponer de medidas de error que permitan evaluar el comportamiento de (2.2) como estimador de la densidad teórica f, tanto desde un enfoque local como global. En este sentido, resulta natural optar por las mismas medidas ya consideradas para el estimador tipo núcleo clásico y presentadas en la Sección 1.2.1: el error cuadrático medio y el error cuadrático medio integrado, que ahora denotaremos por MSE_g y MISE_g, respectivamente. En tal caso,

$$\mathrm{MSE}_g(x) = \mathrm{MSE}\left(\hat{f}_w^g(x)\right) = \mathbb{E}\left[\left(\hat{f}_w^g(x) - f(x)\right)^2\right] = \mathrm{Sesgo}^2\left(\hat{f}_w^g(x)\right) + \mathrm{Var}\left(\hat{f}_w^g(x)\right), \tag{2.3}$$

$$\mathrm{MISE}_g = \mathrm{MISE}\left(\hat{f}_w^g\right) = \int \mathrm{MSE}\left(\hat{f}_w^g(x)\right)\, \mathrm{d}x. \tag{2.4}$$

Tras presentar las hipótesis básicas requeridas para su construcción, en la Sección 2.2.2 se mostrarán las expresiones analíticas de (2.3) y (2.4), que proporcionarán una forma de medir el error de estimación de $\hat{f}_w^g$, tanto a nivel local ($\mathrm{MSE}_g(x)$) como global (MISE_g).

2.2.1. Suposiciones empleadas

Con el fin de obtener expresiones analíticas para las medidas de error relativas a $\hat{f}_w^g$, en lo que sigue se asumirán las siguientes condiciones:

(**S1**): La función de densidad f es una función absolutamente continua, con segunda derivada f'' continua y tal que $\int \left(f''(x)\right)^2 \, dx < \infty$.

(S2): La ventana $h \equiv h_n$ es una sucesión no aleatoria de números positivos tal que $\lim_{n\to\infty} h = 0$ y $\lim_{n\to\infty} nh = \infty$.

Recordemos que esta condición se requiere también para la construcción del estimador de Parzen-Rosenblatt.

(S3): Dado un conjunto de $k \equiv k_n$ intervalos $\{[y_{j-1}, y_j)\}_{j=1}^{k}$, asumiremos que su longitud media, definida como $\bar{l} = \bar{l}_n = \frac{1}{k}\sum_{j=1}^{k} l_j$, verifica las condiciones

$$\lim_{n\to\infty} \bar{l} = 0, \qquad \lim_{n\to\infty} n\bar{l} = \infty, \qquad \bar{l} = o(h).$$

La primera y tercera condición garantizan que, conforme el tamaño muestral aumenta, la longitud media de los intervalos también convergerá a cero y, además, lo hará más rápido de lo que lo hace $h \equiv h_n$. Teniendo en cuenta que la distancia media entre las observaciones es $\bar{l}$, esto implica que la ventana deberá ser siempre mayor que $\bar{l}$, lo cual permite garantizar que los entornos considerados contienen suficiente información.

La condición $\lim_{n\to\infty} n\bar{l} = \infty$ es también importante, puesto que si la longitud de los intervalos convergiese a cero más rápido de lo que aumenta n, llegaría un punto en el que habría más intervalos que observaciones, encontrándose la mayoría de ellos vacíos.

(S4): Supondremos que la longitud del mayor intervalo, que denotaremos $l_{\text{máx}} \equiv l_{\text{máx},n}$, es tal que

$$l_{\text{máx}} = O(\bar{l}), \qquad \max_{1\le j\le k} |l_j - \bar{l}| = o(\bar{l}).$$

Nótese que, si se supone k finito, la segunda condición es consecuencia directa de la primera. Además, esta condición permite controlar la variabilidad de la longitud de los intervalos, en el sentido de que impone que sus longitudes *no pueden ser muy diferentes*. De esta forma, si bien es cierto que el estimador de Cao et al. (2011) se puede emplear en situaciones en las que los intervalos tengan distintas longitudes, estas no podrán diferir *demasiado*.

(S5): La función núcleo L es una función de densidad simétrica y Lipschitziana, con soporte en el intervalo $[-1, 1]$, 6 veces diferenciable y tal que su sexta derivada, $L^{(6)}$, está acotada. Se considerará la notación abreviada

$$\mu_2(L) = \int x^2 L(x)\, dx > 0.$$

Nótese que el hecho de asumir un carácter estrictamente positivo para $\mu_2(L)$ implica que en todo momento se considerarán funciones núcleo de orden dos. Considerar núcleos de orden superior conduciría, quizás, a mejores resultados, pero con el inconveniente de que podrían dar lugar a estimadores de la densidad que fuesen negativos en algún punto del dominio (algo que, en el contexto de la estimación de la densidad, no resulta deseable).

(S6): La función de distribución F es una función continua, Lipschitziana, 7 veces diferenciable y tal que su $j-$ésima derivada — que denotaremos por $F^{(j)}$ — está acotada, para $j = 1, \ldots, 7$.

2.2.2. Medidas de error

Una vez introducidas las hipótesis requeridas, se presentan a continuación las expresiones analíticas de las medidas de error relativas al estimador de Cao et al. (2011), ya obtenidas en Reyes et al. (2017). Para ello, comenzaremos presentando las expresiones de la esperanza y la varianza de $\hat{f}_w^g(x)$ como estimador puntual de $f(x)$ — siendo x un punto del dominio de la densidad teórica — para, a continuación, abordar las expresiones del $\mathrm{MSE}_g(x)$ y MISE_g, que ofrecerán una medida del error cometido por el estimador (2.2).

Por un lado, en Reyes et al. (2016) se prueba que, bajo las suposiciones **(S1)-(S6)**, la expresión analítica de la esperanza de $\hat{f}_w^g$ en un punto x viene dada por

$$\mathbb{E}[\hat{f}_w^g(x)] = f(x) + \left[\frac{\bar{l}^2}{24} + \frac{1}{2}h^2\mu_2(K)\right] f''(x) + o(h^2). \tag{2.5}$$

De esta expresión se deduce que:

- Asintóticamente, la esperanza del estimador de Cao et al. (2011) es idéntica a la del estimador tipo núcleo de Parzen-Rosenblatt. Sin embargo, en segundo orden, esta se ve incrementada por un factor que depende de la longitud media de los intervalos. En efecto, reorganizando los términos involucrados en (2.5) resulta

$$\mathbb{E}[\hat{f}_w^g(x)] = \left[f(x) + \frac{1}{2}h^2 f''(x)\mu_2(K) + o(h^2)\right] + \frac{\bar{l}^2}{24}f''(x) = \left[\mathbb{E}[\hat{f}_n^s(x)] + o(h^2)\right] + O(\bar{l}^2). \tag{2.6}$$

  Observación 2.2. De la expresión (2.6) se deduce que, si asumimos cierta la condición $\bar{l} = o(h)$, entonces el sesgo asintótico del estimador de Cao et al. (2011) resulta idéntico al del estimador de Parzen (1962) y Rosenblatt (1956), tanto en primer como en segundo orden.

- Cuando los intervalos considerados tienen longitud constante (esto es, $l_1 = \cdots = l_k = l$), entonces (2.5) puede reescribirse como

$$\mathbb{E}[\hat{f}_w^g(x)] \approx f(x) + \left[\frac{l^2}{24} + \frac{1}{2}h^2\mu_2(K)\right] f''(x),$$

  lo cual coincide con la expresión asintótica de la esperanza del estimador de Scott y Sheather (1985), tal y como se prueba en Hall (1982).

Por otro lado, bajo las suposiciones **(S1)-(S6)**, la varianza del estimador $\hat{f}_w^g$ en un punto x viene dada por

$$\mathrm{Var}[\hat{f}_w^g(x)] = \frac{1}{nh}R(L)f(x) + o\left((nh)^{-1}\right). \tag{2.7}$$

De esto se concluye lo siguiente:

- Asintóticamente, la varianza del estimador de Cao et al. (2011) es idéntica a la del estimador tipo núcleo de Parzen-Rosenblatt.

- Cuando los intervalos considerados tienen longitud constante (esto es, $l_1, = \cdots = l_k = l$), entonces

$$\mathrm{Var}[\hat{f}_w^g(x)] \approx \frac{1}{nh}R(L)f(x),$$

  coincidiendo, en tal caso, con la expresión asintótica de la varianza del estimador de Scott y Sheather (1985), tal y como se prueba en Hall (1982).

De las expresiones (2.5) y (2.7) se deduce inmediatamente que el error cuadrático medio de $\hat{f}_w^g(x)$ como estimador de $f(x)$ es

$$\mathrm{MSE}_g(x) = \left[\frac{\bar{l}^2}{24} + \frac{1}{2}\,h^2\mu_2(L)\right]^2 (f''(x))^2 + \frac{1}{nh}R(L)f(x) + o(h^4) + o((nh)^{-1}). \tag{2.8}$$

Integrando (2.8) sobre el dominio de definición de f, se obtiene el error cuadrático medio integrado de $\hat{f}_w^g$ (MISE_g), que constituye una medida del error global de $\hat{f}_w^g$ como estimador de la densidad verdadera,

$$\mathrm{MISE}_g = \left[\frac{\bar{l}^2}{24} + \frac{1}{2}h^2\mu_2(K)\right]^2 R(f'') + \frac{1}{nh}R(L) + o(h^4) + o((nh)^{-1}). \tag{2.9}$$

Nótese que, bajo la suposición **(S3)** — en la cual se impone que $\bar{l} = o(h)$ — las expresiones (2.8) y (2.9) pueden reescribirse como

$$\begin{aligned}\mathrm{MSE}_g(x) &= \frac{1}{4}\,h^4\mu_2(L)^2\,(f''(x))^2 + \frac{1}{nh}R(L)f(x) + o(h^4) + o((nh)^{-1}),\\ \mathrm{MISE}_g &= \frac{1}{4}\,h^4\mu_2(L)^2R(f'') + \frac{1}{nh}R(L) + o(h^4) + o((nh)^{-1}).\end{aligned} \tag{2.10}$$

De esta forma, bajo las condiciones **(S1)-(S6)**, tanto el error cuadrático medio como el error cuadrático medio integrado del estimador $\hat{f}_w^g$ adoptan expresiones asintóticas idénticas a las correspondientes al estimador tipo núcleo de Parzen-Rosenblatt.

Siguiendo un procedimiento análogo al caso clásico de datos no agrupados, la minimización de la versión asintótica de (2.10), esta vez denotada por AMISE_g, conduce a la expresión de una ventana global (asintóticamente) óptima,

$$h_{\mathrm{AMISE}_g} = \left[\frac{R(L)}{\mu_2(L)^2R(f'')n}\right]^{\frac{1}{5}}. \tag{2.11}$$

Como era de esperar, la expresión de la ventana asintóticamente óptima es también idéntica a la obtenida con el estimador de Parzen-Rosenblatt (véase (1.4)), con la única diferencia de que esta vez, cuando $R(f'')$ sea desconocido, deberá ser estimado en base a una muestra de datos agrupados.

Observación 2.3. Cuando los intervalos considerados tienen longitud constante l, entonces la versión asintótica de (2.9) — que ahora denotaremos por $\mathrm{AMISE}_{g,\mathrm{ss}}$ — puede reescribirse como

$$\mathrm{AMISE}_{g,\mathrm{ss}} = \left[\frac{l^2}{24} + \frac{1}{2}h^2\mu_2(K)\right]^2 R(f'') + \frac{1}{nh}R(L).$$

En tal caso, denotando $c = l/h\mu_2(K)$ es sencillo comprobar que

$$\mathrm{AMISE}_{g,\mathrm{ss}} = \frac{1}{4}h^4\mu_2(K)^2\left(\frac{c^2}{12} + 1\right)^2 R(f'') + \frac{1}{nh}R(L), \tag{2.12}$$

lo cual coincide con el error cuadrático medio integrado (asintótico) del estimador de Scott y Sheather (1985), tal y como se recoge en su Proposición 2[2].

[2]En la expresión asintótica del error cuadrático medio integrado de $\hat{f}_n^{g,\mathrm{ss}}$ recogido en la Proposición 2 de Scott y Sheather (1985) se incluye también el término $n^{-1}R(f)$, el cual puede obviarse dado que estamos asumiendo que $R(f) < \infty$ (suposición **(S1)**), puesto que en tal caso se tiene que $n^{-1}R(f) = o((nh)^{-1}R(L))$.

Obviamente, la condición $l = o(h)$ — esta vez enunciada en términos de la longitud constante l — permite reescribir (2.12) como

$$\mathrm{AMISE}_{g,\mathrm{ss}} = \frac{1}{4}h^4\mu_2(K)^2R(f'') + \frac{1}{nh}R(L),$$

lo cual conduciría nuevamente a la ventana asintóticamente óptima (2.11).

Sin embargo, comentar que en Scott y Sheather (1985) no se impone la condición $l = o(h)$. En tal caso, minimizando (2.12) se llega a que

$$h_{\mathrm{AMISE}_{g,\mathrm{ss}}} = \left[\frac{R(L)}{n\mu_2(L)^2\left(\frac{c^2}{12}+1\right)R(f'')}\right]^{1/5} = \left(\frac{c^2}{12}+1\right)^{-2/5} h_{\mathrm{AMISE}},$$

donde h_{AMISE} denota la ventana asintóticamente óptima en el caso de datos no agrupados, cuya expresión se recoge en (1.4). De esta forma, si la condición $l = o(h)$ no se verifica, entonces la ventana asintóticamente óptima depende de c, y, por tanto, de la longitud l de los intervalos considerados. En este sentido, y con el objetivo de acotar superiormente — con un cierto valor α — el cociente[3]

$$\frac{\mathrm{MISE}_{g,\mathrm{ss}}(h_{\mathrm{AMISE}_{g,\mathrm{ss}}})}{\mathrm{MISE}\,(h_{\mathrm{AMISE}})} = \left(\frac{c^2}{12}+1\right)^{2/5} + o(1), \tag{2.13}$$

Scott y Sheather (1985) consideran una restricción adicional, bajo la cual se impone que la longitud de los intervalos considerados, l, debe ser elegida de tal forma que $c = l/h\mu_2(K)$ verifique, asintóticamente, la condición

$$c \leq \sqrt{12(1+\alpha)^{5/2}-1} \approx \sqrt{30\alpha}.$$

[3]Nótese que en (2.13) se ha empleado la notación $\mathrm{MISE}_{g,\mathrm{ss}}(h)$ (respectivamente $\mathrm{MISE}(h)$) para hacer referencia al error cuadrático medio integrado del estimador $\hat{f}_n^{g,\mathrm{ss}}$ (respectivamente $\hat{f}_n$) construido en base a la ventana h.

Capítulo 3

Selectores de ventana

Es bien conocido que la construcción de un estimador tipo núcleo requiere de fijar previamente un parámetro ventana h, tal y como se ha mostrado en la Sección 1.2.1. En ocasiones, esta elección se puede realizar de manera subjetiva, construyendo varias estimaciones de la densidad sobre diferentes valores de h y seleccionando aquella que sea más satisfactoria en algún sentido. Sin embargo, en muchas otras ocasiones — y en especial cuando no se tiene ningún conocimiento previo acerca de la verdadera densidad — sería conveniente disponer de algún selector automático que escogiese el *mejor* valor de h en función de la muestra considerada. Esta cuestión, ya presente en la construcción del estimador clásico de Parzen-Rosenblatt, se extiende también al estimador tipo núcleo de Cao et al. (2011), cuya expresión se recoge en (2.2). En lo que sigue, se empleará la notación $\hat{f}^g_{w,h} \equiv \hat{f}^g_w$ para explicitar la dependencia del estimador respecto del parámetro ventana h.

En el Capítulo 2 se ha visto que la ventana global asintóticamente óptima, en el caso de datos agrupados, viene dada por

$$h_{\mathrm{AMISE}_g} = \left[\frac{R(L)}{\mu_2(L)^2 R(f'')n}\right]^{\frac{1}{5}}. \tag{3.1}$$

Esta expresión, que es idéntica a la que se obtiene en el caso de datos no agrupados, depende de una cantidad desconocida en la práctica, $R(f'')$. Algunos selectores de ventana partirán, por tanto, de la expresión (3.1) y tratarán de sustituir $R(f'')$ por algún estimador adecuado. Será precisamente en la forma de estimar esta cantidad en donde residirán las principales diferencias entre el caso de datos agrupados y no agrupados. En esta idea se fundamentan, por ejemplo, los selectores de la regla del pulgar y los selectores plug-in. Por el contrario, y de manera análoga a lo que sucede en el caso clásico, una idea alternativa consiste en intentar aproximar directamente alguna medida de error relativa a $\hat{f}^g_{w,h}$, en cuyo caso nos encontraríamos, por ejemplo, en el contexto de las técnicas de validación cruzada. Finalmente, otros selectores se basarán en procedimientos bootstrap, que presentan la ventaja de no requerir de ninguna hipótesis sobre el mecanismo generador de los datos.

Este capítulo se iniciará proponiendo un nuevo selector de ventana para el caso de datos agrupados, que constituirá una generalización natural de la clásica regla del pulgar de Silverman. A continuación, se introducirán los dos selectores de ventana propuestos en Reyes et al. (2017): un selector plug-in y un selector tipo bootstrap. Asimismo, se propondrá un selector plug-in alternativo, fundamentado en ideas similares a aquellas empleadas en la regla del pulgar; un selector de validación cruzada insesgada y una modificación del selector bootstrap propuesto en Jang y Loh (2010).

3.1. Regla del pulgar para datos agrupados

Silverman propuso uno de los primeros estimadores de la ventana global para el caso de datos no agrupados (ver Sección 3.4 de Silverman (1986)). Para ello, partió de la expresión de h_{AMISE} recogida en (1.4) y propuso reemplazar la cantidad desconocida $R(f'')$ por aquella que se obtiene suponiendo que f sigue una distribución normal $N(\mu, \sigma^2)$. En tal caso, haciendo uso de los polinomios de Hermite, es sencillo comprobar que

$$R(f'') = \int \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right)^2 dx = \frac{3}{8\sqrt{\pi}} \sigma^{-5}, \tag{3.2}$$

donde σ se puede aproximar por un estimador adecuado, siendo el propuesto por Silverman (1986, página 47) uno de los más utilizados. Sustituyendo dicho estimador $\hat{\sigma}$ en (3.2), Silverman obtuvo la ventana clásica

$$\hat{h}_{\text{rp}} = \left[\frac{8\sqrt{\pi} R(L)}{3\mu_2(L)^2 n} \right]^{\frac{1}{5}} \hat{\sigma}. \tag{3.3}$$

A la vista de esto, parece claro que la forma más sencilla y natural de extender la regla del pulgar clásica al contexto de datos agrupados consistiría en considerar un estimador de σ alternativo, construido en base a una muestra de datos agrupados $t_1, \ldots, t_k$, con pesos asociados $w_1, \ldots, w_k$; el cual, sustituido en la expresión (3.2), conduciría a una nueva ventana, que denotaremos por $\hat{h}_{\text{g,rp}}$. Haciendo uso de técnicas de máxima verosimilitud ponderada, en el Apéndice A se propone considerar como estimador de σ a

$$\hat{\sigma}_g = \sqrt{\sum_{j=1}^{k} w_j (t_j - \hat{\mu}_g)^2}, \tag{3.4}$$

donde $\hat{\mu}_g = \sum_{j=1}^{k} w_j t_j$ denota a la media de los valores $t_1, \ldots, t_k$ ponderada por los pesos $w_1, \ldots, w_k$.

Observación 3.1. En el Apéndice A se presenta también un breve estudio de simulación que permite aproximar, en un contexto de datos agrupados, el error cuadrático medio de $\hat{\sigma}_g^2$ como estimador de la varianza poblacional, comparándolo con el correspondiente a la varianza muestral (estimador clásico en el contexto de datos no agrupados), evidenciando en todos los casos una mejor aproximación con el estimador $\hat{\sigma}_g^2$.

Sustituyendo $\hat{\sigma}_g$ en (3.2) se obtiene la siguiente estimación de la ventana global, adaptada, ahora sí, al caso de datos agrupados

$$\hat{h}_{\text{g,rp}} = \left[\frac{8\sqrt{\pi} R(L)}{3\mu_2(L)^2 n} \right]^{\frac{1}{5}} \sqrt{\sum_{j=1}^{k} w_j (t_j - \hat{\mu}_g)^2}. \tag{3.5}$$

De esta forma, y como ya se anticipaba en el Capítulo 2, la construcción de $\hat{h}_{\text{g,rp}}$ requiere de conocer tanto los pesos w_j como el tamaño muestral n, por lo que, en la práctica, únicamente podrá obtenerse en aquellos contextos en los que las frecuencias absolutas $n_1, \ldots, n_k$ (número de observaciones recogidas en cada intervalo) sean conocidas. Como veremos, esto será algo frecuente en la mayoría de los selectores de ventana aquí presentados.

3.1.1. Modificación de la regla del pulgar

De manera análoga a lo que sucede en el caso de datos no agrupados (véase Wand y Jones (1995)), se puede pensar en un estimador de la desviación típica poblacional de la forma

$$\hat{\sigma}_g^* = \text{mín}\left\{\hat{\sigma}_g, \hat{\sigma}_{\text{IQR},g}\right\}, \tag{3.6}$$

donde $\hat{\sigma}_g$ denota el estimador de máxima verosimilitud ponderada de σ (obtenido en el Apéndice A y cuya expresión se recoge en (3.4)) y $\hat{\sigma}_{\text{IQR},g}$ el rango intercuartílico estandarizado muestral adaptado al caso de datos agrupados, definido, por analogía al caso clásico, como

$$\hat{\sigma}_{\text{IQR},g} = \frac{\text{Rango intercuartílico muestral}}{\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)}, \tag{3.7}$$

siendo Φ la función de distribución de una densidad normal estándar. Nótese que el denominador de la cantidad (3.7) no es otra cosa que el rango intercuartílico poblacional de una densidad normal estándar, que puede ser aproximado por 1.34.

La cuestión reside ahora en cómo estimar el rango intercuartílico en base a una muestra de datos agrupados. Una definición natural es

$$\text{Rango intercuartílico muestral} = \hat{Q}^g_{0.75} - \hat{Q}^g_{0.25},$$

donde $\hat{Q}^g_\tau$ denota un estimador del cuantil de orden τ calculado en base a una muestra de datos agrupados. En este sentido, Schmeiser y Deutsch (1977) proponen, para el caso particular de intervalos equiespaciados, un estimador para Q^g_τ, en base al cual se ha propuesto el siguiente, adaptado al caso de intervalos $\{[y_{j-1}, y_j)\}_{j=1}^k$ no necesariamente equiespaciados, de longitudes respectivas $l_1, \ldots, l_k$,

$$\hat{Q}^g_\tau = y_0 + \sum_{j=1}^{q-1} l_i + \frac{l_q}{2} = y_{q-1} + \frac{l_q}{2} = t_q, \tag{3.8}$$

donde q denota al menor entero tal que

$$\sum_{i=1}^{q} n_i \geq \lfloor \tau(n+1) \rfloor,$$

siendo $\lfloor \cdot \rfloor$ la función parte entera por defecto. Nótese que, con esta definición, $\hat{Q}^g_\tau$ coincide con el punto medio del intervalo en el cual se encuentra el $\tau(n+1)-$ésimo estadístico ordenado de la muestra.

Ahora bien, en la práctica, el hecho de identificar los cuantiles muestrales con los puntos medios t_j puede derivar en que $\hat{Q}_{0.25} = \hat{Q}_{0.75} = 0$, siendo esta situación más frecuente cuando el tamaño muestral es *pequeño* (inferior a 100 datos) y la densidad generadora presenta una única moda, concentrando gran parte de la densidad en un pequeño intervalo, lo cual conduciría a un estimador degenerado de la varianza. Por este motivo, consideraremos la siguiente modificación del estimador (3.8),

$$\hat{Q}^{g*}_\tau = y_0 + \sum_{j=1}^{q-1} l_i + \left(\frac{\tau - c_{q-1}}{c_q - c_{q-1}}\right) l_q = y_{q-1} + \left(\frac{\tau - c_{q-1}}{c_q - c_{q-1}}\right) l_q,$$

siendo $c_p = \frac{1}{n}\sum_{j=1}^k \mathbb{I}(t_j < y_p)$, donde $\mathbb{I}$ denota la función indicadora. De esta forma, ahora Q^{g*}_τ se corresponde con un punto proporcional al orden del cuantil del intervalo en el cual se encuentra el $\tau(n+1)-$ésimo estadístico ordenado de la muestra y ya no con su punto medio, como sí hacía $\hat{Q}^g_\tau$.

Volviendo a la expresión (3.6) se tiene que un estimador alternativo de la desviación típica muestral para el caso de datos agrupados es

$$\hat{\sigma}_g^* = \text{mín}\left\{\hat{\sigma}_g, \frac{\hat{Q}_{0.75}^{g*} - \hat{Q}_{0.25}^{g*}}{1.34}\right\}.$$

Esta propuesta ha sido analizada utilizando un estudio de simulación similar al que se describe en el Capítulo 4. Tras comparar los resultados con los obtenidos empleando directamente a $\hat{\sigma}_g$ como estimador de la desviación típica poblacional, se ha concluido que el empleo de $\hat{\sigma}_g^*$ no conduce a ninguna mejora notable — en términos del MISE_g — respecto de $\hat{\sigma}_g$. De hecho, en la mayoría de casos, para valores muestrales pequeños ($n = 50$) se han obtenido mejores resultados con el estimador $\hat{\sigma}_g$. Por este motivo, el cálculo de la ventana proporcionada por la regla del pulgar adaptada al caso de datos agrupados se realizará, de ahora en adelante, empleando a $\hat{\sigma}_g$ como estimador de la desviación típica poblacional.

3.2. Selectores plug-in

Se ha mostrado en el Capítulo 2 que la minimización del AMISE_g permite obtener una ventana global asintóticamente óptima, cuya expresión viene dada por

$$h_{\text{AMISE}_g} = \left[\frac{R(L)}{\mu_2(L)^2 R(f'')n}\right]^{\frac{1}{5}}. \tag{3.9}$$

Los selectores plug-in, tal y como sucedía en el caso clásico, proponen obtener una estimación de la ventana global sustituyendo $R(f'')$, que es una cantidad desconocida en la práctica, por un estimador adecuado. Esta sección se iniciará presentando el selector plug-in propuesto en Reyes et al. (2017), para, a continuación, realizar una nueva propuesta que seguirá las ideas de Ćwik y Koronacki (1997).

3.2.1. Selector plug-in de Reyes *et al.* (2017)

Como ya se ha comentado en anteriores ocasiones, si bien es cierto que la expresión de la ventana asintóticamente óptima en el contexto de datos agrupados es idéntica a la obtenida con el estimador de Parzen-Rosenblatt, la diferencia radica en cómo estimar $R(f'')$, puesto que, en este caso, su estimación debe realizarse en base a una muestra agrupada. En este sentido, en Cao et al. (2011) se propone considerar como estimador a

$$\hat{R}_\eta(f'') = \frac{1}{\eta^5}\sum_{i=1}^{k}\sum_{j=1}^{k} W^{(4)}\left(\frac{t_i - t_j}{\eta}\right) w_i w_j, \tag{3.10}$$

donde $W^{(4)}$ denota la derivada de cuarto orden de una función núcleo, W, posiblemente distinta de L, y donde $\eta > 0$ es otro parámetro de ventana. En lo que sigue, asumiremos que $W = L$, tal y como se propone en Reyes et al. (2017). De esta forma, se obtiene la siguiente estimación de la ventana global asintóticamente óptima,

$$\hat{h}_{\text{R}} = \left[\frac{R(L)}{\mu_2(L)^2 \hat{R}_\eta(f'')n}\right]^{\frac{1}{5}}. \tag{3.11}$$

Elección de la ventana piloto

Como acabamos de ver, la construcción de $\hat{h}_{\mathrm{R}}$ (y, más concretamente, de $\hat{R}_\eta(f'')$) requiere de obtener una ventana piloto η, que debe ser previamente estimada. Para ello, consideraremos el problema general de estimar el funcional $R(f^{(r)})$ para un $r \in \mathbb{N}$ genérico. La idea será escoger como ventana η aquel valor que minimice el error cuadrático medio de $\hat{R}_\eta(f^{(r)})$ como estimador de $R(f^{(r)})$.

Consideremos la cantidad $\psi_r = \mathbb{E}(f^{(r)}(X))$. Bajo ciertas condiciones de suavidad impuestas sobre f, un sencillo argumento de integración por partes permite comprobar que

$$R(f^{(s)}) = \int \left[f^{(s)}(x)\right]^2 \, \mathrm{d}x = (-1)^s \int f^{(2s)}(x) f(x) \, \mathrm{d}x = (-1)^s \, \mathbb{E}(f^{(2s)}(X)) = (-1)^s \psi_{2s}.$$

De esta forma, resulta inmediato que para estimar el funcional $R(f^{(r)})$ es suficiente estudiar la estimación de los funcionales ψ_r para $r \in \mathbb{N}$ par. En este sentido, en Reyes et al. (2017) se considera el estimador

$$\hat{\psi}^g_{r,\eta_r} = \frac{1}{(\eta_r)^{r+1}} \sum_{i=1}^{k} \sum_{j=1}^{k} L^{(r)}\left(\frac{t_i - t_j}{\eta_r}\right) w_i w_j,$$

cuyo caso particular $r = 4$ conduce al estimador de $R(f'')$ propuesto en Cao et al. (2011). Fijado un $r \in \mathbb{N}$ par, se escogerá, pues, aquel valor η_r que minimice el error cuadrático medio de $\hat{\psi}^g_{r,\eta_r}$ como estimador de ψ_r, cuya expresión se recoge en el Teorema 1 de Reyes et al. (2017).

En este sentido, existen dos formas posibles de abordar dicha minimización, donde ninguna ha resultado ser universalmente mejor que la otra (para una discusión más detallada, consultar el Material Complementario de Reyes et al. (2017)). Tras varios estudios de simulación, en Reyes et al. (2017) se propone como ventana piloto, para el caso particular de funciones núcleo de orden dos, a

$$\eta_r^{\mathrm{ópt}} = \left[-\frac{2L^{(r)}(0) R(f) \bar{l}}{\mu_2(L) \psi_{r+2}}\right]^{\frac{1}{r+3}}. \tag{3.12}$$

Observación 3.2. Bajo ciertas condiciones de regularidad allí recogidas, en el Apéndice B se prueba que, asintóticamente, la ventana (3.12) es la única que minimiza el error cuadrático medio de $\hat{\psi}^g_{r,\eta_r}$ como estimador de ψ_r. Este resultado asintótico confirma la conjetura de Reyes et al. (2017) para el caso de funciones núcleo de orden dos, cuando $n \to \infty$.

De esta forma, el estimador (3.10) de $R(f'')$ (correspondiente al caso particular $r = 4$) se construiría en base a la ventana

$$\eta_4^{\mathrm{ópt}} = \left[-\frac{2L^{(4)}(0) R(f) \bar{l}}{\mu_2(L) \psi_6}\right]^{\frac{1}{7}}. \tag{3.13}$$

Sin embargo, esta expresión depende nuevamente de dos cantidades desconocidas, $R(f)$ y ψ_6, que deberán de ser estimadas. Por un lado, la estimación de $R(f)$ se realizará bajo la suposición de que f sigue una distribución normal $N(\mu, \sigma^2)$, en cuyo caso

$$R(f) = \int \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)^2 \, \mathrm{d}x = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{\sigma^2}} \, \mathrm{d}x = \frac{\sigma\sqrt{\pi}}{2\pi\sigma^2} = \frac{1}{2\sigma\sqrt{\pi}}, \tag{3.14}$$

donde σ puede estimarse, por ejemplo, mediante uno de los estimadores propuestos en la Sección 3.1. En este sentido, comentar que en Reyes et al. (2017) se considera el estimador $\hat{\sigma}_g$ que se ha obtenido en el Apéndice A mediante técnicas de máxima verosimilitud ponderada.

Por otro lado, ψ_6 se estimará a través de $\hat{\psi}^g_{6,\eta_6}$, que a su vez depende de una ventana η_6. A partir de este momento se procede de manera iterativa, seleccionando nuevas ventanas piloto en base a la expresión (3.12). Siguiendo el procedimiento marcado en el caso clásico (véase Wand y Jones (1995)), una estrategia habitual es parar el proceso iterativo después de dos pasos[1] y estimar las cantidades desconocidas asumiendo que f sigue una distribución normal $N(\mu, \sigma^2)$. En tal caso, $R(f)$ adopta la expresión (3.14) y, si r es par,

$$\psi_{r+2} = \frac{(-1)^{\frac{r+2}{2}}(r+2)!}{(2\sigma)^{r+3}\left(\frac{r+2}{2}\right)!\sqrt{\pi}}, \tag{3.15}$$

donde σ puede estimarse nuevamente mediante alguno de los estimadores propuestos en la Sección 3.1. La estimación iterativa de las cantidades $R(f)$ y ψ_{r+2} permite la construcción de la ventana $\hat{\eta}_4$ en base a la expresión (3.13). A partir de esta ventana se construye $\hat{\psi}^g_{4,\hat{\eta}_4} = \hat{R}_{\hat{\eta}_4}(f'')$, un estimador de $R(f'')$ que, sustituido en la expresión (3.9), conduce a un valor de la ventana $\hat{h}_{\mathrm{R}}$.

Observación 3.3. En Reyes (2015) se prueba que, bajo ciertas condiciones de regularidad — en las cuales se incluye que $\bar{l} = o(\eta^{r+1})$ — $\hat{\psi}^g_{r,\eta_r}$ es un estimador consistente de $\mathbb{E}(f^{(r)}(t))$, lo cual permite garantizar que la ventana $\hat{h}_{\mathrm{R}}$ se aproxima a la ventana óptima conforme el tamaño muestral aumenta.

3.2.2. Propuesta de selector plug-in

En este apartado se propondrá un selector plug-in alternativo al presentado en Reyes et al. (2017). Para ello, nos basaremos en las ideas de Ćwik y Koronacki (1997), en donde se lleva a cabo la construcción de un selector plug-in bidimiensional para el caso clásico de datos no agrupados.

Recordemos que, en el contexto de datos agrupados, los selectores plug-in parten de la expresión

$$h_{\mathrm{AMISE}_g} = \left[\frac{R(L)}{\mu_2(L)^2 R(f'')n}\right]^{\frac{1}{5}}$$

y tratan de sustituir $R(f'')$ por algún estimador conveniente. La nueva propuesta considerará la expresión que adopta $R(f'')$ cuando f se puede expresar como una mixtura normales. De esta forma, este selector plug-in podrá verse como una extensión natural de la regla del pulgar presentada en la Sección 3.1.

Sea f la función de densidad correspondiente a una mixtura de M densidades gaussianas, esto es,

$$f(x) = \sum_{m=1}^{M} \lambda_m \phi_{\mu_m,\sigma_m}(x), \qquad \forall x \in \mathbb{R}$$

donde ϕ_{μ_m,σ_m} denota la densidad de una distribución normal estándar de media μ_m y desviación típica σ_m y donde los $\lambda_m \in [0,1]$ son tales que $\sum_{m=1}^{M} \lambda_m = 1$. En tal caso, $R(f'')$ se puede reescribir como

$$R(f'') = \int \left(f''(x)\right)^2 \, \mathrm{d}x = \int \left[\frac{\partial^2}{\partial x^2} \sum_{m=1}^{M} \lambda_m \phi_{\mu_m,\sigma_m}(x)\right]^2 \, \mathrm{d}x = \int \left[\sum_{m=1}^{M} \frac{\lambda_m}{\sigma_m} \frac{\partial^2}{\partial x^2} \phi\left(\frac{x-\mu_m}{\sigma_m}\right)\right]^2 \, \mathrm{d}x, \tag{3.16}$$

[1] La función `bw.dens.binned`. de la librería `binnednp` de `R` — la cual implementa el selector plug-in de Reyes et al. (2017) y de la que se hablará con más detalle en la Sección 3.2.2 — realiza la estimación de la ventana $\eta_4^{\text{ópt}}$ mediante un proceso iterativo de tres pasos, sustituyendo en el último de ellos las aproximaciones de $R(f)$ y ψ_{10} recogidas en (3.14) y (3.15)

donde ϕ denota la densidad de una distribución normal estándar. Teniendo en cuenta que la derivada de orden p de la función de densidad normal estándar se puede expresar en términos de los polinomios de Hermite[2] de orden p, He_p, de la forma

$$\frac{\partial^p}{\partial x^p}\,\phi(x) = (-1)^p\,\mathrm{He}_p(x)\,\phi(x), \qquad \forall x \in \mathbb{R},$$

y haciendo uso de la regla de la cadena, se llega a que

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\,\phi\left(\frac{x-\mu_m}{\sigma_m}\right) &= -\frac{\partial}{\partial x}\left[\frac{1}{\sigma_m}\mathrm{He}_1\left(\frac{x-\mu_m}{\sigma_m}\right)\phi\left(\frac{x-\mu_m}{\sigma_m}\right)\right] = -\frac{\partial}{\partial x}\left[\frac{x-\mu_m}{\sigma_m^2}\,\phi\left(\frac{x-\mu_m}{\sigma_m}\right)\right] = \\ &= -\frac{1}{\sigma_m^2}\,\phi\left(\frac{x-\mu_m}{\sigma_m}\right) - \left(\frac{x-\mu_m}{\sigma_m^2}\right)\frac{\partial}{\partial x}\phi\left(\frac{x-\mu_m}{\sigma_m}\right) = \\ &= -\frac{1}{\sigma_m^2}\,\phi\left(\frac{x-\mu_m}{\sigma_m}\right) + \left(\frac{x-\mu_m}{\sigma_m^2}\right)^2\phi\left(\frac{x-\mu_m}{\sigma_m}\right) = \frac{(x-\mu_m)^2-\sigma_m^2}{\sigma_m^4}\;\phi\left(\frac{x-\mu_m}{\sigma_m}\right).\end{aligned}$$

Volviendo a la expresión (3.16), resulta

$$R(f'') = \int\left[\sum_{m=1}^{M}\frac{\lambda_m}{\sigma_m^5}\left[(x-\mu_m)^2-\sigma_m^2\right]\;\phi\left(\frac{x-\mu_m}{\sigma_m}\right)\right]^2\,\mathrm{d}x. \tag{3.17}$$

Nótese que, en la práctica, los vectores de parámetros $\boldsymbol{\lambda}' = (\lambda_1,\ldots,\lambda_M)$, $\boldsymbol{\mu}' = (\mu_1,\ldots,\mu_M)$ y $\boldsymbol{\sigma}' = (\sigma_1,\ldots,\sigma_M)$ serán desconocidos y, por tanto, deberán ser estimados, obteniendo un estimador de $R(f'')$ que denotaremos por $\widehat{R_{\mathrm{M}}}(f'')$. En este sentido, dada una muestra $t_1,\ldots,t_k$ con pesos asociados $w_1,\ldots,w_k$, la función de máxima verosimilitud ponderada adopta la forma

$$\mathcal{L}^w(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\sigma}^2) = \prod_{j=1}^{k} f(t_j)^{w_j} = \prod_{j=1}^{k}\left(\sum_{m=1}^{M}\lambda_m\phi_{\mu_m,\sigma_m}(t_j)\right)^{w_j}. \tag{3.18}$$

Para $M > 1$, las estimaciones de $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{\mu}}$ y $\hat{\boldsymbol{\sigma}}$ se obtendrían a través de un algoritmo de tipo esperanza-maximización (EM, véase Dempster et al. (1977)), puesto que en tal caso no se conoce la expresión explícita de dichos estimadores. La ventana plug-in adoptará, por tanto, la expresión

$$\hat{h}_{\mathrm{g,pi}} = \left[\frac{R(L)}{\mu_2(L)^2\widehat{R_{\mathrm{M}}}(f'')\,n}\right]^{\frac{1}{5}}.$$

Observación 3.4. Comentar que la función `bw.dens.binned` de la librería `binnednp` de `R` — siempre que su argumento `plugin.type` tome un valor distinto a `N` y `A` — implementa un selector de ventana fundamentado en la misma idea que la aquí presentada (véase Barreiro-Ures et al. (2019a) y Barreiro-Ures et al. (2019b)). En tal caso, $R(f'')$ se estima asumiendo que f es una mixtura de normales a través de la expresión (3.17). Sin embargo, la estimación de los vectores de parámetros $\boldsymbol{\lambda}' = (\lambda_1,\ldots,\lambda_M)$, $\boldsymbol{\mu}' = (\mu_1,\ldots,\mu_M)$ y $\boldsymbol{\sigma}' = (\sigma_1,\ldots,\sigma_M)$ se realiza a través de la función `Mclust` de la librería `mclust`, la cual incorpora un algoritmo EM no adaptado al caso de datos agrupados (para más información sobre esta función puede consultarse Scrucca et al. (2016) y Fraley et al. (2016)). Por este motivo, en

[2]Se define el polinomio de Hermite de orden p como el polinomio

$$\mathrm{He}_p(x) = (-1)^p\,e^{\frac{x^2}{2}}\,\frac{d^p}{\mathrm{d}x^p}e^{-\frac{x^2}{2}}.$$

De esta forma, es sencillo probar que, para $p \geq 1$, se verifica la propiedad recursiva $\mathrm{He}_{p+1}(x) = x\mathrm{He}_p(x) - p\mathrm{He}_{p-1}(x)$, siendo $\mathrm{He}_0(x) = 1$ y $\mathrm{He}_1(x) = x$.

los estudios de simulación recogidos en el Capítulo 4 no se presentarán los resultados relativos a esta ventana. Por el contrario, se ha llevado a cabo la implementación en `R` del selector plug-in propuesto en esta sección, teniendo en cuenta tanto el criterio de selección del número de mixturas que se comentará a continuación como el *efecto agrupación* a la hora de realizar la estimación de los parámetros, para lo cual se ha hecho uso de un algoritmo de tipo EM adaptado a este nuevo contexto de datos agrupados, cuya implementación se abordará con más detalle en la Sección 4.1.4.

Elección del parámetro M

El selector plug-in que se acaba de presentar asume que la función f es una mixtura de normales. De esta forma, una cuestión relevante reside en la elección del número M de normales que componen dicha mixtura. Una posible elección sería elegir aquel valor de M que minimice algún criterio de información, como es el criterio de información de Akaike (AIC, ver Akaike (1974)), definido según la expresión

$$\mathrm{AIC} = -2\log(\hat{\mathcal{L}}^w) + 2q,$$

donde $\hat{\mathcal{L}}^w$ denota la verosimilitud ponderada del modelo estimado, cuya expresión se recoge en (3.18), tras sustituir los parámetros $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ y $\boldsymbol{\sigma}$ por sus respectivas estimaciones; y donde q denota el número de parámetros considerados.

Como cada distribución normal tiene asociado tres parámetros (media μ, desviación típica σ y su correspondiente peso en la mixtura, λ) y teniendo en cuenta que $\sum_{m=1}^{M} \lambda_m = 1$, entonces es inmediato que $q = 3M - 1$. De esta forma,

$$\mathrm{AIC} = -2\sum_{j=1}^{k} w_j \log\left(\sum_{m=1}^{M} \hat{\lambda}_m \phi_{\hat{\mu}_m,\hat{\sigma}_m}(t_j)\right) + 6M - 2. \tag{3.19}$$

En la práctica, se considerará un cierto rango de valores de M para los cuales se calculará el AIC en base a la expresión (3.19) y se escogerá aquel valor de M que haya conducido a un menor AIC.

3.3. Propuesta de selector de validación cruzada insesgada

Los métodos de validación cruzada afrontan el problema de selección de la ventana desde una perspectiva diferente a la presentada en las secciones anteriores. En lugar de basarse en las expresiones de las ventanas asintóticamente óptimas, el parámetro de suavizado es escogido en base a aproximaciones de las medidas de error. En este sentido, se propone a continuación un selector que tratará de extender el procedimiento clásico de validación cruzada insesgada o de mínimos cuadrados (UCV o LSCV por sus siglas en inglés) — ya estudiado en Rudemo (1982) y Bowman (1984) — al caso de datos agrupados.

El selector de validación cruzada insesgada parte de la expresión del error cuadrático integrado (ISE_g) de $\hat{f}^g_{w,h}$ como estimador de f,

$$\begin{aligned}\mathrm{ISE}_g(w_1,\ldots,w_k;h) &= \int \left(\hat{f}^g_{w,h}(x) - f(x)\right)^2 \,\mathrm{d}x = \\ &= \int \hat{f}^g_{w,h}(x)^2 \,\mathrm{d}x - 2\int \hat{f}^g_{w,h}(x) f(x) \,\mathrm{d}x + \int f(x)^2 \,\mathrm{d}x.\end{aligned} \tag{3.20}$$

Nótese que, aún bajo diseño fijo, (3.20) es una cantidad aleatoria que depende de los pesos muestrales $w_1, \ldots, w_k$ empleados en la construcción de $\hat{f}^g_{w,h}$. Pues bien, la idea reside en seleccionar la ventana h_{ISE_g} que minimice la expresión (3.20). Teniendo en cuenta que el último sumando no depende de h, es claro que

$$h_{\mathrm{ISE}_g} = \arg\min_{h>0} \left[R\left(\hat{f}^g_{w,h}\right) - 2\int \hat{f}^g_{w,h}(x) f(x)\ \mathrm{d}x \right]. \tag{3.21}$$

Dada una muestra de datos agrupados, el primer sumando de (3.21) es conocido, puesto que no depende de la densidad teórica f. Sin embargo, no sucede lo mismo con el segundo sumando, que sí depende de f y, por tanto, deberá ser aproximado. Teniendo en cuenta que

$$\int \hat{f}^g_{w,h}(x) f(x)\ \mathrm{d}x = \mathbb{E}\left[\hat{f}^g_{w,h}(x)\right],$$

entonces un estimador de esta cantidad viene dado por $\sum_{j=1}^k w_j \hat{f}^g_{w,h}(t_j)$, esto es, por la media muestral de los valores $\hat{f}^g_{w,h}(t_1), \ldots, \hat{f}^g_{w,h}(t_k)$ ponderada por los pesos $w_1, \ldots, w_k$. Sin embargo, con el fin de reducir la dependencia respecto de la muestra — y siguiendo las ideas clásicas de Rudemo (1982) y Bowman (1984) — consideraremos como estimador alternativo de $\int \hat{f}^g_{w,h}(x) f(x)\ \mathrm{d}x$ a

$$\sum_{j=1}^k w_j \hat{f}^{g,-j}_{w,h}(t_j), \tag{3.22}$$

donde

$$\hat{f}^{g,-j}_{w,h}(x) = \frac{n}{n-n_j} \sum_{\substack{i=1 \\ i\neq j}}^k w_i L_h(x-t_i) = \frac{1}{1-w_j} \sum_{\substack{i=1 \\ i\neq j}}^k w_i L_h(x-t_i) \tag{3.23}$$

denota al estimador tipo núcleo construido en base a la muestra $t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_k$, en la cual se excluyen los n_j valores correspondientes al dato t_j, y que denotaremos por estimador *leave-one-group-out*.

Observación 3.5. Con el fin de aproximar el sesgo de (3.22) como estimador de $\int \hat{f}^g_{w,h}(x) f(x)\ \mathrm{d}x$, en el Apéndice C se presenta un pequeño estudio de simulación que permite aproximar las cantidades $\mathbb{E}[\sum_{j=1}^k w_j \hat{f}^{g,-j}_{n,h}(t_j)]$ y $\mathbb{E}[\int \hat{f}^g_{w,h}(x) f(x)\ \mathrm{d}x]$ para diferentes tamaños muestrales y modelos teóricos de referencia. Este estudio parece mostrar que (3.22) es un estimador asintóticamente insesgado de $\mathbb{E}[\int \hat{f}^g_{w,h}(x) f(x)\ \mathrm{d}x]$, aunque dicha convergencia ha resultado ser bastante lenta con respecto al tamaño muestral n.

De esta forma, el selector de validación cruzada insesgada escogerá como ventana aquel valor $\hat{h}_{\mathrm{UCV}_g}$ que minimice la llamada función de validación cruzada insesgada, definida como

$$\mathrm{UCV}_g(h) = R\left(\hat{f}^g_{w,h}\right) - 2\sum_{j=1}^k w_j \hat{f}^{g,-j}_{w,h}(t_j)$$

y que que constituye una aproximación de (3.21). Esto es,

$$\hat{h}_{\mathrm{UCV}_g} = \arg\min_{h>0} \left[R\left(\hat{f}^g_{w,h}\right) - 2\sum_{j=1}^k w_j \hat{f}^{g,-j}_{w,h}(t_j) \right]. \tag{3.24}$$

Una ventaja de este selector es que su puesta en práctica no requiere de conocer las cantidades $n_1, \ldots, n_k$, constituyendo así el único método de selección de ventana — de los aquí presentados — que puede ser empleado en aquellos contextos en los que únicamente se dispone de las proporciones muestrales $w_1, \ldots, w_k$ (*general grouped data case*).

Por otro lado, otra gran diferencia respecto de los selectores anteriormente presentados es que requiere de la optimización numérica para su obtención. Dicha optimización puede no ser sencilla, ya sea por la presencia de varios mínimos locales o por la gran dependencia que presenta la función objetivo respecto de las proporciones muestrales. Como consecuencia, los algoritmos de optimización pueden quedarse "atrapados" en soluciones espúreas[3].

Observación 3.6. Al igual que se ha hecho con el selector UCV, también se podría extender el procedimiento de validación cruzada sesgada (BCV por sus siglas en inglés) al contexto de datos agrupados. Este selector parte de la versión asintótica del MISE_g (AMISE_g) que, como ya se ha comentado anteriormente, presenta la misma expresión que en el caso de datos no agrupados (véase Wand y Jones (1995)),

$$\text{AMISE}_g\left(\hat{f}^g_{w,h}\right) = \frac{R(L)}{nh} + \frac{1}{4}\, h^4 \mu_2(L)^2 R(f''). \tag{3.25}$$

A continuación, y tras sustituir $R(f'')$ por un estimador adecuado, se procede a la minimización de (3.25), lo cual permite obtener la ventana $\hat{h}_{\text{BCV}_g}$. En el contexto clásico de datos no agrupados, Scott y Terrell (1987) proponen como estimador de $R(f'')$ a una modificación de $R(\hat{f}''_{n,h})$ diseñada para reducir su sesgo, siendo $\hat{f}_{n,h}$ el estimador tipo núcleo de Parzen-Rosenblatt. De esta forma, la extensión de este selector al caso de datos agrupados implicaría el cálculo de $\mathbb{E}[R(\hat{f}^{g\,\prime\prime}_{w,h})]$, siendo $\hat{f}^g_{w,h}$ el estimador tipo núcleo propuesto por Cao et al. (2011), para lo cual bastaría seguir un procedimiento análogo al indicado en la Sección 9.2 de Scott y Terrell (1987). Sin embargo, debido a la complejidad de este procedimiento, y teniendo en cuenta los malos resultados que se suelen obtener con el selector de validación cruzada insesgada anteriormente presentado — de los cuales se hablará con más detalle en el Capítulo 4 — dicho desarrollo no se abordará en este trabajo.

3.4. Selectores bootstrap

En esta sección se presentarán los selectores bootstrap propuestos en Jang y Loh (2010) y en Reyes et al. (2017). El primero de ellos combina técnicas de validación cruzada con un suavizado bootstrap llevado a cabo mediante técnicas Monte Carlo, mientras que el segundo propone un método que no requiere, en la práctica, de generar ninguna muestra bootstrap, reduciendo consecuentemente el coste computacional asociado. En ambos casos se abordará el problema de la elección de una ventana inicial. Finalmente, se propondrá también una modificación del selector considerado en Jang y Loh (2010).

[3]Idealmente, se emplearían algoritmos que converjan al óptimo global del problema. Sin embargo, es habitual que estos se diseñen con objetivos menos ambiciosos, de tal manera que terminen cuando se alcancen puntos de un cierto conjunto deseable — conformado, por ejemplo, por los óptimos locales del problema — en cuyo caso podría no llegarse al valor de $\hat{h}_{\text{UCV}_g}$. Además, la implementación práctica de este tipo de algoritmos (método de bisección, método de la sección áurea, método de Newton...) requiere de especificar un intervalo inicial en el cual se llevará a cabo la optimización. En este sentido, elegir intervalos de longitud excesivamente grande podría conducir a soluciones que se encuentran lejos del óptimo, mientras que intervalos demasiado pequeños podrían no contener al óptimo global del problema. En el Capítulo 4 se presentarán algunos detalles referidos a la implementación de este selector en `R`.

3.4.1. Selector bootstrap de Jang y Loh (2010)

Sean $t_1, \ldots, t_k$ los puntos medios de los intervalos considerados, con proporciones muestrales respectivas $w_1, \ldots, w_k$ y frecuencias absolutas (que asumiremos conocidas) $n_1, \ldots, n_k$, de manera que $n = \sum_{j=1}^{k} n_j$. Sean $y_0, \ldots, y_k$ los extremos de dichos intervalos, cuyas longitudes vienen dadas por $l_1, \ldots, l_k$. En lo que sigue, denotaremos por $(T_1, \ldots, T_n) = (t_1, \overset{n_1}{\ldots}, t_1, \ldots, t_k, \overset{n_k}{\ldots}, t_k)$ a la muestra de datos agrupados conformada por los puntos medios repetidos tantas veces como observaciones se hayan recogido en cada intervalo. El proceso de estimación propuesto en Jang y Loh (2010) involucra los siguientes pasos:

1. En cada intervalo $j = 1, \ldots, k$ se genera ruido en base a una distribución uniforme en $[-l_j/2, l_j/2]$ y se añade a las n_j observaciones recogidas en dicho intervalo[4], de manera que estas ya no se superponen. Denotemos por $T_1^U, \ldots, T_n^U$ a la nueva muestra, conformada por datos no agrupados.

2. Usar el selector de validación cruzada insesgada clásico para obtener la ventana óptima en base a la muestra $T_1^U, \ldots, T_n^U$.

3. Repetir los pasos 1 y 2 un número elevado de veces y obtener $\hat{h}_{\text{in}}$ como el valor medio de las ventanas óptimas obtenidas (en Jang y Loh (2010) proponen considerar un total de 1000 repeticiones). Computar un estimador tipo núcleo inicial, $\hat{f}_{n,\hat{h}_{\text{in}}}$, a través del clásico estimador de Parzen-Rosenblatt construido sobre la muestra original $T_1, \ldots, T_n$ y usando a $\hat{h}_{\text{in}}$ como ventana,

$$\hat{f}_{n,\hat{h}_{\text{in}}}(x) = \frac{1}{n\hat{h}_{\text{in}}} \sum_{i=1}^{n} K\left(\frac{x - T_i}{\hat{h}_{\text{in}}}\right).$$

4. A partir de la densidad $\hat{f}_{n,\hat{h}_{\text{in}}}$, generar B remuestras bootstrap suavizadas[5] $\{Y_1^{*(b)}, \ldots, Y_n^{*(b)}\}_{b=1}^{B}$, todas ellas conformadas por datos no agrupados. Sobre cada remuestra, construir la versión bootstrap del estimador de Parzen-Rosenblatt como función de un parámetro de suavizado $h > 0$,

$$\hat{f}_{n,h}^{*(b)}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - Y_i^{*(b)}}{h}\right), \qquad b = 1, \ldots, B.$$

5. Construir el BMISE, una aproximación de la versión bootstrap del MISE definido como

$$\text{BMISE}(h) = \frac{1}{B} \sum_{b=1}^{B} \int \left(\hat{f}_{n,h}^{*(b)}(x; h) - \hat{f}_{n,\hat{h}_{\text{in}}}(x)\right)^2 \, \mathrm{d}x. \tag{3.26}$$

   Se tomará como ventana aquel valor $\hat{h}_{\text{S}}$ que minimice (3.26), esto es,

$$\hat{h}_{\text{S}} = \underset{h>0}{\text{argmin}} \, \text{BMISE}\,(h)\,. \tag{3.27}$$

[4] Aunque en Jang y Loh (2010) se describe este procedimiento tal y como aquí se ha mostrado, en el Script de `R` adjuntado en su material complementario — en el cual se implementa, entre otras cosas, su selector bootstrap — emplean una versión alternativa al paso 1, donde las muestras $T_1^U, \ldots, T_n^U$ se obtienen tras aplicar el ruido uniforme sobre una muestra con reemplazamiento de la muestra original $T_1, \ldots, T_n$.

[5] Este paso puede llevarse a cabo a través del llamado bootstrap suavizado (véase Efron (1982)). Bajo este plan de remuestreo, cada una de las remuestras $Y_1^*, \ldots, Y_n^*$ se construye según el siguiente esquema:

a) Generar $T_1^*, \ldots, T_n^*$ aplicando bootstrap uniforme sobre $T_1, \ldots, T_n$.

b) Simular z_i a partir de la densidad K. Si K es el núcleo Gaussiano, entonces $z_i \sim N(0, 1)$, $i = 1, \ldots, n$.

c) Definir la remuestra suavizada como $Y_i^* = T_i^* + z_i \hat{h}_{\text{in}}$, para cada $i = 1, \ldots, n$.

En resumen, este selector trata de obtener la ventana $\hat{h}_{\mathrm{S}}$ a través de un suavizado bootstrap que emplea a $\hat{h}_{\mathrm{in}}$ como ventana piloto. Esta ventana inicial se obtiene mediante un procedimiento de validación cruzada aplicado sobre una muestra artificial de datos no agrupados procedente de transformar la muestra inicial mediante ruido aleatorio. Comentar que en Jang y Loh (2010) se añade un primer paso adicional consistente en duplicar la muestra inicial, reflejando cada dato respecto del origen, con el fin de reducir el sesgo en la frontera (puesto que allí el objetivo residía en estimar $f(0)$ en base a muestras de observaciones positivas). En nuestro caso, este paso no es realmente necesario, puesto que las variables a considerar no tienen por qué presentar un soporte acotado.

Modificación del selector de Jang y Loh

En el paso 3 del procedimiento anteriormente presentado, Jang y Loh (2010) proponen considerar como estimador inicial — densidad de referencia en el contexto bootstrap — al estimador de Parzen-Rosenblatt construido sobre la muestra $T_1, \ldots, T_n$. Sin embargo, recordemos que esta muestra está conformada por datos agrupados[6], no siendo aconsejable en tal caso el empleo del estimador tipo núcleo clásico, puesto que — como se ha comentado en anteriores ocasiones — no permite incorporar dicho *factor agrupación* al proceso de estimación.

Además, la construcción de la ventana inicial $\hat{h}_{\mathrm{in}}$ como una media de ventanas obtenidas mediante validación cruzada puede presentar problemas en la práctica, derivados de la conocida inestabilidad de este selector en el caso de datos no agrupados. Finalmente, el problema de la selección de la ventana inicial usando la muestra artificial $T_1^U, \ldots, T_n^U$ puede diferir considerablemente del que se tendría con la muestra datos agrupados originales, $t_1, \ldots, t_k$, sobre todo cuando el nivel de agrupación es elevado.

Todo esto podría provocar una selección inapropiada del parámetro de suavizado y, por este motivo, se propone a continuación un procedimiento alternativo:

1. Construir la ventana inicial $\hat{h}_{\mathrm{in}}$ en base a la muestra original $t_1, \ldots, t_k$ a través del selector de validación cruzada insesgada adaptado al caso de datos agrupados (ver Sección 3.3),

$$\hat{h}_{\mathrm{in}} = \arg\min_{h>0} \mathrm{UCV}_g(h).$$

2. En base a la ventana $\hat{h}_{\mathrm{in}}$, computar el estimador tipo núcleo de Cao et al. (2011), $\hat{f}^g_{w,\hat{h}_{\mathrm{in}}}$, construido sobre la muestra de datos agrupados $t_1, \ldots, t_k$,

$$\hat{f}^g_{w,\hat{h}_{\mathrm{in}}}(x) = \frac{1}{\hat{h}_{\mathrm{in}}} \sum_{j=1}^{k} w_j L\left(\frac{x - t_j}{\hat{h}_{\mathrm{in}}}\right),$$

   donde w_j denota la proporción de observaciones recogidas en el $j-$ésimo intervalo.

3. Realizar los pasos 4 y 5 propuestos por Jang y Loh, considerando como estimador inicial a $\hat{f}^g_{w,\hat{h}_{\mathrm{in}}}$, obteniendo así una versión alternativa de (3.27) que denotaremos por $\hat{h}_{\mathrm{S}^*}$.

Será precisamente esta nueva modificación la que se considerará en los estudios de simulación presentados en el Capítulo 4.

[6]En efecto, recordemos que $T_1, \ldots, T_n$ representa la muestra constituida por los puntos medios de los intervalos, $t_1, \ldots, t_k$, repetidos tantas veces como observaciones recogidas en cada uno de ellos, $n_1, \ldots, n_k$. De esta forma, se tendrá que $T_1 = \cdots = T_{n_1}$, $T_{n_1+1} = \cdots = T_{n_1+n_2}$, $\ldots, T_{n-n_k+1} = \cdots = T_n$. Por tanto, si el nivel de agrupación es elevado, el estimador tipo núcleo clásico podría presentar modas artificiales, tal y como sucedía en el ejemplo recogido en la Figura 1.1(b) de la Sección 1.2.2, lo cual conduciría a una interpretación errónea de la densidad teórica.

3.4.2. Selector bootstrap de Reyes *et al.* (2017)

En este apartado se presentará el selector de ventana tipo bootstrap propuesto en Reyes et al. (2017) que, a diferencia del considerado por Jang y Loh (2010), no requerirá de simulación de Monte Carlo, permitiendo así reducir el coste computacional requerido en su implementación.

Sea $t_1, \ldots, t_k$ una muestra de datos agrupados, con proporciones muestrales respectivas $w_1, \ldots, w_k$ y frecuencias absolutas (conocidas) $n_1, \ldots, n_k$, de manera que $n = \sum_{j=1}^{k} n_j$. Denotemos por $y_0, \ldots, y_k$ a los extremos de los intervalos considerados y sean $l_1, \ldots, l_k$ sus respectivas longitudes. En Reyes et al. (2017) se propone considerar el siguiente esquema:

1. Construir el estimador de la densidad de Cao et al. (2011) sobre la muestra $t_1, \ldots, t_k$ en base a una ventana piloto ζ, cuya elección se abordará más adelante,

$$\hat{f}^{g}_{w,\zeta}(x) = \sum_{j=1}^{k} w_j L_\zeta(x - t_j). \tag{3.28}$$

2. A partir de la densidad $\hat{f}^{g}_{n,\zeta}$, generar una muestra bootstrap $X_1^*, \ldots, X_n^*$. En base a ella, construir el análogo bootstrap del estimador de la densidad como función de un parámetro de suavizado $h > 0$,

$$\hat{f}^{g*}_{w,h}(x; h) = \sum_{j=1}^{k} w_j^* L_h(x - t_j),$$

donde los pesos bootstrap se definen como $w_j^* = F_n^*(y_j^-) - F_n^*(y_{j-1}^-)$, para cada $j = 1, \ldots, k$, y donde $F_n^*(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i^* \leq y)$ denota la función de distribución empírica asociada a la remuestra $X_1^*, \ldots, X_n^*$.

3. Definir la versión bootstrap del error cuadrático medio integrado, MISE_g^*, como

$$\mathrm{MISE}_g^*(h) = \mathbb{E}^* \left[\int \left(\hat{f}^{g*}_{w,h}(x; h) - \hat{f}^{g}_{w,\zeta}(x) \right)^2 \, \mathrm{d}x \right], \tag{3.29}$$

donde $\mathbb{E}^*$ denota la esperanza boostrap (con respecto a $\hat{f}^{g}_{w,\zeta}$). Ahora bien, en Reyes et al. (2017) se prueba que (3.29) admite una expresión cerrada de la forma

$$\mathrm{MISE}_g^*(h) = \frac{n-1}{n} \sum_{i=1}^{k}\sum_{j=1}^{k} w_i^\zeta w_j^\zeta (L * L)_h(t_i - t_j) - 2\sum_{i=1}^{k}\sum_{j=1}^{k} w_i^\zeta w_j \times (L_h * L_\zeta)(t_i - t_j) + \\ + \sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j (L * L)_\zeta(t_i - t_j) + \frac{R(L)}{nh}, \tag{3.30}$$

siendo $w_i^\zeta = \mathbb{E}^*(w_i^*)$ y donde la operación $*$ denota la convolución entre dos funciones, de manera que $(L*L)_h(x-y) = \int L_h(v-x)L_h(v-y)\, \mathrm{d}v$. En este sentido, si denotamos $\mathbb{L}(x) = \int_{-\infty}^{x} L(v)\, \mathrm{d}v$, se puede probar que

$$w_i^\zeta = \sum_{j=1}^{k} w_j \left[\mathbb{L}\left(\frac{y_i - t_j}{\zeta} \right) - \mathbb{L}\left(\frac{y_{i-1} - t_j}{\zeta} \right) \right].$$

4. Seleccionar como ventana $\hat{h}_{\mathrm{boot}}$ aquel valor de h que minimice (3.30), esto es,

$$\hat{h}_{\mathrm{boot}} = \underset{h>0}{\mathrm{argmin}}\, \mathrm{MISE}_g^*(h).$$

Observación 3.7. Nótese que la expresión (3.30) permite una evaluación directa del MISE_g^* sobre una malla de valores de h sin necesidad de emplear simulación de Monte Carlo. De esta forma, en la práctica, el selector bootstrap de Reyes et al. (2017) no requiere de generar ninguna remuestra bootstrap, a diferencia de lo que sucede con la mayoría de procedimientos de este tipo, pudiendo prescindir del paso 2 anteriormente presentado.

Comentar finalmente que en Reyes et al. (2017) se llevan a cabo varios estudios de simulación que permiten concluir que, en general, la ventana obtenida mediante este selector bootstrap presenta un comportamiento notablemente mejor que la ventana plug-in propuesta allí mismo (y recogida en la Sección 3.2.1) y, en todos los casos, mejor que la propuesta por Jang y Loh (2010), presentando, además, tiempos de computación considerablemente inferiores a este último. Estas cuestiones se abordarán con más detalle en el Capítulo 4.

Elección de la ventana piloto

Se ha visto que el selector bootstrap de Reyes et al. (2017) depende de una ventana inicial ζ, necesaria para la construcción del estimador tipo núcleo inicial (3.28). Es conocido que en el contexto clásico de datos no agrupados, los selectores tipo bootstrap realizan la estimación de la ventana inicial de tal manera que permitan garantizar buenas propiedades del funcional $R(f'')$, más que de la propia densidad f (véase Cao (1993)). En este sentido, una opción podría ser obtener la ventana inicial ζ a través del mismo procedimiento presentado para el selector plug-in de Reyes et al. (2017) (ver Sección 3.2.1). Sin embargo, la ventana obtenida en (3.13) no tiene en cuenta el tamaño muestral, parámetro que presenta una gran influencia en los procedimientos bootstrap. De hecho, en Reyes et al. (2017) se comenta que, tras haber realizado algunas pruebas, las ventanas obtenidas usando a $\eta_4^{\text{ópt}}$ como ventana inicial suelen ser demasiado grandes.

Por este motivo, en Reyes et al. (2017) se presenta un método alternativo para la selección de ζ, basado en la idea de partir de una ventana *considerablemente buena* para el caso clásico de datos no agrupados y corregirla adecuadamente de manera que pueda ser empleada en el contexto de datos agrupados.

En este sentido — y con el objetivo de encontrar una ventana óptima para el caso de datos no agrupados que permita aproximar correctamente $R(f'')$ — en Cao (1990) se considera el estimador

$$\tilde{R}_\zeta(f'') = \frac{1}{n^2\zeta^6}\sum_{i\neq j}\int K''\left(\frac{x-X_i}{\zeta}\right)K''\left(\frac{x-X_j}{\zeta}\right)\,\mathrm{d}x.$$

Ahora bien, en el Apéndice E.3 de Reyes (2015) se prueba que la ventana ζ que minimiza la cantidad $\mathbb{E}\{[\tilde{R}_\zeta(f'') - R(f'')]^2\}$ viene dada por la expresión

$$\zeta_{\text{ópt}} = \left[\frac{9R(f)R(K''*K'')n^{-2}}{4\mu_2(K)^2R(f''')^2}\right]^{\frac{1}{13}}. \tag{3.31}$$

Para obtener una expresión sencilla de $\zeta_{\text{ópt}}$, si se asume que $f \sim N(\mu,\sigma^2)$ y se considera una función núcleo K gaussiana, es sencillo comprobar que (3.31) se puede reescribir como

$$\zeta_{\text{ópt}} = \left(\frac{11}{200\sqrt{2}}\sigma^{13}n^{-2}\right)^{\frac{1}{13}} \approx 0.78\sigma n^{-\frac{2}{13}},$$

obteniendo así una ventana óptima para el caso clásico de datos no agrupados. Naturalmente, en la práctica σ debe ser sustituido por un estimador adecuado como, por ejemplo, la cuasi-desviación típica.

Basándose en la ventana $\zeta_{\text{ópt}}$ y tras realizar varios estudios de simulación (que pueden consultarse en el Apéndice E.4 de Reyes (2015)), en Reyes et al. (2017) se propone emplear como ventana piloto en el caso de datos agrupados a

$$\zeta_g \approx \begin{cases} 0.78\zeta_{\text{ópt}}, & \text{si} \quad \dfrac{\bar{l}}{r} \leq 0.15, \\ \zeta_{\text{ópt}}\left(4\dfrac{\bar{l}}{r} + 0.4\right), & \text{en otro caso,} \end{cases}$$

si $n < 150$ y

$$\zeta_g \approx \begin{cases} \zeta_{\text{ópt}}, & \text{si} \quad \dfrac{\bar{l}}{r} \leq 0.075, \\ \zeta_{\text{ópt}}\left(7.2\dfrac{\bar{l}}{r} + 0.46\right), & \text{en otro caso,} \end{cases}$$

si $n \geq 150$, donde r denota el rango de la muestra considerada.

3.5. Resumen de nuevas aportaciones

En esta última sección se recoge, a modo de resumen, una breve recopilación de las nuevas aportaciones presentadas a lo largo del capítulo.

- **Regla del pulgar:** en la Sección 3.1 se ha propuesto una extensión natural de la clásica regla de pulgar de Silverman (Silverman (1986)) al contexto de datos agrupados, para lo cual se ha hecho uso de los estimadores de máxima verosimilitud ponderada de los parámetros de una distribución normal (obtenidos previamente en el Apéndice A). En la Sección 3.1.1 se ha abordado también una modificación de este selector siguiendo las ideas clásicas recogidas en la Sección 3.2.1 de Wand y Jones (1995), la cual ha requerido de extender el concepto de rango intercuantílico al contexto de datos agrupados.

- **Selectores plug-in:** en el Apéndice B se recoge una breve discusión sobre la elección de la ventana piloto empleada por el selector plug-in de Reyes et al. (2017) que ha permitido comprobar que, bajo ciertas condiciones de regularidad, la ventana (3.12) es asintóticamente óptima, en el sentido de que minimiza el error cuadrático medio asintótico de $\hat{\psi}^g_{r,\eta_r}$ como estimador de ψ_r, lo cual confirma la conjetura de Reyes et al. (2017) para el caso particular de funciones núcleo de orden dos. Asimismo, en la Sección 3.2.2 se aborda la construcción de un selector plug-in alternativo basado en las ideas de Ćwik y Koronacki (1997).

- **Selector de validación cruzada insesgada:** en la Sección 3.21 se recoge una propuesta de selector de validación cruzada insesgada, obtenido de extender las ideas clásicas de este selector (Rudemo (1982) y Bowman (1984)) al contexto de datos agrupados, el cual constituye el único método de selección de ventana aquí presentado que puede ser empleado en contextos en los que únicamente se conocen las porporciones muestrales $w_1, \ldots, w_k$. En esta línea, en el Apéndice C se presenta un breve estudio de simulación que permite aproximar el sesgo del *leave-one-group-out* como estimador del segundo sumando de la función de validación cruzada, presentando, además, una posible alternativa a este estimador. De igual manera, se comentan las dificultades que surgirían al extender el selector de validación cruzada sesgada de Scott y Terrell (1987) a este nuevo contexto.

- **Selectores bootstrap:** en la Sección 3.4.1 se propone una modificación del selector bootstrap de Jang y Loh (2010) que intenta solventar ciertos problemas relacionados con la elección de la ventana inicial.

Finalmente, se ha llevado a cabo la implementación en `R` de todas estas nuevas propuestas — así como de las ya existentes — y, en el Capítulo, 4 se presentarán los resultados relativos a un estudio de simulación que permitirá comparar, de manera cuantitativa, el comportamiento de las mismas.

Capítulo 4

Estudio de simulación

En este capítulo se abordará el estudio y análisis del estimador de la densidad de Cao et al. (2011) desde un punto de vista práctico. Para ello, tras una primera sección introductoria en la que se presentarán algunos de los aspectos de interés relacionados con su implementación, se procederá a ilustrar su comportamiento sobre muestras aisladas — de datos agrupados — procedentes de diferentes modelos teóricos con distintas características de interés, haciendo especial énfasis en el posible efecto que sobre él tenga la elección del parámetro de ventana, para lo cual consideraremos los selectores presentados en la Sección 3.

Finalmente, en la última sección del capítulo se ofrecerá una comparativa de dichos selectores en base al MISE_g — que será aproximado vía Monte Carlo — considerando diferentes tamaños muestrales. En todos los casos se hará uso del software estadístico `R` 4.0.3 (R Core Team (2020)).

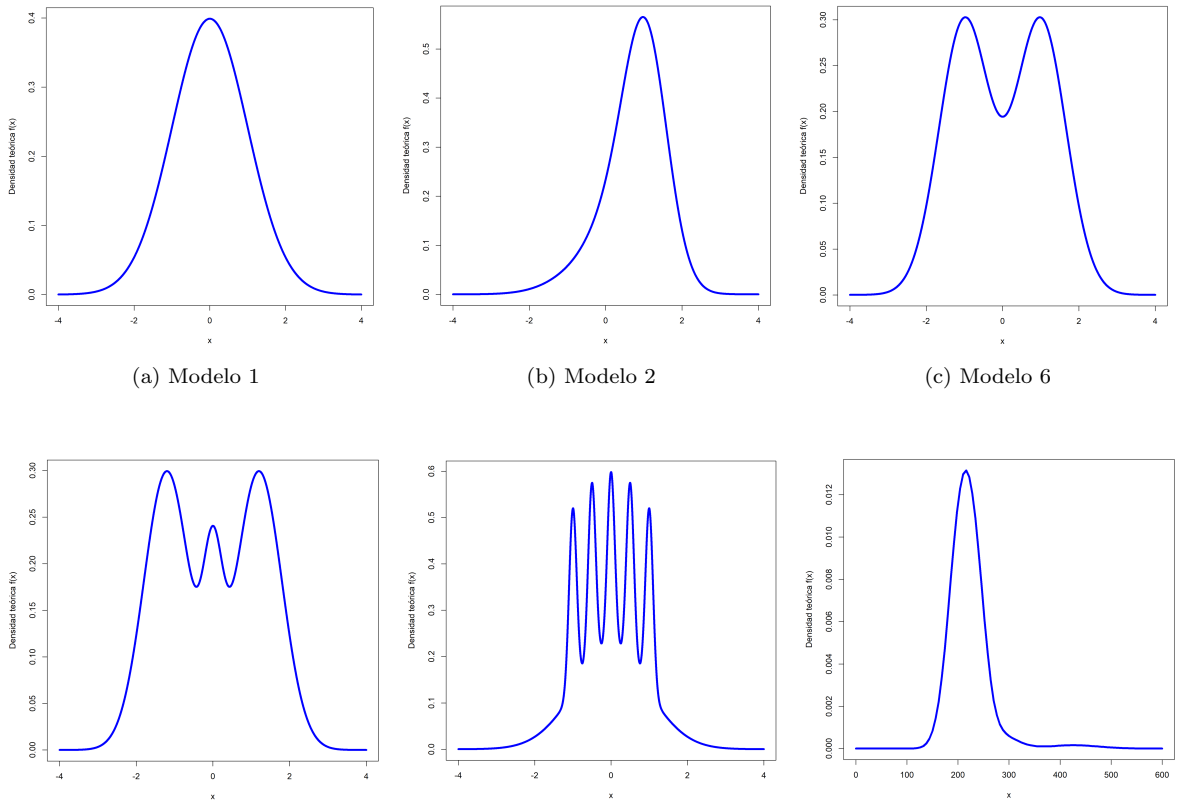
4.1. Detalles de la simulación

En esta primera sección se presentarán algunos aspectos de interés que facilitarán la comprensión de los experimentos computacionales que se llevarán a cabo en este capítulo, así como los modelos teóricos que en ellos se considerarán. Los detalles referidos a la implementación de los diferentes selectores de ventana en `R` se abordarán en las Secciones 4.1.3 y 4.1.4.

4.1.1. Modelos teóricos de referencia

En todos los estudios de simulación aquí presentados se considerarán como densidades teóricas de referencia a los Modelos 1, 2, 6, 9 y 10 introducidos en Marron y Wand (1992), a los cuales también se les añadirá el modelo considerado en Reyes et al. (2016) y Reyes et al. (2017). Todos estos modelos, representados en la Figura 4.1, se corresponden con mixturas de distribuciones normales.

Mientras que el Modelo 1 se corresponde con la densidad normal estándar — una densidad unimodal y simétrica respecto del origen — el Modelo 2 constituye una densidad unimodal pero asimétrica (ligeramente desplazada hacia la derecha), fruto de una mixtura de tres densidades normales. Por el contrario, los Modelos 6, 9 y 10 corresponden a densidades multimodales, con dos, tres y cinco modas, respectivamente (siendo estos modelos mixturas de 2, 3 y 6 distribuciones normales, respectivamente).



(a) Modelo 1 (b) Modelo 2 (c) Modelo 6

(d) Modelo 9 (e) Modelo 10 (f) Modelo R17

Figura 4.1: Densidades teóricas de cada uno de los seis modelos que se considerarán en este capítulo. Obsérvese que la escala empleada en la representación del modelo R17 es diferente a la usada en las demás densidades.

Este último modelo, que presenta características ciertamente poco deseables — como es la presencia de cinco modas distintas, todas ellas muy próximas entre sí — ha sido escogido para mostrar, de alguna manera, las posibles limitaciones del estimador de Cao et al. (2011).

Finalmente, el modelo que se considera en los artículos de Reyes et al. (2016) y Reyes et al. (2017) — representado en la última gráfica de la Figura 4.1 y que a partir de ahora denotaremos por Modelo R17 — se encuentra constituido por una mixtura de cuatro densidades normales. Como se observa en la Figura 4.1(f), a pesar de tratarse de una mixtura de cuatro densidades normales, el Modelo R17 presenta, en principio, características favorables para su estimación, como es un carácter que — prácticamente — se puede asumir unimodal (puesto que la segunda moda, situada en la cola derecha, presenta una magnitud casi despreciable) y una asimetría no demasiado marcada. Destacar además su comportamiento leptocúrtico, que provoca que haya una gran concentración de probabilidad en un pequeño entorno de su primera moda. En este sentido, el Modelo R17 podría considerarse como una extensión del Modelo 1 en la cual se han generado algunos datos atípicos (salvando, claro está, el evidente factor de escala que diferencia a ambos).

En la Tabla 4.1 se recogen, a modo de resumen, las expresiones analíticas que conforman los seis modelos considerados.

Modelo	Densidad
1 (Normal estándar)	$N(0,1)$
2 (Unimodal asimétrica)	$\frac{1}{5}\,N(0,1) + \frac{1}{5}\,N\Big(\frac{1}{2}, \big(\frac{2}{3}\big)^2\Big) + \frac{3}{5}\,N\Big(\frac{13}{12}, \big(\frac{5}{9}\big)^2\Big)$
6 (Bimodal)	$\frac{1}{2}\,N\Big(-1, \big(\frac{2}{3}\big)^2\Big) + \frac{1}{2}\,N\Big(1, \big(\frac{2}{3}\big)^2\Big)$
9 (Trimodal)	$\frac{9}{20}\,N\Big(-\frac{6}{5}, \big(\frac{3}{5}\big)^2\Big) + \frac{9}{20}\,N\Big(\frac{6}{5}, \big(\frac{3}{5}\big)^2\Big) + \frac{1}{10}\,N\Big(0, \big(\frac{1}{4}\big)^2\Big)$
10 (Cinco modas)	$\frac{1}{2}\,N(0,1) + \sum_{i=0}^{4} \frac{1}{10}\,N\Big(\frac{i}{2} - 1, \big(\frac{1}{10}\big)^2\Big)$
R17 (Unimodal asimétrica)	$0.7\,N\big(207, 25^2\big) + 0.22\,N\big(237, 20^2\big) + 0.06\,N\big(277, 35^2\big) + 0.02\,N\big(427, 50^2\big)$

Tabla 4.1: Expresiones analíticas de cada uno de los seis modelos teóricos que se considerarán en este capítulo. Los cinco primeros se corresponden con densidades extraídas de Marron y Wand (1992), mientras que el último se corresponde con la mixtura de cuatro normales considerada en los estudios de simulación que se presentan en Reyes et al. (2016) y Reyes et al. (2017).

4.1.2. Esquema de agrupación de muestras

Naturalmente, la puesta en práctica del estimador de Cao et al. (2011) requerirá de la generación de muestras de datos agrupados. En este sentido, partiremos de muestras continuas — esto es, conformadas por datos no agrupados — generadas a partir de los diferentes modelos teóricos considerados, para, a continuación, aplicarles un cierto mecanismo de agrupación. Pues bien, en todo momento se considerará el esquema de agrupación empleado en los estudios de simulación de Reyes et al. (2016) y Reyes et al. (2017), recogido a continuación:

1. Fijadas unas constantes positivas C, D, α y β, se define $\bar{l} = \bar{l}_n = Cn^{-\alpha}$ y $a_n = Dn^{-\beta}$.
2. Se considera un conjunto base de cinco intervalos $[y_{j-1}, y_j)$ de longitudes l_j, con $j = 1, \ldots, 5$, de tal forma que

$$l_1 = \bar{l} - 4a_n, \quad l_2 = \bar{l} + 0.5a_n, \quad l_3 = \bar{l} - 1.5a_n, \quad l_4 = \bar{l} + 3a_n, \quad l_5 = \bar{l} + 2a_n.$$

   Este conjunto base de intervalos se irá repitiendo[1] hasta cubrir todo el rango de la muestra generada, de manera que, para $j > 5$,

$$l_j = l_{(j-1) \text{ mód } 5+1},$$

   donde mód denota la operación de módulo.

Consideraremos además el escenario recogido en Reyes et al. (2017), en el cual se asume que $n^{\frac{1}{5}}\bar{l} \to 0$. Nótese que esto es equivalente a pedir que $\alpha > \frac{1}{5}$, ya que

$$n^{\frac{1}{5}}\bar{l} = n^{\frac{1}{5}}\left(Cn^{-\alpha}\right) = Cn^{\frac{1}{5}-\alpha} \to 0 \iff \tfrac{1}{5} - \alpha < 0 \iff \alpha > \tfrac{1}{5}.$$

[1] En la práctica se tomará $y_0 = X_{(1)}$, siendo $X_{(1)}$ el mínimo de la muestra continua $X_1, \ldots, X_n$. Los demás extremos se definirán en base a y_0 y a las longitudes $l_1, \ldots, l_k$.

De esta manera, la condición $\bar{l} = o(h)$ estará garantizada para todas las ventanas $h > 0$ tales que $h = O(n^{-1/5})$, que es, precisamente, el orden de las ventanas asintóticamente óptimas, tanto en el caso clásico como en el caso de datos agrupados (y, por tanto, también el orden de las ventanas $\hat{h}_{\mathrm{g,rp}}$, $\hat{h}_{\mathrm{R}}$ y $\hat{h}_{\mathrm{g,pi}}$). En efecto,

$$\bar{l} = O\left(n^{-\alpha}\right) \overset{\alpha>1/5}{=} o\left(n^{-\frac{1}{5}}\right) = o\left(h_{\mathrm{AMISE}_g}\right).$$

Finalmente, es sencillo comprobar que, para que la condición $\max_{1\leq j\leq k} |l_j - \bar{l}| = o(\bar{l})$ se verifique, basta pedir, además, que $\beta > \alpha$, puesto que en tal caso

$$\max_{1\leq j\leq k} |l_j - \bar{l}| = 4a_n = 4Dn^{-\beta} = O(n^{-\beta}) \overset{\beta>\alpha}{=} o(n^{-\alpha}) = o(\bar{l}).$$

De esta forma, en todos los estudios de simulación aquí presentados se considerarán longitudes medias de la forma $\bar{l} = Cn^{-\alpha}$, con $C \in \mathbb{R}^+$ y $\frac{1}{5} < \alpha < \beta$.

Concretamente, para las muestras procedentes del Modelo R17 se considerarán los valores $C = 800$ y $D = 150$, tal y como se propone en Reyes et al. (2017). Por el contrario, para los Modelos 1, 2, 6, 9 y 10 de Marron y Wand (1992) se tomará $C = 18$ y $D = 5$, de manera que la longitud media de los intervalos considerados será mucho menor que la considerada para el Modelo R17 (recordemos que la escala de este último modelo es mucho mayor que la del resto de modelos). En todos los casos fijaremos $\alpha = \frac{4}{5}$ y $\beta = 1$.

4.1.3. Consideraciones previas sobre algunos selectores de ventana

A continuación se realizarán algunas apreciaciones referidas a la implementación en `R` de algunos de los selectores presentados en el Capítulo 3. Comentar que, para la realización de este TFM, se han programado en `R` los selectores de la regla del pulgar (con sus dos respectivas versiones presentadas en la Sección 3.1), el selector plug-in propuesto en la Sección 3.2.2, el de validación cruzada insesgada (Sección 3.21) y la modificación del selector bootstrap de Jang y Loh (2010) recogida en la Sección 3.4.1. En todos los casos se han considerado funciones núcleo de tipo gaussiano, aunque se ha comprobado mediante varias pruebas que los resultados obtenidos son muy similares a los que resultarían empleando núcleos Epanechnikov.

Por un lado, en lo referido al selector de la regla del pulgar — y como ya se ha comentado en la Sección 3.1.1 — se presentarán únicamente los resultados correspondientes a la versión (3.5), en la cual se usa a $\hat{\sigma}_g = \sum_{j=1}^{k} w_j (t_j - \hat{\mu}_g)^2$ como estimador de la desviación típica, puesto que en todos los casos proporciona estimaciones muy similares a las que se obtienen cuando se emplea a $\hat{\sigma}_g^* = \min\{\hat{\sigma}_g, \hat{\sigma}_{\mathrm{IQR},g}\}$, siendo el primero de ellos el que mejores resultados ofrece con tamaños muestrales pequeños.

Por otro lado, los selectores plug-in y bootstrap propuestos en Reyes et al. (2017) (ver Secciones 3.2.1 y 3.4.2, respectivamente) se encuentran ya programados en la función `bw.dens.binned` de la librería `binnednp` (véase Barreiro-Ures et al. (2019a) y Barreiro-Ures et al. (2019b)).

Finalmente, comentar que en la tercera y última parte de este capítulo no se presentarán los resultados relativos a la modificación del selector bootstrap de Jang y Loh (2010) propuesta en la Sección 3.4.1, debido al elevado tiempo de computación que esto requeriría. De todas formas, ya en los estudios de simulación realizados en Reyes et al. (2017) se comenta que los resultados obtenidos con el selector bootstrap propuesto por Jang y Loh (2010) son peores que los proporcionados por su selector bootstrap, sobre todo cuando el nivel de agrupación es elevado.

4.1.4. Consideraciones numéricas

Este último apartado está dedicado a comentar algunas consideraciones de carácter numérico que han surgido al implementar los diferentes estudios computacionales en `R` y que deberán ser tenidas en cuenta si se desea reproducir los experimentos aquí presentados.

En lo referido al selector plug-in propuesto en la Sección 3.2.2, comentar que la estimación de los vectores de parámetros $\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_M)$, $\boldsymbol{\mu}' = (\mu_1, \ldots, \mu_M)$ y $\boldsymbol{\sigma}' = (\sigma_1, \ldots, \sigma_M)$ se ha llevado a cabo a través de un algoritmo de tipo EM adaptado al caso de datos agrupados. Para ello, se ha hecho uso de la función `mixfit` de la librería `mixR` (véase Yu (2021)), la cual ha sido ligeramente modificada para permitir trabajar con un número menor de intervalos al que considera por defecto. Esta función devuelve, en ocasiones, errores por falta de convergencia, siendo este problema más frecuente cuando el tamaño muestral es pequeño o cuando el nivel de agrupación es elevado. En los casos en los que esta convergencia falla, se ha optado por emplear el argumento `ev=TRUE`, el cual fuerza a que todas las densidades normales que conforman la mixtura tengan la misma varianza (disminuyendo, consecuentemente, el número de parámetros a estimar). Además, en aquellos casos en los que la convergencia del algoritmo siga siendo un problema, se adoptará la decisión sistemática de tomar $M = 1$, puesto que en tal caso existen expresiones explícitas para los estimadores de máxima verosimilitud ponderada (ver Sección 3.1).

Ahora bien, en todos aquellos casos en los que se ha requerido de la integración numérica, se ha empleado la función `integrate` de `R`, la cual incorpora el argumento `abs.tol` — cuyo valor por defecto ha sido fijado a $\texttt{1.22}\cdot\texttt{10}^{-4}$ — que permite modificar la precisión con la que se obtienen los resultados, proporcionando una aproximación `Q` de la verdadera integral `I` tal que

$$|\texttt{I} - \texttt{Q}| \leq \text{máx}\left(\texttt{abs.tol},\ \texttt{rel.tol} * |\texttt{Q}|\right),$$

siendo `rel.tol` una medida de error relativo (que ha sido fijada a `rel.tol=abs.tol`). En ocasiones (y, en especial, cuando la función del integrando presenta una expresión *complicada*), modificar este parámetro provoca que se obtengan soluciones ligeramente distintas. En estos casos, la precisión se ha modificado a los valores $\texttt{10}^{-6}$ o incluso a `0`.

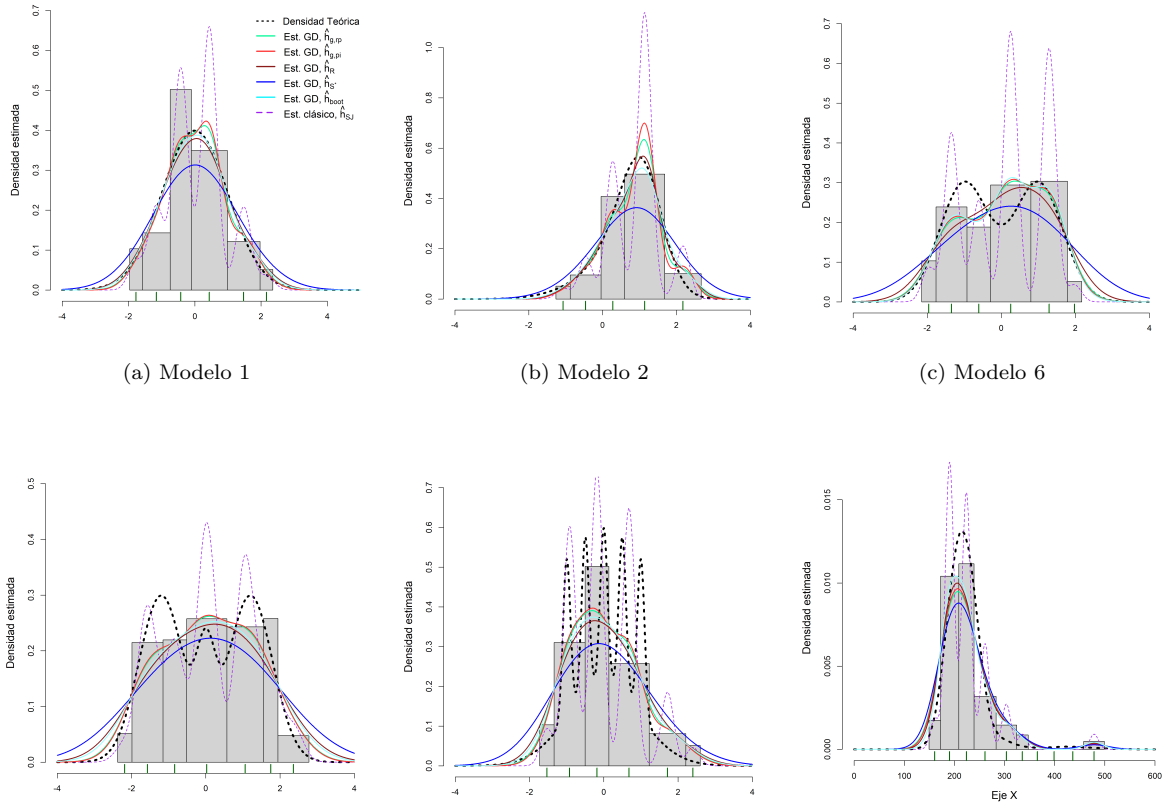
Por otro lado, comentar que en la implementación del selector de validación cruzada insesgada se ha hecho uso de la función `optimize` de la librería `stats` para llevar a cabo la optimización numérica de la función de validación cruzada (cuya expresión se recoge en (3.24)). Esta función combina el método de la sección áurea con una interpolación parabólica sucesiva, garantizando siempre una velocidad de convergencia superior al método de búsqueda de Fibonacci, restringiendo su uso al contexto de funciones continuas (para más información sobre estos métodos de optimización puede consultarse Brent (2013)).

Por último, y con el fin de facilitar la reproducibilidad de los resultados, comentar que todos los resultados que se presentarán en esta sección se han obtenido a través de la versión `4.0.3` de `R` fijando la semilla con el comando `set.seed(123)`.

4.2. Una primera ilustración

Con el fin de ilustrar gráficamente el comportamiento del estimador de Cao et al. (2011) — así como su dependencia respecto del parámetro ventana h — comenzaremos considerando, para cada uno de los modelos teóricos de referencia, una muestra fija de tamaño $n = 50$ conformada por datos previamente agrupados siguiendo el esquema de agrupación recogido en la Sección 4.1.2.

En la Figura 4.2 se recoge la representación gráfica de los correspondientes estimadores tipo núcleo $\hat{f}^{g}_{w,h}$ construidos en base a los selectores de ventana[2] de la regla del pulgar (líneas de color verde), plug-in (tonalidades rojas) y bootstrap (tonalidades azules) presentados en el Capítulo 3. Además, con líneas discontinuas moradas se incorpora la representación gráfica del estimador de Parzen-Rosenblatt construido en base a la ventana plug-in clásica de Sheather y Jones (1991).



(a) Modelo 1 (b) Modelo 2 (c) Modelo 6

(d) Modelo 9 (e) Modelo 10 (f) Modelo R17

Figura 4.2: Histogramas de cada una de las seis muestras (de datos agrupados) de tamaño $n = 50$ generadas a partir de los Modelos 1, 2, 6, 9 y 10 de Marron y Wand (1992) y del Modelo R17. Con líneas continuas se representan los estimadores de la densidad $\hat{f}^{g}_{w,h}$ construidos en base a diferentes ventanas, todas ellas indicadas en la leyenda superior izquierda (verde: regla del pulgar, tonalidades rojas: selectores plug-in, tonalidades azules: selectores bootstrap). Con líneas moradas discontinuas se representa el estimador de Parzen-Rosenblatt construido en base a la clásica ventana plug-in ($\hat{h} = 5\,\hat{h}_{\mathrm{SJ}}$ para los Modelos 1, 2 y R17 y $\hat{h} = 2\,\hat{h}_{\mathrm{SJ}}$ para los Modelos 6, 9 y 10); y con líneas negras punteadas, la correspondiente densidad teórica. Con segmentos verdes se indican los puntos medios de los intervalos considerados, $t_1, \ldots, t_k$.

En este sentido — y para facilitar la interpretación de las gráficas — comentar que para los Modelos 1, 2 y R17, se ha empleado a $\hat{h} = 5\,\hat{h}_{\mathrm{SJ}}$ como parámetro de suavizado en la construcción del estimador de Parzen-Rosenblatt, donde $\hat{h}_{\mathrm{SJ}}$ denota a la ventana plug-in clásica de Sheather y Jones (1991).

[2]En esta primera ilustración no se presentan los estimadores construidos en base al selector de validación cruzada insesgada propuesto en la Sección 3.3, puesto que en todos los casos ha conducido a estimadores con un elevado sesgo. No obstante, el error de estimación cometido por este selector se abordará con más detalle en la Sección 4.3.

El motivo es que la elevada concentración de observaciones entorno a cada valor t_j provoca que las ventanas escogidas por los procedimientos clásicos sean extremadamente pequeñas, conduciendo a estimadores muy fluctuantes que dificultan la comparativa con las demás curvas. Para el resto de modelos (Modelos 6, 9 y 10), en los que la densidad se encuentra repartida en un número mayor de modas, ha sido suficiente tomar $\hat{h} = 2\,\hat{h}_{\mathrm{SJ}}$. Finalmente, las densidades teóricas de referencia se representan con líneas punteadas de color negro.

Las gráficas de la Figura 4.2 hacen evidente la mejora que, para estas muestras, se produce en la estimación al considerar a $\hat{f}^{g}_{w,h}$ como estimador de la densidad, ya que en todos los casos el estimador de Parzen-Rosenblatt se limita a presentar una moda diferente entorno a cada t_j, incluso tras haber multiplicado la ventana empleada en su construcción por 5 (Modelos 1, 2 y R17) o por 2 (Modelos 6, 9 y 10). Además, a pesar del pequeño tamaño muestral considerado, el estimador $\hat{f}^{g}_{w,h}$ parece presentar un buen comportamiento en los Modelos 1, 2 y R17. Recordemos que los dos primeros se corresponden con densidades unimodales, presentando el segundo una cierta asimetría, mientras que el tercero se corresponde con una densidad también próxima a una normal, con una pequeña concentración de masa adicional en su cola derecha (ver Figura 4.1). Ahora bien, la aparición de más de una moda ha conducido, en general, a estimadores con mayor sesgo, como son los obtenidos para los Modelos 6, 9 y 10, siendo este último especialmente malo, puesto que con ningún selector se logra captar correctamente ninguna de las cinco modas teóricas (lo cual era esperable, puesto que únicamente se están considerando seis intervalos).

A modo de resumen podría decirse que, para las muestras consideradas, parece que son los selectores plug-in los que conducen a estimadores con mayor varianza, algo que se hace especialmente notorio con la ventana $\hat{h}_{\mathrm{g,pi}}$. Por el contrario, estimadores con menor variabilidad parecen obtenerse, en general, con la regla del pulgar y el selector bootstrap propuesto en Reyes et al. (2017). Finalmente, la modificación del selector bootstrap de Jang y Loh (2010) presentada en la Sección 3.4.1 parece proporcionar ventanas más amplias, ya que en todos los casos conduce a estimadores que quedan por debajo de la densidad teórica, presentando una mayor dificultad para estimar las zonas de mayor curvatura.

Ahora bien, recordemos que las muestras consideradas presentan un tamaño excesivamente pequeño si lo que se desea es obtener una fiel representación de la densidad teórica, sobre todo si esta presenta características que pueden dificultar su estimación, como puede ser la presencia de multimodalidad, asimetría o de curtosis elevada. A modo de ejemplo, en la Figura 4.3 se recogen las mismas gráficas hasta ahora presentadas pero construidas en base a muestras de tamaño $n = 500$, para los Modelos 6 (izquierda) y 9 (derecha). En este caso, la ventana empleada para la construcción del estimador de Parzen-Rosenblatt ha sido $\hat{h} = \hat{h}_{\mathrm{SJ}}$.

Como vemos, el hecho de haber aumentado el tamaño muestral ha permitido obtener estimaciones mucho más precisas que capturan más satisfactoriamente el carácter multimodal de ambas densidades — especialmente en el Modelo 6 — siendo nuevamente la modificación del selector bootstrap de Jang y Loh (2010) propuesta en la Sección 3.4.1 — y representada con líneas de color azul oscuro — la que proporciona ventanas más amplias. De hecho, el estimador construido en base a este selector es el único que no es capaz de capturar la segunda moda del Modelo 9.

Finalmente, comentar que en ambos casos el estimador clásico de Parzen-Rosenblatt (línea discontinua morada) ha proporcionado una estimación de la densidad muy similar a la ofrecida por el estimador de Cao et al. (2011). Esto se debe a que, a pesar de que la muestra considerada sigue estando constituida por datos agrupados, el hecho de aumentar el tamaño muestral ha disminuido considerablemente el nivel de agrupación (nótese como las cajas que conforman los histogramas son ahora mucho más estrechas). De esta forma, para tamaños muestrales elevados ($n = 500$ o superiores), podría llegarse a justificar el empleo de un estimador tipo núcleo clásico aún cuando la muestra considerada estuviese conformada por datos agrupados.

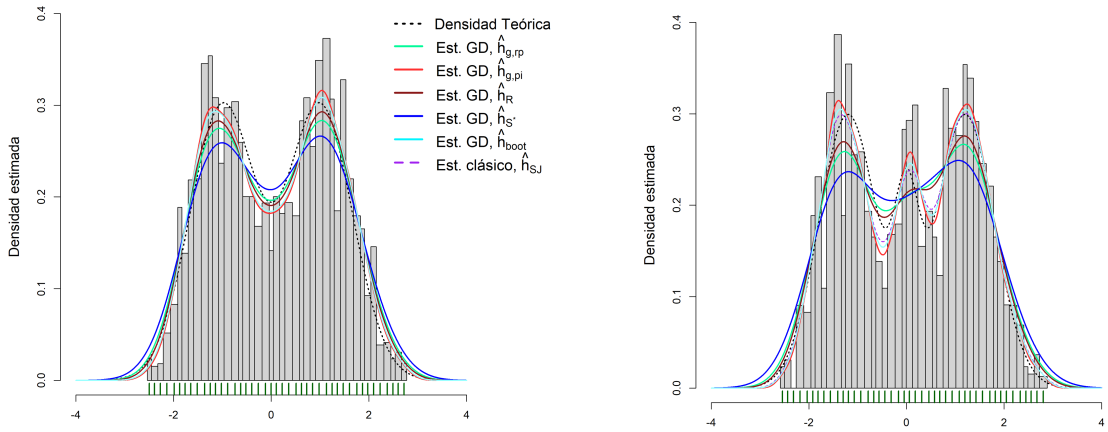


Figura 4.3: Histogramas de cada una de las dos muestras (de datos agrupados) de tamaño $n = 500$ generadas a partir de los Modelos 6 (izquierda) y 9 (derecha). Con líneas continuas se representan los estimadores $\hat{f}^g_{w,h}$ y con líneas moradas discontinuas el estimador de Parzen-Rosenblatt construido en base a la ventana plug-in de Sheather y Jones (1991). Con líneas negras punteadas se representa la correspondiente densidad teórica. Los segmentos verdes indican los puntos medios de los intervalos considerados.

4.3. Aproximación numérica del MISE_g

El estudio realizado en la sección anterior proporciona únicamente una primera ilustración del comportamiento del estimador de la densidad de Cao et al. (2011), pero no permite extraer ninguna conclusión general, puesto que únicamente se ha considerado una muestra procedente de cada modelo. En este sentido, se llevará a cabo continuación un nuevo estudio de simulación que permitirá comparar de manera cuantitativa — a través del cálculo aproximado del MISE_g — la efectividad de los diferentes selectores de ventana presentados en el Capítulo 3. Como ya se ha comentado en la sección introductoria de este capítulo, debido al elevado tiempo de computación que ello requeriría, no se presentarán los resultados correspondientes al selector bootstrap de Jang y Loh (2010) (ni a su modificación propuesta en la Sección 3.4.1). De todas formas, la primera aproximación realizada en la Sección 4.2 parece sugerir que este selector podría dar lugar a estimadores con elevado sesgo.

Consideraremos un total de $B = 1000$ muestras procedentes de cada uno de los seis modelos teóricos de referencia, agrupadas siguiendo el mismo esquema ya indicado en apartados anteriores. A partir de cada conjunto de muestras, se obtendrá una aproximación del MISE_g en base a la siguiente expresión,

$$\widehat{\text{MISE}}(\hat{f}^g_{w,h}) = \frac{1}{B}\sum_{i=1}^{B}\int \left(\hat{f}^{g,(i)}_{w,h}(x) - f(x)\right)^2 \, \mathrm{d}x,$$

donde la anterior integral se aproximará de forma numérica a través de la función `integrate` de `R`, siendo $\hat{f}^{g,(i)}_{w,h}$ el estimador de la densidad de Cao et al. (2011) construido en base a la $i-$ésima muestra generada, $i = 1, \ldots, B$. El experimento se realizará para los tamaños muestrales $n \in \{50, 250, 500\}$. Los resultados, recogidos en la Tabla 4.2, se muestran aproximados a cuatro cifras decimales.

	n	**MISE**$_g$					
		$\hat{f}^{g}_{w,h}$					$\hat{f}_{n,h}$
		$\hat{h}_{\text{rp}}$	$\hat{h}_{\text{g,pi}}$	$\hat{h}_{\text{R}}$	$\hat{h}_{\text{boot}}$	$\hat{h}_{\text{UCV}_g}$	$\hat{h}_{\text{pi}}$
Modelo 1	**50**	1.1666	1.4246	1.2485	**1.1311**	5.6937	95.8994
	250	**0.3097**	0.3302	0.3400	0.3159	1.8489	0.3466
	500	**0.1781**	0.1841	0.1916	0.1803	1.1648	0.1921
Modelo 2	**50**	2.6720	5.0973	3.0737	**2.3984**	11.2279	189.8734
	250	**0.4783**	0.4889	0.5331	0.4849	3.4813	2.2624
	500	0.2880	**0.2759**	0.3068	0.2763	2.2134	0.2856
Modelo 6	**50**	**1.4446**	2.3277	1.6305	1.6127	4.6443	32.7542
	250	0.4964	0.4508	0.5759	0.4329	2.3342	**0.4341**
	500	0.3094	**0.2470**	0.3214	0.2473	1.6160	0.2489
Modelo 9	**50**	**1.6002**	2.4767	1.8225	1.7492	4.8533	23.7123
	250	0.7239	**0.5638**	0.7603	0.5751	2.6545	0.5520
	500	0.4962	**0.3299**	0.4579	0.3458	1.7855	0.3391
Modelo 10	**50**	6.3541	7.0374	6.4193	**6.2426**	12.0316	123.4985
	250	4.8494	**4.6119**	5.0081	4.9101	7.3404	4.6958
	500	4.5715	**3.4925**	4.7297	4.2243	6.4836	3.9621
Modelo R17	**50**	6.6458	6.044	4.0685	**3.1720**	11.6733	46.2383
	250	0.8515	0.8516	0.9035	1.0557	**0.6703**	2.6504
	500	0.5429	0.5436	0.6169	0.6850	**0.4299**	0.7336

Tabla 4.2: Aproximaciones del error cuadrático medio integrado ($\times 100$ en los Modelos 1, 2, 6, 9 y 10; $\times 10^5$ en el Modelo R17) del estimador de Cao et al. (2011) construido en base a algunos de los selectores de ventana presentados en el Capítulo 3 (cinco primeras columnas) y del estimador de Parzen-Rosenblatt construido en base al selector plug-in de Sheather y Jones (1991) (última columna); obtenidas a partir de $B = 1000$ simulaciones. Para cada uno de los seis modelos en cuestión se consideran los tamaños muestrales $n \in \{50, 250, 500\}$.

Para facilitar su comparativa, todos los resultados correspondientes a los modelos de Marron y Wand (1992) se han multiplicado por 100. En este sentido, comentar que la magnitud de los MISE_g obtenidos con el Modelo R17 son del orden de 10^3 veces más pequeños que los obtenidos con el resto de modelos[3]. Es por este motivo por el cual los resultados relativos a este último se mostrarán multiplicados por 10^5 en vez de por 100.

La Tabla 4.2 hace evidente la disminución que sufre el MISE_g conforme aumenta el tamaño muestral, lo cual se traduce en un aumento de la precisión del estimador cuando se consideran muestras de gran tamaño. Además, si nos fijamos en la última columna de la tabla — correspondiente al estimador tipo núcleo de Parzen-Rosenblatt construido en base a la ventana plug-in clásica de Sheather y Jones (1991) — se observa la evidente mejora que se produce, en términos del error cuadrático medio integrado, con el estimador de Cao et al. (2011) frente a dicho estimador. De todas formas, los errores relativos al estimador de Parzen-Rosenblatt se ven significativamente reducidos al aumentar el tamaño muestral, lo cual se debe a algo que ya se ha comentado anteriormente, y es que el hecho de aumentar el tamaño muestral deriva inevitablemente en una disminución del nivel de agrupación, en cuyo caso podría justificarse el empleo del estimador tipo núcleo clásico (obsérvese que los MISE correspondientes al estimador de Parzen-Rosenblatt, cuando $n = 500$, son siempre muy similares a los MISE_g proporcionados por el estimador de Cao et al. (2011)).

Por otro lado, la penúltima columna de la tabla — correspondiente a los MISE_g relativos al estimador de Cao et al. (2011) construido en base a la ventana de validación cruzada insesgada propuesta en la Sección (3.3) — evidencia la elevada magnitud de los errores que se obtienen al emplear este selector. Estos malos resultados pueden deberse a que, como ya se comentó en la Sección 3.3, este método de selección es el único, de los aquí considerados, que no tiene en cuenta el tamaño muestral, empleando únicamente la información proporcionada por los pesos muestrales $w_1, \ldots, w_k$. De todas formas, resaltar que una situación totalmente contraria se obtiene con el Modelo R17, ya que en este caso parece ser que — para tamaños muestrales $n \in \{250, 500\}$ — el selector de validación cruzada insesgada es el que mejores resultados ofrece.

Finalmente, comentar que los errores relativos al Modelo 10 no son demasiado relevantes, puesto que, debido a la elevada complejidad — en términos del número de modas y de la curvatura — de la correspondiente densidad teórica, todas las ventanas han conducido a estimaciones incorrectas. Sin embargo, considerar este tipo de modelos proporciona un buen ejemplo para ilustrar la limitación de los estimadores tipo núcleo, tanto en el contexto de datos agrupados como en el caso clásico.

Centrándonos en los errores relativos al estimador de Cao et al. (2011) y obviando los resultados correspondientes al selector de validación cruzada insesgada, se observa que, para tamaños muestrales pequeños ($n = 50$) parece que es el selector bootstrap de Reyes et al. (2017) — seguido de la regla del pulgar — el que deriva en un menor error (MISE_g), lo cual lo convierte en un selector adecuado cuando se disponen de pocos datos, al menos en aquellas situaciones en las que la densidad teórica (frecuentemente desconocida) presenta características similares a las aquí consideradas. Por el contrario, en estos casos es la ventana plug-in $\hat{h}_{\text{g,pi}}$ la que conduce a errores más elevados, quizás debido a la dificultad que presenta estimar mixturas de normales en base a muestras de tamaño tan pequeño.

Sin embargo, para tamaños muestrales $n \in \{250, 500\}$ parece que es el selector plug-in propuesto en la Sección (3.2.2) (ventana $\hat{h}_{\text{g,pi}}$) el que, en la mayoría de casos, conduce a errores más pequeños. Destacar también los buenos resultados que se obtienen con el selector de la regla del pulgar (adaptada al contexto de datos agrupados) en el Modelo 1. Estos buenos resultados se deben a que este selector obtiene la estimación de $R(f'')$ — necesaria para el cómputo de la ventana h_{AMISE_g} en la práctica, ver

[3]Esta situación puede deberse a la magnitud de las desviaciones típicas que intervienen en la mixtura que conforma el Modelo R17, muy superior a la de las que intervienen en los modelos de Marron y Wand (1992). A modo de ejemplo, si se considera la modificación del Modelo 1 (densidad normal estándar) obtenida reemplazando $\sigma = 1$ por $\sigma = 25$, se obtienen, independientemente del selector escogido, mejores resultados que con el propio Modelo 1, obteniendo valores del MISE_g del orden de cien veces más pequeños.

Sección (3.1) — en base a la suposición de que la densidad teórica se corresponde con la densidad normal estándar, que es precisamente la densidad teórica del Modelo 1. Comentar finalmente la mejora que se produce — en términos del error cuadrático medio — con la ventana plug-in $h_{\text{g,pi}}$ frente al selector plug-in propuesto por Reyes et al. (2017), ya que en todos los casos conduce a MISE_g más pequeños.

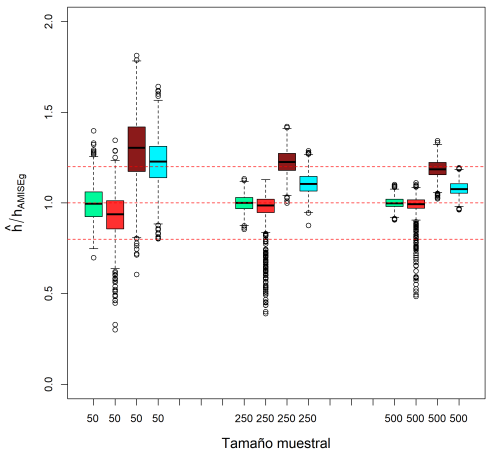
En la Figura 4.4 se recogen, a modo de ejemplo, los diagramas de caja correspondientes a las cantidades $\hat{h}_g/h_{\text{AMISE}_g}$ para cada uno de los seis modelos teóricos aquí considerados, donde $\hat{h}_g$ denota a la ventana estimada a través de un determinado selector adaptado al caso de datos agrupados y donde h_{AMISE_g} denota a la ventana asintóticamente óptima en el contexto de datos agrupados, que ha sido calculada en base a la expresión (2.11). La leyenda de colores, idéntica a la recogida en la Figura 4.2, es la que sigue: verde (regla del pulgar), tonalidades rojas (selectores plug-in: $\hat{h}_{\text{g,pi}}$ y $\hat{h}_{\text{R}}$) y azul (selector bootstrap de Reyes et al. (2017)). Nótese que en este caso se han obviado nuevamente los resultados relativos al selector de validación cruzada insesgada, lo cual ha permitido trabajar con una escala más pequeña en el eje de ordenadas y facilitar así la interpretación de los resultados.

Como vemos, cuando $n = 50$, los diagramas de caja se encuentran más alejados respecto de la recta $y = 1$, indicando así un elevado sesgo de $\hat{h}_g$ como estimador de h_{AMISE_g}. Además, en este caso los diagramas de caja son también más anchos (en el eje vertical), lo cual se traduce en que la precisión de las diferentes ventanas estimadas es todavía bastante escasa. De todas formas, recordemos que h_{AMISE_g} es una ventana asintóticamente óptima, que puede no estar cerca de h_{MISE_g} cuando el tamaño muestral es pequeño.

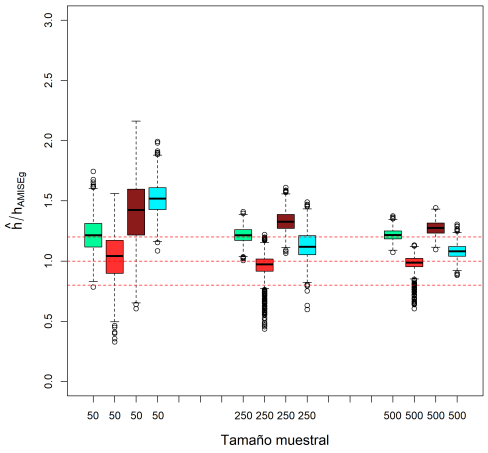
Conforme el tamaño muestral aumenta ($n \in \{250, 500\}$), los diagramas de caja se centran cada vez más próximos a la recta $y = 1$, de manera que las ventanas proporcionadas por los diferentes selectores se aproximan cada vez más a la ventana teórica h_{AMISE_g} (esto es, su sesgo como estimador de h_{AMISE_g} disminuye), disminuyendo simultáneamente su anchura (es decir, disminuye también su variabilidad como estimador de h_{AMISE_g}).

Comentar finalmente que, conforme se consideran modelos más complejos (en términos del número de modas y del grado de simetría), los diagramas de caja tienden a presentar una mayor anchura, situándose, además, a una mayor distancia respecto de la recta $y = 1$. De esto se concluye que, en estos casos — en los que la densidad teórica presenta propiedades que podemos catalogar como *poco deseables* — el proceso de selección de ventana se dificulta, proporcionando ventanas que pueden estar lejos del valor (asintóticamente) óptimo.

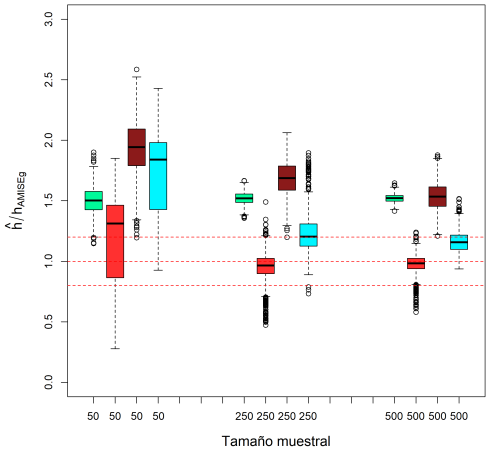
En definitiva, el estudio de simulación aquí presentado ha permitido justificar el empleo de un estimador alternativo al de Parzen-Rosenblatt cuando la muestra observada está conformada por datos agrupados. La mejora que se produce al considerar el estimador de Cao et al. (2011) es muy evidente cuando el tamaño de muestra es pequeño o medio, mientras que, para muestras grandes, esta diferencia es mucho menor, pudiéndose llegar a justificar el empleo del estimador tipo núcleo clásico. Por otro lado, se ha visto que el selector plug-in propuesto en la Sección 3.2.2 debería de ser la primera opción cuando la muestra de la que se dispone es de tamaño medio o grande, superando en todos los casos al selector plug-in propuesto en Reyes et al. (2017). Por el contrario, para muestras de tamaño reducido es preferible emplear el selector bootstrap de Reyes et al. (2017) o, incluso, la regla del pulgar propuesta en la Sección 3.2, puesto que en la mayoría de casos han conducido a errores de estimación más pequeños. Finalmente, el selector de validación cruzada insesgado propuesto en la Sección 3.3 ha mostrado ser, en general, el peor de todos los aquí considerados, posiblemente debido a que es el único que no tiene en cuenta el tamaño muestral, permitiendo trabajar en aquellos contextos más generales en los que únicamente se conocen los pesos muestrales $w_1, \ldots, w_k$.



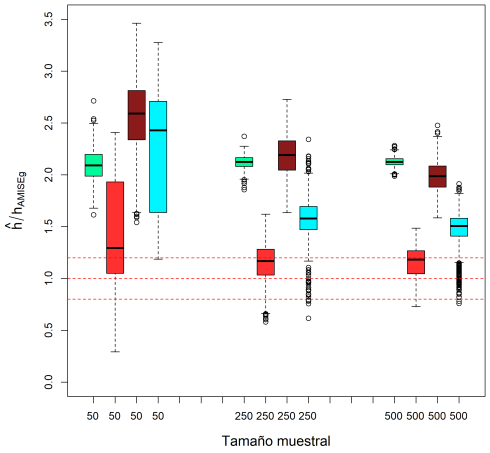
(a) Modelo 1



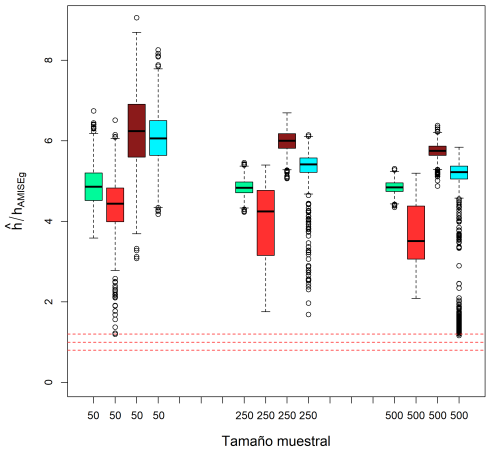
(b) Modelo 2



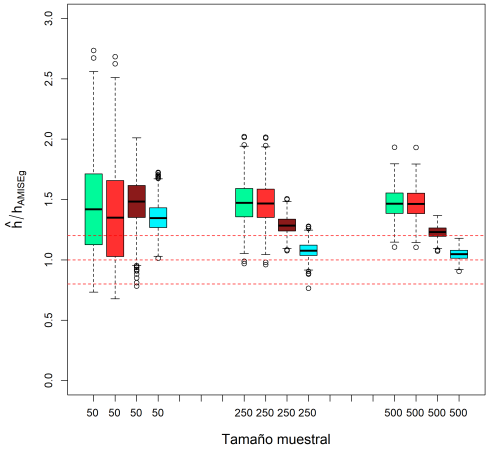
(c) Modelo 6



(d) Modelo 9



(e) Modelo 10



(f) Modelo R17

Figura 4.4: Diagramas de caja de los valores $\hat{h}_g/h_{\mathrm{AMISE}_g}$, construidos en base a $B = 1000$ muestras de datos agrupados, para cada uno de los tres tamaños muestrales considerados.

Capítulo 5

Aplicación a datos reales

Una parte importante de cualquier investigación en técnicas estadísticas es su aplicación a datos reales. Este último capítulo está dedicado, pues, a ilustrar el comportamiento del estimador de la densidad de Cao et al. (2011) sobre tres conjuntos de datos reales diferentes, todos ellos conformados por observaciones agrupadas en intervalos, ya sea debido a una falta de precisión derivada del mecanismo empleado en su recogida o porque realmente no es necesario (o no es factible) obtener medidas con gran precisión. En este sentido, haremos uso, una vez más, de los todos selectores de ventana presentados en el Capítulo 3, reservando el de validación cruzada insesgada (Sección 3.3) para la tercera base de datos, en la que se considera un contexto de agrupación más general, donde únicamente los pesos muestrales son conocidos.

Para ello, se presentará en primer lugar la base de datos `stamps` del paquete `multimode` de `R` (véase Ameijeiras-Alonso et al. (2021)), la cual contiene las medidas de espesor (en milímetros) de 485 sellos pertenecientes a la colección de Hidalgo de 1872 y que fueron analizadas posteriormente en Wilson (1983). En este primer ejemplo se asumirá que — debido a imprecisiones en la recogida de datos — los espesores se encuentran agrupados en intervalos de igual longitud, resultando en un *efecto agrupación* no demasiado elevado.

En segundo lugar, se considerarán los 272 tiempos de espera (en minutos) relativos a los tiempos entre erupciones del géiser Old Faithful (Parque Nacional Yellowstone, Wyoming, Estados Unidos), almacenados en la librería `dataset` de `R`. En este caso se asumirá un mayor *efecto agrupación*, donde los intervalos considerados presentarán longitudes diferentes, lo cual permitirá evaluar el comportamiento del estimador de Cao et al. (2011) en un escenario algo más complejo y realista. Nótese que `faithful` constituye un conjunto de datos clásico en el análisis no paramétrico de curvas, pero del cual se suele ignorar el carácter agrupado de las observaciones (agrupación que, como se comentará más adelante, viene motivada por la falta de precisión en la medición de los tiempos, tanto de espera como de erupción).

Finalmente, se emplearán los datos referidos a los casos de COVID-19 reportados en España durante diferentes periodos de la pandemia y que han sido extraídos de los informes emitidos por la Red Nacional de Vigilancia Epidemiológica (RENAVE, ver `https://cnecovid.isciii.es/`). En este caso, el factor agrupación vendrá definido por los distintos grupos de edad y, además, este será el único ejemplo de los aquí presentados en el que únicamente se dispondrá de las proporciones muestrales (no estando disponible, por tanto, el número concreto de casos relativos a cada grupo de edad), de manera que la construcción del estimador de Cao et al. (2011) únicamente se podrá llevar a cabo a través del selector de ventana de validación cruzada insesgada propuesto en la Sección (3.3), puesto que, como ya se ha comentado en anteriores ocasiones, es el único que permite trabajar en este contexto general de datos agrupados.

5.1. Colección de sellos de Hidalgo (México, 1872)

En esta primera sección — y como ya se anticipó en la introducción del capítulo — se considerarán las medidas de espesor (en milímetros) de 485 sellos pertenecientes a la colección de Hidalgo de 1872, elaborados en México y extraídas de la base de datos `stamps` del paquete `multimode` de `R` (véase Ameijeiras-Alonso et al. (2021)). Dichas medidas fueron analizadas por primera vez en 1983 en Wilson (1983), aunque con algunos errores en el recuento de sellos, puesto que en su estudio omitió un total de 48 medidas, considerando una submuestra conformada por 429 sellos.

En esta colección de sellos — que estuvo en circulación entre 1872 y 1874 — aparecía representado un retrato de Miguel Hidalgo y Costilla (ver Figura 5.1(a)). El valor de los sellos dependía del número de subcolecciones que se podían encontrar en su correspondiente edición y, en este sentido, el espesor era un factor determinante.

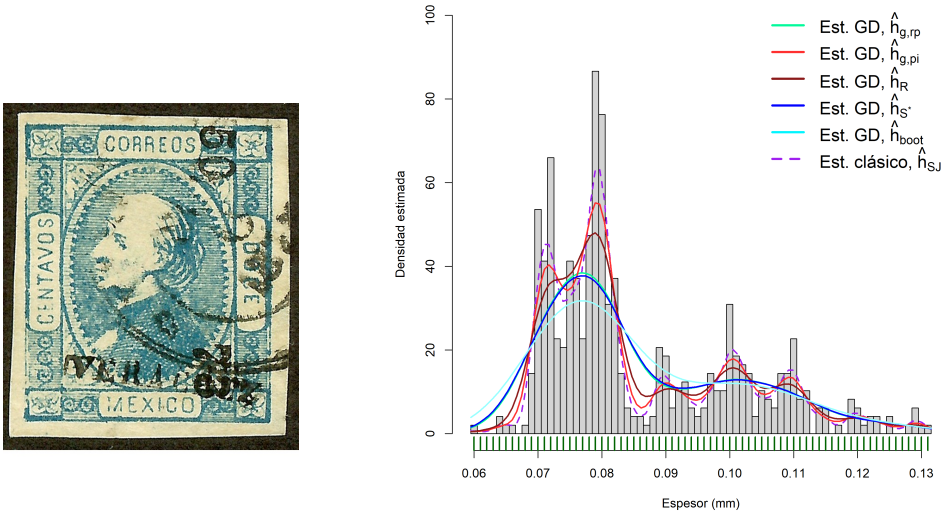
En Izenman y Sommer (1988) se aborda el problema de contabilizar el número de grupos de sellos que se podían diferenciar (siendo el espesor el factor de agrupación), lo cual derivó en el estudio del número de modas a través de estimadores tipo núcleo clásicos y tests de multimodalidad, llegando a encontrar hasta siete modas distintas, siendo las correspondientes a los 0.080 y 0.100 milímetros las dos más sencillas de diferenciar (y que, a su vez, habían sido las dos únicas encontradas en Wilson (1983), lo cual le llevó a la conclusión de que solo se habían empleado dos tipos de papel, el *Papel Sellado* y *La Croix-Freres*). Para ello, estos autores emplearon el test de multimodalidad propuesto por Silverman (1981), un test que ha demostrado presentar un deficiente calibrado en varios estudios de simulación posteriores. Más tarde, este mismo problema se vuelve a abordar en Ameijeiras-Alonso (2017), donde se propone un nuevo test de multimodalidad con un mejor calibrado, incluso para muestras pequeñas, el cual obtiene como conclusión que no se pueden ausmir más de cuatro modas.

Los espesores en cuestión (en milímetros) se encuentran en el intervalo $[\mathcal{L},\mathcal{U}] = [0.060, 0.131]$, presentando una media y desviación típica de 0.086 y 0.015 milímetros, respectivamente. Para tener una idea intuitiva acerca de los valores extremos del grosor del papel, comentar que el grosor de un pañuelo ronda los 0.040 milímetros, mientras que el del cartón, 0.150 milímetros. Además, todos los grosores han sido recogidos con únicamente tres cifras decimales de exactitud, lo cual provoca que haya numerosos empates. A modo de ejemplo, se dispone de un total de 26 observaciones correspondientes a 0.070 milímetros, 20 correspondientes a 0.071 y 32 a 0.072. Por este motivo, parece natural asumir que las verdaderas medidas son desconocidas, encontrándose agrupadas en intervalos de los cuales únicamente se conoce sus puntos medios. En este sentido, y teniendo en cuenta que dichos puntos medios se encuentran — salvo excepciones — equiespaciados a distancia 10^{-3}, consideraremos una sucesión de intervalos consecutivos también equiespaciados y de longitud 10^{-3} de tal forma que recubran el intervalo $[\mathcal{L} - 0.5 \times 10^{-3},\ \mathcal{U} + 0.5 \times 10^{-3}]$. Es decir, consideraremos intervalos de la forma

$$[y_{j-1}, y_j) = [(t_1 - 0.5 \times 10^{-3}) + (j-1) \cdot 10^{-3},\ (t_1 + 0.5 \times 10^{-3}) + (j-1) \cdot 10^{-3}), \qquad j = 1, \ldots, 72,$$

donde $t_1 = 0.060$ denota al primer valor observado, habiendo un total de 10 intervalos en los que no se ha recogido ningún dato. Nótese que la presencia de intervalos vacíos no supone ningún problema, puesto que esto equivaldría a asumir un punto medio adicional con peso asociado nulo.

En este caso, los estimadores tipo núcleo de Cao et al. (2011) — construidos en base a diferentes ventanas — se representan en la Figura 5.1(b). Dado que no se dispone de la densidad teórica, se ha representado también el estimador de Parzen-Rosenblatt (Parzen (1962) y Rosenblatt (1956)), construido en base a la ventana plug-in clásica de Sheather y Jones (1991). Nótese que en este caso, en el cual el factor agrupación no es demasiado elevado — debido a la reducida amplitud de los intervalos considerados, en comparación con el rango muestral — es de esperar que el estimador clásico de la densidad presente un comportamiento razonablemente bueno — como, de hecho, se observa en la Figura 5.1(b) — lo cual nos permitirá tener un estimador de referencia.



(a) (b)

Figura 5.1: Izquierda: Sello de doce centavos de la colección de Hidalgo (México) impreso en 1872, extraído de Wikimedia Commons (2009). Derecha: Histograma de las medidas de espesor de los 485 sellos considerados, junto con los estimadores tipo núcleo de Cao et al. (2011) y el estimador clásico de Parzen-Rosenblatt, construidos en base a diferentes ventanas (cuya leyenda de colores se recoge en la parte superior derecha del gráfico). Con marcas verdes se indican los puntos medios de los intervalos considerados que, en este caso, se han asumido equiespaciados de longitud 10^{-3}.

Como vemos, cada una de las ventanas consideradas ha conducido a estimadores con diferentes grados de suavidad, dando lugar a un número distinto de modas en cada caso[1]. En este sentido, han sido los dos selectores plug-in los que han logrado captar un mayor número de modas. Concretamente, el estimador de la densidad construido en base al selector plug-in propuesto en la Sección 3.2.2, $\hat{h}_{\mathrm{g,pi}}$, presenta hasta un total de siete modas (las mismas que el estimador clásico de Parzen-Rosenblatt), lo cual parece estar en una mayor concordancia con los análisis realizados en Izenman y Sommer (1988); reduciéndose este número hasta cuatro si consideramos el selector plug-in de Reyes et al. (2017), $\hat{h}_{\mathrm{R}}$, coincidiendo en tal caso con el número de modas diferenciadas en Ameijeiras-Alonso (2017). Por el contrario, las ventanas relativas a la regla del pulgar y a la modificación del selector bootstrap de Jang y Loh (2010) presentada en la Sección 3.4.1 ($\hat{h}_{\mathrm{g,rp}}$ y $\hat{h}_{\mathrm{S}^*}$, respectivamente) conducen a estimadores más suaves, capaces de detectar únicamente dos modas (las mismas que se lograron diferenciar en Wilson (1983)). El caso extremo se obtiene con el selector bootstrap de Reyes et al. (2017), $\hat{h}_{\mathrm{boot}}$, el cual ha conducido a una única moda, convirtiéndose en un selector poco adecuado para este contexto, en el cual se sabe que la densidad teórica presenta un carácter multimodal. Finalmente, comentar que los valores de las ventanas que se han empleado para la construcción de estos estimadores, ordenadas de menor a mayor, han sido

$$\hat{h}_{\mathrm{g,pi}} = 1.74 \cdot 10^{-3}, \quad \hat{h}_{\mathrm{R}} = 2.45 \cdot 10^{-3}, \quad \hat{h}_{\mathrm{g,rp}} = 4.59 \cdot 10^{-3}, \quad \hat{h}_{\mathrm{S}^*} = 4.80 \cdot 10^{-3}, \quad \hat{h}_{\mathrm{boot}} = 6.75 \cdot 10^{-3},$$

mientras que la ventana plug-in clásica — en base a la cual se ha construido el estimador tipo núcleo de Parzen-Rosenblatt — ha sido de $\hat{h}_{\mathrm{SJ}} = 1.21 \cdot 10^{-3}$.

[1]En este trabajo no se abordará el problema determinar el número de modas de la densidad teórica subyacente. De todas formas, comentar que en Ameijeiras-Alonso (2017) se puede encontrar una revisión a esta problemática para el caso de datos no agrupados.

5.2. Tiempos de espera entre erupciones (Old Faithful)

En esta segunda sección se considerará la base de datos `faithful` — almacenada en la librería `datasets` de `R`, véase GeyserTimes (2017) — que contiene 272 observaciones correspondientes a los tiempos de espera medidos (en minutos) entre erupciones sucesivas del géiser Old Faithful (Parque Nacional Yellowstone), así como la duración de las mismas. En lo que sigue nos centraremos únicamente en la variable `waiting`, relativa a los tiempos de espera anteriormente mencionados.

En este sentido, comentar que las observaciones han sido recogidas, una vez más, de forma redondeada. De esta forma, se dispone de un total de 12 observaciones correspondientes a 77 minutos, 15 correspondientes a 78 minutos y 10 correspondientes a 79 minutos. Esto justifica, una vez más, el uso de un estimador de la densidad alternativo al de Parzen-Rosenblatt, el cual permita incorporar este *factor agrupación* al proceso de estimación.

Sin embargo, y con el fin de aumentar la magnitud de dicho *factor agrupación*, reagruparemos la muestra observada en intervalos de diferente longitud (cuyos extremos se pueden observar en el histograma de la Figura 5.2). Comentar que esto es lo que se hace, por ejemplo, en Reyes (2015) y en Reyes et al. (2017), aunque definiendo unos intervalos distintos a los aquí considerados. En este caso, los diferentes estimadores tipo núcleo se recogen en la Figura 5.2. Al no disponer de la densidad teórica de la cual proceden los datos, nuevamente se ha optado por construir el estimador tipo núcleo clásico, aunque esta vez en base a la muestra de datos original (sin reagrupar), haciendo uso de la ventana plug-in de Sheather y Jones (1991), que denotaremos por $\hat{h}_{\mathrm{SJ}}$. De esta forma, podremos tomar a dicho estimador como referencia para evaluar el comportamiento del resto de estimadores y, de igual manera, a $\hat{h}_{\mathrm{SJ}}$ como ventana de referencia para el resto de ventanas obtenidas mediante los diferentes selectores adaptados al contexto de datos agrupados.

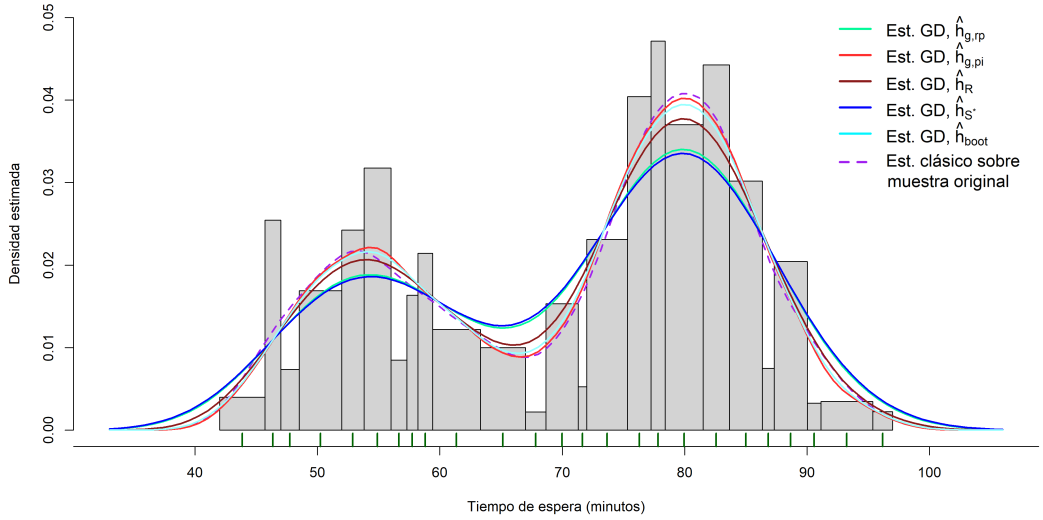


Figura 5.2: Histograma de los tiempos de espera (en minutos) entre erupciones sucesivas del géiser Old Faithful (Parque Nacional de Yellowstone, Wyoming), agrupados en intervalos de diferentes longitudes. Con líneas continuas de colores se representan los estimadores de la densidad de Cao et al. (2011) construidos en base a diferentes ventanas (ver leyenda en la esquina superior derecha). Con una línea morada discontinua se representa el estimador tipo núcleo de Parzen-Rosenblatt, construido en base a la muestra original (sin reagrupar) a través de una ventana plug-in clásica.

Como ya sucedía con la base de datos `stamps`, diferentes ventanas conducen a estimadores con distintos grados de suavidad, siendo nuevamente el selector plug-in propuesto en la Sección (3.2.2) el que logra aproximarse de una manera más satisfactoria al estimador clásico de la densidad construido en base a la muestra sin reagrupar. Destacar también el buen comportamiento que presenta en este caso el selector bootstrap de Reyes et al. (2017), que conduce a un estimador muy similar al obtenido con la ventana $\hat{h}_{\mathrm{g,pi}}$, de manera contraria a lo que sucedía en el ejemplo de la Sección 5.1. Por otro lado, vuelven a ser los selectores de la regla del pulgar y la modificación del selector bootstrap de Jang y Loh (2010) propuesta en la Sección 3.4.1 los que conducen a estimadores más suaves, incapaces de reproducir fielmente las zonas de mayor curvatura del estimador clásico de la densidad. Finalmente, comentar que en este caso los valores de las ventanas estimadas por los diferentes selectores, ordenados de menor a mayor, han sido los siguientes,

$$\hat{h}_{\mathrm{g,pi}} = 2.22, \quad \hat{h}_{\mathrm{boot}} = 2.55, \quad \hat{h}_{\mathrm{R}} = 3.27, \quad \hat{h}_{\mathrm{g,rp}} = 4.64, \quad \hat{h}_{\mathrm{S}^*} = 4.82,$$

mientras que la ventana plug-in clásica empleada en la construcción del estimador de Parzen-Rosenblatt sobre la muestra sin agrupar ha sido $\hat{h}_{\mathrm{SJ}} = 2.50$.

5.3. Casos de COVID-19 en España

En diciembre de 2019 surgió un agrupamiento de casos de neumonía en la ciudad de Wuhan (provincia de Hubei, China), con una exposición común a un mercado mayorista de marisco, pescado y animales vivos. El 7 de enero de 2020, las autoridades chinas identificaron como agente causante del brote un nuevo virus de la familia Coronaviridae, que posteriormente fue denominado SARS-CoV-2, y cuya enfermedad — reconocida como una pandemia global por la Organización Mundial de la Salud (OMS) el 11 de marzo de 2020 — se denominó por consenso internacional COVID-19.

En esta última sección se considerarán los datos relativos a los casos de COVID-19 reportados en España a lo largo de la pandemia y que se encuentran recogidos en los informes semanales emitidos por la Red Nacional de Vigilancia Epidemiológica (RENAVE, ver `https://cnecovid.isciii.es/`). Para ello, se distinguirán las siguientes etapas (ver Figura 5.3):

- **Primer periodo:** desde el inicio de la pandemia hasta el 21 de junio de 2020, fecha en la que se terminó el estado de alarma en España una vez finalizada la primera ola epidémica de COVID-19.
- **Segundo periodo:** desde el 22 de junio hasta el 6 de diciembre de 2020, punto de inflexión de la incidencia acumulada a 14 días de COVID-19, entre el segundo y el tercer periodo epidémico.
- **Tercer periodo:** desde el 7 de diciembre de 2020 hasta el 14 de marzo de 2021, punto de inflexión de la incidencia acumulada a 14 días de casos de COVID-19, entre el tercer y el cuarto periodo epidémico.
- **Cuarto periodo:** desde el 15 de marzo de 2021 hasta el 19 de junio, punto de inflexión de la incidencia acumulada a 14 días de casos de COVID-19, entre el cuarto y el quinto periodo epidémico.
- **Quinto periodo:** desde el 20 de junio de 2021 hasta el 13 de octubre, punto de inflexión de la incidencia acumulada a 14 días de casos de COVID-19, entre el quinto y el sexto periodo epidémico.
- **Sexto periodo:** desde el 14 de octubre de 2021 hasta la actualidad. La última observación recogida data del 5 de enero de 2022, fecha de elaboración del informe analizado (Informe nº 112).

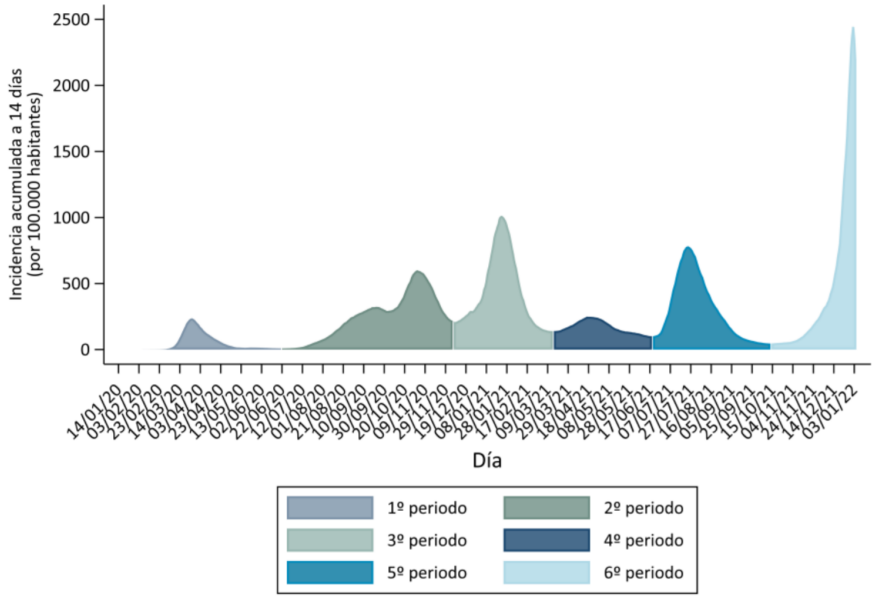


Figura 5.3: Incidencia acumulada a 14 días (por 100.000 habitantes) en España. Fuente: CNE. ISCIII. Red Nacional de Vigilancia Epidemiológica.

En este contexto, la única información de la que se dispone es la proporción de casos registrados para cada uno de los dos sexos y grupos de edad considerados. En este sentido, comentar que en los informes emitidos por la RENAVE se consideran un total de 19 grupos de edad, presentando los dieciocho primeros una longitud constante de 5 años y siendo el último un intervalo no acotado correspondiente a edades superiores a 90 años, el cual identificaremos, por simplicidad, con el intervalo $[90, 113)$[2].

Estos datos constituyen, pues, un ejemplo donde las observaciones vienen recogidas en intervalos de interés — determinados por los distintos grupos de edad — donde únicamente se conocen las proporciones muestrales $w_1, \ldots, w_k$ asociadas, siendo tanto las frecuencias absolutas $n_1, \ldots, n_k$ como el número de casos totales, n, desconocidos; dando lugar a un contexto de datos agrupados en el que la construcción del estimador de Parzen-Rosenblatt no es, ni siquiera, viable. Por este motivo, la estimación de la densidad deberá ser abordada en este caso a través del estimador tipo núcleo de Cao et al. (2011), el cual ha de ser construido en base a la ventana de validación cruzada insesgada propuesta en la Sección 3.21, puesto que es es el único selector de los aquí considerados que puede ser empleado en aquellas situaciones en las que las cantidades $n_1, \ldots, n_k$ son desconocidas.

Consideremos, pues, la colección de intervalos $\{[5(j-1),\, 5j)\}_{j=1}^{18} \cup \{[90, 113)\}$, cuyos puntos medios correspondientes vienen dados por

$$t_j = \begin{cases} \frac{10j-5}{2}, & \text{si} \quad j \in \{1, \ldots, 18\}, \\ 101.5, & \text{si} \quad j = 19. \end{cases}$$

[2]Esta elección se ha hecho en base a la noticia recogida en `https://www.lavanguardia.com/vida/20200511/481096600431/maria-branyas-113-anos-persona-mas-mayor-espana-catalunya-supera-coronavirus.html`.

Fijado un periodo, denotemos por $w_1^{\mathrm{v}}, \ldots, w_{19}^{\mathrm{v}}$ y $w_1^{\mathrm{m}}, \ldots, w_{19}^{\mathrm{m}}$ a los pesos muestrales asociados a hombres y mujeres, respectivamente, tales que verifican la condición $\sum_{j=1}^{19} w_j^{\mathrm{v}} = \sum_{j=1}^{19} w_j^{\mathrm{m}} = 1$ y cuyos valores han sido extraídos del informe nº 112 emitido por la RENAVE. Consideremos, finalmente, los estimadores de Cao et al. (2011) construidos en base a la ventana de validación cruzada insesgada,

$$\hat{f}_w^{g,\mathrm{v}}(x) = \sum_{j=1}^{19} w_j^{\mathrm{v}} L_{\hat{h}^{\mathrm{v}}}(x - t_j), \qquad \hat{f}_w^{g,\mathrm{m}}(x) = \sum_{j=1}^{19} w_j^{\mathrm{m}} L_{\hat{h}^{\mathrm{m}}}(x - t_j), \tag{5.1}$$

donde $\hat{h}^{\mathrm{v}}$ y $\hat{h}^{\mathrm{m}}$ denotan las ventanas estimadas a través del selector de validación cruzada insesgada (adaptado al caso de datos agrupados) en base a las muestras $\{(t_j, w_j^{\mathrm{v}})\}_{j=1}^{19}$ y $\{(t_j, w_j^{\mathrm{m}})\}_{j=1}^{19}$, respectivamente.

En la Figura 5.4 se recogen, para cada sexo y periodo considerado, los histogramas correspondientes a los casos notificados de COVID-19, a los cuales se les han añadido, con líneas continuas de colores, los estimadores (5.1) (azul: hombres, rojo: mujeres). Nótese que en todos los casos se ha empleado la misma escala de representación en el eje de ordenadas, salvo en la figura correspondiente al quinto periodo, donde ha sido necesario aumentarla ligeramente.

Las seis gráficas recogidas en la Figura 5.4 permiten comparar, en cada periodo, la proporción de casos de COVID-19 registrados en hombres con la registrada en mujeres. En líneas generales, podría decirse que los casos asociados a ambos sexos parecen distribuirse de igual manera a lo largo de los diferentes grupos de edad, destacando, quizás, las discrepancias observables entre ambas en el primer y quinto periodo de la pandemia. En el primer caso (primer periodo), las mayores diferencias se encuentran en las edades comprendidas entre los 45 y 90 años, siendo las proporciones de hombres considerablemente mayores en esta franja de edad, invirtiéndose este efecto para el resto de edades. En el segundo caso (quinto periodo), las diferencias radican principalmente en las edades comprendidas entre 10 y 40, siendo nuevamente las proporciones de hombres las que presentan una mayor magnitud. En el resto de periodos, ambas curvas parecen comportarse de manera muy similar, presentando siempre una única moda entorno a los 40-45 años, coincidiendo con la edad media española registrada en 2020.[3]

Observación 5.1. Aunque hasta el momento no se ha comentado nada al respecto, los estimadores de la densidad construidos en la Figura 5.4 evidencian la presencia de un claro efecto frontera — un problema que ya estaba presente en el estimador de Parzen-Rosenblatt — derivado del hecho de que la función de densidad teórica asociada a la variable *edad* presenta un soporte compacto (y, por tanto, acotado). De esta forma, y al igual que sucede en el caso clásico, al considerar la estimación en puntos del intervalo $[0, \hat{h})$, el estimador de Cao et al. (2011) penaliza la no existencia de datos muestrales negativos y, por tanto, tiende a infraestimar el verdadero valor de la densidad teórica en dicha región, asignando, además, masa de probabilidad a valores negativos, algo que, en este contexto, no tiene ningún sentido. De hecho, podría asumirse también un efecto frontera por el lado derecho, puesto que tampoco parece tener demasiado sentido asignar probabilidades positivas a edades excesivamente grandes. Sin embargo, y a la vista de las gráficas presentadas en la Figura 5.4, parece que este último efecto no es demasiado acusado, pudiendo ser incluso ignorado. Ahora bien, sí sería deseable buscar algún procedimiento que permitiese solucionar el efecto frontera por el lado izquierdo. En este sentido, comentar que se ha abordado este problema desde varias perspectivas diferentes, obteniendo resultados en la práctica poco satisfactorios. Por un lado, se ha probado a transformar la muestra original a través de la función $g(x) = \log(x)$ y construir el estimador de la densidad en base a la expresión

$$\hat{f}_w^g(x) = \frac{1}{x} \sum_{j=1}^{k} w_j L_h(\log(x) - \log(t_j)),$$

siendo h la ventana obtenida en base a la muestra transformada $\{(\log(t_j), w_j)\}_{j=1}^{k}$. Sin embargo, las

[3]Información extraída de `https://www.niusdiario.es/sociedad/edad-media-poblacion-espanola-no-para-crecer-situa-43-anos_18_3059445118.html`.

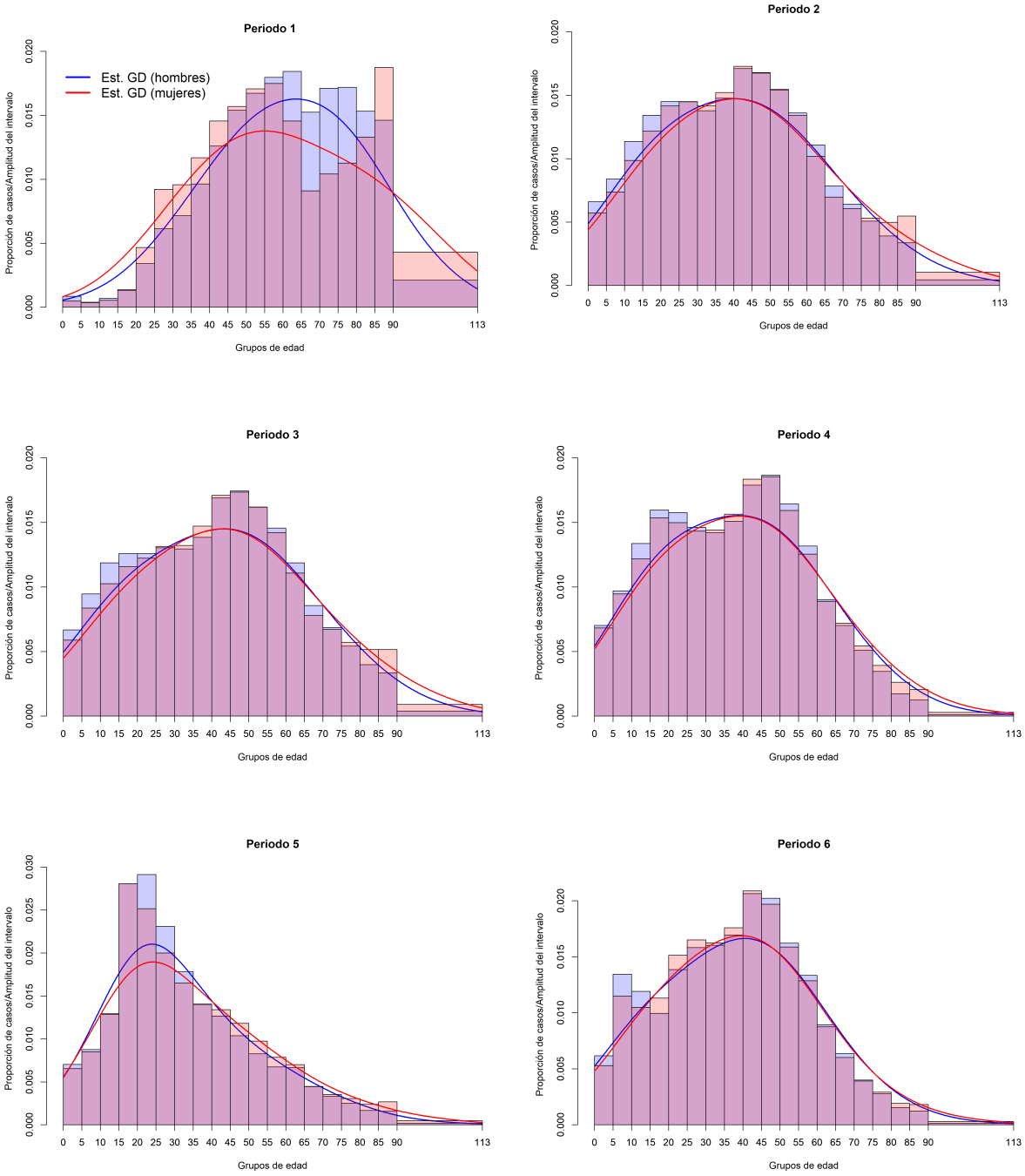


Figura 5.4: Histogramas correspondientes a los casos de COVID-19 notificados a la RENAVE desde el inicio de la pandemia — expresados como proporciones del total de la población y diferenciando por sexos (azul: hombres, rojo: mujeres) — para cada uno de los grupos de edad considerados. Con curvas de colores se representan los correspondientes estimadores de Cao et al. (2011) construidos en base a la ventana de validación cruzada insesgada.

estimaciones obtenidas por este método han resultado ser muy deficientes, puesto que ni siquiera parecían seguir el carácter sugerido por los histogramas, lo cual podría deberse a que al transformar los puntos medios de los intervalos podría estarse incumpliendo alguna de las hipótesis básicas impuestas sobre la longitud de los intervalos (ver suposiciones **(S3)** y **(S4)** del Capítulo 2). Por otro lado — y siguiendo las ideas clásicas de Boneva et al. (1971, páginas 14-17) — se ha empleado también el llamado método de reflexión, consistente en simetrizar[4] la muestra original entorno al origen y construir así el estimador de la densidad

$$\hat{f}^g_{w,h} = \begin{cases} 2\,\hat{g}^g_{w,h}(x), & \text{si } x \geq 0, \\ 0, & \text{en otro caso,} \end{cases}$$

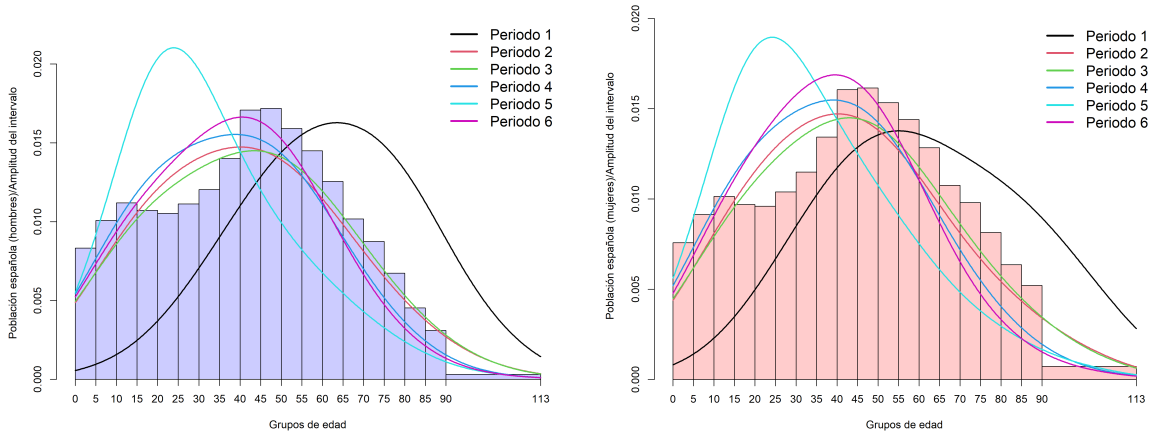
donde $\hat{g}^g_{w,h}$ denota al estimador de Cao et al. (2011) construido en base a la muestra simetrizada $t_1, -t_1, t_2, -t_2, \ldots, t_{19}, -t_{19}$, con pesos $w^{\mathrm{v}}_1/2, \ldots, w^{\mathrm{v}}_{19}/2$ (hombres) y $w^{\mathrm{m}}_1/2, \ldots, w^{\mathrm{m}}_{19}/2$ (mujeres). Sin embargo, en todos los casos este procedimiento ha conducido a estimadores excesivamente sobresuavizados que ocultaban gran parte de la información. También se ha probado — a través de un procedimiento análogo y basándonos en la Sección 2.10 de Silverman (1986) — a considerar una doble simetrización de la muestra — una por cada lado — con el fin de eliminar el efecto frontera en ambos extremos, obteniendo estimadores todavía más sobresuavizados que en el caso anterior. Por este motivo, se ha decidido finalmente ignorar este problema y construir los estimadores de Cao et al. (2011) en base a la muestra agrupada original.

Por último, en la Figura 5.5 se recogen dos gráficas — una para cada sexo — que permiten analizar la evolución de los casos de COVID-19 a lo largo de los diferentes periodos considerados. En cada una de ellas se representan los histogramas correspondientes a la proporción de hombres (Figura 5.5(a)) y mujeres (Figura 5.5(b)) que conforman cada grupo de edad de la población española a 1 de enero de 2021[5], junto con los estimadores de la densidad ya representados en la Figura 5.4 y cuyas expresiones, para un periodo fijo, se recogen en (5.1).

Como era de esperar, ambas gráficas presentan características muy similares, de lo cual se concluye que los diferentes periodos de la pandemia parecen haber afectado a hombres y mujeres de igual manera. A modo de ejemplo, podría decirse que en un inicio (primer periodo) fueron las personas de edad avanzada — tanto hombres como mujeres — las que se vieron más afectadas, puesto que el estimador de la densidad correspondiente (curva negra) presenta una moda entorno a los 60 años. De hecho, si se compara la densidad estimada en el primer periodo con el resto de curvas y con el histograma de las edades de la población española, parece que al inicio de la pandemia se han registrado muchos más casos — en dicho grupo — de los que cabría esperar. Ahora bien, esta moda se ha ido desplazando, en etapas posteriores, hacia la izquierda, concentrando una mayor vulnerabilidad en grupos de edad media, llegando, en el quinto periodo, a concentrarse entorno a los veinticinco años. Este desplazamiento se invierte en la última etapa, en la cual la masa de probabilidad se vuelve a desplazar hacia edades más centrales. Destacar finalmente la drástica reducción que sufre, tras el primer periodo y para edades superiores a 80 años, el área comprendida bajo la curva estimada, lo cual sugiere que los principales esfuerzos se centraron en detectar los casos asociados a este grupo de edad.

[4] En Jones (1993) se presenta, para el caso clásico de datos no agrupados, un análisis del efecto que tiene el método de reflexión sobre la estimación realizada.

[5] Esta información ha sido extraída de la página web del Instituto Nacional de Estadística (INE, `https://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p04/provi/l0/&file=0ccaa003.px&L=0`.)



(a) Hombres

(b) Mujeres

Figura 5.5: Histogramas relativos a la proporción de hombres (Figura 5.5(a)) y mujeres (Figura 5.5(b)) que conforman cada grupo de edad de la población española a 1 de enero de 2021. Con líneas continuas de colores se representan, para cada uno de los seis periodos, los estimadores de Cao et al. (2011) construidos en base a una ventana escogida por un procedimiento de validación cruzada insesgada (ver leyenda correspondiente).

Ahora bien, la Figura 5.5 también permite — en cierta manera — cuantificar el efecto que ha tenido la estrategia de vacunación COVID-19 sobre la evolución de la pandemia, hecho que permite explicar gran parte de los desplazamientos anteriormente mencionados. En este sentido, recordemos que las pautas de vacunación han sido las siguientes: en el tercer periodo se comenzó a distribuir las primeras dosis sobre grupos prioritarios (personal sanitario, residentes y personal en centros de mayores), en el cuarto periodo se incorporaron los grupos de edad superiores a los 50 años y en el quinto, las personas menores de 50 años (para más información puede consultarse `https://www.vacunacovid.gob.es/`). Si bien es cierto que la primera etapa de vacunación (grupos prioritarios) no parece mostrar un efecto notable sobre la evolución de las densidades anteriores, las otras dos sí pueden apreciarse claramente.

De esta forma, en el quinto periodo se observa una drástica reducción — en ambos sexos — de la proporción de casos en las edades superiores a los 50 años, derivada de las dosis administradas en el periodo anterior, encontrándose la moda de la densidad estimada desplazada hacia la izquierda. De igual manera, en el sexto periodo se observa un retorno al patrón de olas anteriores — posiblemente consecuencia de las dosis proporcionadas, a personas menores de 50 años, en el periodo anterior — volviéndose a desplazar la moda de la densidad hacia la edad media de la población española.

Finalmente, la Figura 5.5 evidencia un efecto común a ambos sexos que, exceptuando el primer y quinto periodo, parece perdurar a lo largo de la evolución de la pandemia. En este sentido, entre los 15 y 40 años (intervalo en el cual las densidades estimadas se sitúan por encima del histograma) se observa una proporción de casos considerablemente superior a la que cabría esperar, mientras que entre los 50 y 90 años (región en la cual las curvas se encuentran por debajo del histograma) se observa una proporción siempre inferior a la esperada. Una posible explicación a este efecto, válida para los últimos periodos, radica en el porcentaje de vacunados — con dos dosis — asociado a cada uno de los dos grupos en cuestión, siendo el correspondiente al segundo notablemente superior (un 84 % para edades comprendidas entre los 15 y 40 años frente a un 97 % asociado a edades superiores a los 50 años, ver `https://www.datadista.com/coronavirus/evolucion-de-la-vacunacion-en-espana/`).

Ahora bien, la presencia de este efecto en el segundo y tercer periodo — periodos en los que apenas se administraron vacunas a ningún grupo de edad, ver curvas roja y verde, respectivamente — sugiere la existencia de otros motivos — más allá del porcentaje de vacunados — que justifican la anterior situación. De entre estas razones pueden encontrarse, por ejemplo, los hábitos asociados a cada grupo de edad, que pueden favorecer o disminuir el número de contagios dentro de los respectivos grupos.

Apéndice A

Cálculo de los estimadores de máxima verosimilitud ponderada de una distribución normal

Es bien conocido que la construcción del estimador tipo núcleo de Parzen (1962) y Rosenblatt (1956) depende crucialmente de la elección de un parámetro ventana $h > 0$. En este contexto, la regla del pulgar de Silverman (1986) — recogida al inicio de la Sección 3.1 — parte de la expresión de la ventana asintóticamente óptima en el caso de datos no agrupados

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

y obtiene una estimación de la misma sin más que reemplazar la cantidad $R(f'')$ — desconocida en la práctica — por la que se obtendría asumiendo que f sigue una distribución normal, cuyos parámetros deben ser previamente estimados en base a una muestra dada (frecuentemente a través de técnicas de máxima verosimilitud).

Consideremos ahora un contexto de datos agrupados, en el cual se emplea al estimador tipo núcleo de Cao et al. (2011) como estimador de la función de densidad. En este caso, se ha mostrado en el Capítulo 2 que la ventana asintóticamente óptima adopta la misma expresión que en el caso clásico,

$$h_{\text{AMISE}_g} = \left[\frac{R(L)}{\mu_2(L)^2 R(f'')n}\right]^{\frac{1}{5}},$$

con la única diferencia de que ahora la estimación de $R(f'')$ debe realizarse en base a una muestra de datos agrupados. De esta forma, la extensión del selector de Silverman (1981) al contexto de datos agrupados requeriría de un nuevo mecanismo de estimación — diferente al de máxima verosimilitud, que otorga a todos los datos de la muestra el mismo peso en la estimación — que permita obtener estimaciones de los parámetros de una distribución normal a partir de una muestra de datos agrupados.

En este sentido, se hará uso de la llamada función de verosimilitud ponderada (véase Hu y Zidek (2002)), la cual generaliza el método de estimación clásico de máxima verosimilitud permitiendo que determinados valores de la muestra ejerzan una mayor influencia sobre la estimación de los parámetros (condición que parece natural exigir en un contexto de datos agrupados, en donde cada valor t_j tiene asociado un peso w_j distinto). Comentar que en Shirazi et al. (2021) se aborda también este problema, proponiendo otros métodos de estimación diferentes al aquí presentado, pero en los que se requiere que las frecuencias absolutas $n_1, \ldots, n_k$ sean conocidas.

Con el fin de obtener los estimadores de máxima verosimilitud ponderada de los parámetros de una distribución normal, consideraremos una m.a.s. $X_1, \ldots, X_n$ procedente de una variable aleatoria $X \sim N(\mu, \sigma^2)$. Supongamos que, a partir de estos valores, se observa la muestra $t_1, \overset{n_1}{\ldots}, t_1, \ldots, t_k, \overset{n_k}{\ldots}, t_k$, y definamos los pesos $w_1, \ldots, w_k$ como $w_j = n_j/n$, $j = 1, \ldots, k$, de tal forma que $w_j \geq 0$ para todo $j = 1, \ldots, k$ y $w_1 + \cdots + w_k = 1$. Basándonos en las ideas de Hu y Zidek (2002), se considera la función de verosimilitud ponderada

$$\mathcal{L}_{t,w}(\mu, \sigma^2) = \prod_{j=1}^{k} \phi_{\mu,\sigma}(t_j)^{w_j} = \prod_{j=1}^{k} \left[\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(t_j-\mu)^2}{2\sigma^2}} \right]^{w_j}, \tag{A.1}$$

donde $\phi_{\mu,\sigma}$ denota la función de densidad de una distribución normal de media μ y desviación típica σ. Los estimadores de máxima verosimilitud ponderada, $\hat{\mu}_{\text{MVP}}$ y $\hat{\sigma}^2_{\text{MVP}}$, serán aquellos que maximicen la función (A.1), de manera que

$$\left(\hat{\mu}_{\text{MVP}}, \hat{\sigma}^2_{\text{MVP}}\right) = \underset{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+}{\arg\max}\ \mathcal{L}_{t,w}(\mu, \sigma^2).$$

Teniendo en cuenta que la función logaritmo neperiano — que de ahora en adelante denotaremos por log — es monótona creciente, entonces dichos estimadores serán también máximos de la función de log-verosimilitud ponderada, $\ell_{t,w}(\mu, \sigma^2) = \log\left(\mathcal{L}_{t,w}(\mu, \sigma^2)\right)$. En este caso,

$$\ell_{t,w}(\mu, \sigma^2) = \log\left(\mathcal{L}_{t,w}(\mu, \sigma^2)\right) = \sum_{j=1}^{k} w_j \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(t_j-\mu)^2}{2\sigma^2}} \right) = -\sum_{j=1}^{k} w_j \left[\frac{\log\left(2\pi\sigma^2\right)}{2} + \frac{(t_j-\mu)^2}{2\sigma^2} \right] =$$

$$= -\frac{1}{2}\left[\log\left(2\pi\sigma^2\right) \sum_{j=1}^{k} w_j + \frac{1}{\sigma^2} \sum_{j=1}^{k} w_j (t_j-\mu)^2 \right] = -\frac{1}{2}\left[\log\left(2\pi\sigma^2\right) + \frac{1}{\sigma^2} \sum_{j=1}^{k} w_j (t_j-\mu)^2 \right],$$

donde en la última igualdad se ha empleado que $\sum_{j=1}^{k} w_j = 1$. Derivando la expresión anterior respecto de μ e igualándola a cero (condición necesaria de optimalidad), resulta

$$\frac{\partial \ell_{t,w}(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{j=1}^{k} w_j (t_j - \mu) = \frac{1}{\sigma^2} \left(\sum_{j=1}^{k} w_j t_j - \mu \sum_{j=1}^{k} w_j \right) = \frac{1}{\sigma^2} \left(\sum_{j=1}^{k} w_j t_j - \mu \right) = 0 \iff$$

$$\iff \sum_{j=1}^{k} w_j t_j - \mu = 0 \iff \mu = \sum_{j=1}^{k} w_j t_j,$$

obteniendo así un candidato a estimador de máxima verosimilitud ponderada para μ (que no es más que la media muestral de $t_1, \ldots, t_k$ ponderada por los pesos $w_1, \ldots, w_k$). Siguiendo un procedimiento totalmente análogo para el parámetro σ^2 resulta

$$\frac{\partial \ell_{t,w}(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2}\left[\frac{1}{\sigma^2} - \frac{1}{\sigma^4} \sum_{j=1}^{k} w_j (t_j-\mu)^2 \right] = 0 \iff \frac{1}{\sigma^2} = \frac{1}{\sigma^4} \sum_{j=1}^{k} w_j (t_j-\mu)^2 \iff$$

$$\iff \sigma^2 = \sum_{j=1}^{k} w_j (t_j-\mu)^2.$$

Finalmente, se procede al cálculo de las derivadas parciales de segundo orden, con el fin de comprobar que los valores obtenidos — que denotaremos por $\hat{\mu}_{\text{MVP}}$ y $\hat{\sigma}^2_{\text{MVP}}$, respectivamente — son, efectivamente, máximos de la función de log-verosimilitud (y, por tanto, de la función de verosimilitud) ponderada. En lo referido a la segunda derivada respecto de μ,

$$\frac{\partial^2 \ell_{t,w}(\mu, \sigma^2)}{\partial \mu^2} = \frac{\partial}{\partial \mu}\left[\frac{1}{\sigma^2} \left(\sum_{j=1}^{k} w_j t_j - \mu \right) \right] = -\frac{1}{\sigma^2} < 0, \qquad \forall \mu \in \mathbb{R}.$$

Como $\partial^2 \ell_{t,w}/\partial \mu^2$ resulta ser negativa para cualquier valor del parámetro μ, entonces, en concreto, también lo será para $\mu = \hat{\mu}_{\text{MVP}}$. Por otro lado, la derivada de segundo orden respecto de σ^2 es

$$\frac{\partial^2 \ell_{t,w}(\mu, \sigma^2)}{\partial \sigma^4} = \frac{\partial}{\partial \sigma^2} \left\{ -\frac{1}{2} \left[\frac{1}{\sigma^2} - \frac{1}{\sigma^4} \sum_{j=1}^{k} w_j (t_j - \mu)^2 \right] \right\} = -\frac{1}{2} \left[-\frac{1}{\sigma^4} + \frac{2}{\sigma^6} \sum_{j=1}^{k} w_j (t_j - \mu)^2 \right] =$$

$$= \frac{1}{\sigma^4} \left[\frac{1}{2} - \frac{1}{\sigma^2} \sum_{j=1}^{k} w_j (t_j - \mu)^2 \right].$$

Evaluando la derivada anterior en $\sigma^2 = \hat{\sigma}^2_{\text{MVP}}$ se obtiene

$$\left. \frac{\partial^2 \ell_{t,w}(\mu, \sigma^2)}{\partial \sigma^4} \right|_{(\mu, \hat{\sigma}^2_{\text{MVP}})} = \frac{1}{\left[\sum_{j=1}^{k} w_j (t_j - \mu)^2 \right]^2} \left(\frac{1}{2} - 1 \right) = -\frac{1}{2 \left[\sum_{j=1}^{k} w_j (t_j - \mu)^2 \right]^2} < 0,$$

que, nuevamente, resulta ser estrictamente negativa, de donde se sigue que en $\sigma^2 = \hat{\sigma}^2_{\text{MVP}}$ la función de verosimilitud ponderada alcanza un máximo. Se concluye, pues, que los estimadores de máxima verosimilitud ponderada de una distribución normal de media μ y varianza σ^2 son

$$\hat{\mu}_{\text{MVP}} = \sum_{j=1}^{k} w_j t_j \quad \text{y} \quad \hat{\sigma}^2_{\text{MVP}} = \sum_{j=1}^{k} w_j (t_j - \hat{\mu}_{\text{MVP}})^2, \tag{A.2}$$

de manera que cada t_j ejercerá un peso proporcional a w_j sobre la estimación de los parámetros.

Observación A.1. Teniendo en cuenta que $w_j = n_j/n$, para todo $j = 1, \ldots, k$, resulta inmediato que (A.2) puede reescribirse como

$$\hat{\mu}_{\text{MVP}} = \frac{1}{n} \sum_{j=1}^{k} n_j t_j \quad \text{y} \quad \hat{\sigma}^2_{\text{MVP}} = \frac{1}{n} \sum_{j=1}^{k} n_j (t_j - \hat{\mu}_{\text{MVP}})^2.$$

Esto permite concluir que, cuando las frecuencias absolutas $n_1, \ldots, n_k$ sean conocidas, $\hat{\mu}_{\text{MVP}}$ y $\hat{\sigma}^2_{\text{MVP}}$ coinciden con las expresiones clásicas de la media y varianza muestral (estimadores de máxima verosimilitud) aplicadas sobre la muestra $t_1, \overset{n_1}{\ldots}, t_1, \ldots, t_k, \overset{n_k}{\ldots}, t_k$. Ahora bien, si las cantidades $n_1, \ldots, n_k$ son desconocidas (y, en su defecto, solo se dispone de los pesos $w_1, \ldots, w_k$), entonces el enfoque clásico de máxima verosimilitud se limitaría a asumir que únicamente se observan las k observaciones asociadas a $t_1, \ldots, t_k$, ignorando la información proporcionada por los pesos $w_1, \ldots, w_k$. En tal caso, es de esperar que dichos estimadores no sean consistentes, como, de hecho, se comprobará a continuación a través de un estudio de simulación.

Finalmente, y siguiendo lo comentado en la Observación A.1, se ha llevado a cabo un estudio de simulación que permite comparar — en un contexto de datos agrupados en el cual las cantidades $n_1, \ldots, n_k$ son desconocidas — el comportamiento de los estimadores de máxima verosimilitud ponderada frente a los de máxima verosimilitud clásica. Para ello, consideremos un total de $B = 5000$ muestras de tamaño $n \in \{50, 500, 1000\}$ procedentes de diferentes distribuciones normales $N(\mu, \sigma^2)$ de referencia, las cuales han sido posteriormente agrupadas siguiendo el esquema recogido en la Sección 4.1.2.

Para cada una de estas muestras de datos agrupados, se ha llevado a cabo la estimación de la varianza poblacional[1] σ^2 a través de los estimadores

$$\hat{\sigma}^2_{\text{MVP}} = \sum_{j=1}^{k} w_j (t_j - \hat{\mu}_g)^2 \quad \text{y} \quad S_k^2 = \frac{1}{k} \sum_{j=1}^{k} (t_j - \bar{t}\,)^2,$$

— donde $\bar{t} = \frac{1}{k}\sum_{j=1}^{k} t_j$ denota a la media muestral de los valores $t_1, \ldots, t_k$ — correspondientes a los estimadores de máxima verosimilitud ponderada y de máxima verosimilitud de σ^2, respectivamente. Finalmente, se ha aproximado el error cuadrático medio de ambos estimadores a través de la expresión (enunciada en términos de un estimador genérico $\hat{\theta}$ de θ)

$$\text{ECM}(\hat{\theta}) \approx \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\theta}^{(i)} - \theta \right)^2,$$

donde $\hat{\theta}^{(i)}$ denota al estimador $\hat{\theta}$ evaluado sobre la muestra $i-$ésima. Por último, comentar que como distribuciones de referencia se han considerado densidades normales de parámetros $\mu = 0$ y $\sigma^2 \in \{1, 5^2, 10^2\}$. Los resultados se recogen en la Tabla (A.1), todos ellos aproximados a cuatro cifras decimales.

	$\boldsymbol{N(0,1)}$			$\boldsymbol{N(0,5^2)}$			$\boldsymbol{N(0,10^2)}$		
	50	**500**	**1000**	**50**	**500**	**1000**	**50**	**500**	**1000**
ECM $\left(\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\text{MVP}}\right)$	0.0395	0.0039	0.0019	24.0848	2.4447	1.1860	384.7540	39.1133	18.9746
ECM $\left(\boldsymbol{S}_{\boldsymbol{k}}^{\boldsymbol{2}}\right)$	3.2405	5.3470	7.1269	637.0372	2989.7211	4209.7607	8692.0333	47166.4948	66882.8410

Tabla A.1: Aproximaciones del error cuadrático medio de $\hat{\sigma}^2_{\text{MVP}}$ (primera fila) y de S_k^2 (segunda fila) como estimadores de la varianza poblacional, obtenidas a partir de $B = 5000$ muestras de tamaño $n \in \{50, 500, 1000\}$ agrupadas siguiendo el esquema recogido en la Sección 4.1.2. Como densidades teóricas se han considerado densidades normales de media cero y varianza $\sigma^2 \in \{1, 25, 100\}$.

La Tabla A.1 hace evidente la mejora que se produce en la estimación cuando se considera a $\hat{\sigma}^2_{\text{MVP}}$ como estimador de la varianza poblacional, puesto que en todos los casos ha conducido a un error cuadrático medio muy inferior al proporcionado por S_k^2. De hecho, en las tres densidades de referencia consideradas se observa que el error cuadrático medio de S_k^2 aumenta conforme lo hace el tamaño muestral, lo cual parece indicar que, en un contexto general de datos agrupados, el estimador de máxima verosimilitud de la varianza deja de ser un estimador consistente, lo cual justifica el empleo de un estimador alternativo.

[1] En este caso, el estudio de simulación se ha restringido a la estimación de la varianza poblacional, puesto que es el parámetro que interesa estimar en el selector de la regla del pulgar (ver Sección 3.1). Sin embargo, un procedimiento totalmente análogo podría realizarse para aproximar el error cuadrático medio de $\hat{\mu}_g$ y $\bar{t}$ (estimadores de máxima verosimilitud ponderada y de máxima verosimilitud de μ, respectivamente).

Apéndice B

Sobre la estimación de $R(f'')$

El selector plug-in propuesto por Reyes et al. (2017) (y recogido en la Sección 3.2.1) se fundamenta en reemplazar la cantidad desconocida $R(f'')$ — presente en la expresión teórica de la ventana global asintóticamente óptima — por el estimador $\hat{\psi}^g_{4,\eta_4}$, cuya expresión para un $r \in \mathbb{N}$[1] genérico viene dada por

$$\hat{\psi}^g_{r,\eta_r} = \frac{1}{(\eta_r)^{r+1}} \sum_{i=1}^{k} \sum_{j=1}^{k} L^{(r)}\left(\frac{t_i - t_j}{\eta_r}\right) w_i w_j. \tag{B.1}$$

De esta forma, la construcción de $\hat{\psi}^g_{r,\eta_r}$ requiere de la elección previa de una ventana piloto η_r. Un posible criterio para la selección de esta ventana consiste en escoger aquel valor $\eta_r > 0$ que minimice el error cuadrático medio de $\hat{\psi}^g_{r,\eta_r}$ como estimador de $\psi_r = \mathbb{E}[f^{(r)}(X)]$. Como ya se ha comentado en la Sección 3.2.1, dicha minimización puede ser abordada de dos formas distintas, donde ninguna ha resultado ser universalmente mejor que la otra (para ver una discusión más detallada acerca de este tema puede consultarse el Material Complementario de Reyes et al. (2017)). Tras varios estudios de simulación, en Reyes et al. (2017) se propone considerar como ventana piloto a

$$\eta^{\text{ópt}}_{r,s} = \left[-\frac{2L^{(r)}(0)R(f)\bar{l}}{\mu_2(L)\psi_{r+s}}\right]^{\frac{1}{r+s+1}}, \tag{B.2}$$

donde s denota el orden de la función núcleo L empleada en la construcción de (B.1). En este apéndice se probará que, si L es una función núcleo de orden dos ($s = 2$), entonces (B.2) es la única ventana que minimiza (asintóticamente) el error cuadrático medio de $\hat{\psi}^g_{r,\eta_r}$ como estimador de ψ_r, conduciendo, además, a un estimador consistente en media cuadrática, lo cual permite confirmar la conjetura de Reyes et al. (2017) para el caso particular de funciones núcleo de orden dos, cuando $n \to \infty$. Para ello, en lo que sigue se asumirá que:

(**A1**): La función núcleo L (que supondremos de orden $s = 2$) es una función de densidad simétrica con soporte compacto $[-1, 1]$ y $r + 1$ veces continuamente diferenciable.

(**A2**): La función de distribución F es una función $p + 1$ veces diferenciable, con soporte compacto $[\mathcal{L}, \mathcal{U}]$, tal que $F^{(p+1)}$ es continua para algún $p \geq r$ y tal que $R(F^{(r+1)}) = R(f^{(r)}) < \infty$.

(**A3**): La ventana $\eta \equiv \eta_n$ es una sucesión no aleatoria de números positivos tal que $\lim_{n\to\infty} \eta = 0$ y $\lim_{n\to\infty} n\eta^{2r} = \infty$.

[1] Como ya se ha justificado en la Sección 3.2.1, para estimar $R(f^{(r)})$ es suficiente estudiar la estimación de los funcionales ψ_r para $r \in \mathbb{N}$ par. Por este motivo, en lo que sigue asumiremos que r es un número natural par.

(A4): Dado un conjunto de $k \equiv k_n$ intervalos $\{[y_{j-1}, y_j)\}_{j=1}^k$ tales que $y_0 \leq \mathcal{L}$ e $y_k \geq \mathcal{U}$, asumiremos que su longitud media $\bar{l} = \frac{1}{k}\sum_{j=1}^k l_j$ — donde l_j denota la longitud del $j-$ésimo intervalo — verifica que

$$\lim_{n\to\infty} \bar{l} = 0, \qquad \lim_{n\to\infty} n\bar{l} = \infty, \qquad \bar{l} = o(\eta^{r+1}), \qquad \max_{1\leq j\leq k} |l_j - \bar{l}| = o(\bar{l}).$$

Bajo las condiciones **(A1)-(A4)**, el Teorema 1 de Reyes et al. (2017) recoge las expresiones del sesgo y la varianza de $\hat{\psi}^g_{r,\eta_r}$ que, adaptadas al caso de funciones núcleo de orden $s = 2$, son[2]

$$\text{Sesgo}\left(\hat{\psi}^g_{r,\eta}\right) = S_1(\eta) + S_2\left(\eta, \bar{l}\right) + S_3\left(\eta, \bar{l}, n\right), \tag{B.3}$$

$$\text{Var}\left(\hat{\psi}^g_{r,\eta}\right) = \frac{4}{n}\left[\int f(z)f^{(r)}(z)^2 \, \mathrm{d}z - \left(\int f(z)f^{(r)}(z) \, \mathrm{d}z\right)^2\right] + V_1\left(\eta, \bar{l}, n\right) + o(n^{-1} + \bar{l}n^{-1}\eta^{-2r-1}), \tag{B.4}$$

donde

$$S_1(\eta) = \frac{\eta^2}{2}\mu_2(L)\psi_{r+2}, \qquad S_3\left(\eta, \bar{l}, n\right) = O\left(\eta^4\right) + O\left(\frac{1}{n\eta^{r+1}}\right) + O\left(\frac{\bar{l}^{\,2}}{\eta^{r+1}}\right),$$

$$S_2\left(\eta, \bar{l}\right) = \frac{\bar{l}}{\eta^{r+1}}L^{(r)}(0)R(f), \qquad V_1\left(\eta, \bar{l}, n\right) = \frac{4\bar{l}}{n\eta^{2r+1}}R(L^{(r)})\left(\int f(z)^3 \, \mathrm{d}z\right).$$

Nótese que, gracias a las suposiciones **(A1)** y **(A2)**, las integrales involucradas en la expresión de la varianza son finitas, así como las cantidades $\mu_2(L)$, $L^{(r)}(0)$, $R(f)$ y $R(L^{(r)})$. Las expresiones (B.3) y (B.4) permiten obtener una ventana η asintóticamente óptima en el sentido del error cuadrático medio,

$$\eta^{\text{óptimo}} = \arg\min_{\eta>0}\left[\text{AMSE}\left(\hat{\psi}^g_{r,\eta}\right)\right] = \arg\min_{\eta>0}\left[\text{ASesgo}\left(\hat{\psi}^g_{r,\eta}\right)^2 + \text{AVar}\left(\hat{\psi}^g_{r,\eta}\right)\right],$$

donde $\text{ASesgo}(\hat{\psi}^g_{r,\eta})$ y $\text{AVar}(\hat{\psi}^g_{r,\eta})$ denotan las respectivas versiones asintóticas del sesgo y de la varianza de $\hat{\psi}^g_{r,\eta}$ (obtenidas de sumprimir los términos asintóticos en (B.3) y (B.4), respectivamente). A continuación se recoge la proposición que contiene el resultado principal de este apéndice.

Proposición B.1. *Sea $t_1, \ldots, t_k$ una muestra de datos agrupados con proporciones muestrales respectivas $w_1, \ldots, w_k$. Sea $\psi_r = \mathbb{E}[f^{(r)}(X)]$ y consideremos el estimador*

$$\hat{\psi}^g_{r,\eta} = \frac{1}{\eta^{r+1}}\sum_{i=1}^k\sum_{j=1}^k L^{(r)}\left(\frac{t_i - t_j}{\eta}\right)w_i w_j \tag{B.5}$$

dependiente de una ventana piloto $\eta > 0$, donde L es una función núcleo de orden dos tal que, para todo $r \in \mathbb{N}$ par, verifica

$$(-1)^{\frac{r+2}{2}+1}L^{(r)}(0) > 0.$$

*Bajo las condiciones **(A1)-(A4)**, la única ventana que minimiza el error cuadrático medio (asintótico) de $\hat{\psi}^g_{r,\eta}$ como estimador de ψ_r es la propuesta por Reyes et al. (2017),*

$$\eta^{\acute{o}pt} = \left[-\frac{2\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}}\right]^{\frac{1}{r+3}}.$$

Además, $\hat{\psi}^g_{r,\eta^{\acute{o}pt}}$ es un estimador consistente en media cuadrática.

[2]Por no cargar en exceso la notación, en lo que sigue denotaremos $\eta \equiv \eta_r$.

Demostración. Supongamos, sin pérdida de generalidad, que la longitud media de los intervalos considerados viene dada por $\bar{l} = Cn^{-\alpha} > 0$, para algún $C \in \mathbb{R}^+$ y $\alpha \in \mathbb{R}$. Con esta notación, es inmediato que la hipótesis $\lim_{n\to\infty} \bar{l} = 0$ se verificará si y solo si $\alpha > 0$. Por otro lado, tampoco es difícil comprobar que la condición $\lim_{n\to\infty} n\bar{l} = \infty$ equivale a pedir que $\alpha < 1$. En efecto,

$$\lim_{n\to\infty} n\bar{l} = \infty \iff \lim_{n\to\infty} n(Cn^{-\alpha}) = \infty \iff \lim_{n\to\infty} Cn^{1-\alpha} = \infty \iff 1-\alpha > 0 \iff \alpha < 1.$$

De esta manera, bajo las condiciones impuestas en el enunciado de la proposición, se tiene que $\alpha \in (0,1)$.

Sea $\hat{\psi}^g_{r,\eta}$ el estimador definido en (B.5) y consideremos su error cuadrático medio asintótico (AMSE) como estimador de ψ_r, construido — bajo las condiciones **(A1)-(A4)** — en base a las versiones asintóticas de (B.3) y (B.4),

$$\begin{aligned}\mathrm{AMSE}\left(\hat{\psi}^g_{r,\eta}\right) &= \mathrm{ASesgo}\left(\hat{\psi}^g_{r,\eta}\right)^2 + \mathrm{AVar}\left(\hat{\psi}^g_{r,\eta}\right) = \\ &= \left(S_1(\eta) + S_2\left(\eta,\bar{l}\right)\right)^2 + \frac{4}{n}\left[\int f(z)f^{(r)}(z)^2\,\mathrm{d}z - \left(\int f(z)f^{(r)}(z)\,\mathrm{d}z\right)^2\right] + V_1\left(\eta,\bar{l},n\right).\end{aligned}$$

En lo que sigue se procederá a obtener la ventana $\eta^{\text{ópt}}$ que verifica

$$\eta^{\text{ópt}} = \underset{\eta>0}{\arg\min}\left[\mathrm{AMSE}\left(\hat{\psi}^g_{r,\eta}\right)\right] \overset{(*)}{=} \underset{\eta>0}{\arg\min}\left[\left(S_1(\eta) + S_2\left(\eta,\bar{l}\right)\right)^2 + V_1\left(\eta,\bar{l},n\right)\right], \tag{B.6}$$

donde en $(*)$ se ha tenido en cuenta que el primer sumando de $\mathrm{AVar}\left(\hat{\psi}^g_{r,\eta}\right)$ no depende de η.

Para ello — y siguiendo un procedimiento análogo al marcado en Wand y Jones (1995) para el caso clásico de datos no agrupados — asumiremos inicialmente que $V_1(\eta,\bar{l},n) = o(S_2(\eta,\bar{l})^2)$. En tal caso, la ventana $\eta^{\text{ópt}}$ se puede obtener minimizando únicamente el cuadrado de la suma de los dos términos principales del sesgo de $\hat{\psi}^g_{r,\eta}$[3],

$$\eta^{\text{ópt}} = \underset{\eta>0}{\arg\min}\left[\left(S_1(\eta) + S_2\left(\eta,\bar{l}\right)\right)^2\right] = \underset{\eta>0}{\arg\min}\left[\mathrm{ASesgo}\left(\hat{\psi}^g_{r,\eta}\right)^2\right].$$

En este sentido, las suposiciones **(A1)** y **(A2)** permiten derivar el cuadrado de la versión asintótica del sesgo, $\mathrm{ASesgo}(\hat{\psi}^g_{r,\eta})^2$, respecto de η, resultando

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\eta}\mathrm{ASesgo}\left(\hat{\psi}^g_{r,\eta}\right)^2 &= \frac{\mathrm{d}}{\mathrm{d}\eta}\left[\frac{\eta^4}{4}\mu_2(L)^2\psi^2_{r+2} + \frac{\eta^2}{\eta^{r+1}}\mu_2(L)\psi_{r+2}\,\bar{l}\,L^{(r)}(0)R(f) + \frac{\bar{l}^{\,2}}{\eta^{2r+2}}L^{(r)}(0)^2R(f)^2\right] = \\ &= \eta^3\mu_2(L)^2\psi^2_{r+2} + \frac{(1-r)\bar{l}}{\eta^r}\,\mu_2(L)\psi_{r+2}L^{(r)}(0)R(f) - \frac{(2r+2)\bar{l}^{\,2}}{\eta^{2r+3}}L^{(r)}(0)^2R(f)^2.\end{aligned} \tag{B.7}$$

Igualando (B.7) a cero y multiplicando ambos lados de la igualdad por $\eta^{2r+3} > 0$, se obtiene

$$\eta^{2r+6}\mu_2(L)^2\psi^2_{r+2} + (1-r)\eta^{r+3}\bar{l}\,\mu_2(L)\psi_{r+2}L^{(r)}(0)R(f) - (2r+2)\bar{l}^{\,2}L^{(r)}(0)^2R(f)^2 = 0.$$

[3]Este procedimiento es el que se emplea en Reyes et al. (2017) para obtener la ventana $\eta^{\text{ópt}}$ que allí se propone.

Finalmente, dividiendo ambos lados de la expresión anterior por $\bar{l}\,\mu_2(L)\psi_{r+2}L^{(r)}(0)R(f)^4$ y reorganizando los factores involucrados,

$$\frac{\mu_2(L)\psi_{r+2}}{\bar{l}\,L^{(r)}(0)R(f)}\left(\eta^{r+3}\right)^2+(1-r)\eta^{r+3}-\frac{(2r+2)\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}}=0,$$

resultando así en una ecuación de segundo grado en la variable η^{r+3}, cuyas dos soluciones son

$$\eta_1^{r+3}=-\frac{2\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}} \quad \text{y} \quad \eta_2^{r+3}=\frac{(r+1)\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}},$$

dando lugar a dos posibles mínimos del sesgo asintótico de $\hat{\psi}_{r,\eta}^{g}$ (al cuadrado),

$$\eta_1=\left[-\frac{2\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}}\right]^{\frac{1}{r+3}} \quad \text{y} \quad \eta_2=\left[\frac{(r+1)\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}}\right]^{\frac{1}{r+3}}.$$

A continuación se procede a estudiar el signo de η_1 y η_2. Por un lado, como r es par, entonces $\text{signo}(\psi_{r+2})=\text{signo}(-1)^{\frac{r+2}{2}}$ (ver Sección 3.5 de Wand y Jones (1995)), de donde se sigue que $\text{signo}(-1/\psi_{r+2})=\text{signo}[(-1)^{\frac{r+2}{2}+1}]$. De esta forma, para todo $r\in\mathbb{N}$ par se tendrá que

$$\text{signo}(\eta_1)=\text{signo}\left(\frac{L^{(r)}(0)}{(-1)^{\frac{r+2}{2}+1}}\right)=\text{signo}\left((-1)^{\frac{r+2}{2}+1}L^{(r)}(0)\right).$$

Como por hipótesis $(-1)^{(r+2)/2+1}L^{(r)}(0)>0$, entonces es claro que $\eta_1>0$. Por el contrario, no es difícil comprobar que η_2 es siempre estrictamente negativo, puesto que

$$\eta_2=\frac{(r+1)^{\frac{1}{r+3}}}{(-1)^{\frac{1}{r+3}}}\,\eta_1=\sqrt[r+3]{-1}\,(r+1)^{\frac{1}{r+3}}\,\eta_1=-(r+1)^{\frac{1}{r+3}}\,\eta_1<0,$$

para todo $r\in\mathbb{N}$ par. Descartando entonces la solución negativa, se llega a que

$$\eta^{\text{ópt}}=\left[-\frac{2\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}}\right]^{\frac{1}{r+3}} \tag{B.8}$$

es la única ventana[5] que minimiza $\text{ASesgo}(\hat{\psi}_{r,\eta}^{g})^2$. Nótese que la condición $\lim_{n\to\infty}\eta^{\text{ópt}}=0$ está garantizada para cualquier $r\in\mathbb{N}$ par gracias a la suposición **(A4)**, bajo la cual se impone que la longitud media de los intervalos considerados, $\bar{l}$, converja a cero conforme el tamaño muestral aumenta.

[4]Es sencillo probar que esta cantidad es siempre distinta de cero. Por un lado, dada su definición, es obvio que $\bar{l}$ y $R(f)$ son cantidades estrictamente positivas. Por otro lado, al ser L una función núcleo de orden dos, se tiene que $\mu_2(L)>0$. Finalmente, es conocido que el signo de ψ_s coincide con el signo de $(-1)^{s/2}$ si s es par y que ψ_s vale 0 si s es impar (ver Sección 3.5 de Wand y Jones (1995)), de manera que $\text{signo}(\psi_{r+2})=\text{signo}(-1)^{(r+2)/2}$. De esta forma,

$$\text{signo}\left(\bar{l}\,\mu_2(L)\psi_{r+2}L^{(r)}(0)R(f)\right)=\text{signo}\left((-1)^{\frac{r+2}{2}}L^{(r)}(0)\right).$$

Como por hipótesis $(-1)^{(r+2)/2+1}L^{(r)}(0)>0$, entonces se concluye que la cantidad $\bar{l}\,\mu_2(L)\psi_{r+2}L^{(r)}(0)R(f)$ es estrictamente negativa para todo $r\in\mathbb{N}$ par.

[5]Esta expresión de $\eta^{\text{ópt}}$ es precisamente la que se propone en Reyes et al. (2017).

Acabamos de probar, pues, que la ventana (B.8) es la única que minimiza la cantidad AMSE($\hat{\psi}^g_{r,\eta}$), siempre y cuando la condición $V_1(\eta,\bar{l},n) = o(S_2(\eta,\bar{l})^2)$ se verifique. En este sentido, se probará a continuación que, bajo las condiciones impuestas en el enunciado, esta igualdad será siempre cierta, puesto que no existe ninguna ventana η tal que $S_2(\eta,\bar{l})^2 = O(V_1(\eta,\bar{l},n))$.

Supongamos, por reducción al absurdo, que existe una ventana $\eta > 0$ verificando las condiciones **(A1)-(A4)** tal que $S_2(\eta,\bar{l})^2 = O(V_1(\eta,\bar{l},n))$. Entonces

$$S_2(\eta,\bar{l})^2 = O(V_1(\eta,\bar{l},n)) \Longleftrightarrow \frac{\bar{l}^2}{\eta^{2r+2}} = O\left(\frac{\bar{l}}{n\eta^{2r+1}}\right) \Longleftrightarrow \frac{\eta^{2r+1}\bar{l}^2}{\eta^{2r+2}\bar{l}} = O(n^{-1}) \Longleftrightarrow \frac{\bar{l}}{\eta} = O(n^{-1}).$$

Como $\bar{l} = O(n^{-\alpha})$ y teniendo en cuenta que $\lim_{n\to\infty}\eta = 0$ (condición **(A3)**),

$$S_2(\eta,\bar{l})^2 = O(V_1(\eta,\bar{l},n)) \Longleftrightarrow \frac{n^{-\alpha}}{\eta} = O(n^{-1}) \Longleftrightarrow \frac{n^{1-\alpha}}{\eta} = O(1) \Longleftrightarrow \limsup_{n\to\infty}\left|\frac{n^{1-\alpha}}{\eta}\right| < \infty \Longleftrightarrow$$
$$\Longleftrightarrow \lim_{n\to\infty} n^{1-\alpha} = 0 \Longleftrightarrow 1-\alpha < 0 \Longleftrightarrow \alpha > 1,$$

lo cual contradice la hipótesis **(A4)**, bajo la cual se establece que $\lim_{n\to\infty} n\bar{l} = \infty$ o, equivalentemente, que $\alpha < 1$. De esta forma se concluye que, bajo las condiciones expuestas en el enunciado, $V_1(\eta,\bar{l},n) = o(S_2(\eta,\bar{l})^2)$, lo cual concluye la prueba de que (B.8) es la única ventana que minimiza el error cuadrático medio (asintótico) de $\hat{\psi}^g_{r,\eta}$ como estimador de ψ_r.

Finalmente, probaremos que el estimador $\hat{\psi}^g_{r,\eta^{\text{óptt}}}$ es consistente en media cuadrática, esto es, que

$$\lim_{n\to\infty} \text{MSE}\left(\hat{\psi}^g_{r,\eta^{\text{óptt}}}\right) = \lim_{n\to\infty}\left[\text{Sesgo}\left(\hat{\psi}^g_{r,\eta^{\text{óptt}}}\right)^2 + \text{Var}\left(\hat{\psi}^g_{r,\eta^{\text{óptt}}}\right)\right] = 0,$$

mostrando, además, su orden de convergencia. Para ello, comenzaremos sustituyendo $\eta^{\text{óptt}}$ en las expresiones del sesgo y la varianza de $\hat{\psi}_{r,\eta}$, recogidas en (B.3) y (B.4), respectivamente. Por un lado, el sesgo de $\hat{\psi}^g_{r,\eta^{\text{óptt}}}$ se reduce únicamente al término de segundo orden $S_3\left(\eta^{\text{óptt}},\bar{l},n\right)$. En efecto,

$$\begin{aligned}
\text{ASesgo}\left(\hat{\psi}^g_{r,\eta^{\text{óptt}}}\right) &= S_1(\eta^{\text{óptt}}) + S_2\left(\eta^{\text{óptt}},\bar{l}\right) = \frac{(\eta^{\text{óptt}})^2}{2}\mu_2(L)\psi_{r+2} + \frac{\bar{l}}{(\eta^{\text{óptt}})^{r+1}}L^{(r)}(0)R(f) = \\
&= \frac{\left[-\frac{2\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}}\right]^{\frac{2}{r+3}}}{2}\mu_2(L)\psi_{r+2} + \frac{\bar{l}}{\left[-\frac{2\bar{l}\,L^{(r)}(0)R(f)}{\mu_2(L)\psi_{r+2}}\right]^{\frac{r+1}{r+3}}}L^{(r)}(0)R(f) = \\
&= \frac{(-1)^{\frac{2}{r+3}}\left[\bar{l}\,L^{(r)}(0)R(f)\right]^{\frac{2}{r+3}}\left[\mu_2(L)\psi_{r+2}\right]^{\frac{r+1}{r+3}}}{2^{\frac{r+1}{r+3}}} + \frac{\left[\mu_2(L)\psi_{r+2}\right]^{\frac{r+1}{r+3}}\left[\bar{l}L^{(r)}(0)R(f)\right]^{\frac{2}{r+3}}}{(-1)^{\frac{r+1}{r+3}}\,2^{\frac{r+1}{r+3}}} = \\
&= \frac{\left[\bar{l}\,L^{(r)}(0)R(f)\right]^{\frac{2}{r+3}}\left[\mu_2(L)\psi_{r+2}\right]^{\frac{r+1}{r+3}}}{2^{\frac{r+1}{r+3}}} - \frac{\left[\bar{l}L^{(r)}(0)R(f)\right]^{\frac{2}{r+3}}\left[\mu_2(L)\psi_{r+2}\right]^{\frac{r+1}{r+3}}}{2^{\frac{r+1}{r+3}}} = 0.
\end{aligned}$$

De esta forma, y teniendo en cuenta que $\eta^{\text{óptt}} = O((\bar{l})^{\frac{1}{r+3}})$,

$$\begin{aligned}
\text{Sesgo}\left(\hat{\psi}^g_{r,\eta^{\text{óptt}}}\right)^2 &= S_3\left(\eta^{\text{óptt}},\bar{l},n\right)^2 = O\left((\eta^{\text{óptt}})^8\right) + O\left(\frac{1}{n^2(\eta^{\text{óptt}})^{2r+2}}\right) + O\left(\frac{\bar{l}^4}{(\eta^{\text{óptt}})^{2r+2}}\right) = \\
&= O\left((\bar{l})^{\frac{8}{r+3}}\right) + O\left(n^{-2}(\bar{l})^{-\frac{2r+2}{r+3}}\right) + O\left((\bar{l})^{\frac{2r+10}{r+3}}\right) = O\left((\bar{l})^{\frac{8}{r+3}}\right) + O\left(n^{-2}(\bar{l})^{-\frac{2r+2}{r+3}}\right),
\end{aligned}$$

puesto que $O\left((\bar{l})^{\frac{2r+10}{r+3}}\right) = o\left((\bar{l})^{\frac{8}{r+3}}\right)$ para todo $r \in \mathbb{N}$.

Por otro lado,

$$\begin{aligned}\mathrm{Var}\left(\hat{\psi}^g_{r,\eta^{\text{ópt}}}\right) &= O(n^{-1}) + O\left(\frac{\bar{l}}{n(\eta^{\text{ópt}})^{2r+1}}\right) + o(n^{-1} + \bar{l}n^{-1}(\eta^{\text{ópt}})^{-2r-1}) = \\ &= O(n^{-1}) + O\left(n^{-1}(\bar{l})^{\frac{2-r}{r+3}}\right) + o(n^{-1} + n^{-1}(\bar{l})^{\frac{2-r}{r+3}}) = O(n^{-1}) + O\left(n^{-1}(\bar{l})^{\frac{2-r}{r+3}}\right).\end{aligned}$$

De esta forma,

$$\begin{aligned}\mathrm{MSE}(\hat{\psi}^g_{r,\eta^{\text{ópt}}}) &= \mathrm{Sesgo}^2\left(\hat{\psi}^g_{r,\eta^{\text{ópt}}}\right)^2 + \mathrm{Var}\left(\hat{\psi}^g_{r,\eta^{\text{ópt}}}\right) = \\ &= O\left((\bar{l})^{\frac{8}{r+3}}\right) + O\left(n^{-2}(\bar{l})^{-\frac{2r+2}{r+3}}\right) + O(n^{-1}) + O\left(n^{-1}(\bar{l})^{\frac{2-r}{r+3}}\right).\end{aligned}$$

Teniendo en cuenta que $\bar{l} = Cn^{-\alpha} = O(n^{-\alpha})$, resulta

$$\mathrm{MSE}(\hat{\psi}^g_{r,\eta^{\text{ópt}}}) = O\left(n^{-\frac{8\alpha}{r+3}}\right) + O\left(n^{\frac{2r(\alpha-1)+2\alpha-6}{r+3}}\right) + O(n^{-1}) + O\left(n^{\frac{r(\alpha-1)-2\alpha-3}{r+3}}\right). \tag{B.9}$$

Ahora bien, se ha visto al inicio de la demostración que la condición **(A4)** impone que $\alpha \in (0,1)$. Es sencillo comprobar que, para estos valores de α, todos los exponentes involucrados en (B.9) son estrictamente negativos para cualquier $r \in \mathbb{N}$. En efecto, para todo $r > 0$,

$$-\frac{8\alpha}{r+3} < 0 \Longleftrightarrow -8\alpha < 0 \Longleftrightarrow \alpha > 0.$$

Por otro lado,

$$\frac{2r(\alpha-1)+2\alpha-6}{r+3} < 0 \Longleftrightarrow 2r\alpha - 2r + 2\alpha - 6 < 0 \Longleftrightarrow (2r+2)\alpha < 2r+6 \Longleftrightarrow \alpha < \frac{2r+6}{2r+2}.$$

Como $g_1(r) = \frac{2r+6}{2r+2}$ es una función monótona decreciente tal que $\lim_{r\to\infty} g_1(r) = 1$, entonces para que el exponente en cuestión sea estrictamente negativo para todo $r \in \mathbb{N}$ basta pedir que $\alpha < 1$. Finalmente,

$$\frac{r(\alpha-1)-2\alpha-3}{r+3} < 0 \Longleftrightarrow r\alpha - r - 2\alpha - 3 < 0 \Longleftrightarrow (r-2)\alpha < r+3 \Longleftrightarrow \alpha < \frac{r+3}{r-2}.$$

Siguiendo un razonamiento análogo, se tiene que $g_2(r) = \frac{r+3}{r-2}$ es una función monótona decreciente tal que $\lim_{r\to\infty} g_2(r) = 1$, por lo que la desigualdad anterior será nuevamente satisfecha por cualquier $\alpha < 1$. De esta forma, se concluye que

$$\lim_{n\to\infty} \mathrm{MSE}\left(\hat{\psi}^g_{r,\eta^{\text{ópt}}}\right) = 0$$

como queríamos probar.

□

Apéndice C

Sobre el sesgo del estimador *leave-one-group-out*

Partiendo de las ideas clásicas de Rudemo (1982) y Bowman (1984), en la Sección 3.3 se ha propuesto un selector de validación cruzada insesgada para el estimador de Cao et al. (2011), $\hat{f}^g_{w,h}$, el cual permite obtener una ventana para su construcción en aquellos contextos en los que únicamente los pesos muestrales $w_1, \ldots, w_k$ asociados a las observaciones agrupadas $t_1, \ldots, t_k$ son conocidos. Este selector parte de la expresión del error cuadrático integrado de $\hat{f}^g_{w,h}$,

$$\mathrm{ISE}_g(t_1, \ldots, t_k; h) = \int \hat{f}^g_{w,h}(x)^2 \, \mathrm{d}x - 2 \int \hat{f}^g_{w,h}(x) f(x) \, \mathrm{d}x + \int f(x)^2 \, \mathrm{d}x,$$

para, a continuación, estimar su único sumando desconocido que depende de h, $\int \hat{f}^g_{w,h}(x) f(x) \, \mathrm{d}x$, a través del estimador

$$\sum_{j=1}^{k} w_j \hat{f}^{g,-j}_{w,h}(t_j), \tag{C.1}$$

donde $\hat{f}^{g,-j}_{w,h}$ denota al estimador *leave-one-group-out* definido en (3.23), y construir así la función de validación cruzada

$$\mathrm{UCV}_g(h) = R\left(\hat{f}^g_{w,h}\right) - 2 \sum_{j=1}^{k} w_j \hat{f}^{g,-j}_{w,h}(t_j),$$

cuya minimización conduce a la ventana $\hat{h}_{\mathrm{UCV}_g}$.

En el contexto clásico de datos no agrupados, el estimador análogo a (C.1) — definido a través del estimador tipo núcleo de Parzen-Rosenblatt, $\hat{f}_{n,h}$, y clásicamente basado en estrategias de tipo *leave-one-out*, ver Rudemo (1982) y Bowman (1984) — resulta ser un estimador insesgado de la cantidad $\mathbb{E}[\int \hat{f}_{n,h}(x) f(x) \, \mathrm{d}x]$, verificándose, para una m.a.s $X_1, \ldots, X_n$, la igualdad

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \hat{f}^{-i}_{n,h}(X_i)\right] = \mathbb{E}\left[\int \hat{f}_{n,h}(x) f(x) \, \mathrm{d}x\right],$$

donde $f^{-i}_{n,h}$ denota al estimador *leave-one-out*. Esta misma cuestión, ahora planteada para el caso de datos agrupados, es la que se discute en este apéndice, cuestionando la validez de la igualdad

$$\mathbb{E}\left[\sum_{j=1}^{k} w_j \hat{f}^{g,-j}_{w,h}(t_j)\right] = \mathbb{E}\left[\int \hat{f}^g_{w,h}(x) f(x) \, \mathrm{d}x\right].$$

En este sentido, comentar que el problema se ha abordado desde un enfoque práctico, realizando un breve estudio de simulación consistente en generar $B = 1000$ muestras agrupadas — siguiendo el esquema de agrupación recogido en la Sección 4.1.2 — procedentes de los Modelos 1, 2, 6, 9 y 10 de Marron y Wand (1992) y aproximar el cociente

$$\frac{\mathbb{E}[\sum_{j=1}^{k} w_j \hat{f}_{w,h}^{g,-j}(t_j)]}{\mathbb{E}[\int \hat{f}_{w,h}^{g}(x) f(x)\ \mathrm{d}x]}, \tag{C.2}$$

donde los correspondientes estimadores $\hat{f}_{w,h}^{g}$ han sido construidos en base a la ventana plug-in propuesta en la Sección 3.2.2. Los resultados obtenidos, recogidos en la Tabla C.1, sugieren que, conforme el tamaño muestral aumenta, el cociente entre ambas cantidades se aproxima lentamente a la unidad, lo cual puede ser un indicativo de la insesgadez asintótica de (C.1) (y, por tanto, de la validez de la adaptación del selector de validación cruzada insesgada al contexto de datos agrupados). De todas formas, los resultados obtenidos en este estudio parecen mostrar que dicha convergencia se produce de manera *lenta*.

	Modelo 1	**Modelo 2**	**Modelo 6**	**Modelo 9**	**Modelo 10**
$n = 50$	4.0280	11.2811	4.4015	4.6796	5.7956
$n = 500$	1.1605	1.2718	1.2415	1.2525	1.2441
$n = 5\,000$	1.0372	1.0572	1.0508	1.0526	1.0586
$n = 10\,000$	1.0242	1.0370	1.0332	1.0342	1.0377

Tabla C.1: Aproximaciones de las cantidades $\mathbb{E}[\sum_{j=1}^{k} w_j \hat{f}_{w,h}^{g,-j}(t_j)]/\mathbb{E}[\int \hat{f}_{w,h}^{g}(x) f(x)\ \mathrm{d}x]$ obtenidas a partir de 1000 muestras de tamaño $n \in \{50, 500, 5\,000, 10\,000\}$ generadas en base a los modelos 1, 2, 6, 9 y 10 de Marron y Wand (1992) y agrupadas según el esquema recogido en la Sección 4.1.2. En todos los casos se ha empleado el selector plug-in propuesto en la Sección 3.2.2 para la construcción de los estimadores tipo núcleo involucrados en dicho cociente.

A la vista de estos resultados, se podría pensar en emplear un estimador de $\int \hat{f}_{w,h}^{g}(x) f(x)\ \mathrm{d}x$ alternativo, que proporcione una tasa de convergencia del cociente en cuestión a la unidad más rápida que la proporcionada por (C.1). En este sentido, se ha probado a emplear como estimador — nuevamente para los modelos de Marron y Wand (1992) anteriormente considerados — la media ponderada de los valores $\hat{f}_{w,h}^{g}(t_1), \ldots, \hat{f}_{w,h}^{g}(t_k)$,

$$\sum_{j=1}^{k} w_j \hat{f}_{w,h}^{g}(t_j).$$

En este caso, si bien es cierto que el cociente $\mathbb{E}[\sum_{j=1}^{k} w_j \hat{f}_{w,h}^{g}(t_j)]/\mathbb{E}[\int \hat{f}_{w,h}^{g}(x) f(x)\ \mathrm{d}x]$ parece converger a la unidad de forma más rápida de lo que lo hace (C.2) (puesto que ya con $n = 500$ toma el valor aproximado de 0.994), los resultados obtenidos — a través de un estudio de simulación similar al presentado en el Capítulo 4 — con el estimador $\hat{f}_{w,h}^{g}$ construido en base a esta modificación del selector de validación cruzada insesgada han resultado ser mucho peores — en términos del MISE_g — que los obtenidos con el estimador (C.1).

Bibliografía

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Ameijeiras-Alonso, J. (2017). *Assessing Simplifying Hypotheses in Density Estimation*. Tesis, Universidade de Santiago de Compostela.

Ameijeiras-Alonso, J., Crujeiras, R. M., y Rodríguez-Casal, A. (2021). Multimode: An R package for mode assessment. *Journal of Statistical Software*, 97(9):1–32.

Barreiro-Ures, D., Fraguela, B. B., Doallo, R., Cao, R., Francisco-Fernández, M., y Reyes, M. (2019a). *binnednp: Nonparametric Estimation for Interval-Grouped Data*. Package version 0.4.0. URL `https://CRAN.R-project.org/package=binnednp`.

Barreiro-Ures, D., Francisco-Fernández, M., Cao, R., Fraguela, B. B., Doallo, R., González-Andújar, J. L., y Reyes, M. (2019b). Analysis of interval-grouped data in weed science: The binnednp Rcpp package. *Ecology and evolution*, 9(19):10903–10915.

Blower, G. y Kelsall, J. E. (2002). Nonlinear kernel density estimation for binned data: convergence in entropy. *Bernoulli*, 8(4):423–449.

Boneva, L. I., Kendall, D., y Stefanov, I. (1971). Spline transformations: Three new diagnostic aids for the statistical data-analyst. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(1):1–71.

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

Brent, R. P. (2013). *Algorithms for minimization without derivatives*. Prentice Hall, Englewood Cliffs, NJ.

Cao, R. (1990). *Aplicaciones y nuevos resultados del Método Bootstrap en la estimación no paramétrica de curvas*. Tesis, Universidade de Santiago de Compostela.

Cao, R. (1993). Bootstrapping the mean integrated squared error. *Journal of Multivariate Analysis*, 45(1):137–160.

Cao, R., Francisco-Fernández, M., Anand, A., Bastida, F., y González-Andújar, J. L. (2011). Computing statistical indices for hydrothermal times using weed emergence data. *Journal of Agricultural Science*, 149(6):701–712.

Ćwik, J. y Koronacki, J. (1997). A combined adaptive-mixtures/plug-in estimator of multivariate probability densities. *Computational Statistics & Data Analysis*, 26(2):199–218.

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, 14(1):153–158.

Fix, E. y Hodges, J. L. (1951). *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, Randolph Field, Texas.

Fraley, C., Raftery, A. E., y Scrucca, L. (2016). *mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*. Package version 5.4.8. URL `https://CRAN.R-project.org/package=mclust`.

GeyserTimes (2017). Eruptions of Old Faithful Geyser, May 2014 [Online database]. `https://geysertimes.org`. [Online; consultado el 16 de diciembre, 2021].

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM Journal on Applied Mathematics*, 42(2):390–399.

Hu, F. y Zidek, J. V. (2002). The weighted likelihood. *Canadian Journal of Statistics*, 30(3):347–371.

Izenman, A. J. y Sommer, C. J. (1988). Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, 83(404):941–953.

Jang, W. y Loh, J. M. (2010). Density estimation for grouped data with application to line transect sampling. *The Annals of Applied Statistics*, 4(2):893–915.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and computing*, 3(3):135–146.

Marron, J. S. y Wand, M. P. (1992). Exact mean integrated squared error. *The 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.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, 185:71–110.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Reyes, M. (2015). *Statistical Methods for Studying Emergence Curves in Weed Science*. Tesis, Universidade da Coruña.

Reyes, M., Francisco-Fernández, M., y Cao, R. (2016). Nonparametric kernel density estimation for general grouped data. *Journal of Nonparametric Statistics*, 28(2):235–249.

Reyes, M., Francisco-Fernández, M., y Cao, R. (2017). Bandwidth selection in kernel density estimation for interval-grouped data. *TEST*, 26(3):527–545.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3):832–837.

Rossi, R. J. (2018). *Mathematical Statistics: An Introduction to Likelihood Based Inference*. John Wiley & Sons, USA.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(65):65–78.

Schmeiser, B. W. y Deutsch, S. J. (1977). Quantile estimation from grouped data: The cell midpoint. *Communications in Statistics-Simulation and Computation*, 6(3):221–234.

Scott, D. W. y Sheather, S. J. (1985). Kernel density estimation with binned data. *Communications in Statistics-Theory and Methods*, 14(6):1353–1359.

Scott, D. W. y Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146.

Scrucca, L., Fop, M., Murphy, T. B., y Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.

Sheather, S. J. y Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B*, 53(3):683–690.

Shirazi, Z. A., da Silva, J. P. A. R., y de Souza, C. P. E. (2021). Parameter estimation for grouped data using EM and MCEM algorithms. *arXiv preprint arXiv:2106.02909*.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B*, 43(1):97–99.

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

Sun, X. (2014). Asymmetric kernel density estimation based on grouped data with applications to loss model. *Communications in Statistics-Simulation and Computation*, 43(3):657–672.

Titterington, D. (1983). Kernel-based density estimation using censored, truncated or grouped data. *Communications in Statistics-Theory and Methods*, 12(18):2151–2167.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer, Paris.

Wand, M. P. y Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

Wang, B. y Wertelecki, W. (2013). Density estimation for data with rounding errors. *Computational Statistics and Data Analysis*, 65:4–12.

Wikimedia Commons (2009). Mexico 12c stamp of 1872 with image of Miguel Hidalgo, scanned August 2009 by user: Ecphora. `https://commons.wikimedia.org/wiki/File:Mexico_1872.jpg`. [Online; consultado el 16 de diciembre, 2021].

Wilson, I. G. (1983). Add a new dimension to your philately. *The American Philatelist*, 97:342–349.

Yu, Y. (2021). *mixR: Finite Mixture Modeling for Raw and Binned Data*. Package version 0.2.0. URL `https://CRAN.R-project.org/package=mixR`.