FACULTADE DE MATEMÁTICAS

Trabajo de fin de grado

Métodos núcleo para la estimación de la densidad en la recta real y en el círculo

Sara Dovalo del Río

2020–2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO EN MATEMÁTICAS

Trabajo de fin de grado

Métodos núcleo para la estimación de la densidad en la recta real y en el círculo

Sara Dovalo del Río

2020–2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa

Título: Métodos núcleo para la estimación de la densidad en la recta real y en el círculo

Breve descripción del contenido:

El estimador tipo núcleo de la densidad, propuesto por Parzen Rosenblatt, permite aproximar de manera flexible la densidad de una variable aleatoria con soporte en la recta real sin realizar sobre ella misma suposiciones paramétricas. Además, este estimador se puede extender a otros contextos no euclídeos, como por ejemplo, al círculo. En este TFG se presentará el estimador tipo núcleo de la densidad, tanto en el caso lineal (variables con soporte real) como en el caso circular.

En los dos contextos, será fundamental seleccionar de manera adecuada el parámetro de suavizado, por lo que se revisarán algunos métodos comunes a los dos contextos y se estudiará su comportamiento empírico mediante estudios de simulación en R. Se valorará también de modo crítico las implicaciones que tiene considerar una estimación de la densidad lineal sobre datos observados en el círculo. Las técnicas analizadas se ilustrarán con algunos ejemplos de datos reales.

Agradecimientos

Empezaré expresando mi más sincera gratitud a mi tutora Rosa M. Crujeiras Casais, por acompañarme y guiarme en este trabajo, el cual no hubiese sido capaz de acometer con facilidad sin su buen hacer en la docencia. Gracias por estar ahí siempre que lo he necesitado e implicarte tanto, aún teniendo en cuenta la situación por la que hemos pasado durante estos últimos meses.

También me gustaría agradecer a todos los profesores que me han ido formando durante estos años, y, en especial, a D. Jose Carlos de Miguel Domínguez por inculcarme el amor por las matemáticas en mi primer y último año estudiando Administración y Dirección de Empresas.

Dar las gracias también a mis compañeros de promoción y en especial a mis amigos, que me han servido de apoyo durante esta dura etapa. También a mis amigas de la infancia que me han ayudado a *desconectar* cuando era necesario. En este sentido, tengo que dar las gracias a mi novio, por su compañía y por ser un pilar fundamental durante todos estos años. Siempre ha sabido cómo aguantar todas mis quejas, en especial, en esta parte final de la carrera.

Finalmente, me gustaría dar las gracias a mi familia por todo el apoyo brindado durante toda mi vida, y, en particular a mis padres y hermano, por animarme a seguir en los momentos más duros, apoyarme día a día y por su cariño incondicional.

Por último, ya que en este trabajo se trata el tema de la circularidad, me gustaría cerrar el círculo de agradecimientos con otra Rosa, mi madre, a la que le debo tanto. Por su ayuda incondicional en éste y todos los proyectos en los que me embarco.

Resumen

La estimación de la función de densidad es uno de los principales problemas de la estadística, en particular, de la estadística no paramétrica.

En este contexto, el estimador tipo núcleo de la densidad, introducido por Parzen y Rosenblatt, es un estimador continuo que permite aproximar de manera satisfactoria la densidad de una variable aleatoria con soporte real. En este trabajo, generalizaremos el estimador tipo núcleo lineal (variables con soporte real) al contexto circular (variables con soporte en el círculo unidad).

La continuidad de este estimador fue un gran avance desde la introducción del histograma en cuanto a estimadores no paramétricos de la densidad, sin embargo, su comportamiento es muy sensible a la selección del parámetro ventana o de suavizado: revisaremos los métodos de selección de dicho parámetro más utilizados, tanto en el caso lineal como en el circular, lo cual será determinante para obtener una buena estimación de la función de densidad.

A continuación, realizaremos un estudio de simulación para comparar el funcionamiento de dichos selectores en ambos contextos, cuyo objetivo será buscar aquel que minimice el error cometido con respecto a la densidad teórica de modo local (para un punto fijo) y global.

Para finalizar, se ilustra la aplicación práctica de las técnicas presentadas mediante dos conjuntos de datos reales, uno para cada contexto (lineal y circular), realizando una comparativa con los resultados obtenidos en el estudio de simulación previo.

Abstract

The estimation of the density function is one of the main problems in statistics, in particular in non-parametric statistics.

In this context, the kernel estimator of density, introduced by Parzen and Rosenblatt, is a continuous estimator that allows a satisfactory approximation of the density of a random variable with real support. In this work, we will generalize the linear kernel estimator (variables with real support) to the circular context (variables with support in the unit circle).

The continuity of this estimator was a great advance since the introduction of the histogram in terms of non-parametric density estimators, however, its behavior is very sensitive to the selection of the window or smoothing parameter: we will review the most used selection methods for this parameter, both in the linear and circular cases, which will be decisive in obtaining a good estimate of the density function.

Next, we will carry out a simulation study to compare the operation of these selectors in both contexts, the objective of which will be to find the one that minimizes the error made with respect of the theoretical density locally (for a fixed point) and globally.

Finally, the practical application of the techniques presented is illustrated by means of two sets of real data, one for each context (linear and circular), making a comparison with the results obtained in the previous simulation study.

Índice general

Introducción

La estadística es una ciencia transversal que se encarga del estudio de variables aleatorias y sus propiedades. Dichas variables aleatorias se caracterizan matemáticamente a través de la función de densidad.

Un problema fundamental de la estadística es la estimación de la función de densidad de una variable a partir de una muestra de datos. Dicho problema se puede abordar de dos formas: la primera, consiste en considerar que la función de densidad a estimar pertenece a una determinada familia paramétrica (Normal, Exponencial, etc.) y, por lo tanto, el problema se reduce a determinar los parámetros del modelo a partir de la muestra, lo que se denomina, estimación paramétrica de la densidad. Por otro lado, la segunda forma consiste en dejar que la función de densidad pueda adoptar cualquier forma, imponiendo siempre las propiedades que exigen las funciones de densidad para poder ser consideradas como tales, es decir, no negatividad e integración igual a 1 en su soporte. Esta segunda forma recibe el nombre de estimación no paramétrica de la densidad y en la cúal nos centraremos en este trabajo.

El estimador de densidad no paramétrico más antiguo y más utilizado es el histograma, que no es más que la representación de frecuencias por clases. El histograma es un estimador discontinuo, que depende de la elección de un punto inicial y de un parámetro ventana o parámetro de suavizado. Para solventar este problema, hay que esperar a mediados del s.XX, cuando se desarrolló el denominado estimador naive o histograma móvil, el cuál sigue siendo discontinuo y dependiendo del parámetro ventana. Posteriormente, a finales de los años 50 se introduce el estimador tipo núcleo, que sí es continuo, y que por lo tanto, se ajusta mejor en la mayor parte de las ocasiones a los modelos estudiados, aunque sigue dependiendo de la elección de un parámetro ventana ya que todas las estimaciones de curvas no paramétricas dependen de un parámetro de este tipo que controla la suavidad de la estimación.

En este contexto no paramétrico ha sido muy estudiado el importante papel que juega el parámetro ventana en el estimador tipo núcleo. Este parámetro regula el grado de suavizado del estimador: una mala elección del mismo puede derivar en un estimador

infrasuavizado, o por el contrario, sobresuavizado.

A modo de ejemplo, utilizaremos la librería `nor1mix` del software R para representar cuatro de los quince ejemplos de densidad paramétrica utilizada en el estudio de simulación de Marron y Wand (1992). Los modelos de Marron y Wand son mixturas de normales que, de manera general, se pueden escribir como:

$$f(x) = \sum_{m=1}^{M} p_m f_m(x; \mu_m, \sigma_m^2),$$

donde $p = (p_1, \ldots, p_M), m = 1, ..., M$ ($p_m > 0$ y $\sum_{m=1}^{M} p_m = 1$) son los parámetros de mezcla, $f_m(x)$ es la densidad de una $N(\mu_m, \sigma_m^2)$ para $m = 1, ..., M$, que se denominan densidades componentes de la muestra las cuales dependen del vector de medias $\mu = (\mu_1, \ldots, \mu_M) \in \mathbb{R}^M$ y del vector de varianzas $\sigma^2 = (\sigma_1^2, \ldots, \sigma_M^2) \in (\mathbb{R}^+)^M$.

Nosotros nos centraremos en el estudio de los modelos 1, 6, 8 y 9 los cuales hacen referencia a una normal estándar, bimodal, bimodal asimétrica y trimodal respectivamente.

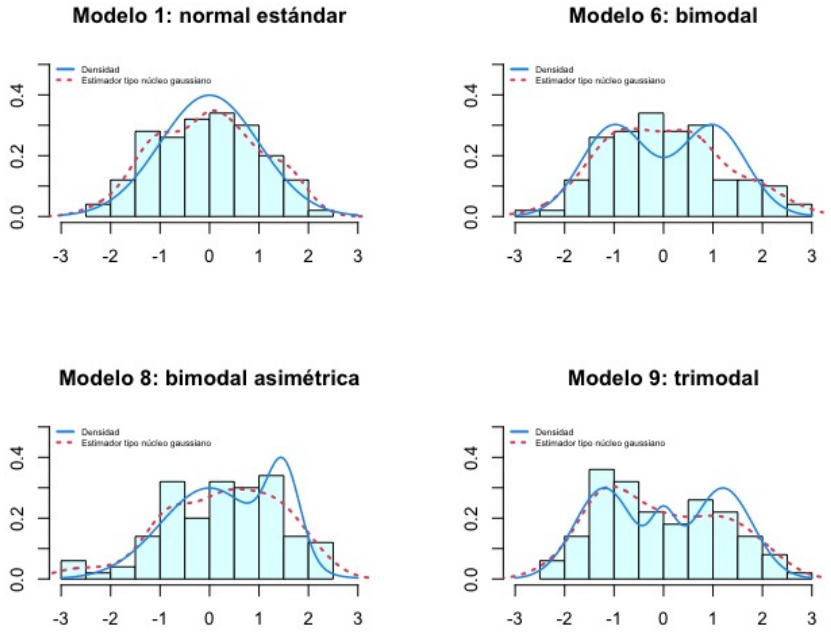


Figura 1: Representación del histograma, curva de densidad parmétrica (línea continua) y estimador tipo núcleo (línea discontinua) con parámetro de ventana escogido mediante la regla del pulgar para los modelos 1, 6, 8 y 9 de Marron y Wand con muestras de datos simuladas con tamaño $n = 100$.

En lo relativo a los histogramas de la Figura 1, se podría decir que, aunque sean esti-

madores discontinuos, muestran la estructura básica de los datos a pesar de que sean muy poco robustos ante modificaciones del parámetro ventana que, en este caso, se corresponde con el ancho de los intervalos de clase. Finalmente, señalar que la estimación mejor ajustada es claramente la que proporciona el estimador tipo núcleo.

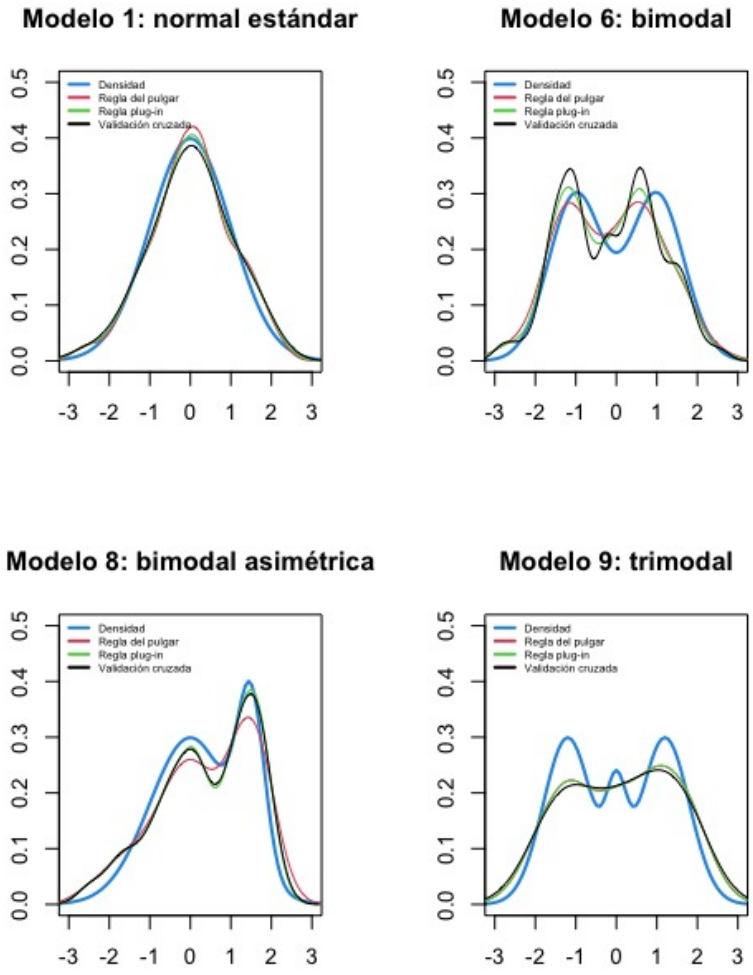


Figura 2: Representación del estimador tipo núcleo con los tres selectores de ventana más utilizados: la regla del pulgar, la regla plug-in y el método de validación cruzada para los modelos $1, 6, 8$ y 9 de Marron y Wand (1992).

A finales del s.XX surgieron diferentes métodos de selección de este parámetro ventana entre las que destacaron: la regla del pulgar, la regla plug-in y el método de validación cruzada, etc (Wand, M. P. y Jones, M. C. (1995), Capítulo 3). Además, en función de la

medida de distancia, el selector puede ser de dos tipos: global o puntual (según se centre en la estimación de toda la curva o en un punto particular). Un aspecto importante sobre los distintos procedimientos de selección del parámetro ventana es que la estimación de la densidad podrían dar resultados muy distintos en función del método escogido como podemos ver en la Figura 2 utilizando de nuevo los modelos antes escogidos de Marron y Wand (1992). Además, también se puede ver en la Figura 2 que el selector introducido por Silverman en 1986, la regla del pulgar, tiende a sobresuavizar ya que sí llega a captar la bimodalidad, sin embargo, no ocurre lo mismo con la trimodalidad.

Por otro lado, en diversas áreas como la biología, geología, paleontología, geografía, meteorología, astronomía, física y medicina surgen problemas estadísticos donde los datos son recogidos mediante medidas angulares dando la orientación o bien ángulos en el plano (datos circulares) o en el espacio (datos esféricos). Los datos circulares constituyen el caso más simple de este tipo de datos llamados datos direccionales los cuales se miden en ángulos o direcciones. El análisis estadístico de los datos circulares difiere de los métodos lineales estándar: las técnicas de inferencia clásicas para datos lineales pueden no proporcionar resultados satisfactorios para datos circulares ya que no tienen en cuenta la naturelaza periódica de este tipo de datos. De aquí surge la necesidad de extender la noción de estimador tipo núcleo a otros contextos como en el de los datos circulares. A modo de ejemplo, utilizaremos el software R para representar tres de los veinte modelos de Oliveira et al. (2012) empleando la librería `NPCirc`. Estos tres modelos son los siguientes:

- Modelo M7. Mixtura de dos von Mises con el mismo peso: $\frac{1}{2}\text{vM}(0,4) + \frac{1}{2}\text{vM}(\pi,4)$

- Modelo M9. Mixtura de dos von Mises con distinto peso: $\frac{1}{4}\text{vM}(0,2) + \frac{3}{4}\text{vM}(\frac{\pi}{\sqrt{3}},2)$.

- Modelo M13. Mixtura de tres von Mises: $\frac{2}{5}\text{vM}(0{,}5,6) + \frac{2}{5}\text{vM}(3,6) + \frac{1}{5}\text{vM}(5,24)$.

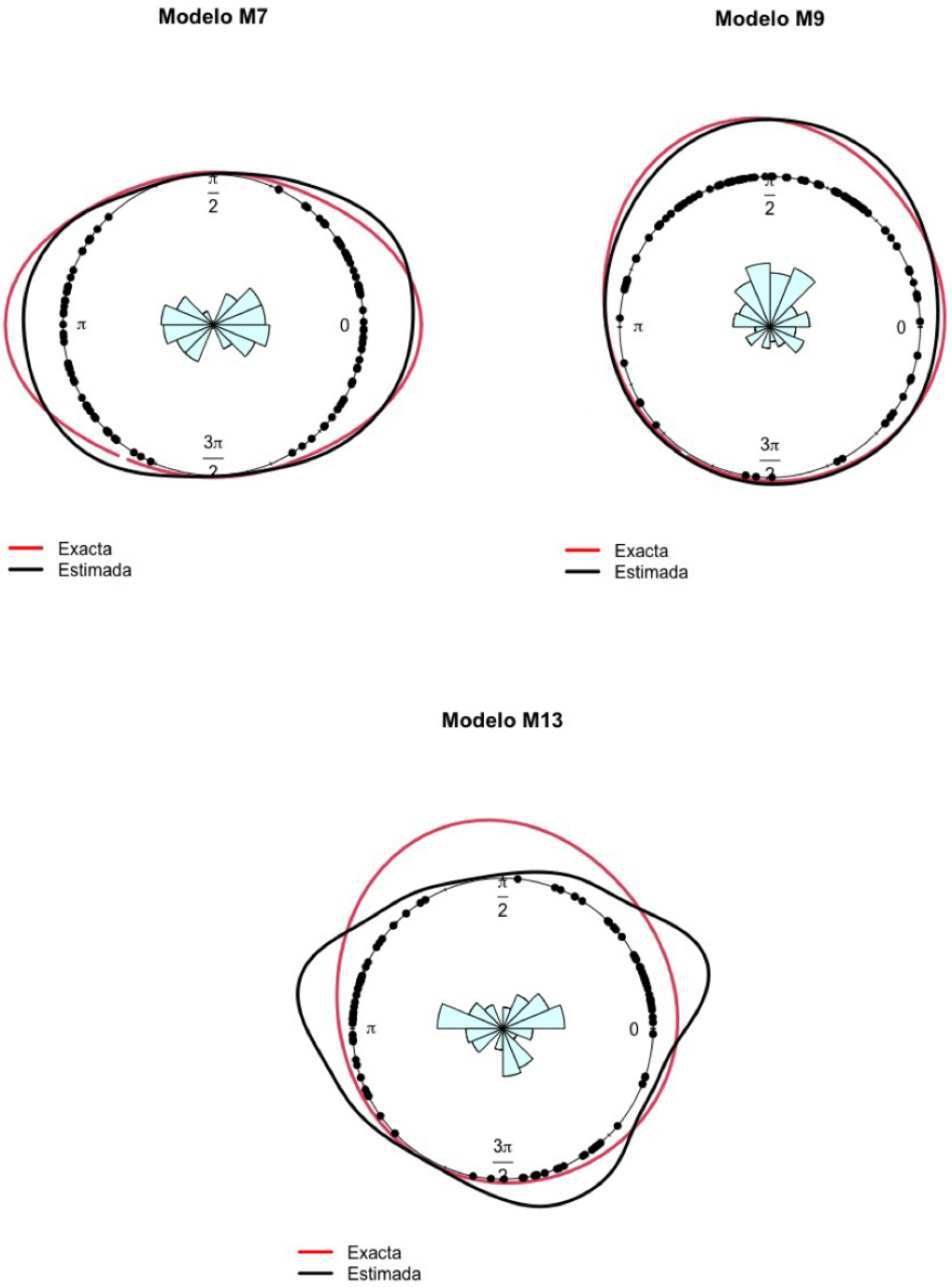


Figura 3: Representación del diagrama de rosas, de la densidad para las distintas distribuciones circulares de los modelos M7, M9 y M13 (línea roja) y estimador tipo núcleo (línea negra) para datos circulares para un ancho de banda dado utilizando la distribución de von Mises como núcleo circular con muestras de datos simuladas de tamaño $n = 100$.

El objetivo de este trabajo es comparar el funcionamiento del estimador tipo núcleo considerando los distintos selectores del parámetro de suavizado, tanto en el caso lineal como en el circular, en la estimación no paramétrica de la densidad. Por otro lado, trataremos de ver lo importante que es estimar datos con soporte en el círculo unidad a través de técnicas no paramétricas para datos circulares, ya que, en caso contrario, no se tienen en cuenta la naturaleza periódica de los mismos.

Este trabajo de fin de grado se organiza de la siguiente manera:

En el Capítulo 1 desarrollaremos las técnicas para estimar de manera no paramétrica la función de densidad para datos lineales a través del estimador tipo núcleo, prestando especial atención a sus propiedades y medidas de error que indiquen cuánto de buena es dicha estimación.

En el Capítulo 2, generalizaremos el estimador tipo núcleo lineal al caso de datos circulares realizando una analogía respecto al capítulo anterior para datos periódicos que se encuentren en el círculo unidad.

En el Capítulo 3, abordaremos la selección del parámetro de suavizado, del cual depende dicho estimador tipo núcleo, tanto para el caso lineal como para el circular, y, estudiaremos el comportamiento de los distintos selectores en el estudio de simulación recogido en el Capítulo 4, donde veremos lo que ocurre si realizamos una estimación no paramétrica de la densidad olvidando que nuestros datos son circulares.

Por último, ilustraremos los resultados obtenidos en dicho estudio a dos conjuntos de datos reales: uno que ejemplifique el comportamiento de los distintos selectores para datos lineales y otro para datos circulares.

Capítulo 1

Revisión del estimador tipo núcleo de la densidad

La estimación de densidad no paramétrica es una herramienta analítica de datos importante, especialmente cuando los modelos paramétricos estándar no son adecuados. Además, la estimación no paramétrica de la densidad proporciona una manera muy eficaz de mostrar la estructura en un conjunto de datos al comienzo de su análisis y, para ello, es necesario un estimador que no asuma que la densidad tiene una forma funcional particular.

Este primer capítulo hace referencia a la definición formal del estimador tipo núcleo de la densidad que debido a su simplicidad nos permite estudiar sus propiedades con bastante detenimiento (detallando su sesgo y varianza) en la Sección 1.1. A continuación en la siguiente Sección 1.3 consideraremos distintas medidas de error (local y global) para el estimador tipo núcleo.

A lo largo de este capítulo se supondrá que tenemos una muestra aleatoria simple (m.a.s) $X_1, \ldots, X_n$ tomada de una densidad f continua y univariante.

1.1. El estimador tipo núcleo

El estimador tipo núcleo de la densidad (KDE, kernel density estimator), fue introducido por Parzen y Rosenblatt a finales de los años 50. Para entender la construcción del estimador tipo núcleo, es necesario conocer la expresión del estimador naive o histograma móvil, el cúal es discontinuo, depende de un parámetro ventana y se construyó bajo la idea de definir la función de densidad mediante el siguiente límite:

$$f(x) = \lim_{h \to 0} \frac{\mathbb{P}(x - h < X < x + h)}{2h}$$

siendo X una v.a con función de densidad f. Dada una m.a.s de X, la probabilidad en el numerador se puede aproximar como la proporción muestral de observaciones de la muestra que caen en el intervalo $(x-h, x+h)$. Así, el estimador naive o histograma móvil viene definido por:

$$\hat{f}_{n,N}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{I}(X_i \in (x-h, x+h)),$$

donde $\mathbb{I}$ es la función indicadora. Si consideramos una función peso ω tal que $\omega(x) = 1/2\mathbb{I}(|x| < 1)$ (que no es más que la densidad de una v.a $U(-1,1)$) podemos reescribir la definición anterior como:

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

donde $h > 0$ es el parámetro de suavizado o ventana y $\omega_h = \frac{1}{h}\omega\left(\frac{\cdot}{h}\right)$ es una función peso reescalada. Cabe observar en la expresión anterior que el estimador se construye colocando una *caja* de base $2h$ y altura $(2nh)^{-1}$ alrededor de cada observación X_i y sumando las alturas, i.e, $(2nh)^{-1}\mathbb{I}(X_i \in (x-h, x+h))$. A pesar de que el estimador naive solo depende del parámetro de suavizado h, sigue siendo una función discontinua, problema que resolverá nuestro estimador tipo núcleo.

Si ahora, en lugar de considerar la densidad uniforme introducimos otra densidad denominada función núcleo, esto es, una función real, no negativa, que por ser una función de densidad es integrable con $\int K(u)du = 1$, unimodal y simétrica con respecto al origen, que denotaremos por K, tendremos la expresión del estimador tipo núcleo, la cuál viene dada por:

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

con núcleo reescalado: $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. Veremos más adelante que la elección de la forma de la función núcleo no es particularmente importante. Sin embargo, la elección del valor del parámetro de suavizado o ventana h es muy importante.

Existen diferentes tipos de funciones núcleo (como caso particular, tomando el núcleo uniforme $1/2\mathbb{I}(|x| < 1)$ tendríamos el estimador naive) pero en la práctica no existen diferencias significativas a la hora de emplear uno u otro tipo. En este trabajo nos quedaremos con una función núcleo de tipo gaussiano:

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

Es sencillo comprobar que si K es la densidad de una v.a W, el núcleo reescalado K_h es la densidad de la variable hW. Por ejemplo, si consideramos un núcleo gaussiano estándar

$N(0,1)$ con una ventana h, el núcleo reescalado se corresponde con la densidad de una $N(0,h^2)$, donde h^2 denota la varianza.

Podríamos concluir que la construcción del estimador tipo núcleo es el promedio en cada punto que se obtiene al colocar centrando en cada valor de la muestra la densidad de una $N(0,h^2)$. Es decir, el KDE es una mixtura de normales donde las medias vienen dadas por los puntos de la muestra y la varianza la determina el parámetro de suavizado o parámetro ventana h seleccionado.

1.2. Propiedades

Para un x fijo, el valor esperado del KDE descrito en la ecuación (1.1) viene dado por:

$$\mathbb{E}(\hat{f}_{n,K}(x)) = \int K(u)f(x-hu)du = (K_h * f)(x)$$

donde $*$ denota la convolución de K_h y f. Esto quiere decir que la curva esperada que se obtiene con un KDE no es la verdadera densidad de f si no la versión suavizada de la misma, dada por $(K_h * f)$. Para obtener una expresión detallada del valor esperado, bajo condiciones de regularidad sobre f, se puede aproximar:

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

Suponiendo que $\mu_2(K) = \int u^2 K(u)du < \infty$ (lo que quiere decir que la varianza de la v.a con densidad K sea finita), entonces:

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

donde se utiliza que K es una función de densidad simétrica:

$$\int K(u)du = 1, \int uK(u)du = 0.$$

En el caso de la varianza, el desarrollo es un poco más tedioso:

$$\mathrm{Var}(\hat{f}_{n,K}(x)) = \frac{1}{n}Var(K_h(x-X_1)) = \frac{1}{n}[\mathbb{E}(K_h^2(x-X_1)) - \mathbb{E}^2(K_h(x-X_1))]$$

$$= \frac{1}{n}\left[\int K_h(x-y)f(y)dy - \left(\int K_h(x-y)f(y)dy\right)^2\right]$$

$$= \frac{1}{nh}\int K^2(u)f(x-hu)du - \frac{1}{n}\left(\int K(u)f(x-hu)du\right)^2$$

$$= \frac{1}{nh}\int K^2(u)(f(x)-o(1))du - \frac{1}{n}(f(x)+o(1))^2 = \frac{1}{nh}R(K)f(x) + o((nh)^{-1})$$

Resumiendo: cuando $h \to 0$:

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

y si $nh \to \infty$

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

Por lo tanto, el estimador tipo núcleo es asintótico, insesgado y consistente.

Una clara interpretación de lo anterior es que para valores pequeños de h se reduce el sesgo pero aumenta la varianza. Sin embargo, para valores grandes de h ocurre lo contrario. Este razonamiento motiva a querer encontrar un equilibrio entre sesgo y varianza, donde juega un papel fundamental la selección del parámetro suavizado óptimo, h_{opt}.

1.3. Medidas del error y ventanas óptimas

Generalmente, para evaluar el comportamiento de un estimador de la densidad, debemos definir mecanismos que midan la bondad de ajuste de los mismos. En particular, podríamos considerar medidas de error locales (para un x fijo) o medidas de error globales (para toda la *curva* estimada). Estas medidas de error nos van a proporcionar funciones objetivo a minimizar para obtener valores óptimos del parámetro de suavizado o ventana h.

Concretamente, si consideramos nuestro estimador tipo núcleo descrito en la Ecuación (1.1), $\hat{f}_{n,K}(x)$, como estimador puntual para $f(x)$ (con un x fijo) podemos considerar una medida de error local como el error cuadrático medio (*Mean Square Error*, MSE) definido de la forma habitual, es decir:

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

Para poder dar una expresión completa de este error local, necesitamos utilizar las expresiones explícitas de sesgo y varianza desarrolladas en la Sección 1.2, de este modo:

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

Minimizar (1.2) no es sencillo ya que aparecen los términos residuales de las aproximaciones asintóticas que hemos utilizado. Para solucionar este problema, será necesario considerar una versión asintótica del MSE para encontrar un valor de h que nos proporcione el error mínimo, la cual se denomina AMSE (*Asymptotic Mean Square Error*)

$$\mathrm{MSE}(\hat{f}_{n,K}(x)) \equiv \mathrm{AMSE}(\hat{f}_{n,K}(x)) + o((nh)^{-1} + h^4) \tag{1.3}$$

donde R es una aplicación que asigna a cualquier función de L^2 la integral de su cuadrado, esto es, si $g \in L^2$ entonces $R(g) = \int g^2(x)dx$; μ_2 es otra aplicación definida como $\mu_2(h) = \int x^2 h(x)dx$, siempre que este valor sea finito. Es sencillo ver que el valor óptimo de h que minimiza la expresión anterior viene dado por:

$$h_{\text{AMSE}}(x) = \left(\frac{R(K)f(x)}{n\mu_2^2(K)(f''(x))^2}\right)^{1/5} \tag{1.4}$$

Obtenemos así una ventana local ya que depende del punto x. Sin embargo, no es directamente aplicable en la práctica, ya que depende de la segunda derivada de la densidad (siendo esta desconocida).

Por consiguiente, pasamos a considrar una medida de error global denominada MISE (*Mean Integrated Square Error*) que podremos obtener de dos maneras: (i) promediando otra medida de error global denominada Error Cuadrático Integrado (*Integrated Square Error*, ISE) que se define como:

$$\text{ISE}(X_1, \ldots, X_n; h) = \int \left(\hat{f}_{n,K}(x) - f(x)\right)^2 dx \tag{1.5}$$

entonces,

$$\text{MISE}(h) = \mathbb{E}\left(\text{ISE}(X_1, \ldots, X_n; h)\right) = \mathbb{E}\left(\int \left(\hat{f}_{n,K}(x) - f(x)\right)^2 dx\right)$$

o bien (ii) integrando el error cuadrático medio descrito en (1.2):

$$\text{MISE}(h) = \int \text{MSE}(\hat{f}_{n,K}(x))dx = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') + o((nh)^{-1} + h^4) \tag{1.6}$$

Al igual que en el caso local, es necesario considerar una versión asintótica de la medida de error global dada en la anterior ecuación (1.6) para poder así encontrar el parámetro ventana h óptimo. Esta versión asintótica recibe el nombre de AMISE (*Asymptotic Mean Integrated Square Error*), la cual se define mediante la expresión:

$$\text{AMISE}(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') \tag{1.7}$$

Minimizando (1.7) con respecto a h, se obtiene:

$$h_{\text{AMISE}} = \left(\frac{R(K)}{\mu_2(K)^2R(f'')n}\right)^{1/5}. \tag{1.8}$$

La anterior expresión se trata de una ventana global donde es sencillo observar que su valor depende del tamaño muestral n: cuanto mayor sea el tamaño de la muestra, más pequeño será el parámetro ventana h y viceversa. Además, a pesar de que $R(K)$ y $\mu_2(K)$ son cantidades conocidas, este parámetro ventana depende de la derivada segunda de la

función de densidad, $R(f'')$, que es la función que deseamos estimar. Por tanto, aunque tengamos una ventana óptima en el sentido teórico, no podremos utilizarla tampoco en la práctica.

El objetivo ahora será encontrar algo aplicable a la práctica que nos resuelva este problema. Lo abordaremos aproximando el parámetro ventana h o bien una medida de error del mismo mediante los selectores del parámetro de suavizado o ventana del estimador tipo núcleo que trataremos en el Capítulo 3.

Capítulo 2

Revisión del estimador tipo núcleo de la densidad en el círculo

2.1. Introducción

La estimación tipo núcleo en el caso lineal (1.1) se extiende fácilmente a datos circulares, aunque se debe tener especial cuidado en la selección de la función tipo núcleo.

En este capítulo presentaremos siguiendo Mardia, K. V. y Jupp, P. E. (2000), la distribución de von Mises, uno de los modelos paramétricos más utilizados en datos circulares y que servirá de base para construir el estimador no paramétrico de la densidad que utilizaremos. Además, definiremos con rigor este estimador, que, al igual que en el caso lineal, la selección del parámetro de suavizado será crucial.

Hablaremos también de algunos criterios de error (locales y globales) que nos ayudarán en la elección del parámetro de suavizado en el caso circular siguiendo principalmente Oliveira, M. (2013), si bien hay contribuciones que vienen de referencias anteriores y que no han sido consultadas para este TFG.

2.2. La distribución de von Mises

Como cada punto en la circunferencia unidad representa una dirección, una distribución de probabilidad circular es una forma de asignar probabilidades a diferentes direcciones concentradas en la circunferencia de un círculo unidad. El caso más simple de una distribución circular es la distribución circular uniforme donde a todos los arcos de igual longitud se le asigna la misma probabilidad.

Una de las distribuciones de probabilidad circular más utilizada es la distribución de von Mises, $vM(\mu, \kappa)$ o distribución Normal Circular (porque los estimadores de máxima

verosimilitud de los parámetros de localización y concentración coinciden con los momentos muestrales). Además, esta distribución de von Mises es simétrica, unimodal y se caracteriza por dos parámetros: la dirección media de la distribución, $\mu \in [0, 2\pi)$, y el parámetro de concentración, $\kappa \geq 0$, los cuales se pueden estimar por el método de máxima verosimilitud.

La expresión de su función de densidad viene dada por:

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\}, \quad 0 \leq \theta < 2\pi \tag{2.1}$$

donde $I_0(\kappa)$ denota la función de Bessel modificada de orden 0, que se define como:

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\kappa \cos\theta) d\theta.$$

El parámetro μ es el punto de simetría de la densidad y donde está localizada la moda. Por otro lado, el parámetro de concentración κ mide la variación de la distribución en relación con una distribución circular uniforme. Además, la antimoda de una distribución de von Mises se encuentra en $\theta = \mu + \pi$.

Como el cociente entre el valor de la densidad en la moda y el valor de la densidad en la antimoda viene dado por $\exp 2\kappa$, cuanto mayor sea el valor de κ, mayor es el agrupamiento cerca de la moda, es decir, la función de densidad estará cada vez más concentrada, como podemos ver en la Figura (2.1).

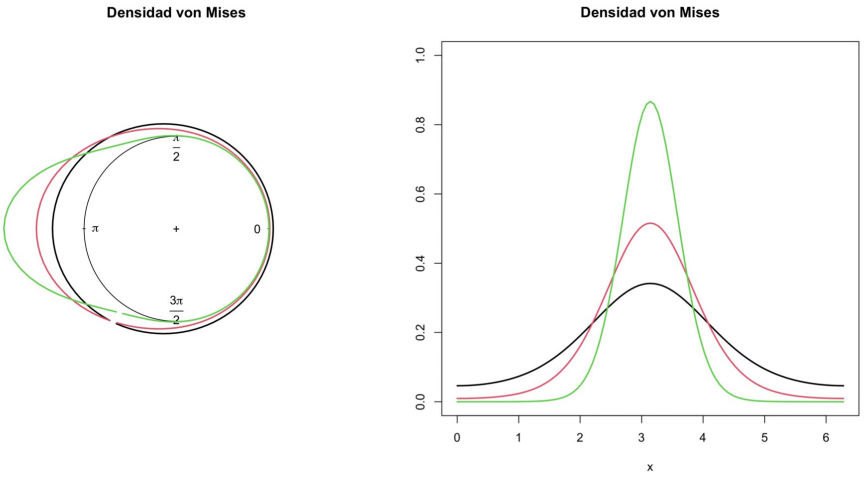


Figura 2.1: Representación circular (izquierda) y representación lineal (derecha) de una densidad de von Mises con $\mu = \pi$ y $\kappa = 1$ (línea negra), 2 (línea roja) y 5 (línea verde). El valor de $\kappa = 0$ se corresponde con la distribución circular uniforme.

La distribución de von Mises juega un papel fundamental en la estimación no paramétrica de la densidad en el campo de los datos circulares. Por ello, es interesante relacionar

y aproximar la distribución de von Mises con otras distribuciones: cuando $\kappa = 0$, como caso particular, se trata de la distribución circular uniforme. Por otro lado, en el caso de que $\kappa \to \infty$ la distribución se concentra en la dirección media: $\theta = \mu$.

Mixtura de von Mises

Del mismo modo que en el contexto de datos lineales, las mixturas finitas de distribuciones von Mises $vM(\mu, \kappa)$, con parámetros de mezcla $p = (p_1, \ldots, p_M), m = 1, ..., M$ ($p_m > 0$ y $\sum_{m=1}^{M} p_m = 1$) proporcionan una clase de modelos circulares. Su función de densidad viene dada por:

$$f(\theta) = \sum_{m=1}^{M} p_m f_m(\theta; \mu_m, \kappa_m) = \sum_{m=1}^{M} p_m \frac{\exp\{\kappa_m \cos(\theta - \mu_m)\})}{2\pi I_0(\kappa_m)}, \quad 0 \leq \theta < 2\pi \tag{2.2}$$

donde $f_m(\theta)$ son densidades de una distribución de von Mises $vM(\mu_m, \kappa_m)$ para $m = 1, \ldots, M$, que se denominan densidades componentes de la mezcla las cuales dependen del vector de medias circulares $\mu = (\mu_1, \ldots, \mu_M) \in [0, 2\pi)^M$ y del vector de concentraciones $\kappa = (\kappa_1, \ldots, \kappa_M) \in (\mathbb{R}^+)^M$.

Por ejemplo, la mixtura de dos von Mises vM(μ_1, κ_1) y vM(μ_2, κ_2) con proporciones de mixtura $p_1 = p$ y $p_2 = 1 - p$ respectivamente, la vamos a denotar por vM(μ_1, μ_2, κ_1, κ_2, p).

Este tipo de modelos circulares nos hace formularnos la siguiente pregunta: ¿Por qué no estimamos la densidad circular a través de estas mixturas de von Mises?

La pregunta no es para nada desacertada ya que sería posible hacerlo, sin embargo, no sabríamos con certeza el número de componentes de la mixtura que dé lugar a la mejor estimación de la densidad circular. Aquí, entra en juego la importancia del número de componentes de la mixtura de von Mises en la selección del parámetro de suavizado, ν, como veremos en los capítulos posteriores.

La selección del número de componentes de la mixtura, M, puede abordarse considerando el Criterio de Información de Akaike (AIC). Se trata de una medida global que tiene en cuenta el ajuste por medio de la función de verosimilitud y a la vez compensa el exceso de parámetros, p, de manera que elige el mejor modelo entre un conjunto de modelos admisibles. Se define como:

$$\text{AIC} = -2\log(L) + 2p, \tag{2.3}$$

donde L es el valor maximizado de la función de verosimilitud para el modelo estimado. Se trata de buscar un modelo cuyo AIC sea pequeño, pues en este caso habría una verosimilitud grande y pocos parámetros. La cuestión es que ambos objetivos, verosimilitud y

número de parámetros, suelen estar contrapuestos. Por tanto, dado un conjunto de modelos, el mejor modelo será, según este criterio, el que tiene el AIC más bajo para que dicho modelo incorpore parámetros realmente útiles para incrementar la verosimilitud. La ilustración del funcionamiento de este método en la práctica lo veremos en el Capítulo 4, concretamente en la Figura 4.3.

2.3. El estimador tipo núcleo de la densidad circular

Dada una muestra aleatoria simple $X_1, \ldots, X_n$, la estimación de la densidad tipo núcleo en el caso lineal dada en (1.1) se extiende fácilmente a datos circulares.

La distancia entre dos puntos en el círculo viene dada por:

$$d_i = ||x - X_i||^2 = 2(1 - x^T X_i) = 2(1 - \cos(\theta - \Theta_i)),$$

donde $x_T = (\cos\theta, \sin\theta)$ y $X_i = (\cos\Theta_i, \sin\Theta_i)^T$.

Considerando la ecuación (1.1) y partiendo de una muestra aleatoria simple de ángulos $\Theta_1, \ldots, \Theta_n \in [0, 2\pi)$ podemos representar la estimación no paramétrica tipo núcleo de la densidad mediante:

$$\hat{f}_{n,K}(\theta) = \sum_{i=1}^{n} \frac{1}{n} K_{i\nu}(\theta) = \frac{1}{2n\pi I_0(\nu)} \sum_{i=1}^{n} \exp\{\nu \cos(\theta - \Theta_i)\} \tag{2.4}$$

Se debe tener especial cuidado en la selección de la función núcleo, ya que en este caso, $K_{i\nu}$ es la von Mises con media Θ_i y parámetro de concentración ν, $vM(\Theta_i, \nu)$. Podemos apreciar, que para el caso circular, el parámetro de concentración ν juega un papel análogo a la ventana h en el caso lineal pero en sentido contrario, es decir, el comportamiento de ν en el caso circular es inverso al de h en el caso lineal: necesariamente $\nu \to \infty$ para que $K_{i\nu}$ esté cada vez más concentrada alrededor del dato θ_i.

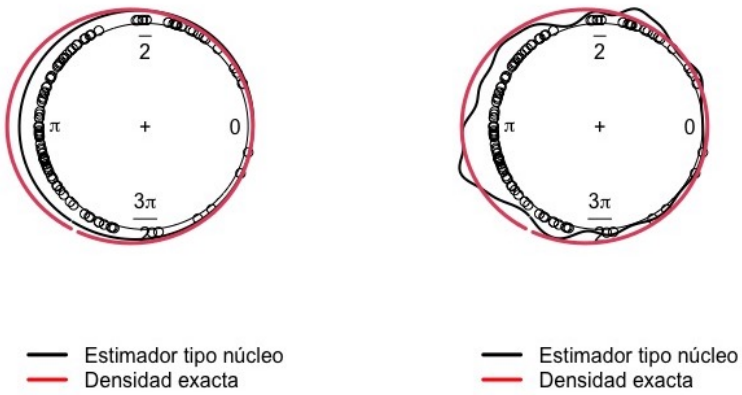


Figura 2.2: Estimador tipo núcleo de la densidad circular (línea negra) con $\nu = 2$ (a la izquierda) y $\nu = 100$ (a la derecha) para 10 muestras aleatorias de tamaño $n = 100$ de una $vM(\pi, 1)$ (línea roja).

Es decir, como podemos ver en la Figura 2.2, valores grandes de ν conducen a estimadores altamente variables (poco suavizados), esto es, estimadores con mucha varianza y poco sesgo. Por el contrario, para valores pequeños de ν proporcionan un exceso de suavizado, es decir, estimadores con gran sesgo y varianza pequeña.

Por esta razón, el estudio de los procedimientos de selección de parámetros de suavizado constituye uno de los problemas más relevantes en la estimación de densidad no paramétrica.

2.4. Medidas del error y ventanas óptimas

Un aspecto crítico a la hora de aplicar este estimador en la práctica es la elección del parámetro de suavizado ν. También en el caso circular, un desafío principal en la estimación no paramétrica de la densidad es el equilibrio entre sesgo y varianza.

En cuanto a los datos lineales, las técnicas más conmúnmente utilizadas para seleccionar el parámetro de suavizado o ventana h se basan en la minimización de algunos criterios de error (locales y globales) que cuantifican la precisión del estimador tipo núcleo de la densidad descrito en (1.1), es decir, qué tan bien el estimador se aproxima a la densidad real.

Al igual que para el caso lineal, podemos considerar distintos criterios de error para el caso circular para medir la calidad como estimador de (2.4). Para un punto θ fijo, $\hat{f}_{n,K}(\theta)$ es una variable aleatoria. Para medir cuánto de bueno es el estimador tipo núcleo de la densidad en el caso circular se puede utilizar una medida de error local denominada Error

Cuadrático Medio (*Mean Squared Error*):

$$\mathrm{MSE}(\theta) = \mathbb{E}\left(\hat{f}_{n,K}(\theta) - f(\theta)\right)^2$$

pero el interés de la estimación no paramétrica radica en obtener una estimación y representación de la densidad completa, por lo tanto es necesario considerar criterios de error globales para abordar dicho problema, como por ejemplo, el Error Cuadrático Integrado (*Integrated Squared Error*):

$$\mathrm{ISE}(\hat{f}) = \int \left(\hat{f}_{n,K}(\theta) - f(\theta)\right)^2 d\theta \tag{2.5}$$

El ISE es una variable aleatoria que depende de la verdadera (y desconocida) densidad del parámetro de concentración ν. Por lo tanto, resulta más realista plantearse como criterio de error un promedio del ISE denominado Error Cuadrático Medio Integrado (*Mean Integrated Squared Error*):

$$\mathrm{MISE}(\hat{f}) = \mathbb{E}(\mathrm{ISE}(\hat{f})) = \mathbb{E}\left[\int \left(\hat{f}_{n,K}(\theta) - f(\theta)\right)^2 d\theta\right] = \int \mathrm{MSE}(\theta) d\theta \tag{2.6}$$

donde en la tercera igualdad basta intercambiar la esperanza con la integral para ver que el MISE no es más que un promedio de los errores cuadráticos medios en cada punto.

Sin embargo, en la práctica, a menudo se utiliza su expresión asintótica AMISE (*Asymptotic Mean Integrated Squared Error*), que, en el caso del estimador tipo núcleo circular (2.4), si f'' es continua e integrable, su expresión cuando $\nu \to \infty$ y $\sqrt{\nu} n^{-1} \to 0$, viene dada por (ver en Oliveira, M. (2013)):

$$\mathrm{AMISE}(\nu) = \left\{ \frac{1}{16}\left[1 - \frac{I_2(\nu)}{I_0(\nu)}\right]^2 \int_0^{2\pi} \left[f''(\theta)\right]^2 d\theta + \frac{I_0(2\nu)}{2n\pi(I_0(\nu))^2} \right\} \tag{2.7}$$

donde f'' denota la derivada de segundo orden de la función de densidad a estimar, que mide la curvatura de f. Suponiendo que los datos siguen una distribución de von Mises con parámetro de concentración κ, podemos escribir la expresión del AMISE como:

$$\mathrm{AMISE}(\nu) = \frac{3\kappa^2 I_2(2\kappa)}{32\pi\nu^2 I_0(\kappa)^2} + \frac{\nu^{1/2}}{2n\pi^{1/2}}. \tag{2.8}$$

Mimizando (2.7) con respecto a ν e igualando a cero obtenemos la denominada regla del pulgar, un método de selección del parámetro de suavizado ν que veremos en el Capítulo siguiente.

Capítulo 3

Selectores del parámetro de suavizado en la recta real y en el círculo

Para estimar de una manera óptima la función de densidad a partir de una muestra de datos, más concretamente, la implementación práctica del estimador tipo núcleo tanto en el caso lineal como en el circular, requiere la especificación del parámetro de suavizado.

Sin embargo, seleccionar este parámetro a simple vista puede llevarnos mucho tiempo, sobre todo si se requieren muchas estimaciones de densidad para un problema dado, o bien, no se tiene conocimiento previo de la estructura de los datos y por tanto resulta imposible considerar un parámetro de suavizado que nos proporcione una estimación cercana a la densidad real.

En este Capítulo presentaremos las ideas principales de los selectores del parámetro de ventana h en el caso lineal (ver más en Wand, M. P. y Jones, M. C. (1995), Capítulo 3) y selectores del parámetro de concentración ν (siguiendo Oliveira, M. (2013)) en el caso de datos circulares. Como vimos, el comportamiento de éste es inverso al de h en el caso lineal: valores altos de ν proporcionan menos suavidad, mientras que valores pequeños de ν dan lugar a una mayor suavidad.

3.1. Selectores de h

Las ventanas óptimas tanto locales como globales descritas en la Sección 1.3 no se pueden utilizar en la práctica ya que dependen de la verdadera función de densidad (que es desconocida) y su derivada segunda.

Sin embargo, resultarán un punto de partida para construir selectores del parámetro ventana h en el caso lineal. Seguiremos fundamentalmente dos estrategias en cuanto a la construcción de los mismos:

- aproximar las ventanas óptimas resultantes de minimizar criterios de error tanto locales como globales, o bien,
- aproximar diferentes medidas de error.

De los primeros podemos decir que su principal objetivo es encontrar un parámetro ventana h que sea *razonable* para una amplia gama de situaciones, pero sin ninguna garantía matemática de estar cerca del valor óptimo de h. Además, suelen ser utilizados en algoritmos que requieren muchos pasos de estimación de curvas y proporcionan un punto de partida para la elección del parámetro de suavizado. De este tipo, en esta sección, veremos la regla del pulgar y la regla plug-in.

Con respecto a los segundos, se basan en argumentos matemáticos más complicados y costosos computacionalmente pero que dan una muy buena respuesta para clases muy generales de funciones subyacentes. Este tipo de selectores suele basarse en minimizar el MISE dado en (1.6), lo cual lograrán pero de manera asintótica. Un selector de ventana h de este tipo será el de validación cruzada.

A lo largo de esta Sección consideraremos una m.a.s. $X_1, \ldots, X_n$, asumiendo que $\hat{f}_{n,K}(\cdot, h)$ es el estimador tipo núcleo de la densidad f.

3.1.1. Regla del pulgar

La regla del pulgar fue introducida por Silverman (finales de los años 80) se basa en utilizar el parámetro de ventana óptimo para el AMISE dado en (1.8).

El objetivo de Silverman era sustutuír el término que depende de la densidad desconocida, $R(f'') = \int (f''(x))^2$ por la densidad de una normal de media 0 y desviación típica σ, $N(0, \sigma)$, en cuyo caso se tiene:

$$h_{\text{AMISE}} = \left(\frac{8\sqrt{\pi} R(K)}{3\mu_2(K)^2 n} \right)^{1/5} \sigma. \tag{3.1}$$

La regla del pulgar se obtiene de (3.1) sin mas que sustituír σ por $\hat{\sigma}$:

$$h_{\text{AMISE}} = \left(\frac{8\sqrt{\pi} R(K)}{3\mu_2(K)^2 n} \right)^{1/5} \hat{\sigma} \tag{3.2}$$

donde $\hat{\sigma}$ es un estimador de σ. A la hora de escoger dicho estimador, $\hat{\sigma}$, las opciones más comunes son $\hat{\sigma} = S$ siendo S la cuasi-desviación típica, o bien, utilizar el rango

intercuartílico estandarizado:

$$\hat{\sigma}_{IQR} = \frac{IQR}{\phi^{-1}(0{,}75)\phi^{-1}(0{,}25)},$$

donde IQR es el rango intercuartílico de la muestra (diferencia entre el tercer y el primer cuartil) y ϕ^{-1} es una función cuantil de una $N(0,1)$. Además, el denominador de $\hat{\sigma}_{IQR}$, es un factor de normalización que se corresponde con el rango intercuartílico de la población de la densidad normal estándar y es aproximadamente igual a 1,349.

Para reducir el riesgo de sobresuavizar, Silverman sugiere utilizar el mínimo entre $\hat{\sigma}_{IQR}$ y S para construír el parámetro ventana h.

Como bien dijimos antes, este selector nos dará una *primera estimación* y se pueden esperar resultados razonables siempre que los datos estén cerca de lo normal. Además, es fácil obtener su adaptación al caso multidimensional. Sin embargo, tienden a sobresuavizar y ocultar características importantes de los datos, por ejemplo, en el caso de densidades asimétricas o con más de una moda.

3.1.2. Regla plug-in

Al igual que la regla del pulgar, la regla plug-in se basa en aproximar el valor óptimo del parámetro ventana h que minimiza el AMISE dado en (1.8).

En este caso, la propuesta de Sheather y Jones (principios de los años 90) se basa en estimar de manera no paramétrica (utilizando estimadores tipo núcleo) la cantidad desconocida $R(f'') = \int (f''(x))^2$.

Denotando por $\psi_r = \mathbb{E}(f^{(r)}(X))$ y aplicando un argumento de integración por partes a $R(f'')$, entonces, se tiene:

$$R(f'') = \int (f''(x))^2 \ \ dx = \int f^{(4)}(x) f(x) \ \ dx = \mathbb{E}(f^{(4)}(X)).$$

Por tanto, el problema se reduce a estimar ψ_4. En general, una aproximación de $\psi_r = \mathbb{E}(f^{(r)}(X)) = \int f^{(r)}(x) f(x) \ \ dx$ viene dada por:

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

donde $\hat{f}_{n,K}^{(r)}$ podría ser un estimador tipo núcleo (para la r-ésima derivada). Por ejemplo, el estimador tipo núcleo con núcleo L y ventana g:

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

Pero, en este caso, necesitamos una nueva ventana g específicamente para $r = 4$:

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

y la ventana g se podría obtener mediante argumentos similares, pero necesitaríamos de $\hat{\psi}_6$. Para $\hat{\psi}_6$ podríamos hacer lo mismo, pero entonces necesitaríamos de $\hat{\psi}_8$.

La estrategia habitual para superar este problema es estimar una función ψ_r con una estimación rápida y sencilla, esto es, como una versión de la regla del pulgar descrita en la Sección 3.1.1.

Esto significa que tenemos una familia de selectores del parámetro ventana h que dependen de un número de etapas de la estimación funcional, ℓ, antes de que se utilice una estimación rápida y simple.

Supongamos que un estimador de h por una regla plug-in implica ℓ estimaciones sucesivas del núcleo con un parámetro ventana inicial elegido mediante un procedimiento rápido y sencillo (como puede ser la regla del pulgar). Llamaremos a esta regla un selector de ancho de banda plug-in directo de etapa ℓ y la denotaremos por $\hat{h}_{DPI,\ell}$. Notar que regla del pulgar (3.2) se puede considerar como un selector de ancho de banda plug-in de etapa 0 ($\ell = 0$).

Surge ahora otro problema de selección: ¿cómo elegir el valor de ℓ, esto es, el número de etapas de la estimación funcional? A medida que ℓ aumenta, vemos que el selector de ancho de banda se vuelve menos sesgado, ya que la dependencia de la regla de estimación rápida y sencilla disminuye. Sin embargo, valores altos de ℓ hace que el selector tenga más varianza, volviendo así al desafío principal en la estimación no paramétrica de la densidad: el equilibrio entre sesgo y varianza. Una opción común es considerar $\ell = 2$.

En este punto, Sheater y Jones sugieren asumir que f es normal, con desviación típica σ:

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

Se procede a estimar ψ_r utilizando $\hat{\sigma}$ y se itera *hacia atrás*.

Como vimos en la sección anterior, $\hat{\sigma}$ es un estimador de σ. A la hora de escoger dicho estimador, $\hat{\sigma}$, las opciones más comunes son $\hat{\sigma} = S$ siendo S la cuasi-desviación típica, o bien, utilizar el rango intercuartílico estandarizado.

Esta ventana proporciona un comportamiento muy satisfactorio en la práctica, como veremos en los capítulos posteriores.

3.1.3. Validación cruzada

Como bien dijimos al inicio de este capítulo, los métodos de validación cruzada afrontan el problema de selección del parámetro ventana desde una perspectiva distinta: en lugar de basarse en las expresiones de las ventanas óptimas, intentarán minimizar medidas de error globales, lo cual se logrará pero de manera asintótica.

El Error Cuadrático Integrado (ISE) de $\hat{f}_{n,K}$ como estimador de f dado en (1.5) puede reescribirse como:

$$\begin{aligned} \text{ISE}(h) &= \int \hat{f}_{n,K}(x)dx - \int \hat{f}_{n,K}(x)f(x)dx + \int f^2(x)dx \\ &= R(\hat{f}_{n,K}) - 2\int \hat{f}_{n,K}(x)f(x)dx + R(f), \end{aligned} \tag{3.3}$$

donde el último sumando no depende de h. Así, la ventana óptima para el ISE es:

$$\begin{aligned} h_{\text{ISE}} &= \arg\min_h \int \hat{f}^2_{n,K}(x)dx - 2\int \hat{f}_{n,K}(x)f(x)dx \\ &= \arg\min_h R(\hat{f}_{n,K}) - 2\int \hat{f}_{n,K}(x)f(x)dx, \end{aligned}$$

donde h aparece dentro del estimador tipo núcleo $\hat{f}_{n,K}$. El proceso ahora trata de aproximar los dos sumandos que conforman la expresión de la ventana. En el caso del primer sumando, este puede escribirse como:

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

de manera que sólo depende de la función núcleo y la ventana h.

Para el segundo sumando, nótese que:

$$\int \hat{f}_{n,K}(x)f(x)dx = \mathbb{E}(\hat{f}_{n,K}).$$

Por tanto, si tuviéramos una muestra $\tilde{X_1}, \ldots, \tilde{X_m}$ de X (independiente de $X_1, \ldots, X_n$) esta cantidad se podría aproximar como:

$$\frac{1}{m}\sum_{k=1}^{m} \hat{f}_{n,K}(\tilde{X_k}),$$

pero no disponemos de tal muestra. Es aquí donde entra la idea de la validación cruzada, ya que únicamente utilizando $X_1, \ldots, X_n$, la cantidad anterior se puede aproximar por:

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

donde $\hat{f}^{-i}_{n,K}$ es el KDE obtenido con toda la muestra excepto el dato i-ésimo.

El método de selección de la ventana por validación cruzada se basa en obtener el h que minimiza la función de validación cruzada, denotada por h_{CV}:

$$\mathrm{CV}(h) = R(\hat{f}_{n,K}) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,K}^{-i}(X_i).$$

Dicha ventana, h_{CV}, se denomina ventana de validación cruzada.

Una de las ventajas de las técnicas de validación cruzada es que son fácilmente exportables a otros contextos, sin embargo, son altamente variables y además tienden a infrasuavizar.

3.2. Selectores de ν

Como mencionamos en el capítulo anterior, al igual que en el caso lineal, una de las técnicas más utilizadas para seleccionar el parámetro de suavizado en el caso circular es la minimización de algunos criterios de error como por ejemplo el MISE, cuya expresión dada en (2.6) se puede reescribir en términos del sesgo y varianza del estimador:

$$\begin{aligned}
\mathrm{MISE}(\nu) &= \mathbb{E}\left[\int_0^{2\pi} \left(\hat{f}_{n,K}(\theta) - f(\theta)\right)^2 d\theta\right] = \int_0^{2\pi} \mathbb{E}\left[\left(\hat{f}_{n,K}(\theta) - f(\theta)\right)^2\right] d\theta \\
&= \int_0^{2\pi} \left[\mathbb{E}\left(\hat{f}_{n,K}(\theta)\right) - f(\theta)\right]^2 d\theta + \int_0^{2\pi} \mathbb{E}\left[\hat{f}_{n,K}(\theta) - \mathbb{E}\left(\hat{f}_{n,K}(\theta)\right)\right]^2 d\theta \\
&= \int_0^{2\pi} \left[\mathrm{Sesgo}\left(\hat{f}_{n,K}(\theta)\right)\right]^2 d\theta + \int_0^{2\pi} \mathrm{Var}\left(\hat{f}_{n,K}(\theta)\right) d\theta.
\end{aligned}$$

A lo largo de este trabajo se hizo especial hincapié en la importancia del equilibrio entre sesgo y varianza dentro de la estimación no paramétrica de la densidad. Sin embargo, en la práctica se suele utilizar su expresión asintótica AMISE (2.7). Por tanto, seleccionar ν minimizando MISE o AMISE se reduce a buscar un equilibrio entre sesgo y varianza simultáneamente.

Considerando como densidad de referencia una von Mises, existen, para datos circulares, diversos selectores del parámetro de suavizado ν como pueden ser la regla del pulgar, regla plug-in y el método de validación cruzada.

3.2.1. Regla del pulgar

La regla del pulgar en el ámbito circular propuesta por Taylor en 2008 consiste en la adaptación de la ideas de Silverman (finales de los años 80) en el caso lineal sin más que tomar como distribución de referencia una von Mises con parámetro de concentración κ. Como vimos en la Sección 2.4, en este caso la expresión del AMISE viene dada por (2.8),

por tanto, el valor del parámetro de suavizado ν que minimiza la expresión anterior puede venir estimado por:

$$\hat{\nu}_{\text{regla del pulgar}} = \left[\frac{3n\hat{\kappa}^2 I_2(2\hat{\kappa})}{4\pi^{1/2} I_0(\hat{\kappa})^2}\right]^{2/5}, \tag{3.4}$$

donde $\hat{\kappa}$ es el estimador del parámetro de concentración de la densidad a estimar obtenido por máxima verosimilitud.

Esta regla de selección funciona satisfactoriamente al ajustar distribuciones simétricas unimodales (ver Figura 3.1). Sin embargo, para distribuciones bimodales y/o sesgadas, la estimación de κ por máxima verosimilitud o por el método de los momentos (ambos métodos coinciden para la estimación de los parámetros de una von Mises), puede ser prácticamente inútil como podemos ver en la Figura 3.2.

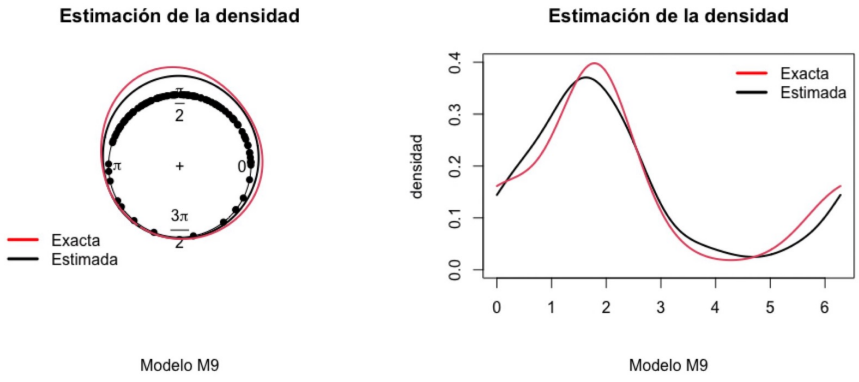


Figura 3.1: Representación circular (izquierda) y representación lineal (derecha) de la estimación tipo núcleo de la densidad de una mixtura de dos von Mises $\frac{1}{4}\text{vM}(0,2) + \frac{3}{4}\text{vM}\left(\frac{\pi}{\sqrt{3}}, 2\right)$ para una muestra de tamaño 100, con parámetro de suavizado ν calculado con la regla del pulgar.

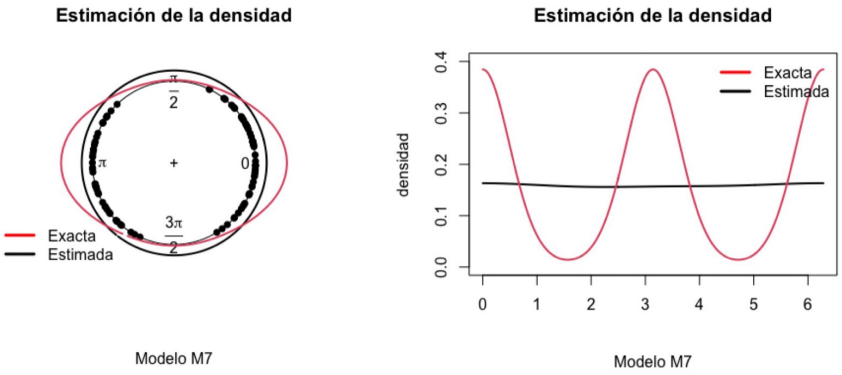


Figura 3.2: Representación circular (izquierda) y representación lineal (derecha) de la estimación tipo núcleo de la densidad de una mixtura de dos von Mises $\frac{1}{2}\text{vM}(0,4) + \frac{1}{2}\text{vM}(\pi,4)$ para una muestra de tamaño 100, con parámetro de suavizado ν calculado con la regla del pulgar.

Este mal funcionamiento a veces se debe a la estimación no robusta por máxima verosimilitud del parámetro de concentración κ, por lo que una posible modificación de (3.4) consiste también en la minimización del AMISE incorporando una familia de distribución más flexible como densidad de referencia en la fórmula de AMISE dada en (2.7).

Esta es la idea de uno de los métodos de selección del parámetro de suavizado en el ámbito de datos circulares que veremos a continuación: la regla plug-in.

3.2.2. Regla plug-in

La regla plug-in como método de selección del parámetro de suavizado ν está basada en las ideas de Oliveira, M. (2013), cuyo objetivo fue buscar una analogía para el caso circular de los trabajos propuestos por Ćwik, J. and Koronacki, J. en 1997 para el caso lineal multivariante. Este selector consiste en considerar una mixtura de von Mises como densidad de referencia. El selector del parámetro de suavizado $\nu_{\text{regla plug-in}}$ se obtiene mediante el siguiente procedimiento:

Paso 1. Seleccionar el número de componentes, M, de la densidad de referencia: una mixtura finita de von Mises (2.2).

Paso 2. Estimar el AMISE dado en (2.7) de la siguiente manera:

Paso 2.1. Estimar los parámetros de la mixtura de von Mises (2.2) (μ_m, κ_m, p_m) para $m = 1, \ldots, M$ por máxima verosimilitud.

Paso 2.2. Calcular numéricamente mediante métodos de cuadratura como la regla de Simpson la integral $R(f'') = \int (\hat{f}'')^2 d\theta$, donde f'' es la segunda derivada de la densidad de una mixtura M de von Mises con los parámetros estimados en el paso anterior.

Paso 2.3. Sustituír la cantidad anterior en AMISE (2.7) para obtener $\widehat{\text{AMISE}}(\nu)$.

Paso 3. Minimizar $\widehat{\text{AMISE}}(\nu)$ mediante un método de optimización y finalmente obtener $\nu_{\text{regla plug-in}}$.

El paso 1 del algoritmo requiere seleccionar el número de componentes de la mixtura para la distribución de referencia (como dijimos, ésta sera siempre una von Mises) lo cual podrá llevarse a cabo utilizando el criterio AIC descrito en (2.3). Es sencillo ver que si $M = 1$, estamos en el caso del selector explicado anteriormente en la Sección 3.2.1: la regla de pulgar introducida por Taylor en 2008.

3.2.3. Validación cruzada

Esta propuesta de selector del parámetro de suavizado introducido por Hall et al. en 1987 utilizando ideas de validación cruzada se basa en minimizar el ISE (2.5) y se denomina validación cruzada de mínimos cuadrados (LSCV). Reescribiendo (2.5, se obtiene:

$$\begin{aligned} \text{ISE}(\hat{f}) &= \int_0^{2\pi} \left(\hat{f}_{n,K}(\theta) - f(\theta)\right)^2 d\theta \\ &= \int_0^{2\pi} \hat{f}_{n,K}^2(\theta) d\theta - 2\int_0^{2\pi} \hat{f}_{n,K} f(\theta) d\theta + \int_0^{2\pi} f^2(\theta) d\theta, \end{aligned} \tag{3.5}$$

donde el último sumando no depende de ν. Así, la minimización de (3.5) involucra sólamente a los dos primeros sumandos:

$$\begin{aligned} \nu_{\text{ISE}} &= \arg\min_{\nu} \int_0^{2\pi} \hat{f}_{n,K}^2(\theta) d\theta - 2\int_0^{2\pi} \hat{f}_{n,K} f(\theta) d\theta \\ &= \arg\min_{\nu} R(\hat{f}_{n,K}) - 2\int_0^{2\pi} \hat{f}_{n,K} f(\theta) d\theta, \end{aligned}$$

donde ν aparece dentro del estimador tipo núcleo $\hat{f}_{n,K}$.

Por tanto, si tuviéramos una muestra $\widetilde{\Theta}_1, \ldots, \widetilde{\Theta}_m$ (independiente de $\Theta_1, \ldots, \Theta_n$) el segundo sumando podría aproximarse

$$\frac{1}{m} \sum_{k=1}^{m} \hat{f}_{n,K}(\widetilde{\Theta}_k),$$

pero como no disponemos de tal muestra utilizamos el método de validación cruzada, que no es más que considerar únicamente nuestra muestra $\Theta_1, \ldots, \Theta_n$ en dicha aproximación, esto es:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{n,K}^{-i}(\Theta_i),$$

donde $\hat{f}_{n,K}^{-i}$ es el estimador tipo núcleo circular obtenido dejando fuera la observación i-ésima. Por tanto, ν_{LSCV} se obtiene como el valor de ν que minimiza:

$$\text{LSCV}(\nu) = \int_0^{2\pi} \hat{f}_{n,K}^2(\theta)d\theta - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,K}^{-i}(\Theta_i).$$

Por otro lado, el parámetro de suavizado de validación cruzada de máxima verosimilitud (LCV) se obtiene maximizando:

$$\text{LCV}(\nu) = \prod_{i=1}^{n} \hat{f}_{n,K}^{-i}(\Theta_i).$$

A pesar de que el método de validación cruzada por mínimos cuadrados sea más estable en el caso circular que en el lineal, ya que en la distribución de referencia no existen colas, en el estudio de simulación realizado en el capítulo siguiente se considerará el método de validación cruzada de máxima verosimilitud pues, en base a diversos estudios de simulación (ver Oliveira, M. (2013), Sección 2.2.2), se concluye que, en el caso circular, parece asintóticamente el más estable.

A continuación, en el siguiente capítulo se realizará un estudio de simulación cuyo objetivo es comparar el funcionamiento de los distintos selectores del parámetro de suavizado para el estimador tipo núcleo, tanto en el caso lineal como en el circular, considerados a lo largo de este capítulo.

También se ilustrarán las técnicas empleadas a través de datos reales.

Capítulo 4

Estudio de simulación e ilustración con datos reales

El objetivo de este capítulo es realizar un estudio de simulación donde se compare el funcionamiento de los distintos selectores del parámetro de suavizado para el estimador tipo núcleo entre sí, los cuales fueron descritos en el capítulo anterior, tanto en el caso lineal como en el circular.

En el caso lineal, los más utilizados y los que emplearemos en este estudio de simulación serán: la regla del pulgar (propuesta por Silverman en 1986), la regla plug-in (propuesta por Sheater and Jones en 1991) y el selector de validación cruzada (propuesto por Browman en 1984). Análogamente, en el caso circular, consideraremos la regla del pulgar (introducida por Taylor en 2008), la regla plug-in (introducida por Oliveira et al. en 2013) y el selector de validación cruzada (introducido por Hall et al. en 1987).

Lo que haremos será elaborar un código en el software R que, dada una muestra, aplica cada uno de los selectores, calcula el estimador de la densidad correspondiente y computa el error global cometido con respecto al modelo teórico inicial y el error local para un punto x_0, el cual será distinto dependiendo del modelo que consideremos.

4.1. Caso lineal

Para el estimador tipo núcleo en el caso lineal dado en (1.1) emplearemos la función `density`, la cual utiliza por defecto el núcleo gaussiano y, en cada situación, escoge el selector del parámetro de suavizado correspondiente, por otro lado, para la estimación de la densidad en un punto de evaluación concreto utilizaremos la función `sm.density`.

Se utilizarán como modelos teóricos alguna de las densidades paramétricas de Marron y Wand (1992): los modelos M1 (normal estándar), M6 (bimodal), M8 (bimodal asimétrica)

y M9 (trimodal) cuya representación se puede ver en la Introducción de este trabajo, concretamente en la Figura 1.

Los tamaños muestrales considerados son $n = 50, 100$ y 500. Para cada una de las $B = 500$ muestras, se obtiene el estimador tipo núcleo. Se comenzará calculando el ISE cometido en cada caso y finalmente se promedian para obtener una aproximación del ISE promedio (AISE: average ISE) asociado a cada selector y modelo, que no es más que una aproximación del MISE teórico (1.6). Posteriormente, analizaremos el error puntual en $x_0 = 0$ utilizando una estimación del Error Cuadrático Medio (MSE).

Es necesario resaltar que para todos los selectores se emplea el mismo conjunto muestral, esto es, que el error cometido se calcula sobre el mismo conjunto de datos, por tanto, la variabilidad de la muestra no repercute. Además, todos los resultados se han redondeado a cuatro cifras decimales.

Average ISE	**M1**	**M6**	**M8**	**M9**
n=50	$5{,}534e-01$	$6{,}182e-01$	$7{,}904e-01$	$7{,}376e-01$
n=100	$3{,}468e-01$	$4{,}021e-01$	$5{,}400e-01$	$4{,}932e-01$
n=500	$9{,}355e-02$	$1{,}298e-01$	$1{,}976e-01$	$1{,}988e-01$

Tabla 4.1: Error global: Average ISE para representar el funcionamiento del selector de la regla del pulgar (propuesta por Silverman en 1986) para el estimador tipo núcleo (1.1).

Average ISE	**M1**	**M6**	**M8**	**M9**
n=50	$5{,}974e-01$	$6{,}922e-01$	$8{,}686e-01$	$7{,}747e-01$
n=100	$3{,}468e-01$	$4{,}362e-01$	$5{,}603e-01$	$4{,}976e-01$
n=500	$9{,}245e-02$	$1{,}277e-01$	$1{,}681e-01$	$1{,}638e-01$

Tabla 4.2: Error global: Average ISE para representar el funcionamiento del selector de la regla plug-in (propuesta por Sheather y Jones en 1991) para el estimador tipo núcleo (1.1).

Average ISE	**M1**	**M6**	**M8**	**M9**
n=50	$7{,}073e-01$	$8{,}757e-01$	$1{,}0898e-00$	$9{,}801e-01$
n=100	$4{,}117e-01$	$5{,}540e-01$	$6{,}621e-01$	$6{,}021e-01$
n=500	$1{,}107e-01$	$1{,}497e-01$	$1{,}873e-01$	$1{,}849e-01$

Tabla 4.3: Error global: Average ISE para representar el funcionamiento del selector de validación cruzada insesgada (unbiased cross-validation, ucv, propuesto por Browman en 1984) para el estimador tipo núcleo (1.1).

Se han reordenado los resultados de las tablas anteriores en las Tablas 4.4 y 4.5 para facilitar la comparación entre los distintos selectores del parámetro de suavizado.

El rango de valores entre los que oscila la regla del pulgar y la regla plug-in observados en las Tablas 4.1 y 4.2 son muy similares entre sí y ligeramente inferiores a los obtenidos con el método de validación cruzada (ver Tabla 4.3). Cabe resaltar también que, cuanto mayor sea el tamaño muestral, el selector óptimo será el construído mediante la regla plug-in (ver Tabla 4.5). Sin embargo, para tamaños muestrales más pequeños la regla plug-in proporciona peores resultados que la regla del pulgar como se puede ver en la Tabla 4.4.

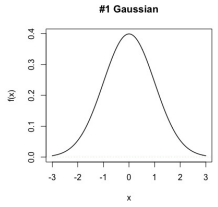
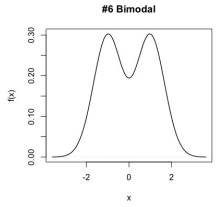
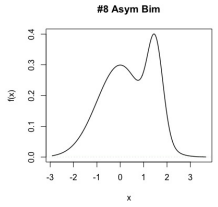
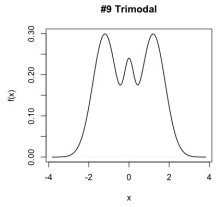
		Regla del pulgar	Regla plug-in	CV
M1		$3{,}327e-01$	$3{,}468e-01$	$4{,}117e-01$
M6		$4{,}021e-01$	$4{,}362e-01$	$5{,}540e-01$
M8		$5{,}400e-01$	$5{,}603e-01$	$6{,}621e-01$
M9		$4{,}932e-01$	$4{,}976e-01$	$6{,}021e-01$

Tabla 4.4: Average ISE para los distintos selectores del parámetro de suavizado del estimador tipo núcleo para densidades lineales (1.1) expuestos en el Capítulo 3 para un tamaño muestral $n = 100$.

Los valores recogidos en la Tabla 4.4 se corresponden con el tamaño muestral $n = 100$. Se espera que a medida que aumentamos el tamaño muestral, los errores disminuyan, veámoslo en la Tabla 4.5.

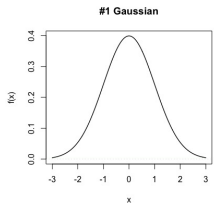
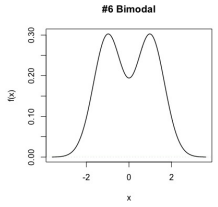
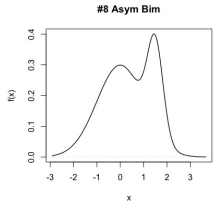
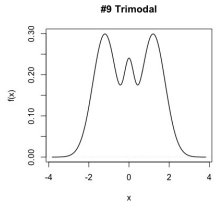
		Regla del pulgar	Regla plug-in	CV
M1		$9{,}355e-02$	$9{,}245e-02$	$1{,}107e-01$
M6		$1{,}298e-01$	$1{,}277e-01$	$1{,}497e-01$
M8		$1{,}976e-01$	$1{,}681e-01$	$1{,}873e-01$
M9		$1{,}988e-01$	$1{,}638e-01$	$1{,}849e-01$

Tabla 4.5: Average ISE para los distintos selectores del parámetro de suavizado del estimador tipo núcleo para densidades lineales (1.1) expuestos en el Capítulo 3 para un tamaño muestral $n = 500$.

Cabe destacar, que para los modelos más complejos (M8 y M9), el parámetro de suavizado construído por el método de validación cruzada funciona mejor que el obtenido mediante la regla del pulgar, lo cual puede deberse a que el método de validación cruzada le da mucho peso a la información muestral, conviertiendo sus irregularidades en características del propio modelo.

Dependiendo del modelo, en $x_0 = 0$ se localizará una moda (como es el caso de los

modelos M1, M8 y M9) o un valle (modelo M6). Sin embargo, su estimación a través del estimador tipo núcleo para cada uno de los tres selectores del parámetro de suavizado en el caso lineal es satisfactoriamente buena como podemos observar en las Tablas 4.6, 4.7 y 4.8. Sin embargo, a continuación, veremos que esto no ocurre en el caso circular.

MSE $x_0 = 0$	**M1**	**M6**	**M8**	**M9**
n=50	$4{,}334e-03$	$2{,}486e-03$	$2{,}443e-03$	$1{,}960e-03$
n=100	$2{,}683e-03$	$1{,}841e-03$	$1{,}671e-03$	$1{,}723e-03$
n=500	$7{,}375e-04$	$7{,}370e-04$	$5{,}252e-04$	$1{,}292e-03$

Tabla 4.6: Error local: MSE para representar el funcionamiento del selector de la regla del pulgar (propuesta por Silverman en 1986) para el estimador tipo núcleo (1.1).

MSE $x_0 = 0$	**M1**	**M6**	**M8**	**M9**
n=50	$5{,}414e-03$	$2{,}603e-03$	$2{,}925e-03$	$3{,}0314e-03$
n=100	$3{,}121e-03$	$1{,}883e-03$	$2{,}180e-03$	$2{,}322e-03$
n=500	$8{,}143e-04$	$6{,}857e-04$	$6{,}446e-04$	$1{,}149e-03$

Tabla 4.7: Error local: MSE para representar el funcionamiento del selector de la regla plug-in (propuesta por Sheather y Jones en 1991) para el estimador tipo núcleo (1.1).

MSE $x_0 = 0$	**M1**	**M6**	**M8**	**M9**
n=50	$6{,}938e-03$	$2{,}869e-03$	$3{,}636e-03$	$3{,}723e-03$
n=100	$3{,}727e-03$	$2{,}190e-03$	$2{,}679e-03$	$2{,}648e-03$
n=500	$9{,}754e-04$	$7{,}734e-04$	$8{,}064e-04$	$1{,}170e-03$

Tabla 4.8: Error local: MSE para representar el funcionamiento del selector de validación cruzada insesgada (unbiased cross-validation, ucv, propuesto por Browman en 1984) para el estimador tipo núcleo (1.1).

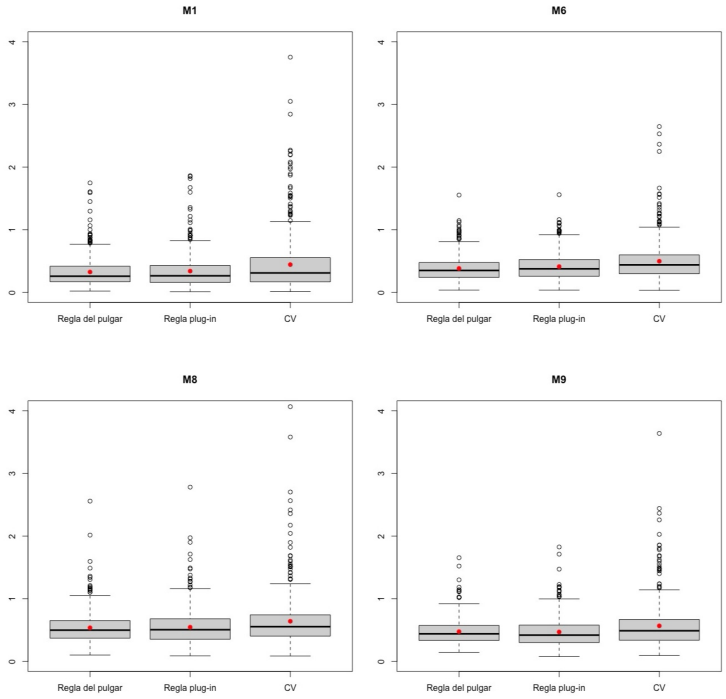


Figura 4.1: Boxplots para los ISE de los modelos M1, M6, M8 y M9 de Marron y Wand (1992) con los selectores del parámetro de suavizado construídos a partir de la regla del pulgar, la regla plug-in y el método de validación cruzada para un tamaño muestral $n = 100$. El punto rojo indica el MISE para cada tipo de selector y modelo.

Como podemos ver en la Figrua 4.1, se presentan los diagramas de cajas de los ISE

para los selectores obtenidos mediante la regla del pulgar, la regla plug-in y el método de validación cruzada para el caso lineal. Además, se puede ver que el selector de validación cruzada genera ISE's atípicos, es decir, tiene mucha variabilidad. Sin embargo, el rango de valores de la regla del pulgar y la regla plug-in son muy similares y ambos tienen menos variabilidad que el selector de validación cruzada.

4.2. Caso circular

Por otro lado, para el estimador tipo núcleo en el caso circular dado en (2.4) emplearemos la función `kern.den.circ`, la cual utiliza por defecto como núcleo circular una distribución de von Mises, y, en cada situación, escoge el selector del parámetro de suavizado correspondiente mediante el argumento `bw` calculado previamente con las funciones `bw.rt`, `bw.pi` o `bw.CV` dependiendo del selector utilizado (implementadas en el paquete `NPCirc`).

Como modelos teóricos se emplearán el conjunto de densidades de mixturas consideradas en la Introducción de este trabajo (ver Figura 3). Es decir, denotando por M el número de componentes de la mixtura de von Mises los modelos que consideraremos serán: modelo M7 ($M = 2$ con el mismo peso), M9 ($M = 2$ con distinto peso) y M13 ($M = 3$) de Oliveira, et al. (2012) sacados de la librería `NPCirc` que muestran multimodalidad, asimetría y/o picos.

Los tamaños muestrales considerados serán, también, $n = 50, 100$ y 500. Además, para los estimadores tipo núcleo se ha empleado la misma rejilla equiespaciada en el intervalo $[0, 2\pi]$ y formada por 250 puntos, por tanto, el paso de dicha rejilla será de 0,02513 aproximadamente.

Al igual que para el caso lineal, para cada una de las muestras, se obtiene el estimador tipo núcleo evaluado en dicha rejilla de puntos, se calcula el ISE cometido en cada caso y finalmente se promedian dichos valores en las muestras para obtener el AISE asociado a cada selector y modelo. También analizaremos el error puntual en $x_0 = \pi/2$ o $x_0 = 0$ dependiendo del modelo, utilizando una estimación del Error Cuadrático Medio (MSE).

A continuación, en las Tablas 4.9, 4.10 y 4.11 se muestra el promedio del ISE del estimador tipo núcleo de la densidad circular (2.4) considerando los diferentes selectores del parámetro de suavizado ν tratados en el capítulo anterior.

Average ISE	**M7**	**M9**	**M13**
n=50	$1{,}055e-01$	$1{,}329e-02$	$1{,}082e-01$
n=100	$1{,}061e-01$	$8{,}110e-03$	$1{,}084e-01$
n=500	$1{,}080e-01$	$2{,}543e-03$	$1{,}085e-01$

Tabla 4.9: Error global: Average ISE para representar el funcionamiento del selector de la regla del pulgar (propuesta por Taylor en 2008) para el estimador tipo núcleo para la densidad circular (2.4).

Como mencionamos en la Sección 3.2.1, en la Tabla 4.9 se refleja el buen funcionamiento de la regla del pulgar en distribuciones circulares unimodales como el modelo M9 (ver Figura 3.1). Sin embargo, este tipo de selector no es adecuado para distribuciones con más de una moda como son los modelos M7 y M13, ya que proporciona errores bastante grandes pues tiende a proporcionar estimaciones uniformes para la densidad circular, que se corresponde con una von Mises de parámetro de concentración $\kappa = 0$ (ver Figura 3.2).

Average ISE	**M7**	**M9**	**M13**
n=50	$2{,}660e-02$	$2{,}306e-02$	$3{,}899e-02$
n=100	$1{,}433e-02$	$1{,}088e-02$	$2{,}098e-02$
n=500	$3{,}440e-03$	$2{,}266e-03$	$5{,}362e-03$

Tabla 4.10: Error global: Average ISE para representar el funcionamiento del selector de la regla plug-in (Oliveira et al., 2013) para el estimador tipo núcleo para la densidad circular (2.4).

Average ISE	**M7**	**M9**	**M13**
n=50	$2{,}119e-02$	$1{,}324e-02$	$3{,}069e-02$
n=100	$1{,}283e-02$	$8{,}073e-03$	$1{,}847e-02$
n=500	$3{,}533e-02$	$2{,}377e-03$	$5{,}320e-03$

Tabla 4.11: Error global: Average ISE para representar el funcionamiento del selector de validación cruzada (propuesto por Hall et al. en 1987) para el estimador tipo núcleo para la densidad circular (2.4).

Se han reordenado los resultados de las tablas anteriores en la Tabla 4.12 para facilitar la comparación entre los distintos selectores del parámetro de suavizado ν.

Como era de esperar, en el caso de modelos simples funciona mejor la regla plug-in mientras que a medida que aumenta la complejidad del modelo, el selector que mejor funciona es el de validación cruzada como podemos ver en la Tabla 4.12. Como excepción, para el modelo M9 ($M = 2$ con distinto peso), la regla del pulgar muestra mejores resultados para tamaños muestrales pequeños o moderados ($n = 50$ ó $n = 100$) que la propia regla plug-in, aunque peores que el que proporciona el método de validación cruzada.

Esto puede deberse a que el método de validación cruzada suele infrasuavizar, es decir, favorece a un buen ajuste en el caso de modelos complejos, sin embargo, en modelos simples presenta demasiadas irregularidades, aunque no tantas como en el caso lineal. De hecho, aquí es bastante estable. Estas irregularidades, el selector de validación cruzada las convierte en características del propio modelo (lo cual es inadecuado ya que le da demasiado peso a la información muestral).

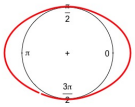
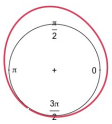
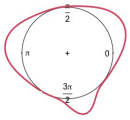
		Regla del pulgar	Regla plug-in	CV
M7		$1{,}061e-01$	$1{,}433e-02$	$1{,}283e-02$
M9		$8{,}110e-03$	$1{,}088e-02$	$8{,}073e-03$
M13		$1{,}084e-01$	$2{,}098e-02$	$1{,}847e-02$

Tabla 4.12: Average ISE para los distintos selectores del parámetro de suavizado del estimador tipo núcleo para densidades circulares (2.4) expuestos en el Capítulo 3 para un tamaño muestral $n = 100$.

A continuación, en las Tablas 4.13, 4.14 y 4.15, analizaremos el error puntual en $x_0 = \pi/2$ o $x_0 = 0$ dependiendo del modelo.

Es claro que a medida que aumenta el tamaño muestral, los errores locales disminuyen. Además, a pesar de que en $x_0 = \pi/2$ en M9 y M13 se localice en una moda y una zona de baja densidad respectivamente (ver Figura 3 de la Introducción), el comportamiento de la estimación de la densidad en dicho punto es igual de bueno. Sin embargo, esto no ocurre en el caso lineal pues en zonas de baja densidad funcionaría peor debido a la presencia de colas en las distribuciones.

MSE x_0	**M7**	**M9**	**M13**
	$x_0 = 0$	$x_0 = \pi/2$	$x_0 = \pi/2$
n=50	$4{,}939e-02$	$5{,}220e-02$	$2{,}055e-02$
n=100	$4{,}997e-02$	$5{,}196e-02$	$2{,}029e-02$
n=500	$5{,}052e-02$	$5{,}106e-02$	$2{,}006e-02$

Tabla 4.13: Error local: Estimación del MSE para representar el funcionamiento del selector de la regla del pulgar (propuesta por Taylor en 2008) para el estimador tipo núcleo para la densidad circular (2.4).

MSE x_0	**M7**	**M9**	**M13**
	$x_0 = 0$	$x_0 = \pi/2$	$x_0 = \pi/2$
n=50	$1{,}021e-02$	$5{,}352e-02$	$3{,}451e-02$
n=100	$5{,}932e-03$	$5{,}251e-02$	$3{,}082e-02$
n=500	$1{,}561e-03$	$5{,}056e-02$	$2{,}983e-02$

Tabla 4.14: Error local: MSE para representar el funcionamiento del selector de la regla plug-in (Oliveira et al., 2013) para el estimador tipo núcleo para la densidad circular (2.4).

MSE x_0	**M7**	**M9**	**M13**
	$x_0 = 0$	$x_0 = \pi/2$	$x_0 = \pi/2$
n=50	$8{,}522e-03$	$5{,}222e-02$	$3{,}311e-02$
n=100	$5{,}781e-03$	$5{,}199e-02$	$3{,}082e-012$
n=500	$1{,}628e-03$	$5{,}061e-02$	$2{,}998e-02$

Tabla 4.15: Error local: MSE para representar el funcionamiento del selector de validación cruzada (propuesto por Hall et al. en 1987) para el estimador tipo núcleo para la densidad circular (2.4).

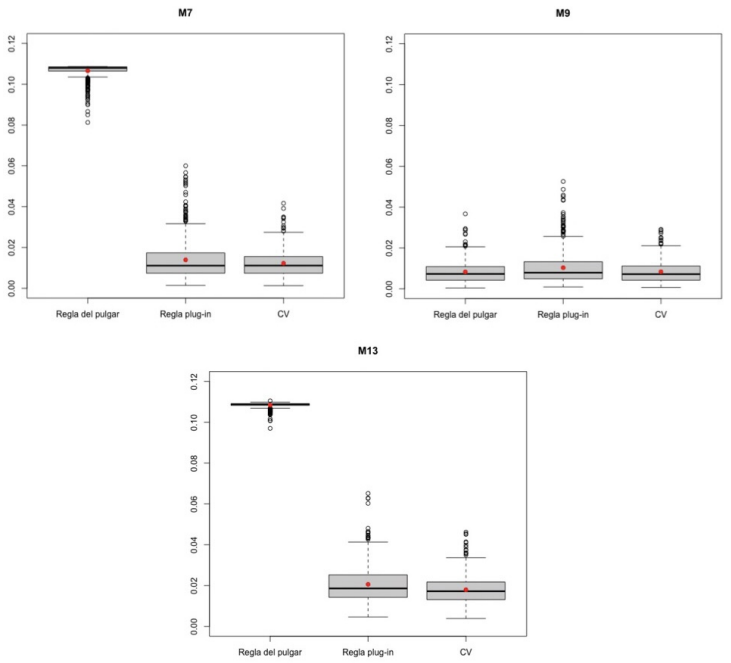


Figura 4.2: Boxplots para los ISE de los modelos M7, M9 y M13 de Oliveira, et al. (2012) sacados de la librería `NPCirc` con los selectores del parámetro de suavizado construídos a partir de la regla del pulgar, la regla plug-in y el método de validación cruzada para un tamaño muestral $n = 100$. El punto rojo indica el MISE para cada tipo de selector y modelo.

En general, valores pequeños de M son adecuados para modelos simples mientras que tamaños grandes M son adecuados para modelos más complejos y tamaños de muestra n grandes. Por tanto, la fijación del número de componentes de la mixtura para la distribución de von Mises (siempre la utilizamos como densidad de referencia en el caso circular, salvo que se indique lo contrario) es esencial para calcular el parámetro de suavizado ν.

Por consiguiente, podemos formular la siguiente pregunta: ¿cuánto afecta el número de componentes de la mixtura, M, en el selector de la regla plug-in?

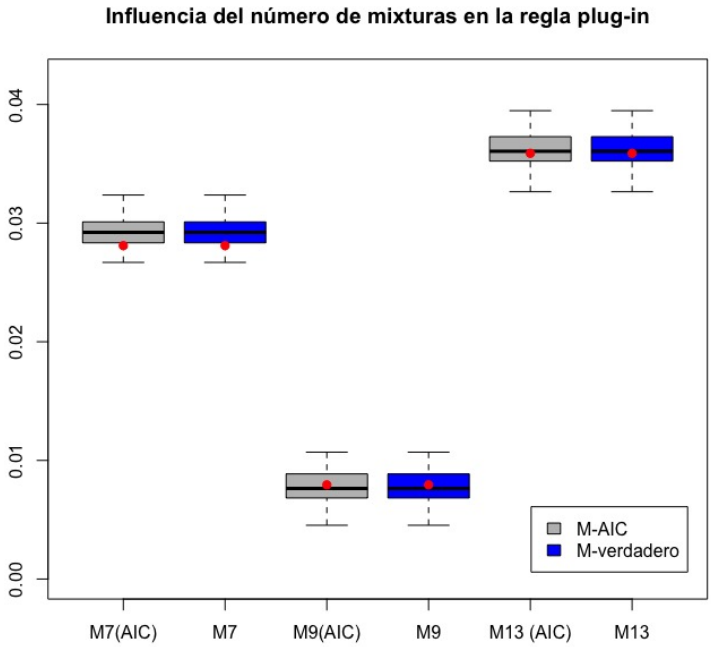


Figura 4.3: Boxplots para los ISE de los modelos M7, M9 y M13 de Oliveira, et al. (2012) sacados de la librería `NPCirc` con el selector del parámetro de suavizado construído a partir de la regla plug-in para un tamaño muestral $n = 500$ considerando el número de componentes de la mezcla relativa a cada modelo M (en azul), y, por otro lado M seleccionada mediante el criterio AIC (en gris) . El punto rojo indica el MISE (promedio del ISE) para cada tipo de selector y modelo.

Para ello emplearemos la función `bw.pi` del paquete `NPCirc` la cual se encarga de calcular el parámetro de suavizado mediante el selector plug-in en el caso circular, que, por defecto, el número de componentes de la mezcla, M es determinado por el criterio AIC (2.3). Este criterio selecciona el mejor modelo entre una mixtura de 2-5 distribuciones de von Mises. Además, mediante el argumento `M` podemos introducir manualmente el número de componentes de la mezcla (por ejemplo: $M = 2$ en el caso de los modelos M7 y M9 y $M = 3$ en el caso del modelo M13).

Lo que haremos primero será generar $B = 100$ muestras de tamaño $n = 500$ y calcular el parámetro de suavizado mediante la regla plug-in considerando, por un lado, el número de componentes de la mixtura, M, escogido según el criterio AIC, y, por otro lado, el M verdadero de cada modelo. Despúes, mediremos la influencia del número de componentes de la mixtura realizando una aproximación del ISE como podemos ver en la Figura 4.3.

Viendo la Figura 4.3 podemos concluír que el criterio AIC funciona satisfactoriamente en la práctica ya que el número de componentes de la mixtura de von Mises obtenida a través de él en la mayoría de los casos coincide con número de componentes de cada uno de los modelos, por lo que los resultados son muy semejantes.

Para finalizar este capítulo de simulación sería interesante formularse la siguiente cuestión: ¿qué pasaría si nos olvidásemos de que nuestros datos están en el círculo?

Para obtener una respuesta a lo anterior lo que haremos será simular $B = 500$ muestras para los diferentes tamaños muestrales y modelos considerados hasta ahora en el caso circular. Posteriormente, pasaremos dichas muestras a tipo numérico mediante la función `as.numeric` y obtendremos el estimador tipo núcleo (1.1) para los diferentes selectores del parámetro de suavizado considerados en el caso lineal.

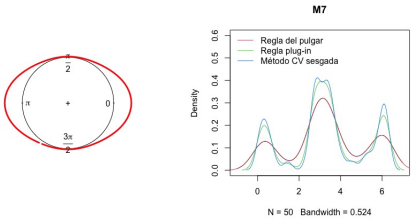
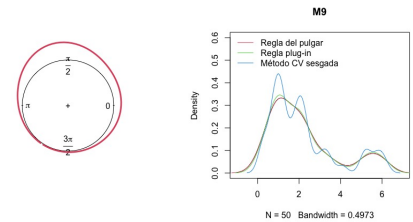
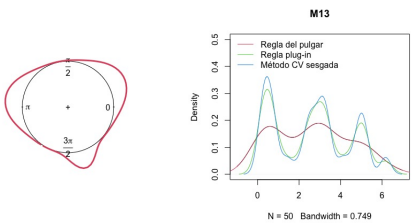
		Regla del pulgar	**Regla plug-in**	**CV**
M7		$7{,}023e-00$	$2,798e-00$	$2,902e-00$
M9		$1,217e-00$	$1,276e-00$	$1,524e-00$
M13		$7,385e-00$	$2,801e-00$	$1,524e-00$

Tabla 4.16: Average ISE para los distintos selectores del parámetro de suavizado del estimador tipo núcleo (1.1) partiendo de datos circulares con un tamaño muestral $n = 50$.

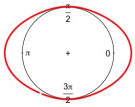
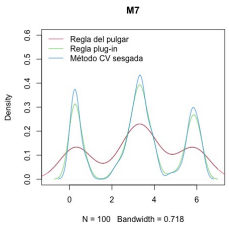
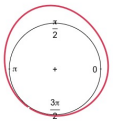
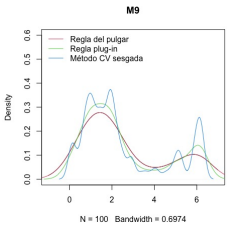
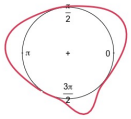
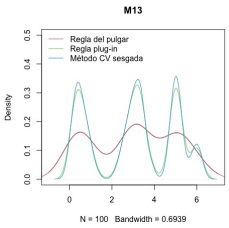
			Regla del pulgar	Regla plug-in	CV
M7			$5,879e-00$	$2,015e-00$	$2,058e-00$
M9			$7,504e-01$	$7,636e-01$	$9,053e-01$
M13			$6,322e-00$	$1,707e-00$	$1,7004e-00$

Tabla 4.17: Average ISE para los distintos selectores del parámetro de suavizado del estimador tipo núcleo (1.1) partiendo de datos circulares con un tamaño muestral $n = 100$.

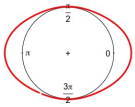
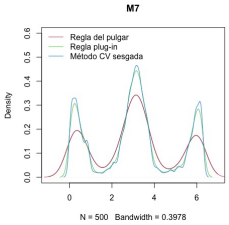
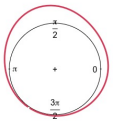
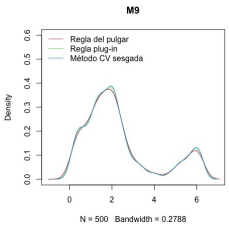
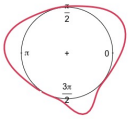
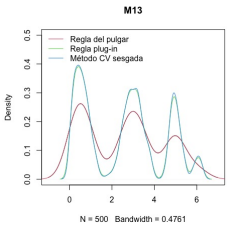
			Regla del pulgar	Regla plug-in	CV
M7			$3,746e-00$	$8{,}569e-01$	$8{,}088e-01$
M9			$3{,}288e-01$	$3{,}162e-01$	$3{,}638e-01$
M13			$3,894e-00$	$5{,}564e-01$	$5{,}569e-01$

Tabla 4.18: Average ISE para los distintos selectores del parámetro de suavizado del estimador tipo núcleo (1.1) partiendo de datos circulares con un tamaño muestral $n = 500$.

Lograremos así, mediante una aproximación del MISE (ver Tablas 4.16, 4.17 y 4.18), plasmar la mala estimación de la densidad obtenida al utilizar la estimación tipo núcleo del caso lineal a datos circulares.

Podemos concluir, por tanto, que las técnicas no paramétricas para datos lineales pueden no proporcionar buenos resultados para datos circulares ya que no tienen en cuenta la periodicidad de los mismos.

4.3. Ilustración con datos reales

En esta sección, se presentarán dos conjuntos de datos: uno para el caso lineal y otro para el caso circular respectivamente con fines ilustrativos los cuales serán analizados como parte de este trabajo. A continuación se realizará una breve descripción de los mismos.

Presentación de los datos y técnicas implementadas

- **La emisión de sellos de México de 1872 en Hidalgo** (caso lineal): Este conjunto de datos surge con motivo del uso de sellos como inversión. Como resume Ameijeiras (2017), antes de 1940 los sellos eran impresos en varios tipos de papel y al no existir un control sobre su calidad, se obtenían distintos grosores, característica que determinaba el valor de cada sello. Por tanto, era más probable que fabricasen sellos delgados que gruesos, y, a pesar de que la marca de agua de algunos sellos (marca que indica dónde se produjo el sello, ver Figura 4.4) puede ayudar a catalogar la emisión de los mismos, no fue suficiente para establecer un criterio de clasificación ya que existía un problema añadido: los diferentes grosores. Ese problema aparece en el catálogo de Hidalgo (1872) donde se ha tomado una muestra de $n = 485$ sellos distintos.

Figura 4.4: Izquierda: *sello de doce centavos* de la edición Hidalgo de México impreso en 1872. Derecha: ejemplo de una marca de agua en la parte superior de un sello de Zululand (región histórica de Sudáfrica) que indica que el sello pertenece a la producción de *Crown CA*. Imágenes extraídas de Wikipedia, bajo licencia Creative Commons.

A continuación, en la Figua 4.5 se ilustra el funcionamiento del estimador tipo núcleo (1.1) con distintos selectores del parámetro de suavizado, h, recogidos en el Capítulo 3 de este trabajo.

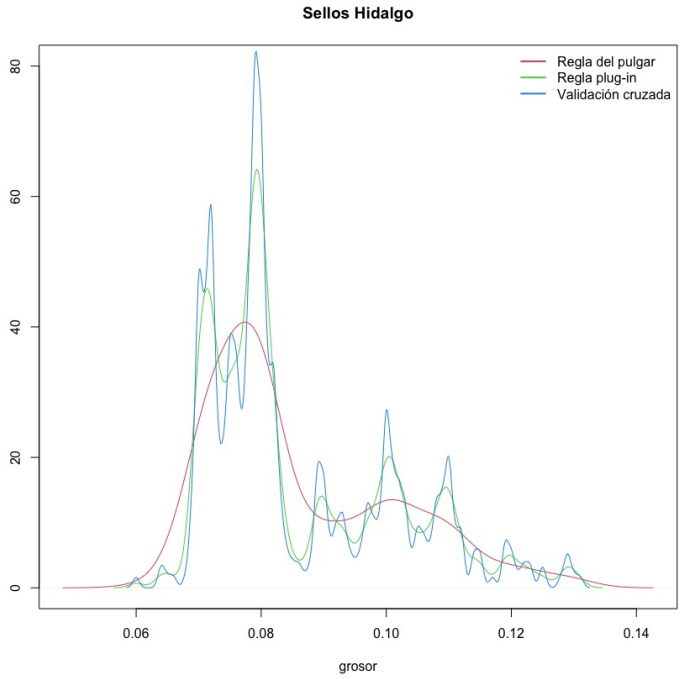


Figura 4.5: Representación lineal de los estimadores tipo núcleo para los ditintos selectores del parámetro de suavizado en el caso lineal para el grosor de los sellos: regla del pulgar (línea roja), regla plug-in (línea verde) y método de validación cruzada (línea azul).

Se han considerado los tres selectores del parámetro de suavizado considerados hasta ahora: la regla del pulgar, la regla plug-in y el de validación cruzada. El estimador con el selector de validación cruzada, en este caso, se comporta de manera similar al de la regla plug-in pero oscila un poco más, ya que, como vimos en el sección anterior, el selector de validación cruzada para el caso lineal tiende a infrasuavizar. También se puede ver en la Figura 4.5 que en las colas de la distribución, el método de validación cruzada detecta muchos grupos mientras que la regla plug-in detecta tres o cuatro. Por otro lado, la regla del pulgar solo detecta un grupo en las colas de la distribución mientras que en zonas con alta densidad detecta dos grupos, es decir, su comportamiento es opuesto al selector de validación cruzada.

Podemos concluir entonces que este conjunto de datos refleja perfectamente los resultados obtenidos en nuestro estudio de simulación: la regla plug-in es el selector que mejor funciona en el caso lineal ya que la regla del pulgar tiende a sobresuavizar, y, por otro lado, el método de validación cruzada proporciona un parámetro de suavizado demasiado grande, lo que resulta en una densidad ajustada poco suavizada.

- **Estrategia de cambio de dirección de presa de peces ante una amenaza** (caso circular): No se sabe a ciencia cierta cuál es la estrategia de escape efectiva entre especies ante una posible amenaza de depredación. La teoría clásica dice que existen dos estrategias de escape principales: por un lado, una estrategia proteica con alta variabilidad en la dirección de escape para evitar que el depredador se anticipe al rumbo de la presa, y, por otro lado, una estrategia óptima que maximiza la distancia del depredador.

  Este conjunto de datos formado por $n = 502$ larvas de peces cebra (Danio rerio, ver Figura 4.6) surge del estudio recogido en Nair et al. (2017) donde se realizan mediciones en su dirección de escape en respuesta a un robot depredador. Por tanto, los datos recogidos son ángulos y pueden ser tratados con técnicas de datos circulares.

Figura 4.6: Pez cebra (Danio rerio). Imágen extraída de Wikipedia, bajo licencia Creative Commons.

  De dicho estudio se concluye que las larvas de pez cebra emplean una estrategia mixta para sobrevivir a los encuentros con depredadores que depende de la dirección en la que se acerque el depredador, según lo detecta el sistema visual. Dicha estrategia consiste en la combinación de una estrategia que maximiza la distancia al depredador mediante una respuesta contralateral como consecuencia de un ataque a través del campo visual central de la presa, y, por otro lado, cuando el depredador aparece por el campo visual periférico de la presa acercándose por los extremos rostral o caudal, utilizará una estrategia de escape con igual probabilidad de una dirección contralateral o ipsilateral, es decir, una combinación entre una estrategia de escape que maximiza la distancia de la amenaza cuando ésta es beneficiosa, y, por el contrario, se utiliza una estrategia de escape aleatoria.

  Destacar también que diversos animales muestran respuestas similares a las amenazas visuales, por tanto, una estrategia mixta puede ofrecer una solución común para evadir a los depredadores. Para más detalles referentes a este conjunto de datos ver Nair et al. (2017).

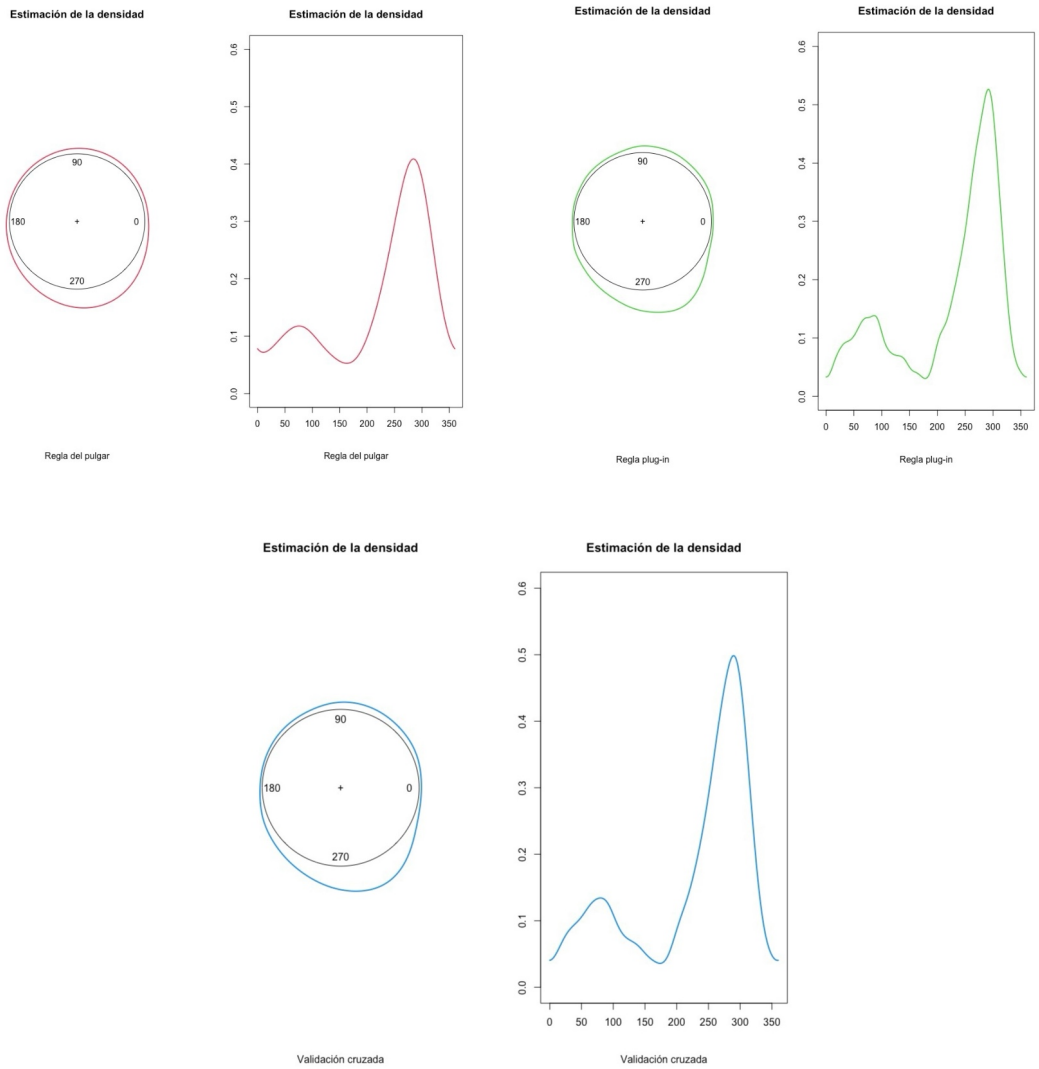


Figura 4.7: Ventana gráfica dividida en tres representaciones, una para cada tipo de selector del parámetro de suavizado, ν: la regla del pulgar (en rojo), la regla plug-in (en verde) y el de validación cruzada por máxima verosimilitud (LCV) (en azul). En cada una de ellas aparece una representación circular (izquierda), y, una representación lineal (derecha) del estimador tipo núcleo para la densidad circular (2.4), utilizando en cada caso el selector del parámetro de suavizado, ν, correspondiente.

El estimador tipo núcleo para la densidad circular (2.4) se ha calculado utilizando los diferentes selectores de parámetros de suavizado, ν, considerados en el Capítulo 3.

En la Figura 4.7, se puede ver que la regla del pulgar, la regla plug-in y el selector de validación cruzada por máxima verosimilitud (LCV) proporcionan curvas ajustadas similares. A pesar de lo que señalaban los resultados del estudio de simulación, en este contexto, no se observa un comportamiento superior de ninguno de los selectores, proporcionando todos ellos estimaciones muy similares y parecidas a una distribución circular bimodal asimétrica.

Bibliografía

[1] Ameijeiras, J. A., *Assessing Simplifying Hypotheses in Density Estimation*, Tesis doctoral USC, 2017.

[2] Mardia, K.V. y Jupp, P. E. *Directional Statistics*, Wiley, 2000.

[3] Nair, A., Changsing, K., Stewart, W. J. y McHenry, M. J. Fish prey change strategy with the direction of a threat. *Proceedings of the Royal Society B: Biological Sciences*, 2017.

[4] Oliveira, M. *Nonparametric Circular Methods for Density and Regression*, Tesis doctoral USC, 2013.

[5] Pewsey, A., Neuhäuser, M. y Ruxton, G. D. *Circular Statistics in R*, 2013.

[6] Wand, M. P. y Jones, M. C. *Kernel Smoothing*, Springer, 1995.