QÜESTIIÓ, Vol 15, 1. pp. 21-45, 1991

TÉCNICAS DE VALIDACIÓN CRUZADA EN LA ESTIMACIÓN DE LA DENSIDAD BAJO CONDICIONES DE DEPENDENCIA

A. QUINTELA Y J.M. VILAR FERNÁNDEZ

Se estudian modificaciones de las técnicas de validación cruzada de Kullback-Leibler y mínimos cuadrados para obtener el parámetro de suavización asociado a un estimador general no paramétrico de la función de densidad, a partir de la muestra, en el supuesto de que los datos verifican alguna condición débil de dependencia.
Se demuestra que los parámetros obtenidos por estas dos técnicas son asintóticamente óptimos. Y se realiza un estudio de simulación.

Cross-Validation Techniques in Density Estimation under Dependence Conditions.

Keywords: Estimadores no paramétricos, selección del parámetro de suavización, condiciones de dependencia.

Clasificación AMS (1980): 62G05, 62G20.

-Dept. Matemáticas. Fac. Informática. Universidad de La Coruña.
-Article rebut el desembre de 1990.

1. INTRODUCCIÓN

1.1 Estimación No Paramétrica de la Densidad

La estimación de curvas asociadas a modelos de probabilidad (densidad, distribución, regresión, razón de fallo, ...) desde un enfoque no paramétrico, esto es, asumiendo solamente hipótesis generales de regularidad de la curva pero no una forma funcional específica, ha sido ampliamente estudiada desde los primeros trabajos de Rosenblatt (1956) y Parzen (1962), por ser una técnica más flexible y de fácil cálculo que la metodología paramétrica clásica, que puede utilizarse en un conjunto más amplio de situaciones y, en cualquier caso, siempre es válido como un estudio inicial exploratorio.

En este trabajo se estudian problemas relacionados con la estimación no paramétrica de la función de densidad, $f(x)$, asociada a una variable aleatoria real, X, a partir de una muestra de datos $X_1, X_2, \cdots, X_n$, no necesariamente independientes. En concreto, se estudian dos técnicas de obtención del parámetro de suavización asociado a un estimador no paramétrico de $f(x)$. En el Apartado 1 se expone el problema, en el 2 se definen las técnicas de validación cruzada, en el 3 se obtiene la bondad asintótica de las técnicas definidas y en el 4 se realiza un estudio de simulación.

La mayoría de los estimadores no paramétricos de la función de densidad pueden escribirse en la forma:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_h(x, X_i) \tag{1}$$

como han indicado entre otros Watson y Leadbetter (1964).

Siendo $\{\delta_h(x,u)\}$ una sucesión de funciones de ponderación definidas en $\mathbb{R}^2$, con valores en $\mathbb{R}$, y que nos indican el peso del dato X_i en la estimación de $f(x)$, tomando valores altos si x está próximo a X_i y valores próximos a cero, o cero, en caso contrario. El parámetro $h = h(n)$, llamado parámetro de suavización, ventana o banda, nos indica el entorno de x en el que tomamos puntos muestrales que influyen en la estimación de $f(x)$ y su elección es fundamental para obtener buenas estimaciones como discutiremos más adelante.

Entre los estimadores no paramétricos más utilizados que se ajustan al modelo (1) se pueden citar los siguientes:

(i) EL HISTOGRAMA: sea $P_n = \{A_j : j \in \mathbf{N}\}$ una partición en $\mathbb{R}$ con $\|P_n\| = h_n \to 0$ e I_j la función indicadora de A_j, entonces

(2) $$\delta_h(x,u) = \frac{1}{h} \sum_{j=1}^{\infty} I_j(x)\, I_j(u)$$

(ii) EL ESTIMADOR NÚCLEO (Kernel): sea $K(u)$ una función real de variable real con $\int K(u)\, d(u) = 1$, entonces

(3) $$\delta_h(x,u) = \frac{1}{h} K\left(\frac{x-u}{h}\right)$$

(iii) EL ESTIMADOR DE SERIES ORTOGONALES: sea $\{\Psi_k(x)\}$ una sucesión de funciones, ortonormal y completa respecto al producto interior:

$$\langle \Psi_i, \Psi_j \rangle = \int \Psi_i(x)\, \Psi_j(x)\, \omega(x)\, d(x)$$

entonces definimos las siguientes funciones de ponderación:

(4) $$\delta_h(x,u) = \sum_{i=1}^{m(n)} \Psi_i(x)\, \Psi_i(u)\, \omega(u)$$

siendo $m(n) = Ch_n^{-1}$ y $\omega(u)$ una función de ponderación.

1.2 El Problema de la elección del parámetro de suavización

Para utilizar el estimador (1) y, en general, cualquier estimador no paramétrico de una curva, un problema básico es la elección de la cantidad de suavización que se requiere, la cual está controlada por el parámetro de suavización: h_n, ya que tomar una banda grande nos lleva a utilizar muchas observaciones en la ponderación obteniendo estimaciones sobresuavizadas, con poca varianza pero mucho sesgo. Por el contrario, si utilizamos una "banda estrecha" no se utilizaría el número necesario de observaciones en la estimación con lo que se obtienen estimaciones poco suavizadas, con poco sesgo pero mucha varianza.

Este hecho se refleja en las figuras 1-2-3, en las que se ha simulado la estimación no paramétrica de la función de densidad asociada a un proceso AR (1) generado a partir de ruido normal (0,1) y, por tanto, sigue una distribución normal (0, 1'1547). La estimación se ha obtenido a partir de una muestra de 400 datos, utilizando el estimador núcleo, siendo $K(u) = 3/4\,(1-u^2)$ si $|u| \leq 1$ (Kernel de Epanechnikov), con bandas $h(n) = Cn^{-1/5}$, obteniéndose:

Figura 1, $C = 1$	(poca suavización)	E.C.M. = 552'134	E-6
Figura 2, $C = 2.5$	(suavización adecuada)	E.C.M. = **154'610**	**E-6**
Figura 3, $C = 5$	(mucha suavización)	E.C.M. = 1195'650	E-6

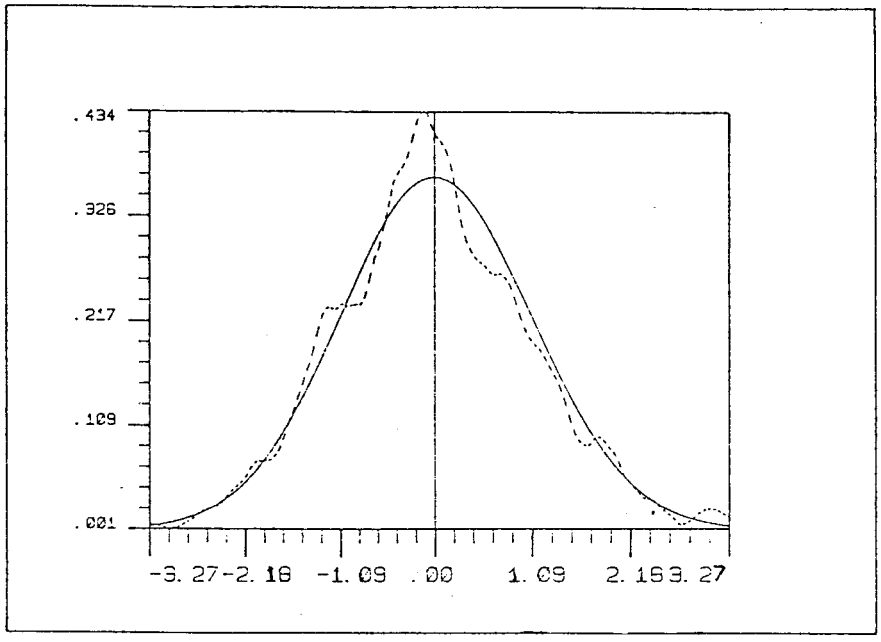

Figura 1. $C = 1$ Poca Suavización

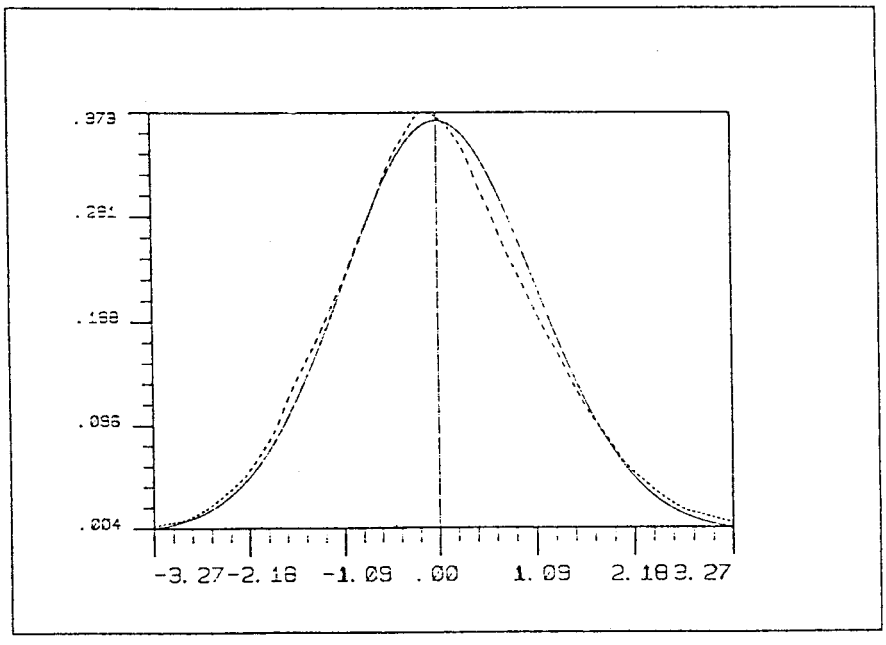

Figura 2. $C = 2'5$ Suavización adecuada

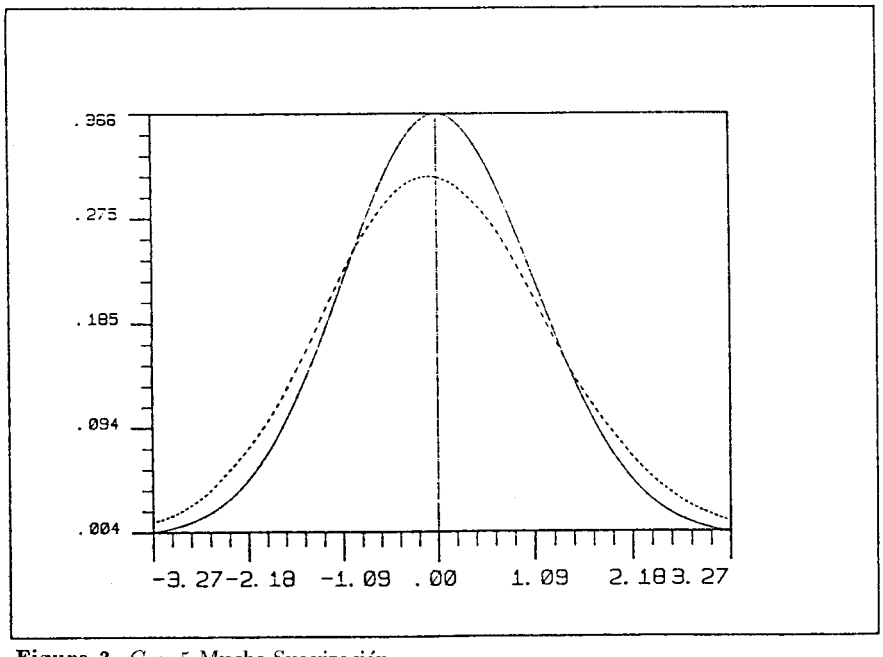

Figura 3. $C = 5$ Mucha Suavización

Un parámetro de suavización es bueno u óptimo respecto a un criterio de error elegido previamente, éste puede ser puntual o global según estemos interesados en la estimación en un punto o de toda la curva. En este trabajo se utilizan criterios globales basados en la norma L^2 y cuya definición es la siguiente:

$$\text{Error Cuadrático Ponderado,} \quad \text{ECP} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_h(X_i) - f(X_i)\right)^2 \omega(X_i)$$

$$\text{Error Cuadrático Integrado,} \quad \text{ECI} = \int \left(\hat{f}_h(x) - f(x)\right)^2 \omega(x)\, d(x)$$

$$\text{Media del Error Cuadrático Integrado,} \quad \text{MECI} = \text{E(ECI)}$$

siendo $0 \leq \omega(x)$ una función de ponderación.

Estos criterios son los más estudiados, aunque Devroye-Gyorfi (1984) han utilizado criterios basados en la norma L^1, indicando que hay diferencias importantes respecto a los primeros, los cuales son más fáciles de manejar matemáticamente.

Un parámetro de suavización, h, será óptimo respecto a una medida de error si minimiza la expresión de ésta. Si se desea obtener el óptimo respecto al MECI, como éste es díficil de calcular se utiliza una aproximación, el AMECI, que para el estimador núcleo viene dado por:

$$\text{(5)} \qquad \text{AMECI}(h) = (nh)^{-1} \left(\int K^2\right) + h^4 \left(\int x^2 K\right)^2 \left(\int (f'')^2\right)/4$$

el primer sumando es debido a la varianza y el segundo al sesgo (ver Silverman (1986)). Y minimizando esta expresión respecto a h se obtiene el óptimo cuya expresión es:

$$\text{(6)} \qquad h_{\text{AMECI}} = n^{-1/5} \left(4 \int K^2\right)^{1/5} \left(\left(\int x^2 K\right)^2 \left(\int (f'')^2\right)\right)^{-1/5}$$

que desafortunadamente no se puede calcular ya que por el término $\left(\int (f'')^2\right)$ depende de la función teórica que deseamos estimar.

1.3 Métodos de Validación Cruzada

El problema anterior ha hecho que se estudien métodos para obtener el parámetro h que minimice algún criterio de error a partir de los datos muestrales.

El primer método que se desarrolla es el propuesto por Habema-Hermans-Van der Broek (1974), que basándose en la idea de verosimilitud elige el h que maximiza la función:

$$L_1(h) = \prod_{i=1}^{n} \hat{f}_n(X_i; h) \tag{7}$$

que trivialmente es cero, ello es porque se está utilizando el dato muestral X_i para estimar la densidad en dicho punto, por ello se ha modificado la función anterior como sigue:

$$L_2(h) = \prod_{i=1}^{n} \hat{f}_n^i(X_i; h) \tag{8}$$

donde $\hat{f}_n^i(X_i; h) = \dfrac{1}{n-1} \sum_{j \pm i} \delta_h(X_i, X_j)$ que es la estimación de $f(X_i)$ utilizando la muestra de la que hemos eliminado el dato X_i.

Bowman (1984) muestra que este criterio equivale a minimizar la función:

$$L_3(h) = \frac{1}{n} \sum_{i=1}^{n} \log \left(f(X_i) \Big/ \hat{f}_n^i(X_i; h) \right) \tag{9}$$

que puede interpretarse como la distancia de Kullback-Leibler entre $\hat{f}_n$ y f.

Este criterio llamado de "validación-cruzada de Kullback-Leibler" en un contexto de datos independientes ha sido estudiado por Chow-Geman-Wu (1983) y Marron (1985) que propone una modificación que elimina algunos problemas que se presentan en la formulación inicial.

Una segunda técnica de validación cruzada es la introducida por Rudemo (1982) y Bowman (1984) que se basan en utilizar como criterio de error el ECI, en su desarrollo se observa

$$\text{ECI} = \int \left(\hat{f}_h(x) - f(x) \right)^2 \omega(x)\, d(x) = \int \hat{f}_n^2 \omega - 2 \int f \hat{f}_h \omega + \int f^2 \omega$$

que el último sumando no depende de h, por lo que se elegiría el h que minimice una estimación de la diferencia entre los dos primeros que viene dada en la siguiente expresión:

$$\text{(10)} \qquad \mathrm{CV_{MC}}(h) = \int \hat{f}_h^2(x)\omega(x)d(x) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_n^i(X_i)\omega(X_i)$$

Este criterio llamado de "validación-cruzada de mínimos cuadrados" ha sido estudiado entre otros por Hall (1983) para datos independientes y por Hart-Vieu (1989) para datos dependientes.

Existen otras técnicas para obtener el parámetro de suavización a partir de la muestra, de las que destacamos los métodos "plug-in" que consisten en minimizar el AMICE calculando previamente una estimación del término $\left(\int (f'')^2\right)$ que a su vez depende de una banda pero cuya influencia en la estimación es menor. La técnica de "validación cruzada sesgada" propuesta por Scott y Terrell (1987) que es una mezcla de la técnica de validación cruzada mínimo cuadrática y el "plug-in". Y una de las más recientes es la "validación-cruzada particionada", Marron (1987), que consiste en particionar la muestra, calcular por validación cruzada la banda en cada uno de los elementos de la partición y tomar como parámetro de suavización para la muestra la media de estas bandas reescalada.

2. DEFINICIONES

Si los datos muestrales son dependientes las técnicas de validación cruzada expuestas en el apartado 1 no son buenas, ya que si los datos están altamente correlacionados positivamente, al estimar $f(X_i)$ por $\hat{f}_n(X_i, h)$, los datos $X_{i-l}, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{i+l}$, proporcionan información sobre X_i que viene dada por la dependencia de la muestra y no por su estructura probabilística, lo que lleva a tomar bandas pequeñas y por tanto estimaciones poco suavizadas. Si los datos están altamente correlacionados negativamente el razonamiento es el inverso y se obtienen estimaciones sobresuavizadas.

Siguiendo las ideas de Hart-Vieu (1989) se modifican las técnicas de validación cruzada para eliminar el problema anterior en la siguiente forma:

1. **Validación cruzada de Kullback-Leibler.**

   Se elige el h que haga máxima la siguiente expresión:

$$\text{(11)} \qquad \mathrm{KL}(h) = \prod_{j=1}^{n} \left(\hat{f}_j(X_j)^{u(X_j)}\, e^{-p(h)}\right)$$

siendo:

- $\hat{f}_j^+(x) = \max\left(\hat{f}_j(x), 0\right)$, $\hat{f}_j(x) = \frac{1}{\gamma}\sum_{i=1}^{n}\delta_h(x, X_i)\gamma(i-j)$ con $\gamma(u)$ una función que verifica $\gamma(0) = 0$, $\gamma(u) = 1$ si $u > l_n$ (l_n una sucesión de enteros positivos) $0 \leq \gamma(u) \leq 1$ si $u \leq l_n$ $\overline{\gamma} = \sum_i \gamma(i-j)$

- $u(x)$ es una función de pesos no negativa, con soporte en un conjunto en el que f está acotada sobre cero. (Por ejemplo, la función indicadora de dicho conjunto)

- $p(h) = \int \hat{f}_h(x)\, u(x)\, d(x)$

El papel de la función $\gamma(u)$ es similar al de la $\delta(x, u)$. Mientras ésta se usa para clasificar los datos en función de su cercanía en el espacio, γ los clasifica en función de su cercanía en el tiempo. La sucesión $\{l_n\}$ determina la distancia (temporal) a partir de la cual dos datos pueden ser tratados como si fueran independientes.

2. **Validación Cruzada de Mínimos Cuadrados.**

Se elige el h que minimice la siguiente expresión:

$$\text{MC}(h) = \int \hat{f}_h(x)^2 \omega(x) d(x) - 2n^{-1}\sum_{j=1}^{n}\hat{f}_j(X_j)\omega(X_j) \tag{12}$$

donde $\omega(x)$ es una función de pesos no negativa.

A efectos de comparación entre los dos métodos se establece la siguiente relación entre las dos funciones de pesos $u(x)$ y$\omega(x)$ (Ver Marron, 1987):

$$u(x) = \omega(x)\, f(x) \tag{13}$$

3. RESULTADOS

En este apartado se probará que elegir el parámetro h por cualquiera de las dos técnicas expuestas en el apartado 2, en un contexto de dependencia de la muestra, es asintóticamente equivalente a minimizar cualquiera de las tres medidas de error cometido al aproximar la función de densidad f teórica por el estimador $\hat{f}_h$: ECP, ECI y MECI, definidas en el apartado 1.

Estas tres medidas son asintóticamente equivalentes en el siguiente sentido:

$$\limsup_{n\to\infty} \sup_{h\in H_n'} \left| \frac{\mathrm{ECP}(h) - \mathrm{MECI}(h)}{\mathrm{MECI}(h)} \right| = 0 \quad \text{casi seguro} \tag{14}$$

$$\limsup_{n\to\infty} \sup_{h\in H_n'} \left| \frac{\mathrm{ECI}(h) - \mathrm{MECI}(h)}{\mathrm{MECI}(h)} \right| = 0 \quad \text{casi seguro} \tag{15}$$

siendo H_n' un conjunto finito de posibles parámetros h.

Los resultados (14) y (15) han sido probados por Hardle-Marron (1986) para datos independientes. En un contexto de datos debilmente dependientes, Vieu (1989) los ha demostrado trabajando con el estimador núcleo. Y su generalización cuando se trabaja con estimadores basados en funciones δ, se obtiene de forma análoga, bajo algunas de las hipótesis que detallamos a continuación y que serán utilizadas para la obtención de los resultados.

Sobre la estructura de dependencia: se supondrá que los datos verifican la condición de dependencia "α-mixing" (fuertemente mixing) introducida por Rosenblatt (1956) y cuya definición es la siguiente:

> "Sea $\{X_i : i \in \mathbf{Z}\}$ una sucesión de v.a. y sea F_i^{i+n} la σ-álgebra generada por las variables $X_i, X_{i+1}, \ldots, X_{i+n}$, se define $\alpha(n) = \sup \left\{ |P(A\cap B) - P(A)P(B)| : A \in F_{-\infty}^{i}, B \in F_{i+n}^{+\infty} \right\}$. Entonces se dice que $\{X_n\}$ es α-mixing si $\lim_{n\to\infty} \alpha(n) = 0$."

Esta condición es muy débil y la verifican muchos procesos Gaussianos o los procesos ARMA generados a partir de ruido continuo. (Para más detalles ver Bradley, 1986)

Hipótesis sobre la sucesión δ :

H1. $\sup_u \delta_h(x,u) = 0(h^{-1})$, y para todo n existe $\epsilon_n \in \mathbb{R}^+$ tal que si $|x-u| > C\epsilon_n$ $(C \in \mathbb{R}^+, C > 1)$ entonces $\delta_h(x,u) = 0$.

H2. $\delta_h(x,u) = \delta_h(0,u-x)$
$\delta_h(0,u) = \delta_h(0,-u)$

H3. $\delta_h(x,-)$ es una densidad con función característica absolutamente integrable.

H4. Para cada $k = 2,3,\ldots$ hay una constante C_k de modo que si $m = 2,\ldots,k$ se verifica $\int\cdots\int \delta_h(x_{i1},x_{j1})\ldots\delta_h(x_{ik},x_{jk})\,dx_1\ldots dx_m| \leq C_k h^{-k+m/2}$ donde $i_1,j_1,\ldots i_k,j_k$ varían de 1 hasta m y sujetos a que $i_1 \pm j_1 \ldots i_k \pm j_k$, y cada valor de 1 hasta m aparece al menos dos veces en la lista $i_1,j_1,\ldots,i_k,j_k$.

Hipótesis sobre la densidad f:

H5. f tiene k derivadas continuas, $k \geq 1$

H6. $\max\{f(x),\, f(-x)\} \to 0$ si $x \to \infty$.

H7. Existe $M_1 \in (0,+\infty)$ tal que $f(x) \leq M_1$.

H8. Para cualquier j, la variable bidimensional (X_j, X_{j+1}) posee una densidad f_j respecto a la medida de Lebesgue.

Otras hipótesis son:

H9. El parámetro h se elegirá dentro de un conjunto finito $H_n' \subset H_n = [An^{-a'}; bn^{-b'}]$, con $0 < b' \leq \frac{1}{2k+1} \leq a' < \frac{1}{1+4k}$ $A,B > 0$ y $\#(H_n') \leq Cn^\rho$ $C,\rho > 0$.

H10. Sobre la sucesión $\{l_n\}$ y los coeficientes α-mixing supondremos: $l_n \leq l_n^* = n^{r_1}$ para algún $0 < r_1 < \frac{2-a'(1+4k)}{2}$ y $\sup_{j \geq l_n^*} \alpha(j) = 0(n^{-r_2})$ con

$$\begin{aligned} r_2 &= U + V + (2a' + 4ka')(2 + U/V) \\ U &= 1 + 2a' + 2ka' - b' \\ V &= 2 - a'(1 + 4k) - 2r_1 \end{aligned}$$

H11. La función ω está acotada y tiene soporte compacto.

H12. Si denotamos por $B(x) = \int \delta_h(x, y) f(y)\, dy - f(x)$ el sesgo cuadrático del estimador, se supondrá que $\int B^2(x)\omega(x)dx = 0(h^{2k})$.

H13. $\int \text{var}(\delta_h(x, X_i))\omega(x)dx \geq C'h^{-1}$.

H14. $f(x) \geq C' > 0$, para $x \in$ Soporte (ω).

H15. $\sup_{j,h,x} |\hat{f}_{j,h}(x) - f(x)| \to 0$ casi seguro, con $j = 1, \ldots, n$ $h \in H_n'$, $x \in$ Soporte (ω).

Comentarios a las Hipótesis:

- La hipótesis 4, de enunciado complejo, la verifican los estimadores más utilizados, descritos en el apartado 1, como han probado Hardle-Marron (1986).

- En la hipótesis 9, haciendo $k = 2$, se puede elegir $a' = b' = 1/5$, con lo que $h = Cn^{-1/5}$, que es la forma del parámetro de suavización que minimiza el MECI.

- Las condiciones impuestas en la hipótesis 10 son muy técnicas, pero las cumplen los coeficientes α-mixing de tipo exponencial o geométrico, es decir, $\alpha(k) = a\rho^k$ $(0 < \rho < 1)$ ó $\alpha(k) = Ck^{-\gamma}$.
  Por otra parte, la acotación $l_n \leq n^{r_1}$ (donde r_1 va a ser menor que uno), resulta lógica ya que la dependencia entre los datos va disminuyendo según la distancia temporal en que van siendo observados y por tanto no debe de suprimirse un número grande de ellos. Téngase en cuenta que es posible elegir $l_n = 0$, que daría lugar a las técnicas de validación cruzada ordinarias, utilizadas en un contexto de datos independientes.

- Las hipótesis 14 y 15 son necesarias únicamente en la validación cruzada de Kullback-Leibler. La hipótesis 15 se establece así porque las condiciones para que se dé dicha convergencia varían según el tipo de estimador que se esté usando. (Ver Vilar Fernández, 1989).

Teorema

Bajo las hipótesis anteriores si $\hat{h}$ es el parámetro que maximiza $\mathrm{KL}(h)$, definido en (11), o minimiza $\mathrm{MC}(h)$, definido en (12) se tiene que:

$$\lim_{n\to\infty} \frac{\mathrm{ERR}(\hat{h})}{\inf_{h\in H_n'} \mathrm{ERR}(h)} \longrightarrow 1 \qquad \text{casi seguro}$$

donde ERR = ECP, ECI o MECI.

4. SIMULACIONES

Con el fin de observar la conducta de los algoritmos de validación cruzada definidos en el apartado 2, se ha realizado el siguiente estudio de simulación: Se han generado 50 muestras de 100 datos de un proceso autorregresivo de orden 1:

$$X_t = \rho X_{t-1} + e_t \qquad \text{con } e_t \in \mathrm{N}(0,1),$$

siendo $\rho = 0, 0.3, 0.6, 0.9$ y se ha estudiado la estimación no paramétrica de la función de densidad $f(x)$ asociada al proceso (una Normal de media cero y varianza $\sigma^2 = 1/(1-\rho^2)$) en el intervalo $[-\sigma, +\sigma]$, haciendo el estudio en los puntos $X = -\sigma + 0.02k\sigma$, con k variando de 0 a 100.

Para cada una de las muestras se ha calculado:

- la banda teórica h_T (que minimiza el AMECI) es un valor orientativo, ya que para su cálculo no se tiene en cuenta el término de la varianza del estimador dado por la dependencia de la muestra, que asintóticamente es de orden inferior al valor del AMECI, pero para un tamaño muestral concreto tiene un valor positivo relativamente alto cuando la dependencia es fuerte. En este supuesto, la banda óptima será, por tanto, mayor que h_{T}.
- la banda que minimiza el ECP h_{E}.
- la banda que se obtiene por validación cruzada de Kullback-Leibler, definida en (11) h_{KL}, eligiendo el intervalo $[a,b] = [-\sigma, +\sigma]$.

- la banda que se obtiene por validación cruzada de mínimos cuadrados, definida en (12) h_{MC}, eligiendo como función de peso $\omega(x) = 1_{[-\sigma,+\sigma]}(x)$.

y para cada una de ellas se ha calculado el ECP asociado.

En todos los casos se ha utilizado la estimación no paramétrica núcleo, con

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{si } u \in [-1,1] \\ 0 & \text{en otro caso} \end{cases}$$

$$\text{y} \quad \gamma(u) = \begin{cases} 1 & \text{si } |u| > l_n \\ 0 & \text{si } |u| \leq l_n \end{cases}$$

En las tablas 1–4 se exponen los resultados obtenidos. Para cada supuesto se ha calculado la media sobre las cincuenta muestras de los parámetros h_{KL} y h_{MC}, la media de sus ECP, y la desviación típica correspondiente de cada una de las bandas.

Tabla 1

$\rho = 0$ (datos independientes)
$h_{\mathrm{T}} = 0.9331$
$h_{\mathrm{E}} = 0.8906 \quad \sigma(h_{\mathrm{E}}) = 0.1683 \quad \mathrm{ECP}(h_{\mathrm{E}}) = 0.0010$

l_n	E(h_{KL})	σ	ECP	E(h_{MC})	σ	ECP
0	0.904	0.330	0.0022	0.866	0.360	0.0035
1	0.810	0.333	0.0023	0.802	0.344	0.0025
2	0.855	-0.318	0.0021	0.842	0.354	0.0024
3	0.806	0.326	0.0022	0.800	0.339	0.0023
4	0.859	0.338	0.0020	0.825	0.347	0.0023
5	0.810	0.361	0.0025	0.791	0.361	0.0026
6	0.789	0.351	0.0024	0.777	0.346	0.0026
7	0.796	0.345	0.0023	0.752	0.349	0.0027
8	0.792	0.363	0.0024	0.752	0.364	0.0028
9	0.779	0.342	0.0022	0.755	0.349	0.0024
10	0.796	0.355	0.0022	0.731	0.362	0.0027

Tabla 2

$\rho = 0.3$
$h_{\mathrm{T}} = 0.9782$
$h_{\mathrm{E}} = 0.9690 \quad \sigma(h_{\mathrm{E}}) = 0.1870 \quad \mathrm{ECP}(h_{\mathrm{E}}) = 0.00084$

l_n	$\mathrm{E}(h_{\mathrm{KL}})$	σ	ECP	$\mathrm{E}(h_{\mathrm{MC}})$	σ	ECP
0	0.882	0.345	0.0025	0.854	0.351	0.0027
1	0.902	0.366	0.0027	0.907	0.348	0.0025
2	0.903	0.376	0.0026	0.892	0.374	0.0028
3	0.892	0.376	0.0026	0.862	0.359	0.0028
4	0.895	0.389	0.0027	0.881	0.386	0.0028
5	0.896	0.408	0.0030	0.870	0.400	0.0030
6	0.862	0.431	0.0032	0.872	0.414	0.0031
7	0.896	0.427	0.0030	0.838	0.414	0.0033
8	0.876	0.407	0.0029	0.824	0.409	0.0033
9	0.893	0.419	0.0028	0.822	0.407	0.0033
10	0.915	0.413	0.0028	0.823	0.387	0.0031

Tabla 3

$\rho = 0.6$
$h_{\mathrm{T}} = 1.1664$
$h_{\mathrm{E}} = 1.1778 \quad \sigma(h_{\mathrm{E}}) = 0.3260 \quad \mathrm{ECP}(h_{\mathrm{E}}) = 0.0013$

l_n	$\mathrm{E}(h_{\mathrm{KL}})$	σ	ECP	$\mathrm{E}(h_{\mathrm{MC}})$	σ	ECP
0	1.011	0.446	0.0025	1.029	0.441	0.0026
1	1.066	0.486	0.0025	1.142	0.466	0.0024
2	1.151	0.495	0.0025	1.144	0.497	0.0026
3	1.160	0.508	0.0025	1.192	0.496	0.0025
4	1.195	0.509	0.0025	1.184	0.487	0.0025
5	1.273	0.489	0.0023	1.223	0.498	0.0025
6	1.293	0.482	0.0022	1.221	0.502	0.0025
7	1.221	0.497	0.0024	1.202	0.516	0.0026
8	1.243	0.514	0.0024	1.203	0.513	0.0026
9	1.275	0.521	0.0024	1.199	0.526	0.0026
10	1.331	0.475	0.0026	1.212	0.506	0.0026

Tabla 4

$\rho = 0.9$
$h_{\mathrm{T}} = 2.1407$
$h_{\mathrm{E}} = 3.1736 \quad \sigma(h_{\mathrm{E}}) = 0.6482 \quad \mathrm{ECP}(h_{\mathrm{E}}) = 0.0008$

l_n	$\mathrm{E}(h_{\mathrm{KL}})$	σ	ECP	$\mathrm{E}(h_{\mathrm{MC}})$	σ	ECP
0	3.413	0.789	0.0014	1.716	0.710	0.0028
1	3.604	0.466	0.0011	2.239	0.943	0.0024
2	3.608	0.464	0.0011	2.486	0.936	0.0022
3	3.620	0.443	0.0011	2.722	0.852	0.0020
4	3.617	0.458	0.0011	2.709	0.919	0.0021
5	3.624	0.432	0.0011	2.751	0.886	0.0018
6	3.631	0.410	0.0010	2.739	0.934	0.0019
7	3.614	0.541	0.0011	2.738	0.991	0.0020
8	3.617	0.541	0.0011	2.773	0.961	0.0020
9	3.616	0.544	0.0011	2.800	0.997	0.0020
10	3.616	0.544	0.0011	2.824	0.978	0.0019

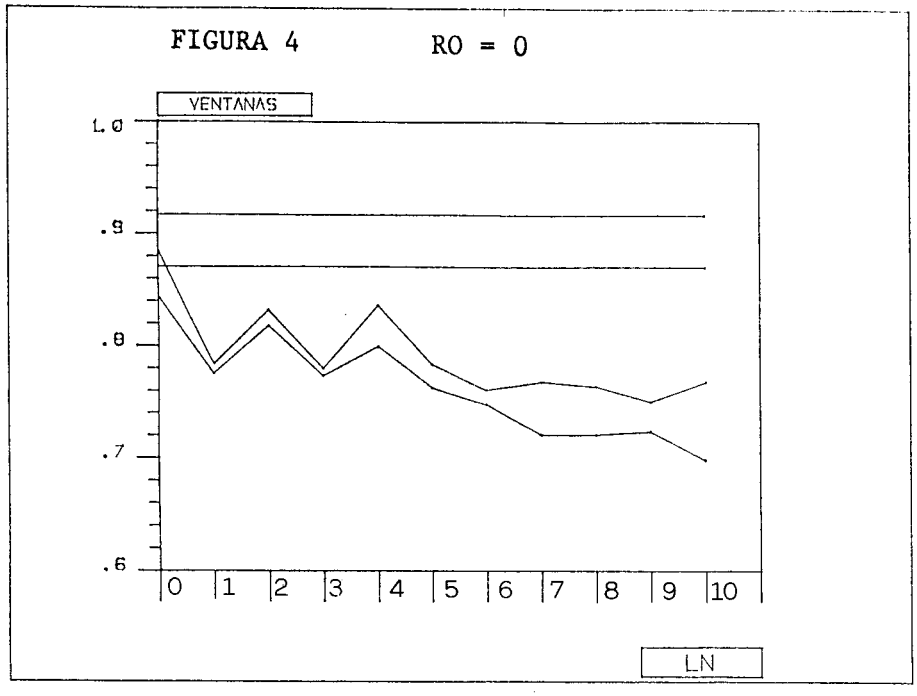

Figura 4.

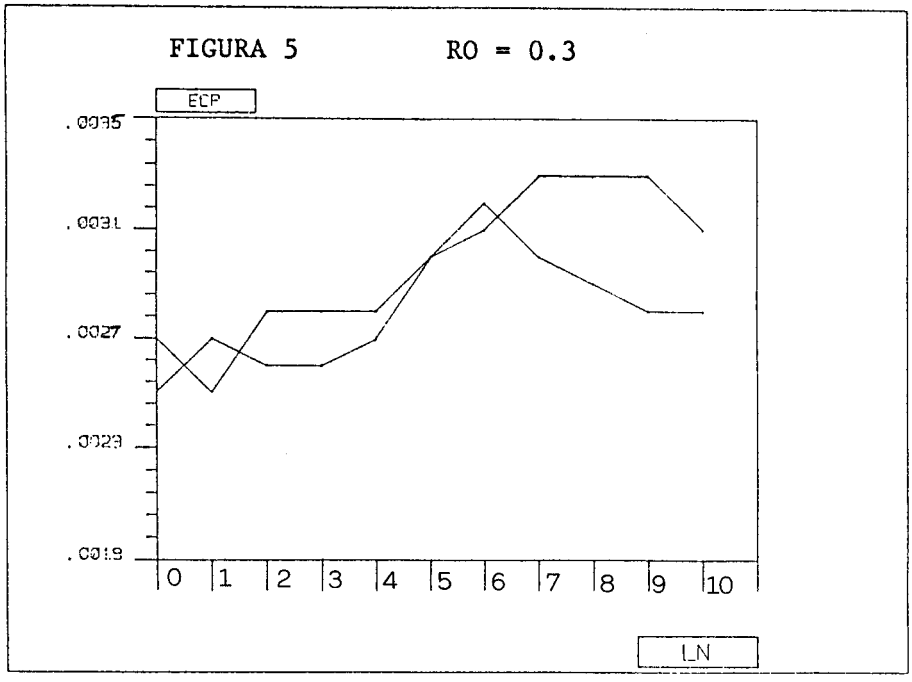

Figura 5.

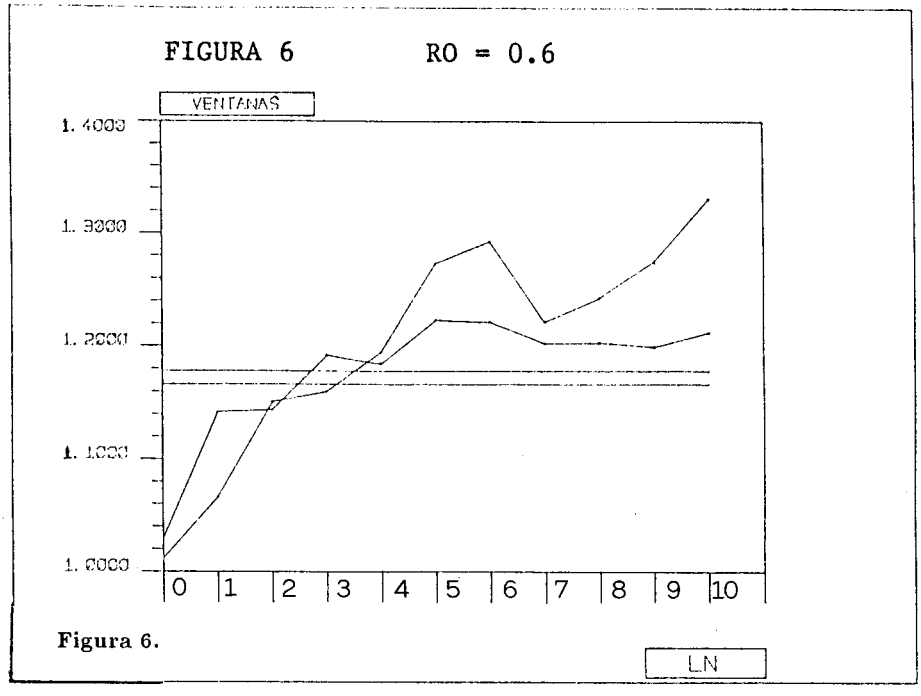

Figura 6.

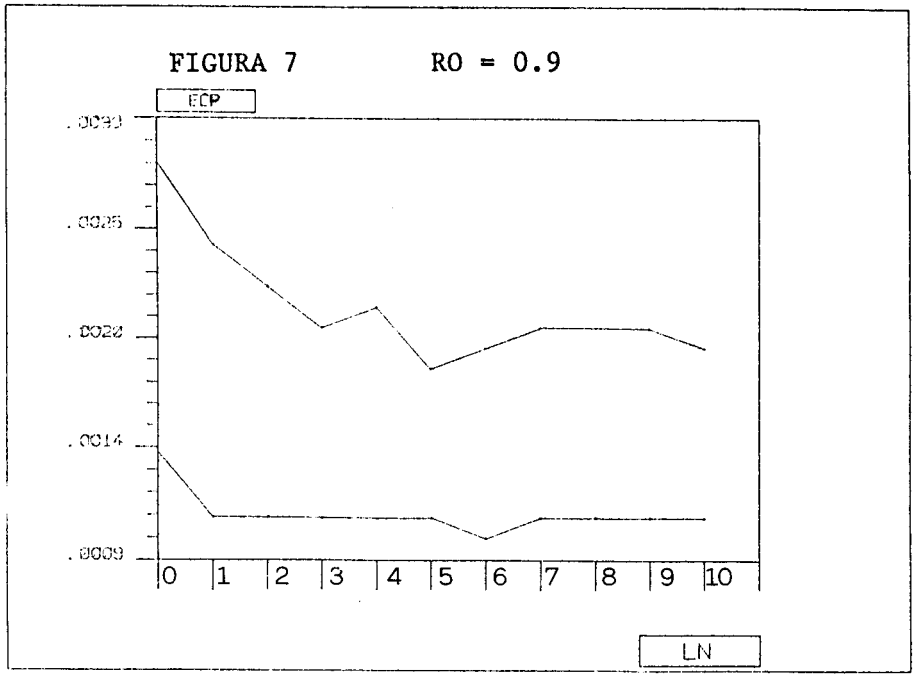

Figura 7.

Comentarios:

1) Respecto a la elección del parámetro l_n, se observa que en el supuesto de independencia ($\rho = 0$) debe tomarse $l_n = 0$, que proporciona las bandas más próximas a las teóricas. Al aumentar la dependencia ($\rho = 0.3/0.6$) se obtienen mejores resultados eligiendo $l_n = 2$ ó 3 en el primer caso y 3 en el segundo, ya que las bandas obtenidas son más próximas a las teóricas aunque la mejora en el ECP es pequeña. En el supuesto de fuerte dependencia ($\rho = 0.9$) es donde se aprecia claramente que eligiendo un l_n alto (5 ó 6) disminuye el ECP y aunque las bandas obtenidas no son próximas a h_T, este valor es menor que el de la banda óptima al no tener en cuenta en su cálculo el efecto de la dependencia, como se comentó anteriormente.
Será interesante estudiar métodos en los que no se tenga que elegir explícitamente la sucesión l_n, sino que ésta venga dada a partir de los datos. En este sentido proponemos trabajar con funciones del tipo

$$\gamma(u) = 1 - |r(u)| \qquad \text{si } r(u) > 2/\sqrt{n}$$

2) Comparando los dos métodos de validación cruzada se observa que proporcionan resultados parecidos. En el caso de fuerte dependencia ($\rho = 0.9$), la técnica de Kullback-Leibler es mejor, pero con valores de ρ menores las diferencias son mínimas, incluso si se tiene en cuenta la dispersión de las bandas obtenidas, relativamente alta con ambas técnicas.

3) Hemos realizado simulaciones con este y otros modelos de procesos MA y ARMA, también con muestras de tamaño 100 y con muestras de tamaño 25 y 50, obteniéndose resultados parecidos. Trabajando con otras funciones de densidad, como una bimodal obtenida como mezcla de dos normales, las técnicas descritas proporcionan buenos resultados; sin embargo, al trabajar con procesos estacionarios con densidad exponencial los resultados son bastante peores que los descritos en el caso gaussiano.

5. APÉNDICE

Demostración del Teorema

Consideremos

$$\mathrm{KL}(h) = \prod_{j=1}^{n} \left(\hat{f}_j^+(X_j)^{u(X_j)}\, e^{-p(h)} \right) \tag{16}$$

Elegir h que maximice $\mathrm{KL}(h)$ es lo mismo que elegir el h que maximice la expresión $(1/n)\log \mathrm{KL}(h) + R$, donde R, que no depende de h, viene dado por la expresión

$$R = \int f(x)u(x)dx - \frac{1}{n}\sum_{i=1}^{n} (\log f(X_j))\; u(X_j) \tag{17}$$

Se define

$$\Delta_j = \frac{\hat{f}_j(X_j) - f(X_j)}{f(X_j)} \quad \text{y} \quad \Delta_j^+ = \frac{\hat{f}_j^+(X_j) - f(X_j)}{f(X_j)} \tag{18}$$

y para $n = 1, 2, \ldots$ el suceso:

$$U_n = \left\{ \Delta_j \, 1_{\mathrm{sop}(\omega)}(X_j) = \Delta_j^+ \, 1_{\mathrm{sop}(\omega)}(X_j),\; h \in H_n',\; j = 1, \ldots, n \right\}$$

De las hipótesis 14 y 15 se sigue que:

$$\sup_{j,h} \mid \Delta_j^+ \mid \leq \sup_{j,h} \mid \Delta_j \mid \rightarrow 0 \quad \text{casi seguro}$$

$$\text{y} \quad \lim_n P(U_n) = 1$$

Así, para cada $h \in H_n'$, y sobre el suceso U_n se obtiene:

$$\frac{1}{n} \log\left(\mathrm{KL}(h)\right) + R = \tfrac{1}{n} \sum_{j=1}^n \left(u(X_j) \log\left(\hat{f}_j(X_j)\right) - p(h)\right) + I -$$

$$(19) \qquad -n^{-1} \sum_{j=1}^n u(X_j) \log(f(X_j)) = n^{-1} \sum_{j=1}^n (u(X_j) \log(1+\Delta_j) - p(h) + I)$$

Siendo $I = \int f(x)u(x)dx$

Haciendo un desarrollo de Taylor en la función logaritmo, (19) toma la siguiente expresión:

$$(20) \qquad n^{-1} \sum_{j=1}^n (u(X_j)\ \Delta_j - p(h) + I) - \frac{1}{2}\, \widehat{\mathrm{ECP}}(h) + \frac{1}{n} \sum_{j=1}^n r_j u(X_j)$$

siendo

$$(21) \qquad \widehat{\mathrm{ASE}}(h) = n^{-1} \sum_{j=1}^n \left(\hat{f}_j(X_j) - f(X_j)\right)^2 f(X_j)^{-2}\, u(X_j)$$

y r_j es el resto en el desarrollo de Taylor del logaritmo.

El primer y el tercer término de (20) son despreciables en virtud de los siguientes lemas:

Lema 1.

Bajo las hipótesis del teorema se verifica

$$(22) \qquad \sup_{h \in H'_n} \frac{\left| n^{-1} \sum_{j=1}^{n} (u(X_j)\ \triangle_j - p(h) + I) \right|}{\mathrm{MECI}(h)} \to 0 \quad \text{casi seguro}$$

Lema 2.

Bajo las hipótesis del teorema se verifica

$$(23) \qquad \sup_{h \in H'_n} \frac{\left| n^{-1} \sum_{j=1}^{n} (u(X_j)\, r_j) \right|}{\mathrm{MECI}(h)} \to 0 \quad \text{casi seguro}$$

Como ya se comentó en (14) y (15) las medidas del error ECP, ECI y MECI son asintóticamente equivalentes, y por el siguiente lema el $\widehat{\mathrm{ECP}}$ también se puede incluir dentro de estas equivalencias:

Lema 3.

Bajo las hipótesis del teorema se verifica:

$$(24) \qquad \sup_{h \in H'_n} \frac{\left| \widehat{\mathrm{ECP}}(h) - \mathrm{MECI}(h) \right|}{\mathrm{MECI}(h)} \to 0 \quad \text{casi seguro}$$

De la expresión (20) se sigue que maximizar $\mathrm{KL}(h)$ es equivalente a minimizar el error $\widehat{\mathrm{ECP}}(h)$ y por las equivalencias asintóticas a minimizar cualquiera de los errores definidos anteriormente.

Por otra parte la función $\mathrm{MC}(h)$ puede escribirse como sigue:

(25)

$$\mathrm{MC}(h) = \mathrm{ECI}(h) - \frac{2}{n} \sum_{j=1}^{n} \hat{f}_j(X_j)\omega(X_j) + 2 \int \hat{f}_h(x) f(x) \omega(x) dx - \int f(x)^2 \omega(x) dx =$$

$$= \mathrm{ECI}(h) - \int f(x)^2 \omega(x) dx + 2 \left(\int \hat{f}_h(x)\omega(x)dx - n^{-1} \sum_{j=1}^{n} \hat{f}_j(X_j)\omega(X_j) \right)$$

El segundo sumando no depende de h, veamos que el tercero es despreciable en el siguiente sentido:

$$\limsup_{n\to\infty} \sup_{h\in H'_n} \frac{\left|\int \hat{f}_h(x)f(x)\omega(x)dx - n^{-1}\sum_{j=1}^{n}\hat{f}_j(X_j)\omega(X_j)\right|}{\mathrm{MECI}(h)} \longrightarrow 0 \quad \text{casi seguro}$$

(26)

o lo que es equivalente

$$\sup_{h\in H'_n} \frac{\left|\int \hat{f}_h(x)f(x)\omega(x)dx - n^{-1}\sum_{j=1}^{n}\hat{f}_j(X_j)\omega(X_j) + n^{-1}\sum_{j=1}^{n}f(X_j)\omega(X_j) - \right.}{\mathrm{MECI}(h)}$$

(27)

$$\left. - \int f(x)^2\omega(x)dx\right| \longrightarrow 0$$

Los dos últimos sumandos del numerador de la expresión (27) no dependen de h y desarrollando el numerador de la expresión del Lema 1 se obtiene la misma expresión que en (27), por tanto hemos obtenido que, salvo una constante, minimizar la función $\mathrm{MC}(h)$ es equivalente a minimizar el ECI.

Además se observa que obtener el h que maximiza $\mathrm{KL}(h)$ es equivalente a obtener el h que minimiza $\mathrm{MC}(h)$.

A partir de la expresión (20), utilizando los lemas 1, 2, 3 se obtiene fácilmente la relación:

$$\sup_{h,h'\in H_n} \frac{|\mathrm{ERR}(h) - \mathrm{ERR}(h') - (\mathrm{KL}(h') - \mathrm{KL}(h))|}{\mathrm{ERR}(h) + \mathrm{ERR}(h')} \longrightarrow 0 \quad \text{casi seguro}$$

y de aquí se deduce que si, $\hat{h}$ maximiza $\mathrm{KL}(h)$ entonces

$$\frac{\mathrm{ERR}(\hat{h})}{\inf_{h\in H'_n}\mathrm{ERR}(h)} \longrightarrow 1 \quad \text{casi seguro}$$

■

Demostración del Lema 1.

Puede verse en Hart-Vieu (1988) que lo desarrollan para el caso particular de estimadores núcleo. Ver también Gyorfi y otros (1989). La demostración se basa en la utilización de teoremas sobre la acotación de momentos de sumas de variables aleatorias α-mixing.

Demostración del Lema 2.

Al ser r_j el desarrollo en serie de Taylor del logaritmo, es de la forma:

$$(28) \quad r_j = \frac{2}{\xi^3}\left(\hat{f}_j(X_j) - f(X_j)\right)^3 f(X_j)^{-3} - \frac{2}{\xi^3}\Delta_j^3, \text{ con } \xi \in (1, 1+\Delta_j)$$

entonces

$$(29) \qquad \frac{\left|n^{-1}\sum_{j=1}^n r_j u(X_j)\right|}{\text{MECI}(h)} = \frac{\left|n^{-1}\sum_{j=1}^n \Delta_j^3 \frac{2}{\xi^3} u(X_j)\right|}{\text{MECI}(h)}$$

y para demostrar que esta expresión tiende a cero, llega con probar que

$$(30) \qquad \frac{\left|n^{-1}\sum_{j=1}^n \Delta_j^3 \frac{2}{\xi^3} u(X_j)\right|}{\widehat{\text{ECP}}(h)} \longrightarrow 0$$

en virtud de la equivalencia asintótica de los errores (Lema 3).

Como $\widehat{\text{ECP}}(h) = \frac{1}{n}\sum_{j=1}^n \Delta_j^2 u(X_j)$ y de las hipótesis H14 y H15 se sigue que $\Delta_j \to 0$ y por tanto, se verifica (30).

■

Demostración del Lema 3.

Se tiene la siguiente relación:

$$\hat{f}_j(x) - \hat{f}_h(x) = \frac{n-\overline{\gamma}}{n\overline{\gamma}} \sum_{|i-j|>l_n} \delta_h(x, X_i) + \frac{1}{\overline{\gamma}} \sum_{|i-j|\leq l_n} \delta_h(x, X_i)\gamma(i-j) -$$

$$- \frac{1}{n} \sum_{|i-j|>l_n} \delta_h(x, X_i) \tag{31}$$

De esta expresión y el hecho de que $\widehat{\mathrm{ECP}} = ECP$+A+B, siendo

$$A = n^{-1} \sum_{j=1}^{n} \left(\hat{f}_j(X_j) - \hat{f}_h(X_j)\right)^2 f(X_j)^{-2} u(X_j)$$

$$B = 2n^{-1} \sum_{j=1}^{n} \left(\hat{f}_j(X_j) - \hat{f}_h(X_j)\right) \left(\hat{f}_h(X_j) - f(X_j)\right) f(X_j)^{-2} u(X_j)$$

se sigue que:

$$\frac{\left|\widehat{\mathrm{ECP}} - \mathrm{MECI}\right|}{\mathrm{MECI}} \leq \frac{|\mathrm{ECP} - \mathrm{MECI}|}{\mathrm{MECI}} + \frac{|A|}{\mathrm{MECI}} + \frac{|B|}{\mathrm{MECI}}$$

el primer sumando de la derecha de la desigualdad anterior tiende a cero por (14). La demostración de que los dos sumandos restantes convergen a cero es muy compleja y técnica, se basa en utilizar el hecho de que $\overline{\gamma} = 0(n)$, la expresión (31), las hipótesis H1, H10–H13, H15, y la siguiente igualdad:

$$\mathrm{MECI}(h) = \int B_i^2(x)\omega(x)dx + \int n^{-1} \mathrm{var}\left(\delta_h(x, X_i)\right)\omega(x)dx.$$

■

6. BIBLIOGRAFÍA

[1] **Bowman, A.** (1984). "An alternative method of cross-validation for the smoothing of density estimates". *Biometrika,* 65, 521-528.

[2] **Bradley, R.** (1986). "Basic Properties of strong mixing conditions". Ed. Birkhäuser.

[3] **Chow-Geman-Wu** (1983). "Consistent cross-validated density estimation". *Annals of Statistics,* 11, 25-38.

[4] **Gyorfi-Hardle-Sarda-Vieu** (1989). "Nomparametric curve. Estimation from time series". *Lecture Notes in Statistics,* vol. 60.

[5] **Habbema-Hermans-Van der Broek** (1974). "A stepwise discriminant analysis program usig density estimation". In Compstat, Ed. G. Bruckmann, 101-10. Vienna: Phisica-Verlag.

[6] **Hall, P.** (1983). "Large sample optimality of least square cross-validation in density estimation". *Annals of Statistics,* 11, 1156-74.

[7] **Hardle-Marron** (1986). "Random aproximations to some measures of accuracy in nonparametric curve estimation". *Journal of Multivariate Analysis,* 20, 91-113.

[8] **Hart-Vieu** (1988). "Data-driven bandwidth choice for density estimation based on dependent data". Preprint.

[9] **Marron, J.S.** (1985). "An asymptotically efficient solution to the bandwidth problem of kernel density estimation". *Annals of Statistics,* 13, 1011-23.

[10] **Marron, J.S.** (1987). "A comparison of cross-validation techniques in density estimation". *Annals of Statistics,* 15, 152-162.

[11] **Marron, J.S.** (1987). "Partitioned cross-validation". Norht-Carolina Institute of Statistics, Mimeo series # 1721.

[12] **Parzen, E.** (1962). "On estimation of a probability density function and mode". *Ann. Math. Statist.,* 33, pp. 1065-76.

[13] **Rosenblatt** (1956). "A central limit theorem and strong mixing condition". *Proc. Nat. Acad. Sci.,* 42, 43-47.

[14] **Rudemo** (1982). "Empirical choice of histogramas and kernel density estimators". *Scand. J. Statis.,* 9, 65-78.

[15] **Scott-Terrell** (1987). "Biased and unbiased cross-validation in density estimation". JASA, 82, 1131-1146.

[16] **Silverman** (1986). Density estimation. Chapman and Hall.

[17] **Vieu** (1989). "Quadratic errors for nonparametric estimates under dependence". Preprint.

[18] **Vilar Fernández** (1989). "Estimación no paramétrica de curvas notables para datos dependientes". *Trabajos de Estadística,* v. 4, n. 2, 69-88.

[19] **Watson-Leadbetter** (1964). Hazard analysis I, *Biometrika,* 51, 175-184.