QÜESTIIÓ, Vol 16, 1,2,3 pp. 3-26, 1992

ESTUDIO COMPARATIVO DE DISTINTAS FUNCIONES NÚCLEO PARA LA OBTENCIÓN DEL MEJOR AJUSTE SEGÚN EL TIPO DE DATOS

J.E. MARTÍNEZ FALERO*, E. AYUGA TÉLLEZ†, C. GONZÁLEZ GARCÍA*

En este artículo se presenta una contribución a la selección de la función núcleo y del parámetro de alisado que mejor se adaptan a las características muestrales en muestras de tamaño pequeño (25). Para ello se obtuvieron 200 realizaciones muestrales procedentes de 5 distribuciones continuas, prácticamente todas ellas con soporte [0,1]; y se agruparon en función de sus características muestrales. En cada grupo de los obtenidos se ajustaron funciones de densidad correspondientes a 8 núcleos diferentes, con anchos de ventana variables entre 0'2 y 4'8, calculando posteriormente un ancho de ventana medio mejor para cada grupo. Este ancho de ventana se comparó con los anchos de ventana óptimos para cada realización muestral, obtenidos por minimización del error cuadrático medio integrado y por validación cruzada. El análisis del sesgo y la eficiencia de los valores del estadístico "ancho de ventana correspondiente al error óptimo medio por grupo menos ancho de ventana óptimo de cada muestra", y de la bondad del ajuste de las funciones estimadas a las distribuciones de partida, permite determinar la función núcleo y el ancho de ventana que mejor se adaptan a las características muestrales.

A comparative study of different kernel functions according to data type.

Keywords: Funciones núcleo, parámetro de alisado, agrupamiento.

---

*Profesor Titular de Universidad. Estadística e I.O.
†Profesor Titular de Escuela Universitaria. Matemática Aplicada.
-Departamento de Economía y Gestión de las Explotaciones e Industrias Forestales.
-E.T.S.I. de Montes. Universidad Politécnica de Madrid. Ciudad Universitaria s/n. 28040 — Madrid.
-Article rebut el juliol de 1991.
-Acceptat el febrer de 1992.

1. INTRODUCCIÓN

A partir de que Glivenko (1934) demostrara la convergencia casi segura del histograma de frecuencias a la densidad, los estudios de estimación no paramétrica de esta función han sido numerosos, adquiriendo especial relevancia durante los años 80, paralelamente al desarrollo de la capacidad de procesamiento de los ordenadores, y como consecuencia de los numerosos cálculos que precisa cualquiera de los estimadores propuestos.

Entre los estimadores no paramétricos de las funciones de densidad, se pueden destacar: los estimadores basados en la definición de una función núcleo o "kernel" (Rosenblatt, 1956; Parzen, 1962; Nadaraya, 1989); los derivados de la "verosimilitud penalizada" (Scott, *et al.*, 1980; Silverman, 1986); las series de estimadores ortogonales (Cencov, 1962); el método "PPDE" (Friedman, *et al.*, 1984); y los métodos basados en los $k-$puntos más próximos (Loftsgaarden y Quesenberry, 1965).

Tanto estos trabajos, como la mayoría de las referencias disponibles se centran en la definición de distintos estimadores y en la comprobación de la convergencia teórica a la función de densidad; sin embargo, los estudios relativos a la especificidad de los distintos métodos en cuanto a precisión, adecuación a la estructura de los datos analizados y al tiempo de CPU requerido para la estimación son escasos. En este sentido se pueden señalar los trabajos de Epanechnikov (1969) y Deheuvels (1977) referidos a la precisión y los de Scott y Factor (1981), Silverman (1982) y Härdle (1991) en cuanto al tiempo de CPU.

De los diferentes procedimientos de optimización, los mejor estudiados matemáticamente (y aquellos para los que existe un mayor número de aplicaciones a datos reales), son los basados en la definición de una función núcleo. Para su empleo es necesario elegir tanto el "núcleo" como un valor del parámetro de alisado, ambos determinarán la expresión final de la función de densidad estimada.

El núcleo es una función $K(x)$, a partir de la cual se puede establecer el siguiente estimador no paramétrico de cualquier función de densidad $f(x)$ (Rosenblatt, 1956):

$$f_n(x, h_n) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

donde h_n es el parámetro de alisado y $X_1, \ldots, X_n$ los datos observados.

Nadaraya (1989) establece las siguientes condiciones para la función núcleo:

a) $K(x) = K(-x)$

b) $\int K(x)\,dx = 1$

c) $\sup_{-\infty<x<\infty} |K(x)| \leq A < \infty$

d) $\int x^i K(x)\,dx = 0, \quad i = \overline{1, s-1}$

con s par y mayor o igual que 2;

e) $\int x^s K(x)\,dx \neq 0$

f) $\int x^s |K(x)|\,dx < \infty$

Respecto a las propiedades estadísticas de estos estimadores, Parzen (1962) demostró que si el parámetro de alisado, h_n, tiende a cero cuando $n \rightarrow \infty$ entonces el estimador núcleo es asintóticamente insesgado y asintóticamente normal. Estudios más completos sobre la convergencia de este estimador son los realizados por Devroye (1983) y Devroye y Penrod (1984), que establecieron la convergencia de los estimadores núcleo en el espacio de funciones integrables y en el de funciones de cuadrado integrable.

El parámetro de alisado h_n, también llamado "ancho de ventana" o "ancho de banda", es un número positivo que cumple la condición de Parzen. Scott y Factor (1981) y Marron (1987) presentan recopilaciones de los métodos o algoritmos de cálculo más empleados. En general, el ancho de ventana se determina de forma que se minimice algún tipo de error, Hall y Marron (1988) proponen minimizar la integral del error cuadrático sobre el rango de variación de la variable aleatoria:

$$ECI(h_n) = \int [f_n(x, h_n) - f(x)]^2 \, dx$$

y Devroye y Gyorfi (1985) minimizar la integral de las diferencias en valor absoluto entre el estimador y la función:

$$ABSEMC(h_n) = \int |f_n(x, h_n) - f(x)|, dx$$

Sin embargo, el proceso de minimización de estas medidas es substancialmente más complejo (ver por ejemplo Izenman, 1991) que la optimización de la medida propuesta por Rosenblatt (1956), la cual se define como el error cuadrático medio

integrado (o el riesgo medio de la función de pérdida cuadrática entre la función y su estimador); y toma la siguiente expresión:

$$U(h_n) = \int E\left[f_n(x;h_n) - f(x)\right]^2 \, dx$$

En este artículo se realiza una comparación por simulación en ordenador, del ajuste de distintas funciones núcleo, con diferentes anchos de ventana, a muestras pequeñas (tamaño 25) caracterizadas por distintos parámetros muestrales.

La comparación de la eficiencia de algoritmos por simulación en ordenador presenta algunas dificultades: en primer lugar, deberán elegirse los algoritmos a comparar; en nuestro caso ocho funciones núcleo y dos valores del ancho de banda que pueden obtenerse sin conocer la distribución de los datos (uno derivado de la minimización del error cuadrático medio integrado y otro obtenido por validación cruzada). En segundo lugar deberá seleccionarse un conjunto representativo de problemas para la comparación de los algoritmos, para lo cual se simularon 200 realizaciones muestrales procedentes de cinco distribuciones contínuas, que se agruparon en 10 clases en función de sus características muestrales. Finalmente, ha de definirse la "bondad o calidad" de los algoritmos a comparar para elegir el mejor, en nuestro caso la función núcleo y el ancho de ventana que mejor se adapten a cada grupo muestral.

Para conseguir este objetivo, en cada una de las realizaciones muestrales se calculó una estimación de la función de densidad con ocho funciones núcleo diferentes, y con anchos de banda comprendidos entre 0'2 y 4'8. Posteriormente se calculó el ancho de ventana con menores desviaciones medias en cada grupo. Este ancho de ventana se comparó con los dos anchos de ventana calculados para cada realización muestral. El análisis del sesgo y la eficiencia de los valores del estadístico: "*ancho de ventana correspondiente al mínimo error medio por grupo de muestras* **menos** *ancho de ventana óptimo correspondiente a cada muestra*" y de la bondad del ajuste de las funciones estimadas a las distribuciones de partida, permite determinar la mejor función núcleo a partir de las características muestrales.

2. FUNCIONES NÚCLEO Y ANCHOS DE BANDA EMPLEADOS

Existen numerosas funciones que cumplen las propiedades necesarias para ser funciones núcleo. Para el presente estudio se han escogido ocho funciones $K(x)$ entre las más conocidas, y que permiten el cálculo del parámetro de alisado óptimo. Estas funciones son las siguientes

$$
\begin{aligned}
K_1 &= \begin{cases} \frac{4}{3} - 8x^2 + 8|x|^3, & |x| < \frac{1}{2} \\ \frac{8}{3}(1-|x|)^3, & \frac{1}{2} \leq |x| \leq 1 \\ 0, & |x| > 1 \end{cases} \\
K_2 &= \frac{1}{2}e^{-|x|} \\
K_3 &= \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3x^2}{20\sqrt{5}}, & |x| \leq \sqrt{5} \\ 0, & |x| > \sqrt{5} \end{cases} \\
K_4 &= \begin{cases} 0'54 + 0'46\cos \pi x, & |x| \leq 1 \\ 0, & |x| > 1 \end{cases} \\
K_5 &= \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \\
K_6 &= \frac{3}{2}\left(1 - \frac{x^2}{3}\right)\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
K_7 &= \frac{15}{8}\left(1 - \frac{2}{3}x^2 + \frac{1}{15}x^4\right)\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
K_8 &= \begin{cases} \frac{15}{16}(1-x^2)^2, & |x| \leq 1 \\ 0, & |x| > 1 \end{cases}
\end{aligned}
$$

La función Gaussiana (K_5) y la propuesta por Epanechnikov (1969) (K_3) han sido las más utilizadas en los estudios realizados con estimadores núcleo. La popularidad de la función Gaussiana se debe al conocimiento de sus propiedades y el frecuente empleo de la función de Epanechnnikov a la economía en cálculos que representa su uso.

Respecto a los parámetros de alisado se han considerado dos valores para la simulación: el que se denomina en el presente trabajo Optimo.-1, que es un estimador del ancho de ventana que proporciona un menor error medio integrado; y un segundo valor, que denominamos Optimo.-2, y que se obtiene por el método de validación cruzada máximo verosimil.

El error medio integrado para los estimadores núcleo de una función de densidad puede expresarse como:

$$
U(h_n) = \frac{1}{nh_n}\int K^2(x)dx + h_n^{2s}\frac{\left(\int x^s K(x)dx\right)^2}{(s!)^2}\int \left|f^{(s)}(x)\right|^2 dx + O\left[\frac{1}{nh_n} + h_n^{2s}\right]
$$

(Nadaraya 1989). Minimizando esta expresión se encuentra el óptimo asintótico del parámetro de alisado que depende de la densidad desconocida f, de la función núcleo elegida y de la muestra. El valor óptimo viene dado por la fórmula:

$$h_n = A(K)\, B(f)\, n^{-\frac{1}{(2s+1)}}$$

siendo:

$$A(K) = \left[\frac{\int K^2(x)dx}{2s\left(\frac{\int x^s K(x)dx}{s!} \right)^2} \right]^{\frac{1}{(2s+1)}}$$

$$B(f) = \left[\int \left| f^{(s)}(x) \right|^2 dx \right]^{-\frac{1}{(2s+1)}}$$

Esta valor no puede obtenerse directamente en el caso de estimar una densidad desconocida, ya que depende de la derivada de orden s de dicha función. Se hace preciso, por tanto, estimar el óptimo. Para ello se suele utilizar el algoritmo iterativo de Scott *et al.*, (1977) que utiliza $f_n^{(s)}$ (derivada s–ésima de la función estimada) como estimador asintóticamente insesgado (Nadaraya, 1989) de $f^{(s)}$, procediendo de la siguiente forma:

Primero se toma una estima inicial de h_n, por ejemplo el rango muestral (h_n^0).

Segundo, con este valor se estima la derivada de la función de densidad, $f_n^{(s)0}$.

Tercero, se calcula el valor h_n^1 con la expresión dada para el valor óptimo y empleando la estimación anterior.

Cuarto, se repiten los pasos anteriores de tal forma que para la iteración i–ésima tendremos:

$$h_n^{i+1} = A(K)\, B\left(f_n^i\right)\, n^{-\frac{1}{(2s+1)}},$$

lo que proporciona una secuencia de valores h_n^i que convergen a la solución del algoritmo, que tomaremos como estimación del óptimo.

El segundo valor del ancho de ventana se obtiene por el método de validación cruzada máximo verosimil (Duin, 1976 y Hermans y Habbema, 1976), procediendo de la siguiente forma:

Primero se obtiene el estimador núcleo de la función de densidad en los valores muestrales según la ecuación:

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

Segundo se determina el valor del parámetro que optimiza la función de pseudoverosimilitud:

$$L(h) = \prod_i \hat{f}_{h_n,i}(X_i)$$

El algoritmo de optimización de la función anterior ha sido el método adaptativo aplicado como procedimiento de optimización global bayesiana (ver apéndice).

3. SELECCIÓN DEL CONJUNTO DE PROBLEMAS PARA COMPARACIÓN

Una simulación extensiva por el método de Monte Carlo requiere encontrar algoritmos más rápidos que los actuales para el cálculo de la estimación. Con objeto de comparar los distintos núcleos se obtuvieron 200 realizaciones muestrales sobre el espacio de funciones continuas, con las siguientes características:

1º Se definieron cinco tipos de funciones de densidad que se consideraron representativas del conjunto de funciones continuas con soporte aproximado [0,1] (ver figura 1):

   a) Distribución uniforme.
   b) Distribución N(0'5, 0'2).
   c) Distribución N(0'5, 0'4).
   d) Una distribución asimétrica a la izquierda, $\beta(3,6)$.
   e) Una distribución bimodal, cuya función de densidad es 0'5 por la densidad de una N(0'25, 0'125) más la de la otra N(0'75, 0'125).

2º Se obtuvieron 200 realizaciones muestrales de tamaño 25 sobre las mencionadas distribuciones de la siguiente manera: primero se generaron 200 muestras de tamaño 25 de una distribución uniforme (0,1) con diferente RUN-TIME; a continuación se generaron otros 200 números aleatorios truncados a enteros de 1 a 5 que representan a las cinco distribuciones de partida; finalmente los datos uniformes se transformaron en las muestras correspondientes a las cinco distribuciones utilizando el GPSS (Chisman, 1992).

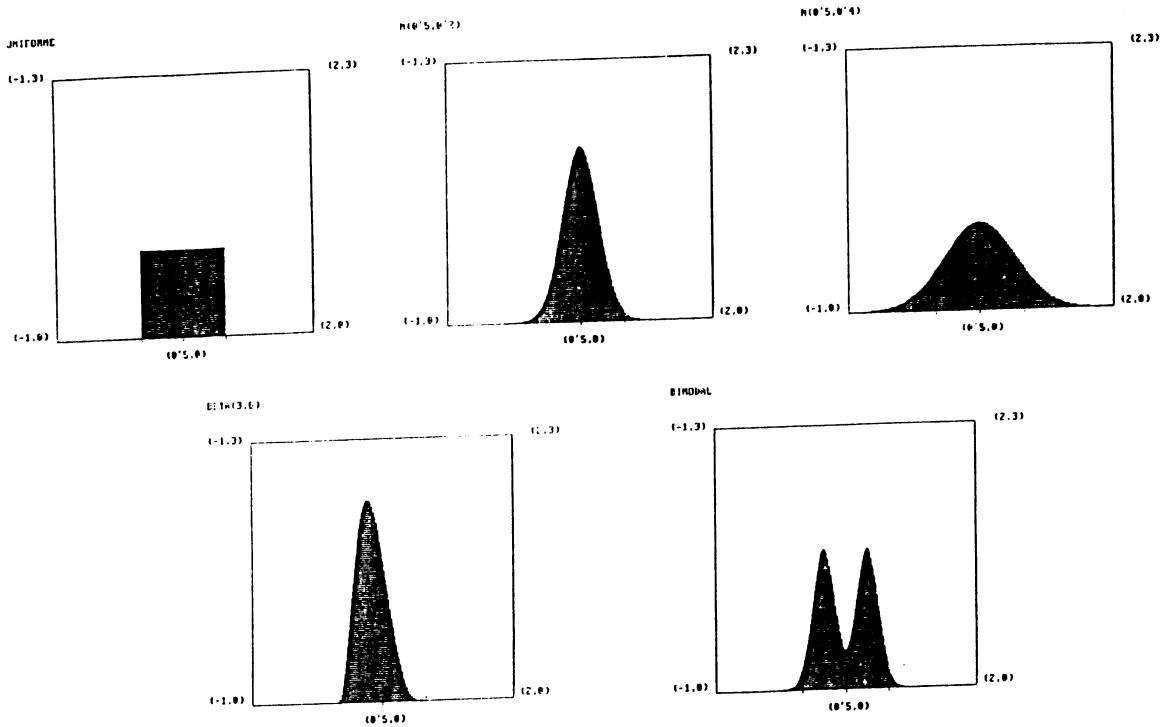
UNIFORME
(-1,3)
(2,3)
(-1,0)
(2,0)
(0'5,0)
N(0'5,0'2)
(-1,3)
(2,3)
(-1,0)
(2,0)
(0'5,0)
N(0'5,0'4)
(-1,3)
(2,3)
(-1,0)
(2,0)
(0'5,0)
BETA(3,0)
(-1,3)
(2,3)
(-1,0)
(2,0)
(0'5,0)
BIMODAL
(-1,3)
(2,3)
(-1,0)
(2,0)
(0'5,0)

Figura 1.

Para efectuar la comparación se ha procedido a agrupar los 200 conjuntos de datos, utilizando un método divisivo y politético (Hill, 1975), en función de sus características muestrales. Para ello se obtuvieron la varianza, el coeficiente de asimetría de Fisher, el coeficiente de apuntamiento y el número de modas de las 200 muestras. La finalidad de esta agrupación es determinar el comportamiento de las distintas funciones núcleo en grupos homogéneos por características muestrales, de forma que puedan establecerse pautas de ajuste a partir, simplemente, de estas características.

Para aplicar el método de agrupación se requiere partir de una matriz presencia-ausencia de distintos descriptores para cada uno de los 200 datos analizados. Los descriptores seleccionados fueron:

Descriptor 1: Varianza alta, cuando la varianza de la muestra es mayor que la varianza media muestral más la mitad de la desviación típica muestral, $(\mathrm{V} > \overline{V} + .5\overline{S})$

Descriptor 2: Varianza media, $(\mathrm{V} \in [\overline{V} - .5\overline{S}, \overline{V} + .5\overline{S}])$

Descriptor 3: Varianza baja $(\mathrm{V} < \overline{V} - .5\overline{S})$

Descriptor 4: Asimetría alta, $(\mathrm{As} > \overline{As} + .5\overline{S}_{As})$

Descriptor 5: Asimetría media $(\mathrm{As} \in [\overline{As} - .5\overline{S}_{As}, \overline{As} + .5\overline{S}_{As}])$

Descriptor 6: Asimetría baja $(\mathrm{As} < \overline{As} - .5\overline{S}_{As})$

Descriptor 7: Apuntamiento alto, $(\mathrm{Ap} > \overline{Ap} + .5\overline{S}_{Ap})$

Descriptor 8: Apuntamiento medio $(\mathrm{Ap} \in [\overline{Ap} - .5\overline{S}_{Ap}, \overline{Ap} + .5\overline{S}_{Ap}])$

Descriptor 9: Apuntamiento bajo $(\mathrm{Ap} < \overline{Ap} - .5\overline{S}_{Ap})$

Descriptor 10: Una moda

Descriptor 11: Más de una moda

Los descriptores definidos, se han seleccionado de forma que permitan establecer la pertenencia de cualquier realización muestral a los parámetros definidos. Incluso desde un punto de vista cualitativo, es fácil establecer si una realización muestral tiene varianza alta, media o baja, y proceder de idéntica forma con la simetría y el apuntamiento; la determinación del número de modas también es posible a la vista del histograma de frecuencias.

A partir de esta matriz se procede a la clasificación dicotómica de los datos de partida mediante la definición de dos gradientes. El primer gradiente es el factor que absorbe máxima variación, obtenido mediante el análisis factorial de correspondencias, y define una ordenación de las muestras seleccionadas en función de todos los parámetros descriptores. El segundo gradiente se forma a partir de los descriptores más significativos en la ordenación del primer gradiente, y permite la clasificación dicotómica de las muestras. Al mismo tiempo este segundo gradiente facilita la definición de un umbral que permite la clasificación automática

de cualquier otra realización muestral en alguno de los dos grupos definidos a partir de los valores cuantitativos de la varianza, el coeficiente de asimetría, el de apuntamiento y del número de modas. El proceso se repite iterativamente, en cada uno de los dos grupos establecidos, hasta definir el número de grupos deseados (en nuestro caso 10).

El resultado de aplicar este método de clasificación se presenta en la figura 2. En el apéndice se presenta, también, la pertenencia de las muestras analizadas a cada uno de los grupos obtenidos.

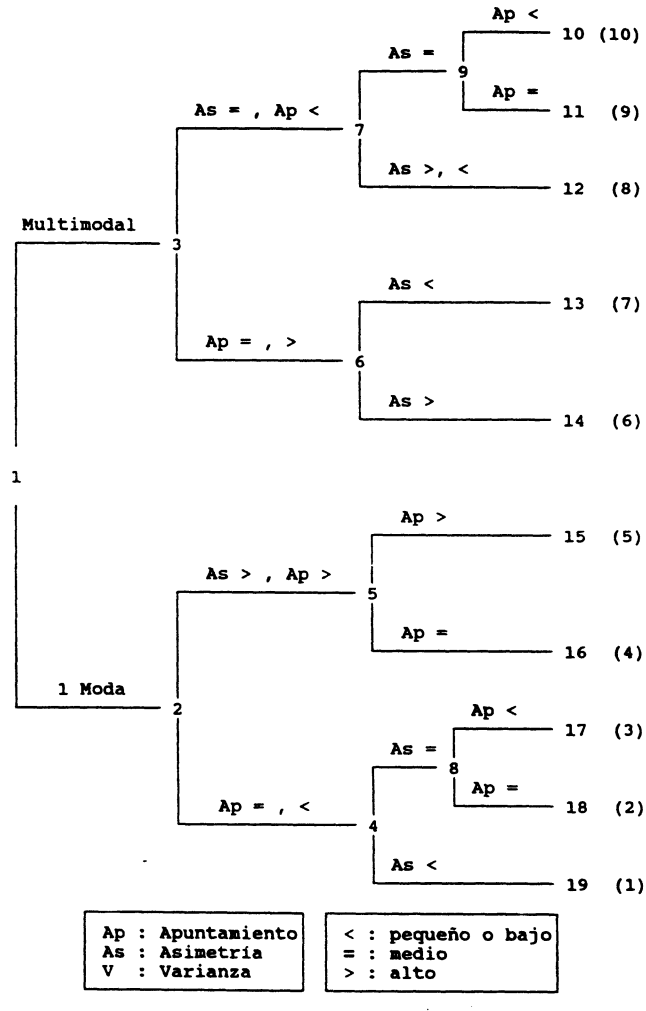

Figura 2. Agrupación de muestras.
Entre paréntesis aparece la denominación de grupos adoptada para las figuras 3, 4 y 5.

4. COMPARACIÓN DE LOS NÚCLEOS Y ANCHOS DE VENTANA

Para proceder a la comparación de algoritmos se requiere disponer de algún índice del error que se produce en cada grupo por el hecho de aplicar diferentes funciones núcleo con distintos anchos de ventana. Más adelante, se formulan algunas medidas en este sentido. Por otra parte, la aplicación de estos índices no siempre permite determinar unívocamente la mejor función núcleo; por este motivo, también se presentan medidas generales de la bondad del ajuste de las funciones núcleo.

La principal dificultad para la obtención de una medida del error en cada grupo muestral procede de que los grupos están formados por muestras procedentes de diferentes distribuciones de probabilidad; por tanto debe prescindirse de la utilización del error medio integrado. Por otra parte, no basta con determinar la función núcleo que mejor se adapte a las muestras del grupo, es necesario definir un ancho de ventana adecuado. Estas consideraciones aconsejan acudir a alguna medida del error propia de cada realización muestral y establecer un promedio de esa medida para todo el grupo, que sea función del valor del parámetro de alisado. En este sentido, para cada una de las realizaciones muestrales, y cada función núcleo, se obtuvo el error cuadrático integrado, para valores de h_n comprendidos entre 0'2 y 4'8:

$$ECI(h_n) = \int [f_n(x, h_n) - f(x)]^2 \, dx$$

La media del ECI para las muestras de cada grupo, y cada núcleo, que denominaremos MECIGN, es una función del ancho de banda y proporciona una estimación del error por grupo al aplicar las distintas función núcleo:

$$MECIGN_{ij}(h_n) = \frac{1}{n_i} \sum_{k \in G_i} \int [f_{kjn}(x, h_n) - f_k(x)]^2 \, dx$$

donde G_i es el grupo i
j representa al núcleo j
n_i es el número de muestras en el grupo i.
$f_k(x)$ es la función de densidad a partir de la que se generó la muestra k, y
k representa a las distintas realizaciones muestrales.

En la figura 3 se muestran los valores del MECIGN para el grupo 1 y el núcleo 1, así como los valores del ancho de banda para las muestras del grupo obtenidas por los procedimientos óptimo-1 y óptimo-2.

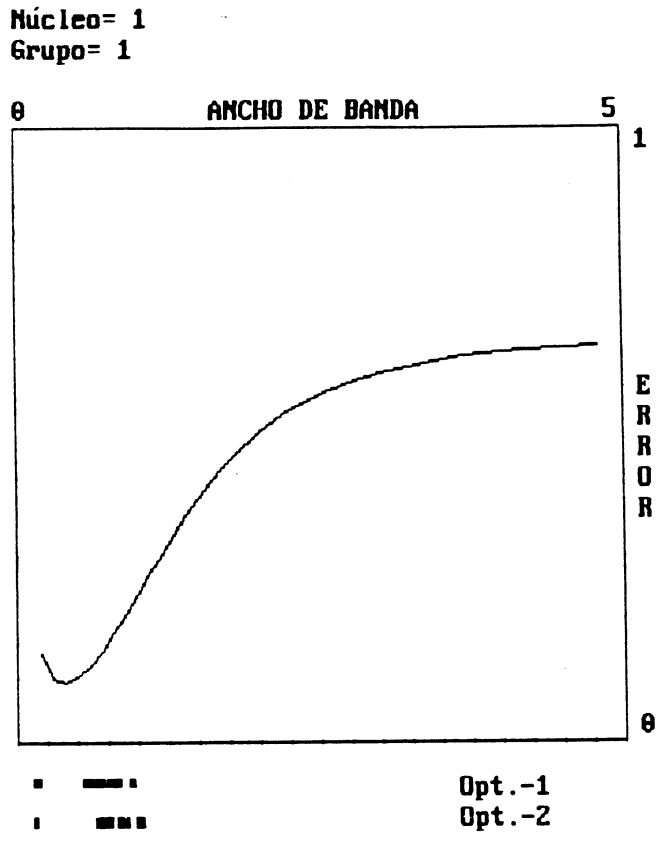

Figura 3.

El MECIGN también permite determinar un mejor valor medio del parámetro de alisado para cada grupo y cada función núcleo (h^*_{G-N}). La comparación de este valor con los anchos de ventana óptimo-1 y óptimo-2, en las realizaciones muestrales de cada grupo, puede proporcionar otras medidas del error. En este sentido se definió una nueva variable que, para cada muestra del grupo, representa la diferencia entre el mejor ancho de ventana del grupo y el ancho óptimo definido por los dos procedimientos independientes de la distribución de partida:

$$\begin{aligned} d_{k1} &= h^*_{Gi-Nj} - h_{k,01} \quad (k1 = 1, \ldots, n_{Gi}) \\ d_{k2} &= h^*_{Gi-Nj} - h_{k,02} \quad (k2 = 1, \ldots, n_{Gi}) \end{aligned}$$

donde n_{Gi} es el número de muestras en el grupo i
N_j representa el núcleo j.
$h_{k,01}$ es el valor del ancho de ventana para la muestra k,
obtenido por el procedimiento óptimo-1
$h_{k,02}$ es el valor del ancho de ventana para la muestra k,
obtenido por el procedimiento óptimo-2

La media y la varianza de esta variable, en cada grupo, representa el sesgo y la eficiencia, respecto del mejor h_n del grupo al calcular automáticamente el ancho de ventana. Las figuras 4 y 5 muestran estos valores presentando las gráficas de los diez grupos y en cada una los valores para los ocho núcleos.

Con objeto de comprobar la existencia de algún núcleo que proporcione sistemáticamente mejores ajustes en muestras de tamaño pequeño se ha estimado el error cuadrático medio integrado para cada núcleo:

$$U(h_n) = \int \mathrm{E}\left[\hat{f}_n(x, h_n) - f(x)\right]^2 dx$$

a partir de las muestras procedentes de la misma distribución de partida:

$$\hat{U}_{ij}(h_n) = \int \left\{ (1/n_i) \sum_{k \in F_i} \left[\hat{f}_{kjn}(x, h_n) - f_{F_i}(x)\right]^2 \right\} dx$$

donde $f_{F_i}(x)$ es la función de densidad de la distribución F_i,
$F_i = 1, \ldots, 5$
j representa al núcleo j
n_i es el número de muestras generadas de la distribución F_i,
k representa a las distintas realizaciones muestrales.

En las figuras 6.a y 6.b se muestran los resultados obtenidos para la distribución uniforme y los ocho núcleos estudiados. En el eje de abcisas se representan también los intervalos de confianza, al 95%, de los anchos de banda óptimos obtenidos por los dos procedimientos de optimización. La tabla 1 muestra los valores del error $U(h_n)$ obtenidos para un ancho de banda correspondiente a la media de los óptimos en muestras procedentes de la misma distribución de partida.

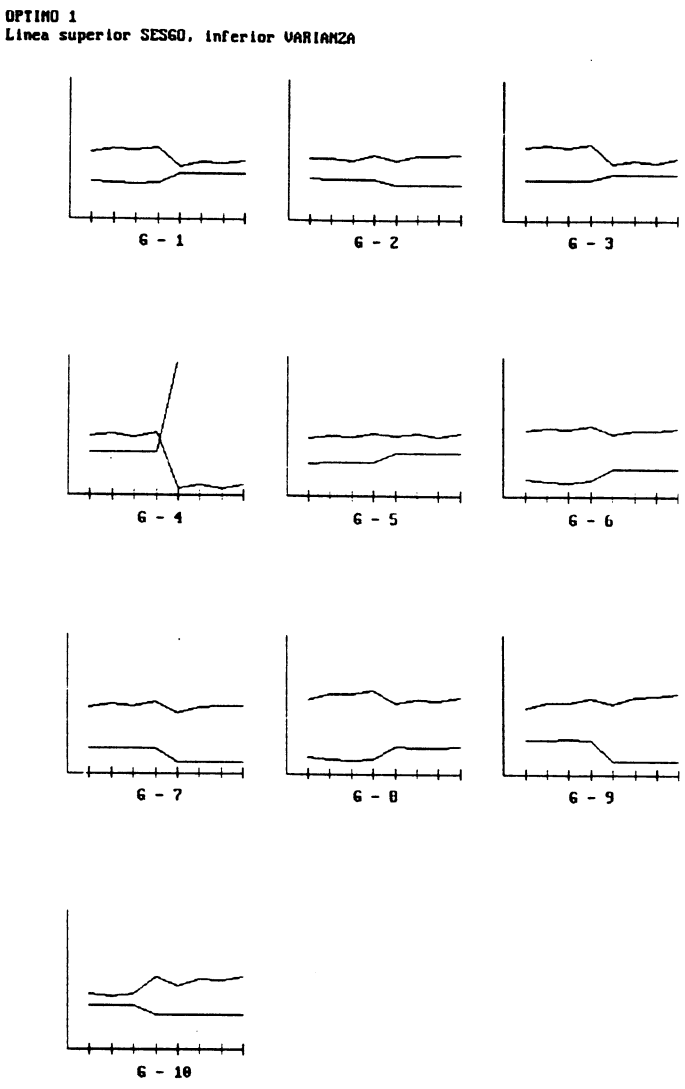
OPTIMO 1
Linea superior SESGO, inferior VARIANZA
G - 1
G - 2
G - 3
G - 4
G - 5
G - 6
G - 7
G - 8
G - 9
G - 10

Figura 4.

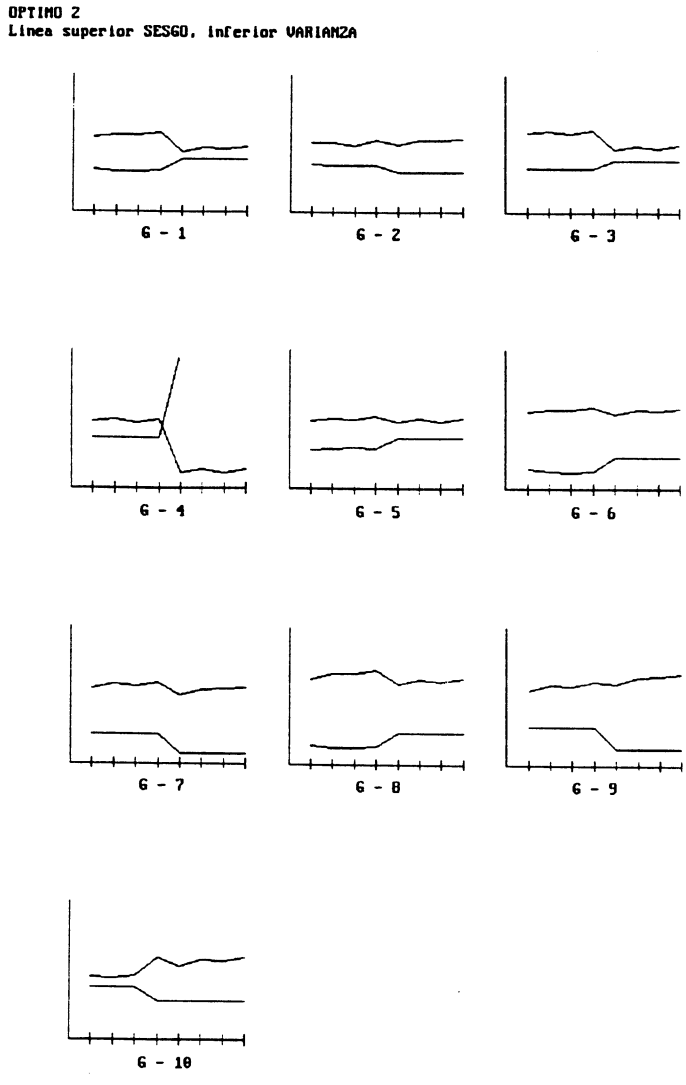
OPTIMO 2
Linea superior SESGO, inferior VARIANZA
G - 1
G - 2
G - 3
G - 4
G - 5
G - 6
G - 7
G - 8
G - 9
G - 10

Figura 5.

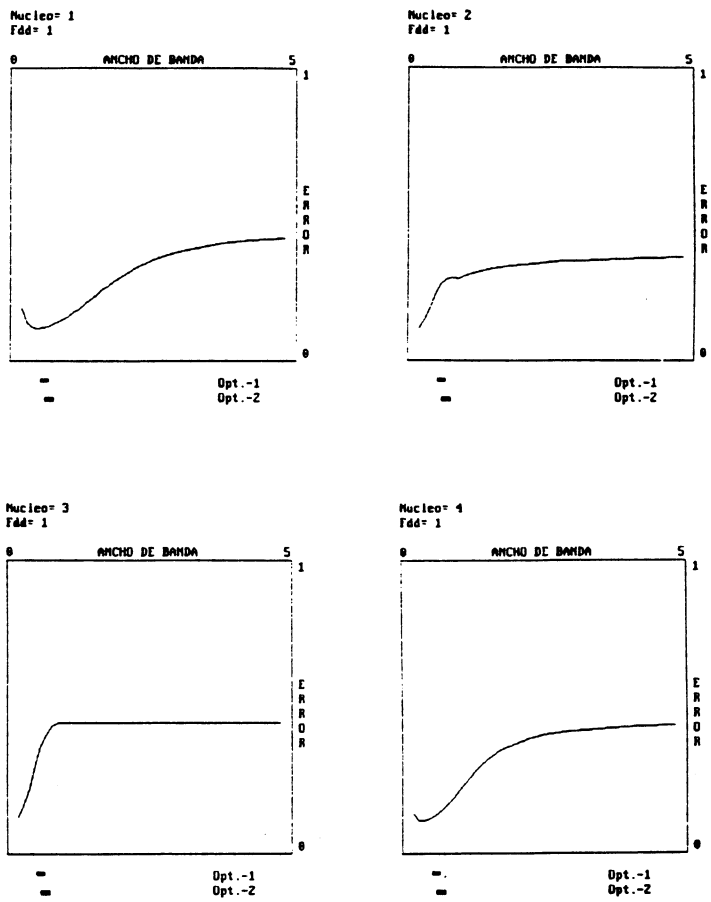
Nucleo= 1
Fdd= 1
0
ANCHO DE BANDA
5
1
ERROR
0
Opt.-1
Opt.-2
Nucleo= 2
Fdd= 1
0
ANCHO DE BANDA
5
1
ERROR
0
Opt.-1
Opt.-2
Nucleo= 3
Fdd= 1
0
ANCHO DE BANDA
5
1
ERROR
0
Opt.-1
Opt.-2
Nucleo= 4
Fdd= 1
0
ANCHO DE BANDA
5
1
ERROR
0
Opt.-1
Opt.-2

Figura 6.a

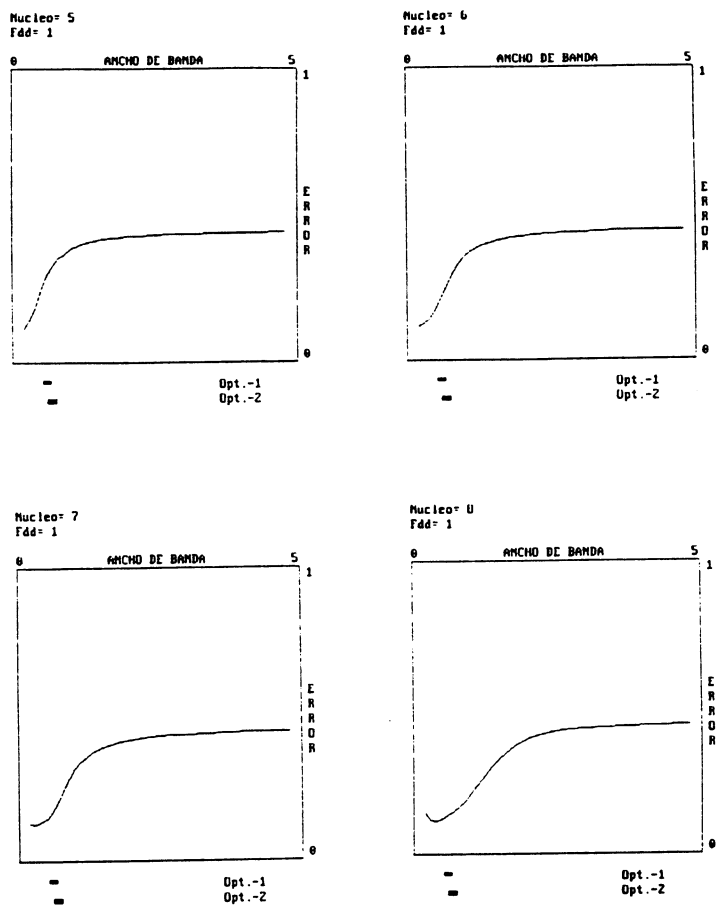
Nucleo= 5
Fdd= 1
0
ANCHO DE BANDA
5
1
E
R
R
O
R
0
Opt.-1
Opt.-2
Nucleo= 6
Fdd= 1
0
ANCHO DE BANDA
5
1
E
R
R
O
R
0
Opt.-1
Opt.-2
Nucleo= 7
Fdd= 1
0
ANCHO DE BANDA
5
1
E
R
R
O
R
0
Opt.-1
Opt.-2
Nucleo= 8
Fdd= 1
0
ANCHO DE BANDA
5
1
E
R
R
O
R
0
Opt.-1
Opt.-2

Figura 6.b

Tabla 1

Núcleo	Distribc.	U (Ópt.-1)	U (Ópt.-2)
K_1	1	0,1125	0,1875
	2	0,1250	0,1500
	3	0,1250	0,1250
	4	0,3750	0,4875
	5	0,1875	0,2000
K_2	1	0,2625	0,2750
	2	0,5500	0,5438
	3	0,0500	0,0563
	4	0,8250	0,8563
	5	0,3875	0,4130
K_3	1	0,3750	0,4125
	2	0,8125	0,8375
	3	0,0375	0,0375
	4	>1	>1
	5	0,5000	0,5375
K_4	1	0,1375	0,1500
	2	0,2500	0,3125
	3	0,1375	0,1375
	4	0,6250	0,7500
	5	0,2500	0,2625
K_5	1	0,3125	0,3313
	2	0,6875	0,7125
	3	0,0500	0,0500
	4	>1	>1
	5	0,4250	0,4500
K_6	1	0,2250	0,2500
	2	0,5625	0,6500
	3	0,0875	0,0875
	4	>1	>1
	5	0,3250	0,3625
K_7	1	0,1750	0,2000
	2	0,5000	0,5625
	3	0,1250	0,1250
	4	0,8875	>1
	5	0,3000	0,3250
K_8	1	0,1500	0,1375
	2	0,2375	0,3000
	3	0,3250	0,3250
	4	0,6125	0,6875
	5	0,2500	0,2625

Estos valores se han empleado para seleccionar entre las ocho funciones núcleo las que, independientemente de la función de partida, originen estimaciones con menos error. Esta selección no debe depender de las distribuciones originarias ya que, en la realidad son desconocidas, disponiendo sólo de los valores muestrales para tomar decisiones respecto a los algoritmos investigados.

5. RESULTADOS

El estudio de los errores MECIGN frente a distintos anchos de ventana permite la selección de tres de los núcleos investigados: K_1, K_4 y K_8, en los cuales se obtienen menores errores que con los restantes para todos los grupos de muestras. Ninguno de ellos, sin embargo, proporciona mejores estimas, respecto al MECIGN, que los otros para todos los posibles valores del parámetro de alisado. Los errores MECIGN correspondientes a los núcleos K_4 y K_8 son prácticamente iguales, pero difieren del núcleo K_1.

Es necesario, por tanto, utilizar la información que proporcionan los gráficos 4 y 5, respecto al sesgo y la varianza de los estimadores del parámetro de alisado óptimo. Entre los tres núcleos, se seleccionará aquel que posea una varianza y sesgo mínimos para el grupo. Con este criterio se elige el núcleo K_1, para los grupos 1, 5, 6, 8 y 9 y el K_8 para los grupos 7 y 10. En los grupos restantes tampoco se puede determinar el mejor núcleo con este criterio.

En estos casos, para elegir la mejor función núcleo se recurre a las medidas del error medio cuadrático integrado, según las cuales, en cualquier caso, la elección de K_1 supone un mínimo error (tabla 1).

Como conclusión, la decisión más adecuada, sería elegir:

Para el grupo 1, el núcleo K_1.
Para el grupo 2, el núcleo K_1.
Para el grupo 3, el núcleo K_1.
Para el grupo 4, el núcleo K_1.
Para el grupo 5, el núcleo K_1.
Para el grupo 6, el núcleo K_1.
Para el grupo 7, el núcleo K_8.
Para el grupo 8, el núcleo K_1.
Para el grupo 9, el núcleo K_1.
Para el grupo 10, el núcleo K_8.

Los dos estimadores del ancho de banda óptimo son estimadores sesgados para un tamaño de muestra pequeño, obteniéndose valores superiores al ancho óptimo, lo que origina estimas sobrealisadas de la densidad; en general el ancho de banda obtenido por el procedimiento 1 presenta mejores aproximaciones que el obtenido por el procedimiento 2. Sin embargo, las diferencias entre los anchos de banda obtenidos por ambos procedimientos no son significativas. Los intervalos de confianza al 99% para la razón de varianzas y para la diferencia de medias entre los dos estimadores, son (para las 200 realizaciones muestrales y los 8 anchos de banda):

Para la razón de varianzas, (0.51797, 1.07847).
Para la diferencia de medias con varianzas iguales, (-0.139309, 0.006769).

Puede aceptarse, por tanto, la igualdad en ambos casos. En cambio no puede decirse lo mismo del tiempo de CPU requerido para su cálculo.

La obtención de un ancho de ventana minimizando el riesgo de las pérdidas cuadráticas (Óptimo.-1) supone un procedimiento iterativo, cuya convergencia, en los problemas analizados conlleva un promedio de 14 iteraciones. En cada una de estas iteraciones se consume $n(n-1)$ veces más tiempo (n es el tamaño muestral), que en la obtención de la función de densidad estimada. La determinación del parámetro de alisado que maximiza la función de pseudoverosimilitud (Óptimo.-2) representa un consumo de CPU mucho menor (un promedio de 7 iteraciones con un tiempo por iteración del orden de n). Por todo esto, creemos recomendable el empleo del estimador por validación cruzada máximo verosimil y corregirlo, si fuera necesario, a la vista de la función de densidad estimada, al menos para tamaños muestrales pequeños.

APÉNDICE

Determinación del ancho de banda por aplicación del modelo adaptativo.

Los procedimientos bayesianos para optimizar una función $f(x, w)$, continua en x y medible en w, donde:

$$x \in A \subset \mathbb{R}^m \text{ (}\mathbb{R}\text{ conjunto de números reales), y}$$
$$w \subset \Omega \text{ (espacio muestral).}$$

a partir de una serie de muestras (x_i, y_i), $y_i = f(x_i)$, suponen definir una regla de decisión (d) que, a partir de un vector de observaciones $z_n = (x_i, y_i\ i = 1, \ldots, n)$,

permita determinar un x_{n+1} más próximo al óptimo. Esta regla de decisión se calcula para minimizar el riesgo de la decisión $R_0(d)$:

$$R_0(d) = \int_\Omega f(x_{n+1}(d), w)P(dw) - \int_\Omega \underset{x \in A}{opt}\, f(x, w)P(dw)$$

El segundo término de la expresión anterior es constante, por tanto la minimización del riesgo supone minimizar el primer término, que llamaremos $k(x)$; con lo cual:

$$x_{n+1} \in \underset{x \in A}{arg\, opt}\, k(x)$$

La aplicación del modelo adaptativo es posible al sustituir la condición de consistencia de Kolmogorov por las de continuidad de las funciones muestrales, homogeneidad de P e independencia en las diferencias parciales (ver Mockus, 1988 o Benveniste *et al.*, 1990). Bajo estas hipótesis se induce una distribución P gaussiana con media μ constante y matriz de covarianzas dependiente sólo de la distancia entre valores muestrales; en estas condiciones:

$$k(x) = \sigma^2/(\mu - c)$$

La definición de un modelo adaptativo supone referir la función a optimizar a funciones con soporte en intervalos convexos de los valores muestrales:

$$f(x) = f_i(x);\ x \in A_i;\ \bigcup_{i=1}^{n} A_i = A;\ A_i \cap A_j = \emptyset;\ i \neq j;\ i = 1, \ldots, n$$

si $f_i(x)$ es una función estocástica gaussiana, μ_x^i será el valor observado y_i en A_i, y la varianza una función de la distancia entre cualquier x y la realización muestral x_i; por tanto:

$$k_i(x) = \sigma_x^i/(\mu_x^i - c)$$

donde $\sigma_x^i = \sigma_0^2\, g(\|x - x_i\|)$ y

$$\mu_x^i = y_i$$

De los resultados anteriores se puede deducir:

$$x_{n+1} \in \underset{x \in A}{arg\, opt}\, \sigma_x^i/(\mu_x^i - c)$$

con la única restricción de continuidad de las funciones $k_i(x)$:

$$A_i = \{x : k_i(x) \leq k_j(x),\ j = 1, \ldots, n\}$$

Desde un punto de vista operativo, dado un conjunto de n realizaciones muestrales (x_i, y_i), y en el caso de que x pertenezca a R^1, para el cálculo de x_{n+1} basta determinar los x_k tales que:

$$(1) \qquad \frac{g(\|x_k - x_i\|)}{y_i - c} = \frac{g(\|x_j - x_k\|)}{y_j - c}, \quad (x_i \geq x_j)$$

El valor x_k con menor y_k, determinará el siguiente punto de muestreo x_{n+1}. La convergencia de este algoritmo está asegurada (Ljung, 1978).

La determinación del ancho de banda por maximización de la función de pseudo-verosimilidad es inmediata aplicando este algoritmo. Es suficiente determinar dos realizaciones muestrales de partida h_1 y h_2, que correspondan a los valores mínimo y máximo posibles del ancho de banda (0'2 y 4'8, para nuestras muestras). Estos valores se corresponden con x_1 y x_2, de forma que $y_1 = L(h_1)$ e $y_2 = L(h_2)$. A continuación, se puede calcular h_3 mediante la aplicación de la expresión (1) y continuar hasta que el algoritmo converja. El número medio de iteraciones requeridas para la determinación del ancho de banda fue de 7.

APÉNDICE

Pertenencia de las muestras analizadas a cada uno de los grupos obtenidos en la clasificación.

MUESTRAS DEL GRUPO 1:
12, 15, 38, 39, 48, 49, 61, 63, 65, 72, 85, 89, 125, 129, 140, 144, 146, 163, 166, 173, 187, 188, 191.
MUESTRAS DEL GRUPO 2:
14, 16, 17, 25, 44, 47, 54, 57, 75, 78, 90, 97, 107, 110, 111, 117, 118, 126, 133, 156, 200.
MUESTRAS DEL GRUPO 3:
3, 41, 50, 53, 71, 88, 102, 112, 121, 138, 151, 194.
MUESTRAS DEL GRUPO 4:
1, 4, 13, 22, 28, 52, 55, 59, 62, 68, 74, 82, 92, 98, 103, 108, 130, 137, 150, 152, 165, 167, 172, 179, 180, 181, 183, 185, 186, 198.
MUESTRAS DEL GRUPO 5:
9, 11, 20, 21, 24, 26, 31, 35, 42, 45, 60, 80, 81, 84, 91, 104, 114, 123, 132, 143, 149, 162, 171, 175, 177, 182, 195.
MUESTRAS DEL GRUPO 6:
8, 10, 27, 56, 58, 83, 86, 93, 100, 105, 131, 135, 145, 161, 170, 176, 184, 193, 199.

MUESTRAS DEL GRUPO 7:
19, 34, 66, 76, 77, 96, 99, 124, 134, 141, 142, 155, 158, 174.
MUESTRAS DEL GRUPO 8:
2, 6, 7, 33, 37, 43, 46, 69, 106, 109, 120, 127, 128, 136, 164, 169.
MUESTRAS DEL GRUPO 9:
5, 32, 70, 95, 101, 115, 122, 147, 153, 154, 157, 160, 196, 197.
MUESTRAS DEL GRUPO 10:
18, 23, 29, 30, 36, 40, 51, 64, 67, 73, 79, 87, 94, 113, 116, 119, 139, 148, 159, 168, 178, 189, 190, 192.

6. REFERENCIAS

[1] **Benveniste, A; Metivier, M y Priouret, P.** (1990). "*Adaptive Algorithms and Stochastics Approximations*". Springer-Verlag. Berlín.

[2] **Cencov, N.N.** (1962). "Evaluation of an unknown distribution density from observations". *Soviet Math.*, **3**, 1559–1562.

[3] **Chisman, J.A.** (1992). "*Introduction to Simulation Modeling Using GPSS*". Prentice Hall. Nueva Jersey. Englewood Cliffs.

[4] **Deheuvels, P.** (1977). "Estimation nonparamétrique de la densité par histogrammes généralisés". *Revue de Statistique Appliquée*, 25/3, 5–42.

[5] **Devroye, L.** (1983). "The equivalence of weak, strong, and complete convergence in L_1 kernel density estimates". *Annl. Stat.*, **11**, 896–904.

[6] **Devroye, L. y Gyorfi, L.** (1985). "*Nonparametric Density Estimation: the L_1 view*". John Wiley. Nueva York.

[7] **Devroye, L. y Penrod, C.S.** (1984). "The consistency of automatic kernel density estimates". *Annl. Stat.*, **12**, 1231–1249.

[8] **Duin, R.P.W.** (1976). "On the choice of smoothing parameters for Parzen estimators of probability denstiy functions". *IEEE Transactions on Computers*, C-25, 1175–1179.

[9] **Epanechnikov, V.A.** (1969). "Nonparametric estimates of a multivariate probability density". *Th. Prob. Applic.*, **14**, 153–158.

[10] **Friedman, J.H., Stuetzle, W. y Schroeder, A.** (1984). "Projection pursuit density estimation (PPDE)". *J.A.S.A.*, **79**, 599–608.

[11] **Glivenko, V.I.** (1934). "*Course in Probability Theory*". Moscú.

[12] **Hall, P. y Marron, J.S.** (1988). "Choice of kernel order in density estimation". *Th. Ann. of Stats.*, **16**, 161–173.

[13] **Hermans, J. y Habbema, J.D.F.** (1976). "*Manual for the ALLOC discriminant analysis programs*". Universidad de Leiden, Dept. de Estadística Médica.

[14] **HÄRDLE, W:** (1991). "*Snoothing Techniques with Implementation in S*". Springer-Verlag, Nueva York.

[15] **Hill, M. O.** *et al.* (1975). "Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Sctoland". *Journal Ecology,* **63**, 597–613.

[16] **Izeman, A.J.** (1991). "Recent developments in nonparametric density estimation". *JASA.* Vol. 86, Nº 413, 205–224.

[17] **Ljung, L.** (1978). "Convergence of an adaptive filter algorithms". *Int. J. Control,* **27**, 673–693.

[18] **Loftsgaarden, D.O. y Quesenberry, C.P.** (1965). "A nonparametric estimate of a multivariate density function". *Ann. Math. Statist.,* **36**, 1049–1051.

[19] **Marron, J.S.** (1987). "Automatic smoothing parameter selection: a survey". *Empirical Economics,* **13**, 187–208.

[20] **Mockus, J.** (1988). "*Bayesian Approach to Global Optimization*". Kluwer. Dordrecht.

[21] **Nadaraya, E.A.** (1989). "*Nonparametric Estimation of Probability Densities and Regression Curves*". Kluwer Academic Publishers, Dordrecht.

[22] **Parzen, E.** (1962). "On a estimation of a probability density function and mode". *Annals Math. Statist.,* **33**, 1065–1076.

[23] **Rosenblatt, M.** (1956). "Remarks on some non-parametrics estimates of a density function". *Annals Math. Statist.,* **27**, 832–837.

[24] **Scott, D.W.; Tapia, R.A. y Thompson, J.R.** (1977). "Kernel density estimation revisited". *Nonlinear Analysis, Th., Met. Appl.,* **1**, 339–372.

[25] **Scott, D.W.; Tapia, R.A. y Thompson, J.R.** (1980). "Nonparametric probability density estimation by discrete maximum penalized —likelihood criteria". *The Annals of Statistics,* **8**, 820–832.

[26] **Scott, D.W. y Factor, L.E.** (1981). "Monte Carlo study of three data-based nonparametric probability density estimators". *JASA,* **76**, 9–15.

[27] **Silverman, B.W.** (1982). "Kernel density estimation using the fast fourier transformation". *Appl. Stat.,* **31**, 93–97.

[28] **Silverman, B.W.** (1986). "*Density Estimation for Statistics and Data Analysis*". Chapman and Hall, London.