Comparación en la elección de una ventana óptima para algunos estimadores de densidad

Comparison of optimal bandwidths for some density estimators

Mathias Bourel[1]

Recibido: Junio 2013 Aprobado: Setiembre 2013

Resumen.- La estimación de una densidad es un problema clásico y muy estudiado en Estadística. En este artículo repasamos brevemente algunos estimadores usuales de una densidad unidimensional como el histograma, el Averaged Shifted Histograms (ASH) y el estimador por núcleo. Nos interesamos en particular a la elección de la ventana, parámetro determinante en cuanto a la predicción del modelo. Terminamos nuestro trabajo con una simulación comparando los métodos presentados sobre un conjunto común de densidades.

Palabras clave: Estimación densidad; Histograma; Estimación por núcleo.

Summary.- *Density estimation is a classic problem and has been extensively studied in Statistics. In this paper we briefly review some usual density estimators as the histogram, the Averaged Shifted Histograms (ASH) and the kernel density estimator (Kde). We care in particular on the choice of the bandwidth of Kde which is a fundamental parameter of the model and greatly influences the prediction. We finished our work with a simulation comparing the described methods on a common set of densities.*

Keywords: *Density Estimation; Histogram; Kernel Density Estimation.*

1. Introducción.- La distribución de una variable aleatoria absolutamente continua X puede ser descrita a través de su función de densidad f. En efecto, si X es una variable aleatoria con densidad f y A un subconjunto medible entonces

$$P(X \in A) = \int_A f(x)dx$$

Uno de los principales problemas de la Estadística es, dada una muestra aleatoria simple $\{x_1, \dots, x_n\}$, estimar la función de densidad f de la variable aleatoria *X* de la cual provienen. Un primer enfoque consiste en suponer que la distribución de *X* pertenece a cierta clase paramétrica de funciones de distribuciones y estimar los parámetros de la misma. Por ejemplo suponer que $\{x_1, \dots, x_n\}$ son *n* realizaciones independientes de una variable aleatoria $X \sim N(\mu, \sigma^2)$ y estimar la media μ y la desviación estándar σ a partir de las observaciones $x_1, \dots, x_n$. Este enfoque paramétrico tiene la desventaja de ser poco flexible pues, al suponer que pertenece a alguna familia de funciones, impone cierta restricción sobre la densidad f. La idea del enfoque no paramétrico se basa en no hacer ninguna suposición sobre f y estimarla directamente a partir de los datos. Supongamos que X ϵ *A*, siendo *A* un intervalo pequeño donde f no varía demasiado. Entonces:

[1] M. Bourel, Magister en Matemática, Profesor Area Matemática, Universidad de Montevideo, mbourel@um.edu.uy

ISSN 2301-1092 • ISSN (en línea) 2301-1106

$$P(X \epsilon A) = \int_A f(x)dx \approx f(x)l(A) \; con \; x \epsilon A$$

donde *l(A)* es la longitud del intervalo *A*, por lo cual

$$f(x) \approx \frac{P(X \epsilon A)}{l(A)}$$

Por otro lado, como la probabilidad de que *k* puntos de las *n* observaciones caigan dentro del intervalo *A* es $\binom{n}{k} P(X \in A)^k (1 - P(X \in A))^{n-k}$, el estimador de máxima verosimilitud de *P (X ϵ A)* es $\frac{k}{n}$, entonces sustituyendo en la expresión anterior se tiene que:

$$f(x) \approx \frac{k}{nl(A)}$$

El problema consiste entonces en elegir adecuadamente el conjunto *A* para poder estimar $f(x)$. En este trabajo repasamos varios estimadores no paramétricos clásicos para una densidad tales como el histograma y los estimadores basados en núcleos. Estos estimadores han dado lugar a numerosos trabajos con la finalidad de elegir *A* de manera óptima. Repasaremos dichos enfoques y algún otro método para después comparar sus performances sobre un conjunto de densidades con diversas dificultades.

2. Histograma y Averaged Shifted Histograms (ASH)

2.1. Histograma.- El histograma es seguramente el estimador de densidad más sencillo e intuitivo. En una dimensión, el mismo se construye eligiendo un origen x_0 y subdividiendo el eje real en varios intervalos o celdas B_j de longitud h, obteniendo la partición $\{B_j\}_j \epsilon Z$ siendo:

$$B_j = [x_0 + (j-1)h, x_0 + jh), j \epsilon Z$$

Para cada *j* se cuenta la cantidad de observaciones de la muestra que caen en el intervalo B_j, es decir si $x \epsilon B_j$, asignaremos a x el valor

$$\widehat{f_h}(x) = \frac{1}{n} \frac{cantidad \; de \; observaciones \; en \; B_j}{l(B_j)}$$

donde $l(B_j)$ es la longitud del intervalo B_j.

Si suponemos que cada intervalo tiene la misma longitud h, podemos entonces definir el histograma como:

$$\widehat{f_h}(x) = \frac{1}{nh} \sum_{j \epsilon Z} \sum_{i=1} \mathbb{1}_{\{x_i \in B_j\}} \mathbb{1}_{\{x \in B_j\}}$$

Siendo $\mathbb{1}_A$ la función indicatriz definida por $\mathbb{1}_A(x) = \begin{cases} 1, & si \; x \in A \\ 0, & si \; x \notin A \end{cases}$. Obsérvese que el término $\frac{1}{nh}$ hace que $\widehat{f_h}$ sea una densidad, es decir que $\int \widehat{f_h}(x)dx = 1$.

ISSN 2301-1092 • ISSN (en línea) 2301-1106

El histograma depende de dos parámetros: el largo de la ventana h y el origen x_0. Para distintas elecciones del origen los histogramas construidos pueden ser completamente diferentes. La Figura I muestra el histograma que se obtiene al considerar 50 datos provenientes de una densidad *N* (0, 1) cambiando los valores de estos dos parámetros.

Para x fijo, del cálculo del sesgo $Sesgo\left(\widehat{f_h}(x)\right) = \mathbf{E}(\widehat{f_h}(x) - f(x))$, y de la varianza de $\widehat{f_h}$, se observa que si h es grande, y por lo tanto el histograma tiene pocas celdas, la varianza disminuye pero el sesgo es grande (ver [1]). En cambio si h es pequeño, el histograma resultante tiene muchas celdas y se produce el fenómeno inverso: el sesgo es pequeño pero la varianza es grande. Hay que encontrar por lo tanto un compromiso entre el sesgo y la varianza de manera a que ambos sean relativamente pequeños. Existe una vasta literatura que propone algunos métodos de manera a optimizar la ventana h: regla de Sturges, regla de Scott, etc ([2], [3]).

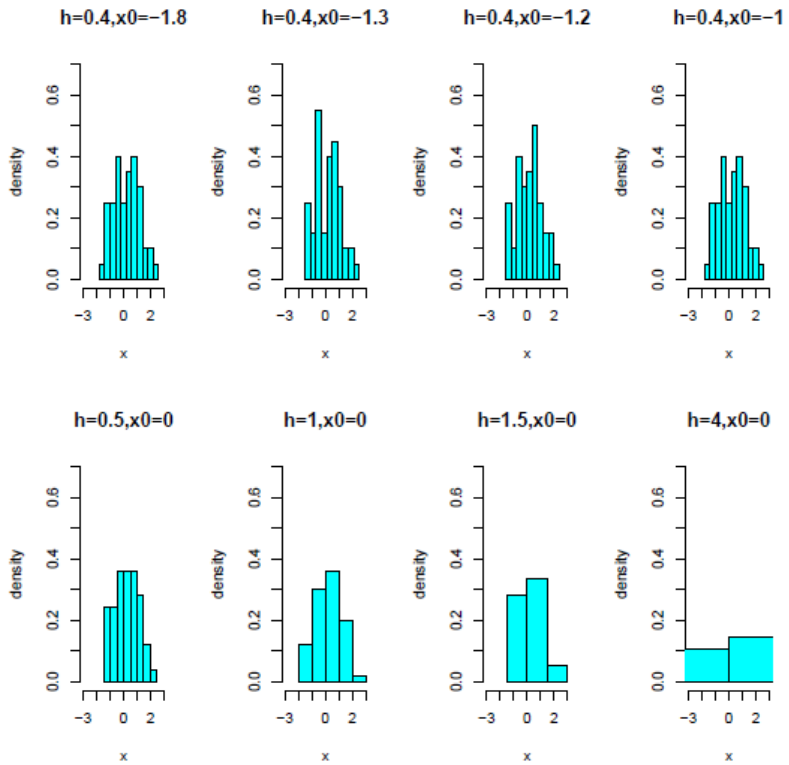

Figura I.- Histogramas obtenidos a partir de 50 datos provenientes de una N (0,1) al cambiar los valores del origen x0 y la longitud de la ventana h.

Otra manera de encontrar un valor optimal para h es considerar para el histograma el error cuadrático medio integrado, en inglés *Mean Integrated Square Error*, *MISE*, definido por

$$MISE\left(\widehat{f_h}\right) = \mathbf{E}\left(\int \left(\widehat{f_h}(x) - f(x)\right)^2 dx\right)$$

donde $\mathbf{E}$ denota la esperanza y f es la densidad a estimar. Este error se puede aproximar (ver [1]) por el *Asymptotic Mean Integrated Square Error*, AMISE:

$$AMISE\left(\widehat{f_h}\right) = \frac{1}{nh} + \frac{1}{12}h^2 R(f')$$

ISSN 2301-1092 • ISSN (en línea) 2301-1106

donde $R(f') = \int (f'(x))^2 dx$.

Un cálculo sencillo muestra que el valor de h que minimiza la expresión anterior es:

$$h_{opt} = \left(\frac{6}{n \int \left(f'(x)\right)^2 dx} \right)^{\frac{1}{3}}$$

Si bien este valor en la práctica es difícil de obtener, ya que al desconocer f, tampoco se conoce f', muchas veces se suele suponer que la distribución es normal para calcularlo y en este caso:

$$h_{opt} = 3.5\hat{\sigma} n^{\frac{1}{3}}$$

2.2 Averaged Shifted Histograms (ASH).- En la Figura I podemos ver como distintos valores para el origen x_0 modifican el histograma que se obtiene al considerar un mismo conjunto de datos, aun considerando la misma ventana h. Una manera de "independizarse" de la elección del origen fue propuesto por Scott en 1985 en [2] y consiste en promediar varios histogramas con el mismo valor de h pero con distintos orígenes y por lo tanto tomando varias particiones del eje real.

Consideramos un histograma con ancho de ventana h, origen en 0, y grilla $\{[jh, (j+1)h)\}_{j \in Z}$ con celdas de largo h. Sea $M > 0$, definimos $\delta = \frac{h}{M}$ y dividimos cada celda $[jh, (j+1)h)$ en M nuevas celdas $B_k = [k\delta, (k+1)\delta)$ obteniendo de esta manera una grilla más fina. Denotamos por v_k las observaciones que pertenecen a la celda B_k. Por ejemplo si k = 0, la celda original es $[0, h]$ y la dividimos en M subceldas:

$$B_0 = [0, \delta),\ B_1 = [\delta, 2\,\delta),\ \ldots,\ B_{M-1} = [(M-1)\delta, M\delta) = [(M-1)\delta, h).$$

Supongamos que $x \in B_0 = [0, \delta)$. Entonces hay M histogramas de ancho de ventana $h = M\delta$ que cubren a B_0.

El primer histograma, evaluado en x, vale

$$\widehat{f_1}(x) = \frac{v_{1-M} + v_{2-M} + \cdots + v_0}{nh}$$

El segundo histograma, evaluado en x, vale

$$\widehat{f_2}(x) = \frac{v_{2-M} + v_{3-M} + \cdots + v_0 + v_1}{nh}$$

Y así sucesivamente hasta obtener el último histograma, evaluado en x:

$$\widehat{f_M}(x) = \frac{v_0 + \cdots + v_{M-1}}{nh}$$

El estimador *Averaged Shifted Histograms*, *ASH* se obtiene promediando estos M estimadores, es decir,

$$\hat{f}_{ASH}(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

Observar que en nuestro ejemplo, en la suma final la frecuencia de v_{1-M} es $\frac{1}{M}$, la de v_{2-M} es $\frac{2}{M}$,

ISSN 2301-1092 • ISSN (en línea) 2301-1106

…, la de v_{M-1-M} es $\frac{M-1}{M}$, la de v_0 es $\frac{M}{M}=1$, la de v_1 es $\frac{M-1}{M}$, la de v_{M-1} es $\frac{1}{M}$. Entonces podemos reescribir en general $\hat{f}_{ASH}(x)$ como

$$\hat{f}_{ASH}(x) = \frac{1}{M}\sum_{j=1-M}^{M-1}\left(\frac{M-|j|}{nh}\right)v_{k+j} \qquad \forall\, x \in B_k$$

y simplificando:

$$\hat{f}_{ASH}(x) = \frac{1}{nh}\sum_{j=1-M}^{M-1}\left(1-\frac{|j|}{M}\right)v_{k+j} \qquad si\ x \in B_k$$

Es importante notar que *ASH* no es un simple histograma con un ancho de ventana más pequeño como lo muestra la Figura II. Si consideramos un conjunto de 100 datos provenientes de una densidad $N(0,1)$ y utilizamos para *ASH* M = 8 histogramas con largo de ventana $h = 0.8$, el histograma obtenido con largo de ventana $\frac{h}{M} = 0.1$ es diferente al estimador obtenido por *ASH*.

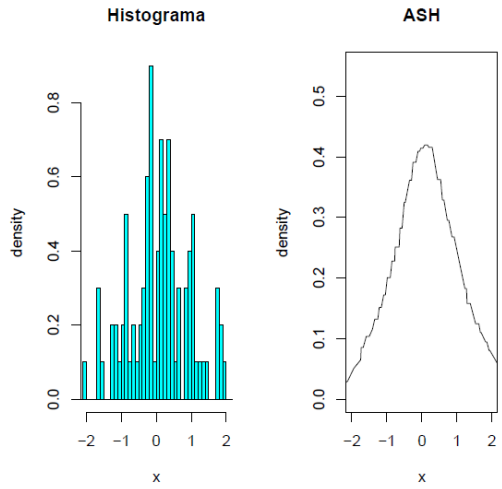

Figura II.- Histograma con ancho de ventana δ = 0.1 y ASH promediando M = 8 histogramas con ancho de ventana 0.8 obtenidos a partir de un conjunto de 100 datos provenientes de una densidad N (0, 1).

En [2], se prueba que si la derivada segunda de la densidad a estimar es uniformemente continua en **R** y la derivada tercera pertenece a L^2, entonces tomando $M = +\infty$ y para valores de *n* grande, el MISE del estimador que se obtiene por *ASH* es mínimo. En la figura III mostramos el ajuste del estimador obtenido por *ASH* para una mezcla de tres gaussianas incrementando la cantidad de histogramas que se promedia.

Por otro lado, fijados *h* y *n*, si $x \in B_k$ y la observación $x_i \in B_{k+j}$, tomando $M \to +\infty$, la cantidad de celdas entre $x \in B_k$ y $x_i \in B_{k+j}$ es aproximadamente *j* y entonces $|x - x_i| \approx |j|\delta$ por lo cual:

$$1-\frac{|j|}{M} = 1-\frac{|j|\delta}{M\delta} \approx 1-\frac{|x-x_i|}{h} \quad si\ |x-x_i| < h$$

Al ser *M* grande, se tiene que la cantidad de observaciones en cada celda es 0 o 1. Por lo tanto, si x_i no pertenece al intervalo $(x-h, x+h)$ entonces x_i no tiene influencia en la estimación de x en *ASH*. Podemos entonces escribir:

$$\lim_{M\to\infty} \hat{f}_{ASH}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left(1-\left|\frac{x-x_i}{h}\right|\right)\mathbf{1}_{(-1,1)}\left(\frac{x-x_i}{h}\right)$$

Donde esta vez la suma se hace sobre los datos. Esta última expresión nos permite considerar una clase más importante de estimadores para una densidad, llamados estimadores por núcleo que estudiaremos en la sección siguiente.

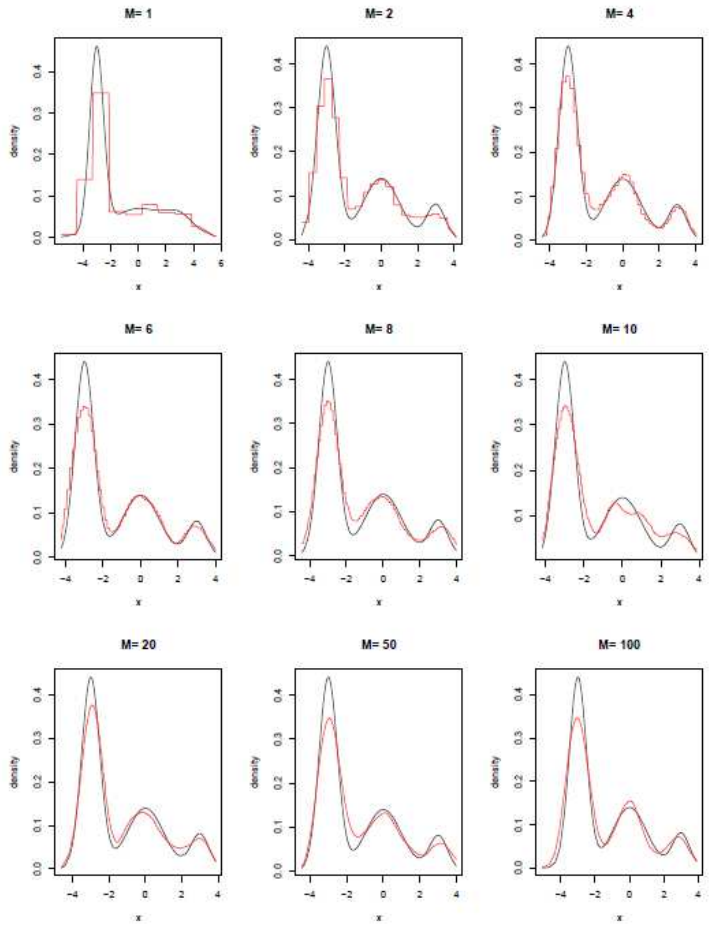

Figura III.- El estimador ASH con distintos valores para M para una mezcla de tres gaussianas.

3. El estimador por núcleo.- Además de depender de dos parámetros, la ventana *h* y el origen x_0, una de las características del histograma que parece por lo menos restrictiva es que se asigna a cada uno de los *x* de una misma celda B_j el mismo valor estimado $\hat{f}_h(x)$. Esto hace también que el histograma no sea una función continua, tenga varios saltos y por lo tanto no sea derivable en estos saltos y tenga derivada nula en los restantes puntos, lo cual no es del todo conveniente si se quiere estimar una función de densidad *f* que sea por lo menos continua.

En vez de subdividir el eje real en celdas de largo *h* determinadas por el origen que se elija y contar la cantidad de observaciones que caen en cada de ellas, en 1956 Rosenblatt propone en [4] considerar, para cada *x*, un intervalo centrado en *x* de radio *h*. Esta manera de abordar el problema de la estimación de una densidad *f* es coherente con aplicar el teorema de Glivenko-Cantelli. En efecto, este teorema asegura que un buen estimador de la función de distribución *F* de una muestra aleatoria simple es la distribución empírica

$$F_n(x) = \frac{\#\{i : x_i \in (-\infty, x]\}}{n}$$

ya que, casi seguramente,

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0$$

Como la función de densidad *f* es la derivada de la distribución *F*, la misma se puede obtener como

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

Y podemos entonces hacer una estimación de *f(x)* mediante

$$\hat{f}_h(x) = \frac{F_n(x+h) - F(x-h)}{2h} = \frac{\frac{\#\{x_i : x_i \in (x-h, x+h]\}}{n}}{2h}$$

siendo F_n la distribución empírica de *X*. Por lo cual

$$\hat{f}_h(x) = \frac{1}{n} \frac{cantidad\ de\ observaciones\ en\ (x-h, x+h]}{l((x-h, x+h])}$$

Es un estimador muy parecido al histograma definido en la sección 2.1. Este estimador se suele llamar histograma con ventana móvil. Esta fórmula puede ser reescrita como:

$$\hat{f}_h(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{(x-h, x+h]}(x_i) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{(-1, 1]}\left(\frac{x - x_i}{h}\right)$$

Formalizando un poco más, si definimos $K: \mathbb{R} \to \mathbb{R}$ por $K(x) = \frac{1}{2} 1_{(-1, 1]}(x)$ entonces $\hat{f}_h(x)$ puede escribirse como:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Esta manera de escribir el estimador $\hat{f}_h$ de *f* permite ver claramente cómo cuando una observación x_i se encuentra dentro del intervalo $(x-h, x+h]$ la misma contribuye a la frecuencia calculada. Sin embargo la estimación no toma en cuenta que tan alejada o tan cerca de *x* se encuentra x_i. La idea del estimador por núcleo consiste en ponderar y darle mayor peso a las observaciones que se encuentran cerca de *x* en $(x-h, x+h]$.

3.1 Kernel Density Estimator (Kde).- Un *estimador por núcleo* o *Kernel Density Estimator*, *Kde*, de una densidad f es la función:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)$$

ISSN 2301-1092 • ISSN (en línea) 2301-1106

donde $K_h: \mathbb{R} \to \mathbb{R}$ está definida por $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$. A *K* se la llama *función núcle*o y se pide en general que sea no negativa, simétrica, unimodal y tal que $\int K(u)du = 1$. Al parámetro *h* se le llama *ventana* de la estimación. Es importante remarcar que $\hat{f}_h$ es una densidad; depende de la ventana h que se elija (ver sección 3.2); es una variable aleatoria pues depende de las observaciones; y que hereda todas las propiedades de la función *K* (continuidad, diferenciabilidad, etc.). La función K_h indica el peso que tiene la observación x_i de la estimación de $f(x)$. Existen varios ejemplos de funciones núcleos, por ejemplo, para citar los más importantes:

- El núcleo uniforme $K(t) = \frac{1}{2}\mathbf{1}_{\{|t|\leq 1\}}$,
- El núcleo triangular $K(t) = (1 - |t|)\ \mathbf{1}_{\{|t|\leq 1\}}$,
- El núcleo de Epanechnikov $K(t) = \frac{3}{4}(1 - t^2)\mathbf{1}_{\{|t|\leq 1\}}$,
- El núcleo gaussiano $K(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$

Observar que el estimador que se obtiene utilizando un histograma con ventana móvil es un estimador por núcleo con un núcleo uniforme, y el estimador que se obtiene haciendo tender $M \to \infty$ en *ASH* es un estimador por núcleo con núcleo triangular. Queda claro que en el caso del núcleo uniforme todas las observaciones que caen en el intervalo $(x - h, x + h]$ tienen el mismo peso, pero que tanto con el núcleo triangular, el de Epanechnikov y el Gaussiano, las observaciones más cercanas a *x* están afectadas de un peso mayor.

A pesar de poder elegir dentro de una vasta gama de funciones núcleos *K*, la elección del mismo no es tan determinante como la elección de la ventana *h*. Al igual que para el histograma, la ventana *h* juega un papel crucial en qué tan "suave" será la estimación. Al tomar un valor grande para *h*, muchas observaciones serán tenidas en cuenta en el cálculo de *f*, "suavizando" de esta manera la estimación obtenida. Por el contrario, si se toma una ventana *h* demasiada pequeña, pocas observaciones o ninguna pertenecerán al intervalo que contiene a *x* y la estimación será poca suave. La Figura IV muestra la estimación por núcleo de una gaussiana tomando como valores para *h* = 3, 1, 0.01 y 0.001. Este problema será abordado en la sección siguiente.

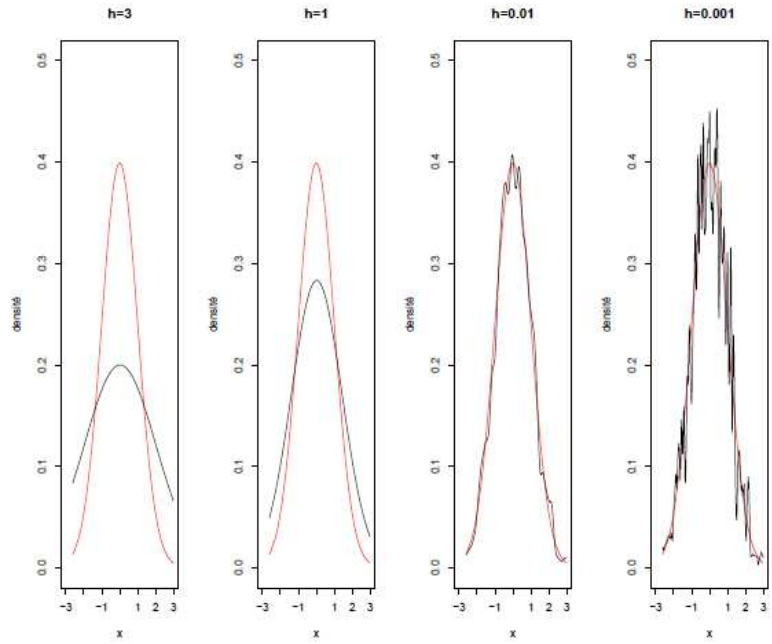

Figura IV.- Estimadores por núcleos de una densidad gaussiana con diferentes valores para la ventana h

Finalmente es de destacar que el estimador por núcleo puede ser visto como la suma de *n* funciones núcleo reescaladas, cada una de las cuales esta´ centrada sobre una observación. Es decir

$$\hat{f}_h(x) = \sum_{i=1}^{n} \underbrace{\frac{1}{n} K_k(x - x_i)}_{función\ núcleo\ reescalada}$$

donde cada $\frac{1}{n} K_k(x - x_i)$ integra $\frac{1}{n}$ (Figura V).

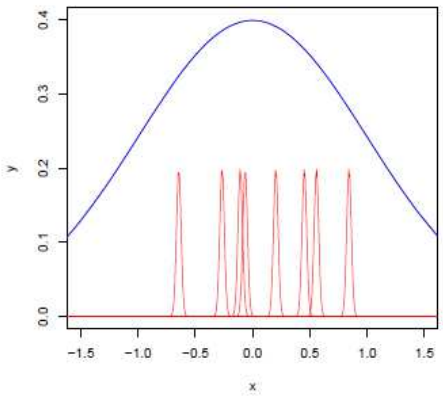

Figura V.- Estimador por núcleo como suma de funciones núcleos reescaladas

3.2 Elección de la venta h.- Calculando el sesgo y la varianza de $\hat{f}_h(x)$ se observa, al igual que para el histograma, que si *h* es grande la varianza disminuye pero el sesgo es grande. En cambio si *h* es pequeño, el sesgo es pequeño pero la varianza es grande. Una manera de optimizar el valor de la ventana *h* consiste en, como en el histograma, encontrar un compromiso entre el sesgo y la varianza, minimizando el error cuadrático medio integrado. Un cálculo sencillo (ver [1]) permite mostrar que si $\hat{f}_h$ es un estimador con núcleo entonces

$$MISE(\hat{f}_h) = \int Sesgo(\hat{f}_h(x))dx + \int Var(\hat{f}_h(x))dx$$

Es importante notar que el sesgo del estimador con núcleo es del orden de h^2 mientras que el sesgo del histograma es del orden de *h* (ver [1]).

Por otro lado se puede probar que

$$Sesgo(\hat{f}_h(x)) = \frac{f''(x)}{2}\mu_2(K)h^2 + o(h^2)$$

y que

$$Var(\hat{f}_h(x)) = \frac{R(K)f(x)}{nh} + o\left(\frac{1}{nh}\right)$$

siendo $R(t) = \int (t(x))^2 dx$ y $\mu_r(K) = \int x^r K(x)dx$. Por lo tanto es inmediato escribir que

$$MISE\left(\hat{f}_h\right) = \frac{1}{nh}R(K) + \frac{h^4}{4}\mu_2(K)^2 R(f'') + o\left(\frac{1}{nh}\right) + o(h^4)$$

Entonces si $h \to 0, nh \to +\infty$ y $n \to \infty$ se puede aproximar el MISE por:

$$AMISE\left(\hat{f}_h\right) = \frac{1}{nh}R(K) + \frac{h^4}{4}\mu_2(K)^2 R(f'')$$

Y el valor de *h* que minimiza la expresión anterior es

$$h^* = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

Sin embargo este valor nuevamente depende de f, pues depende de f'', la cual es desconocida.

Para intentar paliar este problema se utilizan algunos de los métodos descritos a continuación.

1. *El método de Silverman.*

   Este método supone que la verdadera densidad f es una normal $N(\mu, \sigma^2)$. Si notamos por φ a la densidad normal $N(0,1)$ entonces

   $$R(f'') = \|f''(x)\|_2^2 = \sigma^{-5}\int \varphi(x)^2 dx = \sigma^{-5}\frac{3}{8\sqrt{\pi}} \approx 0.212\,\sigma^{-5}$$

   En la práctica se sustituye σ por el estimador $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$. Si K es el núcleo

ISSN 2301-1092 • ISSN (en línea) 2301-1106

gaussiano $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ tenemos entonces que sustituyendo en la expresión de h^*:

$$h^* = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}$$

Para evitar los outliers que impliquen una varianza importante reemplazamos $\hat{\sigma}$ por la distancia intercuartile $\hat{R}$ de los datos. Es fácil ver que en este caro $\hat{R} \approx \sigma(2 \times 0.67) = 1.34\sigma$ y que por lo tanto $\hat{\sigma} = \frac{\hat{R}}{1.34}$. Sustituyendo se tiene que

$$\hat{h}^* = 1.06.\frac{\hat{R}}{1.34}n^{-\frac{1}{5}} = 0.79\hat{R}n^{-\frac{1}{5}}$$

Finalmente elegiremos

$$\hat{h}_{NRD} = 1.06\, min\left\{\hat{\sigma}, \frac{\hat{R}}{1.34}\right\}n^{-\frac{1}{5}}$$

Una variante sugerida por el propio Silverman consiste en tomar como constante 0.9 (Nrd0) en vez de 1.06 (Nrd).

2. Validación cruzada

Otra manera de elegir la ventana consiste en realizar un proceso de validación cruzada ([5]).Nos interesamos esta vez al error cuadrático medio (*Integrated squared error, ISE*) dado por

$$ISE\left(\hat{f}_h\right) = \int(\hat{f}_h(x) - f(x))^2 dx = \int(\hat{f}_h(x))^2 dx - 2\int \hat{f}_h(x)f(x)dx + \int(f(x))^2 dx \quad (2)$$

Observar que el último termino en la expresión (2) no depende de *h*. Consideramos entonces el *Least Square Cross* de *h*, $LSCV(h)$, definido por

$$LSCV(h) = \int(\hat{f}_h(x))^2\, dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i)$$

Donde $\hat{f}_{-i}$ es el estimador por núcleo hallado a partir de todos los datos sacando la observación *i*. En ([1]), se muestra que

$$\int(\hat{f}_h(x))^2\, dx = \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n} K * K\left(\frac{x_j - x_i}{h}\right)$$

Siendo $K * K$ la convolución de *K*. Podemos entonces escribir $LSCV(h)$como

$$LSCV(h) = \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n} K * K\left(\frac{x_j - x_i}{h}\right) - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)$$

Por lo cual $LSCV(h)$es un estimador de $ISE\left(\hat{f}_h\right) - \int f^2(x)dx$ y como

ISSN 2301-1092 • ISSN (en línea) 2301-1106

$$\mathbf{E}\big(LSCV(h)\big) = MISE\big(\hat{f}_h\big) - \int f^2(x)dx$$

resulta que $LSCV(h)$ es insesgado a menos de $\int f^2(x)dx$. Denotamos por h_{UCV} (*UCV* por unbiased cross-validation) el valor de h que minimiza $LSCV(h)$.

Stone prueba en ([6]) que encontrar h por este método es asintóticamente óptimo bajo condiciones muy débiles sobre la densidad f.

3. Método plug-in de Sheater y Jones

Nuevamente se quiere estimar $R(f'')$ en la expresión

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

Esta vez observamos como $R(f'') = \int (f''(x))^2 dx = \int f^{(4)}(x) f(x) dx = \mathrm{E}(f^{(4)}(X))$, podemos entonces estimar $R(f'')$ por

$$S(g) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{ng}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{g}\right)\right)^{(4)} = \frac{1}{n^2 g^5}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(4)}\left(\frac{x_i - x_j}{g}\right)$$

donde g es una nueva ventana para el estimador por núcleo considerado.

Sheater y Jones ([7]) y [8]) proponen como valor para g a

$$g = g(h) = \left(\frac{2K^{(4)}(0)\mu_2(K)S(g_1)}{R(K)T(g_2)}\right)^{\frac{1}{7}} h^{\frac{5}{7}}$$

donde

$$S(g_1) = \frac{1}{n^2 g_1^5}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(4)}\left(\frac{x_i - x_j}{g_1}\right)$$

y

$$T(g_2) = \frac{1}{n^2 g_1^7}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(6)}\left(\frac{x_i - x_j}{g_2}\right)$$

los estimadores por núcleo $S(g_1)$ y $S(g_2)$ son estimadores de $R(f'')$ y $R(f''')$ con ventanas g_1 y g_2 respectivamente. Estas dos ventanas se calculan mediante la Regla de Silverman, asumiendo únicamente en esta etapa que la distribución es normal. Se usa entonces $S(g(h))$ como estimador de $R(f'')$. Por lo tanto la ventana h_{SJ} obtenida por el método de Sheater y Jones es la solución de la ecuación

$$h = \left[\frac{R(K)}{\mu_2(K)^2 S(g(h))}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

4. Agregación de estimadores por núcleos.- Los métodos de agregación de modelos en aprendizaje automático combinan varios estimadores construidos sobre un mismo conjunto de datos con el fin de obtener un modelo predictivo con una mejor performance. Los mismos han sido ampliamente estudiados y han dado lugar a numerosos trabajos tanto experimentales como teóricos principalmente en clasificación y en regresión, en un contexto supervisado. Algunos trabajos han aparecidos también aplicando estos métodos al problema de la estimación de una densidad. Invitamos al lector interesado en la misma en consultar [9], [10] y [11] para un desarrollo más detallado de estos enfoques. En este trabajo proponemos incluir el algoritmo secuencial de M. Di Marzio y C. Taylor, BoostKde, inspirado del Boosting (ver [9]). Para este algoritmo, se inician uniformemente los pesos de cada observación y se elige un valor para la ventana h. En cada etapa *m* se calcula un estimador de la siguiente manera:

$$\hat{f}_m(x) = \sum_{i=1}^{n} \frac{w_m(i)}{h} K\left(\frac{x - x_i}{h}\right)$$

siendo *K* una función núcleo, y $w_m(i)$ el peso de la observación *i* de la etapa *m*. Observar que en la etapa 1, $\hat{f}_1$ coincide con el estimador por núcleo y que en las etapas siguientes podemos pensar, a pesar que no sea una densidad, el estimador como una especie de estimador por núcleo ponderado. Al igual que en Boosting, el peso de cada observación se actualiza en cada etapa de la siguiente manera:

$$w_{m+1}(i) = w_m(i) + \log\left(\frac{\hat{f}_m(x_i)}{\hat{f}_m^{(-i)}(x_i)}\right)$$

Donde $\hat{f}_m^{(-i)}(x_i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{w_m(j)}{h} K\left(\frac{x_j - x_i}{h}\right)$. El estimador final es $\hat{f}_M(x) = C \prod_{m=1}^{M} \hat{f}_m(x)$, donde C es un factor tal que $\hat{f}_M$ sea una densidad.

En [12], los autores muestran para $M = 2$, una reducción del sesgo del orden de h^4(recordar que el sesgo en *Kde* es del orden de h^2). No obstante, no es claro cómo se comporta el algoritmo para valores de $M > 2$. En las simulaciones utilizan varios modelos, y hallan distintos valores para la ventana h minimizando el MISE para diferentes valores de *M*.

5. Simulación.- En esta sección comparamos los estimadores presentados anteriormente sobre un variado conjunto de densidades que hemos encontradas en la bibliografía especializada (ver Figura VI):

ISSN 2301-1092 • ISSN (en línea) 2301-1106

$\mathcal{M}1$: Normal Estandar
$\mathcal{M}2$: Exponencial
$\mathcal{M}3$: Chi cuadrado
$\mathcal{M}4$: t-Student
$\mathcal{M}5$: Mezcla de dos normales
$\mathcal{M}6$ y $\mathcal{M}7$: Mezcla de tres normales
$\mathcal{M}8$: Claw density
$\mathcal{M}9$: Smooth Comb Density
$\mathcal{M}10$: Mezcla normal y uniformes
$\mathcal{M}11$: Densidad triangular
$\mathcal{M}12$: Beta

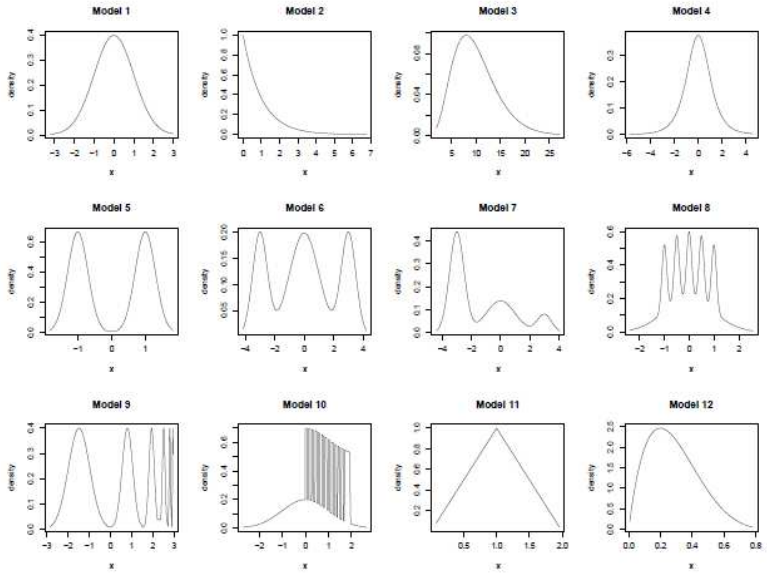

Figura VI.- Densidades utilizadas para la simulación.

Comparamos sobre estos modelos los siguientes estimadores de densidad: *Hist* (histograma), *ASH*, *BoostKde* y *Kde*, realizando los siguientes ajustes de sus parámetros:

- para el histograma y *ASH* utilizamos una grilla de 10, 20 y 50 intervalos y, para cada densidad considerada, retuvimos la cantidad de intervalos que maximiza la log-verosimilitud del modelo realizando 100 simulaciones de Monte Carlo independientes;
- tanto para *BoostKde* que para *ASH* tomamos $M = 5$ como la cantidad de estimadores intermediarios que consideramos dentro del algoritmo;
- para *Kde*, utilizamos un núcleo gaussiano y optimizamos h con los cuatro métodos vistos: *Nrd*, *Nrd0*, *UCV* y *SJ* ;
- corrimos el algoritmo *BoostKde* con un núcleo gaussiano para $M=5$ y con la ventana h obtenida por los métodos *Nrd, Nrd0, UCV y SJ*.

En las tablas siguientes, para las 12 densidades, presentamos la performance de cada método para $n = 100$, 500 y 1000 habiendo medido el error por el criterio *MISE*. La misma se calcula como un promedio del error cuadrático medio sobre 100 simulaciones Monte Carlo. En el caso de Kde y de BoostKde indicamos entre paréntesis el método de elección de la ventana óptima con la que se obtuvo el mejor resultado. Por lo general, para $n = 100$ o $n = 500$, BoostKde o Kde mejoran los resultados obtenidos por el histograma. No es claro el desempeño de BoostKde para $n = 1000$ con esta elección de M. Por otro lado, observamos que a medida que va creciendo $n = 1000$, elegir la ventana en Kde por el método UCV es cada vez más efectivo, verificando de esta manera el

ISSN 2301-1092 • ISSN (en línea) 2301-1106

teorema de Stone. Si bien muchas veces *ASH* mejora el valor obtenido por el histograma, para algunos modelos, los resultados obtenidos van en el sentido de las hipótesis del teorema obtenido por Scott [13], pues por ejemplo, para aquellas densidades que no son uniformemente continua como la exponencial, al incrementar el valor de *M* pudimos comprobar que esto no necesariamente implica una disminución del *MISE*.

	Hist	ASH	BoostKde	Kde
$\mathcal{M}1$	2.91	1.97	0.455(nrd)	0.189(nrd0)
$\mathcal{M}2$	5.14	4.98	7.34(nrd)	4.3(ucv)
$\mathcal{M}3$	0.156	0.102	0.0358(nrd)	0.0107(nrd)
$\mathcal{M}4$	1.44	0.949	1.17(nrd)	0.191(nrd0)
$\mathcal{M}5$	6.33	4.2	0.519(nrd)	3(ucv)
$\mathcal{M}6$	0.81	0.543	0.0803(nrd0)	0.156(ucv)
$\mathcal{M}7$	1.39	0.81	0.14(nrd0)	0.508(ucv)
$\mathcal{M}8$	3.93	2.48	3.35(nrd)	2.17(ucv)
$\mathcal{M}9$	2.68	1.69	1.31(nrd0)	1.36(ucv)
$\mathcal{M}10$	6.21	4.71	5.88(nrd)	6.65(nrd0)
$\mathcal{M}11$	18.7	1.98	1.84(nrd)	2.24(nrd0)
$\mathcal{M}12$	130	34	14.7(nrd)	58.2(nrd0)

Table 1: $100\times$ MISE con $n = 100$ y $M = 5$

	Hist	ASH	BoostKde	Kde
$\mathcal{M}1$	0.454	0.116	0.127(nrd)	0.0825(nrd0)
$\mathcal{M}2$	0.75	3.18	6.25(nrd)	3.02(ucv)
$\mathcal{M}3$	0.0236	0.0156	0.0104(ucv)	0.00291(nrd)
$\mathcal{M}4$	0.165	0.0935	1.88(nrd)	0.0938(nrd0)
$\mathcal{M}5$	1.29	0.818	0.115(nrd)	1.98(ucv)
$\mathcal{M}6$	0.0686	0.0308	0.019(nrd0)	0.0883(ucv)
$\mathcal{M}7$	0.234	0.0656	0.034(nrd)	0.309(ucv)
$\mathcal{M}8$	0.928	0.517	2.08(nrd)	1.77(ucv)
$\mathcal{M}9$	0.623	0.424	0.901(nrd0)	0.877(ucv)
$\mathcal{M}10$	4.78	4.45	5.48(nrd)	6.14(ucv)
$\mathcal{M}11$	0.875	0.444	0.542(nrd)	1.53(nrd0)
$\mathcal{M}12$	9.35	5.5	4.18(nrd)	42.4(ucv)

Table 2: $100\times$ MISE con $n = 500$ y $M = 5$

	Hist	ASH	BoostKde	Kde
$\mathcal{M}1$	0.221	0.0584	0.0628(nrd)	0.0745(nrd0)
$\mathcal{M}2$	0.365	3.24	5.22(nrd)	2.51(ucv)
$\mathcal{M}3$	0.0114	0.00757	0.0124(nrd)	0.00178(nrd)
$\mathcal{M}4$	0.0897	0.0431	1.58(sj)	0.073(nrd0)
$\mathcal{M}5$	0.639	0.4	0.0663(nrd)	1.7(ucv)
$\mathcal{M}6$	0.0537	0.0195	0.0111(nrd)	0.0648(ucv)
$\mathcal{M}7$	0.136	0.0361	0.0165(nrd0)	0.231(ucv)
$\mathcal{M}8$	0.606	0.291	1.54(nrd0)	1.7(ucv)
$\mathcal{M}9$	0.394	0.285	0.633(sj)	0.757(ucv)
$\mathcal{M}10$	4.8	4.59	5.35(nrd)	5.55(ucv)
$\mathcal{M}11$	0.726	0.438	0.314(nrd)	1.18(nrd0)
$\mathcal{M}12$	5	2.68	2.91(nrd)	33.3(ucv)

Tabla 3.- 100× MISE con n = 1000 y M = 5

6. Conclusión.- La estimación de una densidad es un problema muy estudiado en Estadística. En el correr de este trabajo repasamos algunos estimadores clásicos de una densidad unidimensional, algunas estrategias para optimizar sus parámetros y los comparamos sobre un conjunto variados de

densidades que se encuentran frecuentemente en la literatura especializada sobre este tema. Si bien abordamos únicamente el caso unidimensional, es importante saber que todos estos métodos se generalizan naturalmente en varias dimensiones. Recientemente, debido al desarrollo del Aprendizaje Estadístico, varios métodos provenientes de esta área intentaron mejorar, complementar o englobar las técnicas existentes. En los artículos [9], [10] y [11] proponemos varios estimadores de densidad obtenidos por métodos de agregación. Estos trabajos forman parte de la tesis [14] y buscan generalizarse al caso multidimensional.

7. Referencias

[1] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. Nonparametric and Semiparametric Models.Springer Verlag, Heidelberg, 2004.

[2] D.W. Scott. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1992.

[3] B. W. Silverman. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, 1986.

[4] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27:832–837, 1956.

[5] A. Bowman. An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71:353–36, 1984.

[6] C. Stone. An asymptotically optimal window selection rule for kernel density estimates. The Annals of Statisitics, 12:1285–1297, 1984.

[7] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B, Methodological, 53:683–690, 1991.

[8] S.J. Sheather. Density estimation. Statistical Science, 19(4):588–597, 2004.

[9] M. Bourel. Metodos de agregacion de modelos y aplicaciones. Memoria de trabajos de difusion cientifica y tecnica, 10:19–32, 2012.

[10] M. Bourel, R. Fraiman, and B. Ghattas. Random average shifted histogram. To appear, 2013.

[11] M. Bourel and B. Ghattas. Aggregating density estimators: an empirical study. Open Journal of Statistics, 3(5), 2013.

[12] M. Di Marzio and C.C. Taylor. Boosting kernel density estimates: A bias reduction technique? Biometrika, 91(1):226–233, 2004.

[13] D.W Scott. Averaged shifted histogram: Effective nonparametric density estimators inseveral dimensions. The Annals of Statistics, 13(3):1024–1040, 1985.

[14] M. Bourel. Apprentissage statistique par agrégation de modèles pour l'estimation de la densité et la classification multiclasse. Tesis Université Aix-Marseille, France, 2013.