Estimación no paramétrica de la función de densidad

Antonio Miñarro

Barcelona, Enero 1998

Índice general

Capítulo 1

Introducción

1.1. Introducción

Es difícil concebir la estadística actual sin el concepto de *distribución de probabilidad* de una variable aleatoria, entendiéndolo como un modelo matemático que describe el comportamiento probabilístico de la misma. Cualquier utilización posterior de la variable aleatoria: cálculo de probabilidades, inferencia estadística o las técnicas de análisis de datos multidimensionales, utilizan de una u otra forma y son dependientes de la distribución de probabilidad que se presupone para la variable. La representación matemática más tangible de la distribución de una variable aleatoria se corresponde con las denominadas funciones de distribución y de densidad de probabilidad de la variable aleatoria, íntimamente relacionadas entre sí. Conocer la función de densidad de una variable aleatoria implica tener una completa descripción de la misma. Es por tanto un problema fundamental de la estadística la estimación de la función de densidad de una variable o vector aleatorio a partir de la información proporcionada por una muestra.

Un posible enfoque consiste en considerar que la función de densidad que deseamos estimar pertenece a una determinada clase de funciones paramétricas, por ejemplo a algunas de las clásicas distribuciones: normal, exponencial, Poisson, etc. Dicha suposición usualmente se basa en informaciones sobre la variable que son externas a la muestra, pero cuya validez puede ser comprobada con posterioridad mediante pruebas de bondad de ajuste. Bajo esta suposición la estimación se reduce a determinar el valor de los parámetros del modelo a partir de la muestra. Esta estimación es la que denominaremos *estimación paramétrica* de la densidad. La posibilidad alternativa es no predeterminar a priori ningún modelo para la distribución de probabilidad de la variable y dejar que la función de densidad pueda adoptar cualquier forma, sin más límites que los impuestos por las propiedades que se exigen a las funciones de densidad para ser consideradas como tales. Este enfoque, en el que se centra el presente trabajo, es el que denominaremos *estimación no paramétrica* de la densidad, y tiene uno de sus orígenes más comunmente aceptado en los trabajos de Fix y Hodges (1951) que buscaban

una alternativa a las técnicas clásicas de análisis discriminante que permitiera liberarse de las rígidas restricciones sobre la distribución de las variables implicadas. En cierta manera el enfoque no paramétrico permite que los datos determinen de forma totalmente libre, sin restricciones, la forma de la densidad que los ha de representar.

La controversia sobre la utilización de una estimació paramétrica o no paramétrica no ha cesado a lo largo de los años, a la eficiencia en la estimación que proporciona la estimación paramétrica se contrapone el riesgo que suponen desviaciones de las suposiciones que determinan el modelo y que pueden conducir a errores de interpretación que supongan mayor pérdida que la ganancia proporcionada por la eficacia estimadora.

Entre las principales situaciones en las cuales la estimación no paramétrica de la densidad ha resultado ser de especial interés podemos destacar:

- Análisis exploratorio: Diversas características descriptivas de la densidad, tales como multimodalidad, asimetrías, comportamiento en las colas, etc., enfocadas desde un punto de vista no paramétrico, y por tanto más flexible, pueden ser más reveladoras y no quedar enmascaradas por suposiciones más rígidas. Como ejemplo presentamos los resultados de un estudio realizado por Park y Marron (1990) donde se han estudiado los ingresos netos familiares a lo largo de varios años, obteniéndose una secuencia de estimaciones de la densidad; en la Figura 1 (a) se ha supuesto una densidad lognormal para los ingresos netos, observándose que todas las poblaciones son unimodales y que esencialmente no hay cambio a lo largo de los años, en la Figura 1 (b) se ha obtenido una estimación no paramétrica por el método de las funciones núcleo, observándose un mínimo de dos modas en todas las poblaciones, así como un gran cambio en la estructura a lo largo del tiempo.

- Presentación de datos: La presentación gráfica de los resultados obtenidos en una estimación no paramétrica de la densidad es fácilmente comprensible e intuitivo para aquellas personas no especialistas en estadística que muy a menudo son los clientes de los servicios de estadística. Como ejemplo en la Figura 2 presentamos, tomado de Izenman (1991), los resultados de estimación mediante funciones núcleo, de las frecuencias cardiacas en reposo y la máxima, para un grupo de varones que sufren una enfermedad coronaria y otro grupo de varones normales.

- Técnicas multivariantes: Estimaciones no paramétricas de la densidad son utilizadas en problemas de discriminación, clasificación, contrastes sobre las modas, etc. Es ilustrativo el ejemplo presentado en Silverman (1986) sobre discriminación entre instituciones médicas y no médicas de la Universidad de Londres basado en el tiempo de utilización de dos sistemas operativos.

- Regresión: Estimaciones no paramétricas de la densidad permiten estimar la *Curva de Regresión de la Media*, que sabemos que es la que minimiza la esperanza del error

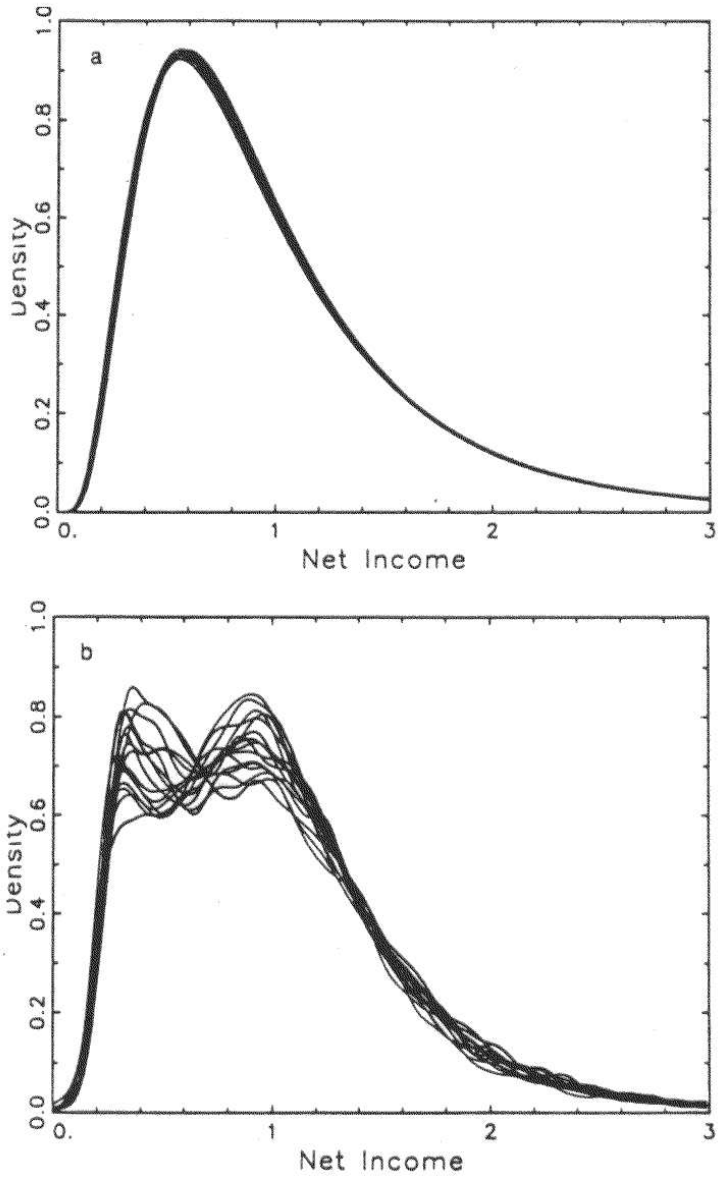

Figura 1.1: Estimaciones de los ingresos netos: (a) Ajuste Lognormal; (b) Estimación tipo núcleo. De Park y Marron (1990).

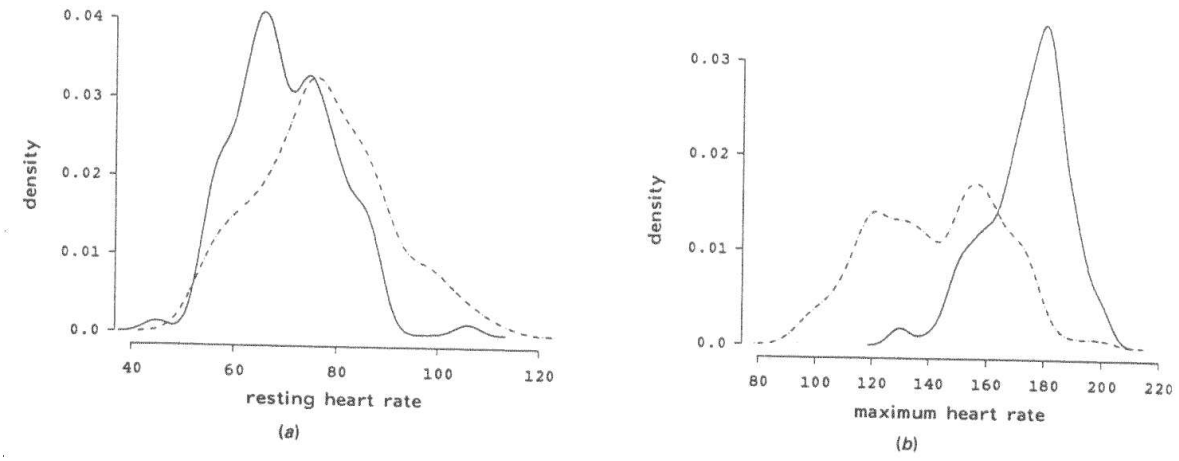

Figura 1.2: Estimaciones de las frecuencias cardiacas: (a) en reposo, y (b) máxima, para un grupo de 117 enfermos de corazón (línea de puntos) y otro grupo de 117 varones normales (línea sólida). De Izenman (1991).

cuadrático y se obtendría a partir de

$$r(x) = \mathrm{E}(Y|X = x) = \frac{\int y f(x, y)\, dy}{\int f(x, y)\, dy}$$

Hemos de destacar finalmente que si en los últimos años se ha producido un gran desarrollo de las técnicas de estimación no paramétrica, dicho desarrollo ha sido paralelo al de la informática y su aplicación a la estadística, acceso a nuevos y potentes ordenadores, y aparición de una gran gama de software estadístico y facilidades gráficas de alto nivel.

1.2. Propiedades estadísticas de los estimadores

Es necesario que consideremos qué propiedades deben verificar las funciones utilizadas como estimadores.

1.2.1. Propiedades de verdadera densidad

Entendemos por propiedades de verdadera densidad que la estimación no sea nunca negativa y su integral sea uno

$$f(x) \geq 0, \qquad \int f(x)dx = 1 \tag{1.1}$$

En algunos casos ciertos métodos proporcionan estimaciones que pueden dar valores negativos debido a la dispersión de los datos (Boneva, Kendall y Stefanov 1971) o a un relajamiento de

las condiciones exigidas con vistas a aumentar la tasa de convergencia de las estimaciones a la verdadera densidad. Este aumento de la convergencia se puede lograr también relajando la condicion de la integral de la densidad. De cualquier forma estos problemas pueden solventarse por ejemplo: truncando la densidad a su parte positiva y renormalizando, estimando una versión transformada de la densidad, por ejemplo $f^{1/2}, \log f$ y transformando posteriormente para obtener densidades no negativas. Gajek(1986) proporciona un esquema por el cual cualquier estimador que no de una auténtica densidad puede hacerse converger a una densidad real.

1.2.2. Sesgo

Un estimador $\hat{f}$ de una función de densidad f es insesgado para f si $\forall x \in \mathbf{R}^d, E_f[\hat{f}(x)] = f(x)$. Aunque existen estimadores insesgados conocidos para muchas densidades paramétricas, Rosenblatt (1956) demuestra que no es posible la existencia de un estimador de la densidad que verifique (1.1) y que sea insesgado para todas las densidades continuas. Esto ha motivado que se centre la atención en secuencias de estimadores no paramétricos $\{\hat{f}_n\}$ que sean asintóticamente insesgadas, es decir

$$E_f[\hat{f}(x)] \to f(x) \text{ si } n \to \infty \tag{1.2}$$

1.2.3. Consistencia

La noción intuitiva de consistencia es que el estimador se aproxime a la verdadera densidad cuando aumenta el tamaño de la muestra. Diremos que un estimador de la densidad $\hat{f}$ es débilment consistente en forma puntual si $\hat{f}(x) \to f(x)$ en probabilidad para todo $x \in \mathbf{R}$, y es fuertemente consistente en forma puntual si la convergencia es casi segura. Otros tipos de convergencia están relacionados con los criterios de error que veremos más adelante.

Capítulo 2

Estimación directa

El Teorema de Glivenko-Cantelli es uno de los resultados fundamentales de la estadística, en dicho teorema se demuestra la convergencia uniforme de la función de distribución empírica de una muestra a la verdadera función de distribución de la variable aleatoria,

Teorema 1 (Glivenko-Cantelli) *$F_n(x)$ converge uniformemente a $F(x)$, es decir $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} P\{\sup_{-\infty<x<\infty} |F_n(x) - F(x)| > \epsilon\} = 0$$

donde

$$F_n(x) = \frac{\sharp\{X_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ \frac{k}{n} & \text{si } X_{(k)} \leq x < X_{(k+1)} \\ 1 & \text{si } x \geq X_{(n)} \end{cases} \tag{2.1}$$

y $X_1, \ldots, X_n$ son variables i.i.d. que corresponden a una muestra aleatoria de la variable $X \sim F$, y donde $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ es la correspondiente muestra ordenada (*order statistics*).

La función de distribución empírica se convierte así en un estimador inmediato para la función de distribución de la variable aleatoria. Entre sus propiedades podemos destacar: su forma escalonada, el ser un estimador insesgado para cada x

$$\mathrm{E}F_n(x) = \mathrm{E}I_{(-\infty,x]}(X) = 1 \times P(X \in (-\infty, x])) = F(x) \tag{2.2}$$

y que no existe ningún otro estimador insesgado con menor varianza. Esto último debido a que la muestra ordenada es un estadístico suficiente completo y $F_n(x)$ es insesgado y función del estadístico suficiente. El gran incoveniente que encontramos es que mientras la función de distribución de la varible puede ser continua, $F_n(x)$ siempre es discontinua.

A partir de $F_n(x)$ podemos construir una estimación directa de la función de densidad a través de la relación $f(x) = F(x)'$, resultando la denominada *función de densidad de probabilidad empírica*

$$f_n(x) = \frac{d}{dx} F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - X_i) \tag{2.3}$$

donde $\delta(x)$ es la función delta de Dirac, definida por

$$\int_{-\infty}^{\infty} \delta(y)dy = 1, \qquad \delta(y) = 0 \text{ si } y \neq 0$$

Resulta por tanto una densidad uniforme discreta en los puntos de la muestra con masa de probabilidad n^{-1} en cada punto, poco útil tanto desde el punto de vista del análisis gráfico como en aplicaciones derivadas.

Se hace por tanto imprescindible la introducción de versiones modificadas de la estimación. Una de las alternativas es la utilización del clásico histograma como estimador de la densidad, ya desde el punto de vista gráfico resulta más adecuado que la función de densidad empírica y es el estimador por el que de forma tradicional comienzan su estudio los tratados sobre estimación no paramétrica. Es obligado mencionar también modificaciones de (2.3) como la introducida sin demasiados comentarios por Rosenblatt (1956)

$$f_n(x) = \frac{F_n(x + h) - F_n(x - h)}{2h} \tag{2.4}$$

donde $h = h_n$ verifica $h_n \to 0$ si $n \to \infty$, y que a través de generalizaciones ha dado lugar a la popular *estimación kernel* de la densidad. De hecho (2.4) puede ser considerada una modificación del histograma donde cada punto es el centro de un intervalo de anchura $2h$, obviando el problema que supone escoger los puntos de origen de los intervalos del histograma tradicional; volveremos sobre este tema y este estimador más adelante.

Capítulo 3

Criterios de error para estimaciones de la densidad

Es inevitable la selección de criterios que nos permitan comparar entre varios estimadores en la búsqueda del estimador óptimo para un problema determinado. Hasta el presente, y debido a la subjetividad de la elección de los criterios de error, no se ha llegado a un consenso entre los diversos investigadores del área existiendo dos grandes líneas que optan por criterios que minimizan el error absoluto o el error cuadrático de la estimación.

Cuando utilizamos estimadores sesgados en una estimación paramétrica, el criterio de minimizar la varianza es, a veces, substituido por el criterio de minimizar el error cuadrático medio (MSE), que es la suma de la varianza y del sesgo al cuadrado. Cuando trabajamos con estimaciones de la función de densidad el criterio es:

$$MSE\{\hat{f}(x)\} = \mathrm{E}[\hat{f}(x) - f(x)]^2 = \ \mathrm{Var}\{\hat{f}(x)\} + \mathrm{Sesgo}^2\{\hat{f}(x)\} \tag{3.1}$$

donde $\mathrm{Sesgo}^2\{\hat{f}(x)\} = E[\hat{f}(x)] - f(x)$. Esta ecuación afronta el problema de la estimación no paramétrica de una forma puntual standard con paramétro desconocido $\theta = f(x)$. Sin embargo el interés de la estimación no paramétrica radica en obtener una estimación y representación de la densidad completa, por tanto se hace necesario recurrir a criterios de error globales como pueden ser:

- Norma L_∞

$$\sup_x |\hat{f}(x) - f(x)| \tag{3.2}$$

- Norma L_1, también denominada *error absoluto integrado*

$$IAE = \int |\hat{f}(x) - f(x)| dx \tag{3.3}$$

- Norma L_2, también denominada *error cuadrático integrado* (ISE)

$$ISE\{\hat{f}(x)\} = \int [\hat{f}(x) - f(x)]^2 \tag{3.4}$$

Nos centraremos en este último, por ser el más ampliamente utilizado y por su facilidad de manipulación. Sin embargo existen detractores de la utilización del error L_2 debido a que no refleja de modo adecuado la proximidad entre $\hat{f}(x)$ y $f(x)$ y propugnan la utilización de la norma de L_1 (Devroye y Gyorfi (1985)). Más adelante comentaremos alguna diferencia entre el enfoque L_1 y el L_2.

El ISE es una variable aleatoria que depende de la verdadera (y desconocida) densidad, el estimador utilizado y el tamaño muestral. Incluso con estas tres cantidades fijadas, el ISE es una función de una realización particular de n puntos. Es más realista plantearse como criterio de error un promedio del ISE sobre las diversas realizaciones

$$MISE\{\hat{f}(x)\} = \mathrm{E}[ISE\{\hat{f}(x)\}] = \int \mathrm{E}[\hat{f}(x) - f(x)]^2 \tag{3.5}$$

De todos modos no existe una diferencia práctica importante entre el ISE y el MISE. Los métodos basados en el MISE se comportan también bien desde el punto de vista del ISE (Grund, Hall y Marron (1994)).

Dado que el integrando es no-negativo, el orden de la integración y la esperanza pueden intercambiarse aplicando el teorema de Fubini, para dar lugar a las formas alternativas

$$\begin{aligned} MISE\{\hat{f}(x)\} &= \int \mathrm{E}[\hat{f}(x) - f(x)]^2 dx = \int \mathrm{E}[\hat{f}(x) - f(x)]^2 dx = \int MSE\{\hat{f}(x)\} dx = \\ &\int \mathrm{Var}\{\hat{f}(x)\} dx + \int \mathrm{Sesgo}^2\{\hat{f}(x)\} dx \equiv IMSE\{\hat{f}(x)\} \end{aligned} \tag{3.6}$$

De esta forma el MISE tiene dos interpretaciones diferentes pero equivalentes: es una medida del error global promedio y de error puntual acumulado. Este criterio podria se modificado introduciendo un peso que por ejemplo diera más énfasis a las colas o a un determinado intervalo.

$$MISE_w\{\hat{f}(x)\} = \int \mathrm{E}[\hat{f}(x) - f(x)]^2 w(x) dx \tag{3.7}$$

Análogamente es posible considerar la esperanza del IAE denominado MIAE = E[IAE]. En los trabajos de Devroye se presentan algunos resultados importantes el comportamiento asintótico de IAE y MIAE. Hall y Wand (1988) obtienen una expresión asintótica general para el MIAE y muestran que su minimización se reduce a resolver numéricamente una determinada ecuación. De todos modos la labor técnica para obtener resultados en L_1 es de mayor dificultad que para resultados análogos en L_2.

Existen otras alternativas como criterios de error, destaquemos entre ellas:

- El criterio de Kullback-Leibler

$$\int \hat{f} \log(\hat{f}/f) \tag{3.8}$$

- La distancia de Hellinger

$$[\int (\hat{f}^{1/p} - f^{1/p})^p]^{1/p} \tag{3.9}$$

- La variación total

$$TV(\hat{f}, f) = \sup_A |\int_A \hat{f} - \int_A f| \tag{3.10}$$

- Otras normas L_p

$$\left\{ \int |\hat{f} - f|^p \right\}^{1/p} \quad 0 < p < \infty \tag{3.11}$$

Finalmente algunos comentarios sobre la utilización de la norma L_1 frente a la L_2.

- El criterio L_1 da más importancia a las colas de la densidad que el criterio L_2, este último al elevar al cuadrado resta importancia a valores pequeños de la densidad.

- Otra diferencia es que la norma L_1 es una cantidad adimensional, ya que la densidad tiene como unidades el inverso de la longitud. Mientras que la norma L_2 mantiene como unidades las unidades de la densidad, aunque existen métodos de adimensionalizarla no son totalmente satisfactorios (Scott (1992)).

- Una tercera diferencia entre ambas es que L_1 es invariante frente a cambios de escala monótonos y contínuos. Supongamos que $X \sim f$ y que $Y \sim g$ y definamos $X^* = h(X)$ $Y^* = h(Y)$; entonces $f^* = f[h^{-1}(x^*)]|J|$ y una expresión similar para g^*, un cambio de variable produce

$$\int_u |f^*(u) - g^*(u)| du = \int_u |f[h^{-1}(u)] - g[h^{-1}(u)]||J| du = \int_v |f(v) - g(v)| dv \tag{3.12}$$

  Todo lo anterior hace que L_1 sea más fácil de interpretar que L_2, en particular es posible comparar la dificultad de estimar diferentes densidades.

- Cabe destacar también que $0 \leq L_1 \leq 2$ mientras que $0 \leq L_2 \leq \infty$.

$$\int |\hat{f}(x) - f(x)| dx \leq \int |\hat{f}(x)| dx + \int |f(x)| dx \leq 2$$

- También existe una conexión entre la norma L_1 y las técnicas de clasificación y discriminación, puede demostrarse (Scott (1992)) que si un dato puede provenir al azar de dos densidades f y g y utilizamos una regla bayesiana de la forma: asignar x a f si $x \in A$ para algún conjunto A y a g en caso contrario, la probabilidad de clasificación errónea es:

$$\Pr(\text{error}) = \frac{1}{2}\int_{A^c} f + \frac{1}{2}\int_A g = \frac{1}{2}\left(1 - \int_A f\right) + \frac{1}{2}\int_A g = \frac{1}{2} - \frac{1}{2}\int_A (f-g). \quad (3.13)$$

Y si elegimos A de forma que se minimize el error $A = B = \{x : f(x) > g(x)\}$ y se llega a

$$\Pr(\text{error}) = \frac{1}{2} - \frac{1}{2}\int_B (f-g) = \frac{1}{2} - \frac{1}{4}\int |f-g| \quad (3.14)$$

Donde en el último paso hemos utilizado el lema de Scheffé

$$\int |f(x) - g(x)|dx = 2TV(f,g) = 2\int (f(x) - g(x))_+ dx \quad (3.15)$$

$$= 2\int (g(x) - f(x))_+ dx \quad (3.16)$$

donde el subíndice + indica parte positiva.

La demostración de este último lema la descomponemos en dos partes

1.

$$\int (f-g)_+ = \int_B f - g = \int_B f - \int_B g = 1 - \int_{B^c} f - 1 + \int_{B^c} g = \int_{B^c} g - f = \int (g-f)_+$$

2.

$$\int |f-g| = \int_B f - g + \int_{B^c} g - f = \int_B f - g + \int_B f - g = 2\int_B (f-g)$$

De modo que minimizar la distancia entre $g = \hat{f}$ y f es equivalente a maximizar la probabilidad de error anterior buscando la mayor *confusión* entre f y $\hat{f}$, que es precisamente lo que deseamos.

Por otro lado, en situaciones prácticas los estimadores que optimizan estos criterios son similares. Devroye y Györfi (1985) han realizado un abundante tratamiento teórico basado en L_1, sin embargo por la simplicidad analítica del error cuadrático y su utilidad en situaciones prácticas es el preferido por muchos autores. Algunos resultados asintóticos de Hall y Wand (1988) y Scott y Wand (1991) refuerzan la idea de que las diferencias prácticas entre los dos criterios son razonablemente pequeñas excepto en situaciones extremas.

Capítulo 4

Histogramas

Es el más sencillo y mejor conocido de los estimadores no paramétricos de la densidad. Muchos autores distinguen la utilización del histograma como técnica de representación de datos o como estimador de la densidad, la diferencia básica es que en este último caso debe estar normalizado para integrar 1.

Supongamos que f tiene soporte en $[a, b]$ generalmente deducido de los datos, efectuamos una partición en k intervalos no solapados $B_i = [t_i, t_{i+1})$ $i = 1, \ldots, k$ donde $a = t_1 < t_2 < \ldots < t_{k+1} = b$, el histograma viene definido por

$$\hat{f}(x) = \sum_{i=1}^{k} \frac{N_i/n}{t_{i+1} - t_i} I_{B_i}(x) \tag{4.1}$$

donde N_i es el número de datos dentro de B_i. Si la longitud de los intervalos es siempre la misma $h_n = t_{i+1} - t_i$, valor que denominaremos anchura del intervalo o *ancho de ventana*,la expresión resulta

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{k} N_i I_{B_i}(x) \tag{4.2}$$

o en forma equivalente

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} I_{B_i}(x_i) = \frac{N_i}{nh_n} \quad x \in B_i \tag{4.3}$$

4.1. Regla de Sturges

Aplicada por defecto en muchos paquetes estadísticos. Tomemos una distribución binomial de parámetros $B(k-1, 1/2)$

$$1 = \sum_{i=0}^{k-1} \left(\begin{array}{c} k-1 \\ i \end{array} \right) \frac{1}{2}^i \frac{1}{2}^{k-1-i} \Rightarrow \sum_{i=0}^{k-1} \left(\begin{array}{c} k-1 \\ i \end{array} \right) = 2^{k-1},$$

supongamos que el número de individuos para cada valor es $N_i = \begin{pmatrix} k-1 \\ i \end{pmatrix}$, tenemos entonces que

$$n = \sum_{i=0}^{k-1} \begin{pmatrix} k-1 \\ i \end{pmatrix} = 2^{k-1} \Rightarrow k = 1 + \log_2 n$$

4.2. Propiedades estadísticas

4.2.1. Error cuadrático medio y consistencia

De las definiciones

$$N_i \sim B(n, p_i) \qquad p_i = \int_{B_i} f(t)dt$$

Vamos a calcular $MSE(\hat{f}(x)) = \mathrm{Var}(\hat{f}(x)) + \mathrm{sesgo}^2(\hat{f}(x))$ tomando $\hat{f}(x) = \frac{N_i}{nh_n}$ para $x \in B_i$.

$$\mathrm{Var}(\hat{f}(x)) = \frac{\mathrm{Var}(N_i)}{n^2h^2} = \frac{p_i(1-p_i)}{nh^2} \tag{4.4}$$

$$\mathrm{sesgo}(\hat{f}(x)) = \mathrm{E}(\hat{f}(x)) - f(x) = \mathrm{E}\left(\frac{N_i}{nh}\right) - f(x) = \frac{p_i}{h} - f(x) \tag{4.5}$$

y por tanto

$$MSE = \frac{p_i(1-p_i)}{nh^2} + \left(\frac{p_i}{h} - f(x)\right)^2 \tag{4.6}$$

Teniendo presente que por el teorema del valor medio

$$p_i = \int_{B_i} f(t)dt = h \cdot f(\xi_i) \quad \xi_i \in B_i$$

deducimos que

$$\frac{p_i(1-p_i)}{nh^2} \leq \frac{p_i}{nh^2} = \frac{f(\xi_i)}{nh}$$

y si f es Lipschitz en B_i, es decir

$$|\frac{p_i}{h} - f(x)| \leq \gamma_i|\xi_i - x| \leq \gamma_i h \quad \gamma_i > 0$$

se llega a

$$MSE \leq \frac{f(\xi_i)}{nh} + \gamma_i^2 h^2 \tag{4.7}$$

y consecuentemente se verifica el siguiente

Teorema 2 *Dado $x \in B_i$ y si f es Lipschitz en B_i, entonces $\hat{f}(x)$ es consistente en media cuadrática si*

$$n \to \infty \Rightarrow h \to 0 \quad i \quad nh \to \infty$$

Minimización de (4.7) Vamos a buscar el valor de h que minimice la expresión del error cuadrático medio

$$\frac{-nf(\xi_i)}{n^2h^2} + 2\gamma_i^2 h = 0 \Rightarrow h^* = \left(\frac{f(\xi_i)}{2\gamma_i^2 n}\right)^{1/3}$$

y por tanto

$$MSE^*(h^*) = \frac{f(\xi_i)}{n}\left(\frac{2\gamma_i^2 n}{f(\xi_i)}\right)^{1/3} + \gamma_i^2\left(\frac{f(\xi_i)}{2\gamma_i^2 n}\right)^{2/3}$$

Como vemos la anchura de intervalo óptima decrece a un ritmo proporcional a $n^{-1/3}$ y el MSE es $O(n^{-2/3})$, sin alcanzar la tasa de la cota de Cramer-Rao en estimadores paramétricos $O(n^{-1})$.

4.2.2. Obtención del MISE exacto

$$MISE = \int \mathrm{E}[\hat{f}(x) - f(x)]^2 = \int_{-\infty}^{\infty} \mathrm{Var}[\hat{f}(x)]dx + \int_{-\infty}^{\infty} \mathrm{Sesgo}^2[\hat{f}(x)]dx = IV + ISB \quad (4.8)$$

$$IV = \int_{-\infty}^{\infty} \mathrm{Var}[\hat{f}(x)]dx = \sum_i \int_{B_i} \mathrm{Var}[\hat{f}(x)]dx = \sum_i \int_{B_i} \frac{p_i(1-p_i)}{nh^2}dx \quad (4.9)$$

$$= \sum_i \frac{p_i(1-p_i)}{nh} = \frac{1}{nh}\left[\sum_i p_i - \sum_i p_i^2\right] \quad (4.10)$$

y teniendo en cuenta

$$\sum_i p_i = \int_{-\infty}^{\infty} f(t)dt = 1$$

y

$$\sum_i p_i^2 = \sum_i \left(\int_{B_i} f(t)dt\right)^2 = \sum_i (hf(\xi_i))^2 \quad (4.11)$$

$$= \sum_i h^2 f^2(\xi_i) = h\left[\int f^2(x)dx + o(1)\right] \quad (4.12)$$

se llega a

$$IV = \frac{1}{nh} - \frac{1}{n}R(f) + o(n^{-1}) \quad (4.13)$$

donde

$$R(f) = \int f^2 dx \tag{4.14}$$

o también podemos considerar

$$IV = \frac{1}{nh} - \frac{\sum_i p_i^2}{nh} \tag{4.15}$$

La expresión para el sesgo es

$$ISB = \int_{-\infty}^{\infty} \text{Sesgo}^2[\hat{f}(x)]dx = \int_{\Re} \frac{p_i^2}{h^2} - 2\frac{p_i}{h}f(x) + f(x)^2 dx \tag{4.16}$$

$$= \sum_i \int_{B_i} \frac{p_i^2}{h^2} - 2\frac{p_i f(x)}{h} + f(x)^2 dx \tag{4.17}$$

$$= \sum_i \frac{p_i^2}{h^2}h - \frac{2}{h}\sum_i p_i \int_{B_i} f(t)dt + \int_{\Re} f(x)^2 dx = R(f) - \frac{\sum_i p_i^2}{h} \tag{4.18}$$

y finalmente de (4.15) y (4.18) se llega a

$$MISE = IV + ISB = \frac{1}{nh} - \frac{\sum_i p_i^2}{nh} + R(f) - \frac{\sum_i p_i^2}{h} \tag{4.19}$$

$$= \frac{1}{nh} - \frac{n+1}{nh}\sum_i p_i^2 + R(f) \tag{4.20}$$

4.2.3. Obtención del MISE asintótico

De (4.13)

$$IV = \frac{1}{nh} - \frac{1}{n}R(f) + o(n^{-1})$$

y por tanto

$$AIV = \frac{1}{nh} \tag{4.21}$$

Por otro lado

$$\text{Sesgo}[\hat{f}(x)] = \frac{p_i}{h} - f(x) \tag{4.22}$$

Tengamos presente que $h \to 0$ y hagamos el siguiente desarrollo para p_k

$$p_k = \int_{t_k}^{t_k+h} f(t)dt = \int_{t_k}^{t_k+h} f(x) + f'(x)(t-x) + \frac{1}{2}f''(x)(t-x)^2 + \cdots dt$$

$$= hf(x) + f'(x)\left(\frac{h^2}{2} + (t_k - x)h\right) + O(h^3) \tag{4.23}$$

y por tanto

$$\text{Sesgo}[\hat{f}(x)] = \frac{p_k}{h} - f(x) = \left(\frac{h}{2} + (t_k - x)\right) f'(x) + O(h^2) \quad x \in B_k$$

$$\int_{B_k} \left(\frac{h}{2} + (t_k - x)\right)^2 f'^2(x)dx = f'^2(\eta_k) \int_{B_k} \left(\frac{h}{2} + (t_k - x)\right)^2 dx = f'^2(\eta_k)\frac{h^3}{12}$$

esta última aproximación por el teorema del valor medio generalizado. Resulta finalmente

$$ISB = \frac{h^2}{12} \sum_i f'^2(\eta_i)h = \frac{h^2}{12} \int f'^2(x)dx + o(h^2) \tag{4.24}$$

y utilizando la nomenclatura introducida en (4.14)

$$AISB = \frac{h^2}{12} R(f') \tag{4.25}$$

La expresión aproximada para el error cuadrático medio integrado es finalmente

$$AMISE = \frac{1}{nh} + \frac{h^2}{12} R(f') \tag{4.26}$$

El valor de h que minimiza (4.26) es

$$h^* = \left(\frac{6}{R(f')}\right)^{1/3} n^{-1/3} \tag{4.27}$$

y el AMISE resultante es

$$AMISE^* = (3/4)^{2/3}[R(f')]^{1/3} n^{-2/3} \tag{4.28}$$

4.2.4. Influencia de la anchura de ventana en el MISE

Consideremos $h = ch^*$, tenemos

$$\frac{AMISE(ch^*)}{AMISE(h^*)} = \frac{n^{-2/3}R(f')^{1/3}6^{-1/3}(1/c + c^2/2)}{(3/4)^{2/3}R(f')^{1/3}n^{-2/3}} = \frac{2 + c^3}{3c} \tag{4.29}$$

En la Tabla 4.1 tenemos algunos valores obtenidos de la expresión anterior.

De los resultados de la tabla se observa que una oscilación de aproximadamente un 33 % del valor óptimo produce un incremento bastante aceptable del AMISE. También se observa que el AMISE es más sensible a anchuras de ventana superiores a la óptima (c=2 vs. c=1/2), es decir el AMISE es más sensible a valores superiores del sesgo que a valores superiores de la varianza.

c	$\frac{2+c^3}{3c}$
1/2	1,42
3/4	1,08
1	1
4/3	1,09
2	1,67

Cuadro 4.1: Sensibilidad del AMISE a la anchura de ventana.

4.3. Elección del ancho de ventana

De (4.27) el valor óptimo del ancho de ventana para minimizar el AMISE es

$$h^* = \left(\frac{6}{R(f')}\right)^{1/3} n^{-1/3} \tag{4.30}$$

Al intervenir el factor desconocido $R(f')$ es necesario introducir algún algoritmo que permita escoger h de forma práctica.

4.3.1. Referencia a la distribución Normal

Consideremos $f \sim N(\mu, \sigma)$, entonces

$$R(f') = \frac{1}{4\sqrt{\pi}\sigma^3} \tag{4.31}$$

resultando

$$h^* = \left(\frac{24\sqrt{\pi}\sigma^3}{n}\right)^{1/3} \simeq 3,5\sigma n^{-1/3} \tag{4.32}$$

Sustituyendo σ por su estimación obtenemos

$$\hat{h}_1 = 3,5\hat{s}n^{-1/3} \tag{4.33}$$

Esta regla es muy estable puesto que $\hat{s} \to \sigma$ a un ritmo superior a $O(n^{-1/3})$. Una regla algo más robusta es la propuesta por Freedman y Diaconis (1981)

$$\hat{h}_2 = 2(IQ)n^{-1/3} \tag{4.34}$$

donde IQ es el rango intercuartílico. Si la distribución es realmente Normal, $\hat{h}_2$ es un 77 % de $\hat{h}_1$. En la Tabla 4.2 mostramos el número de intervalos que determinan la regla de Sturges y las dos estimaciones anteriores suponiendo datos normales y un histograma construido entre (-3,3)

n	Sturges	$\hat{h}_1$	$\hat{h}_2$
50	5,6	6,3	8,5
100	7,6	8	10,8
500	10	13,6	18,3
1000	11	17,2	23,2
100000	17,6	79,8	107,6

Cuadro 4.2: Número de intervalos con tres diferentes métodos.

4.3.2. Cota superior para el ancho de ventana

De (4.27) se observa que cualquier límite inferior para $R(f')$ conduce a un límite superior para el ancho de ventana, nos planteamos el siguiente problema de optimización:

$$\min_f \int_{\Re} f'^2(x)dx \quad s/t \quad \text{Var de } f = \sigma^2 \tag{4.35}$$

Terrell (1990) encuentra como solución

$$f_2(x) = \frac{15}{16\sqrt{7}\sigma}\left(1 - \frac{x^2}{7\sigma^2}\right)^2 I_{[-\sqrt{7}\sigma,\sqrt{7}\sigma]}$$

Resultando

$$R(f_2') = 15\sqrt{7}/343\sigma^3$$

y

$$h^* \leq \left(\frac{686\sigma^3}{5\sqrt{7}n}\right)^{1/3} \simeq 3,729\sigma n^{-1/3} = h_{os} \tag{4.36}$$

reemplazando σ por su estimación, o considerando

$$h_{os} = 2,603(IQ)n^{-1/3} \tag{4.37}$$

4.4. Propiedades estadísticas con la norma L_1

No existe una descomposición tan simple en sesgo-varianza. Hjort (1986) muestra

$$\mathrm{E}\int|\hat{f}-f| \leq \mathrm{E}\int|\hat{f}-\mathrm{E}\hat{f}| + \int|\mathrm{E}\hat{f}-f| \simeq \sqrt{\frac{2}{\pi nh}}\int\sqrt{f} + \frac{1}{4}h\int|f'| \tag{4.38}$$

y minimizando

$$h^* = 2\pi^{-1/3}\left[\int f^{1/2} \div \int|f'|\right]^{2/3} n^{-1/3} = 2,717\sigma n^{-1/3} \tag{4.39}$$

Criterio de Error	Ancho óptimo	Error esperado
L_1 límite superior	$2,72n^{-1/3}$	$1,6258n^{-1/3}$
L_1 simulación numérica	$3,37n^{-1/3}$	$1,1896n^{-1/3}$
L_2	$3,49n^{-1/3}$	$(0,655n^{-1/3})^2$

Cuadro 4.3: Anchos de ventana óptimos para varios criterios de error con datos N(0,1).

para datos normales. Por simulación numérica se ha encontrado como valor óptimo

$$h^* = 3,37\sigma n^{-1/3}$$

Presentamos en la tabla 4.3 una comparación de los anchos de ventana con diferentes criterios para datos N(0,1).

4.5. Influencia del origen de los intervalos

No existen demasiados estudios sobre el efecto del cambio de origen de los intervalos. El MISE es prácticamente insensible a la posición de origen (*anchor position*) excepto si una discontinuidad de la densidad es cruzada por un intervalo en lugar de coincidir con un extremo del mismo, vease Simonoff (1995) o Scott (1992). El efecto sí que se hace notar en el aspecto del histograma para muestras finitas, por ejemplo en el número de modas. Según datos de Scott (1992) para una distribución N(0,1) con ancho de ventana óptimo el MISE se minimiza si x=0 está en el centro de un intervalo, sin embargo si estuviera en un extremo, la diferencia en el MISE sería del orden del 1,09 % para un tamaño muestral de 25 y menor de 10^{-5} para un tamaño muestral superior a 100.

Los principales problemas de la estimación por histogramas: la discontinuidad de la estimación y la dependencia del origen de los intervalos, han motivado la aparición de métodos alternativos como los que estudiamos en la siguiente sección: los *Polígonos de frecuencias* y los **ASH .ªveraged Shifted histograms"**.

4.6. Problemas

1. Suponiendo datos provenientes de una distribución N(0,1), calcular el MISE exacto del estimador $N(\bar{x}, 1)$. Comparar para diversos tamaños muestrales con el AMISE del estimador histograma.

2. Demostrar que cuando $h = h^*$ en el histograma, la contribución al AMISE de la IV y del ISB está en proporción 2:1.

3. Construir un programa en MATLAB que tras leer unos datos calcule el ancho de ventana óptimo según la referencia a la distribución Normal y efectúe la representación gráfica del histograma resultante.

4. Examinar con ayuda de MATLAB el efecto que, sobre un histograma con ancho de ventana fijo, tiene el cambio del origen de los intervalos.

Capítulo 5

Polígonos de Frecuencia

Representan una estimación continua de la densidad derivada de los histogramas mediante una interpolación a partir de los puntos medios de los intervalos de igual longitud. Aun cuando muchos autores los consideran equivalentes, incluso con confusión en la terminologia aplicando el término histograma a ambas estimaciones, podremos observar como las propiedades estadísticas son notablemente diferentes a las que presentan los histogramas convencionales.

5.1. Definición

El polígono de frecuencias (PF) conecta dos valores adyacentes del histograma entre sus puntos medios del intervalo.

$$\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)\hat{f}_1 \qquad -\frac{h}{2} \leq x < \frac{h}{2} \tag{5.1}$$

Véase la Figura 5.1

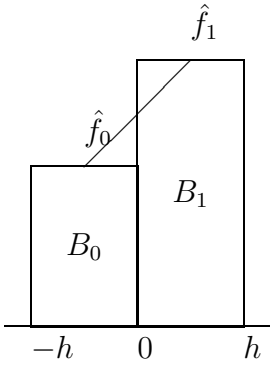

Figura 5.1: Polígono de frecuencias derivado de un histograma.

5.2. MISE

Recordemos que $\hat{f}_i = \frac{N_i}{nh}$ por tanto $\mathrm{E}\hat{f}_i = \frac{p_i}{h}$. Desarrollemos por Taylor

$$f(x) \approx f(0) + f'(0)x + \frac{1}{2}f''(0)x^2$$

y las aproximaciones para p_0 y p_1 se pueden obtener de

$$p_0 = \int_{-h}^{0} f(s)ds \approx hf(0) - \frac{1}{2}h^2 f'(0) + \frac{1}{6}h^3 f''(0) \tag{5.2}$$

$$p_1 = \int_{-h}^{0} f(s)ds \approx hf(0) + \frac{1}{2}h^2 f'(0) + \frac{1}{6}h^3 f''(0) \tag{5.3}$$

resultando, recordando que x no es aleatorio,

$$\mathrm{E}(\hat{f}(x)) = \left(\frac{1}{2} - \frac{x}{h}\right)\frac{p_0}{h} + \left(\frac{1}{2} + \frac{x}{h}\right)\frac{p_1}{h} \approx f(0) + xf'(0) + \frac{h^2 f''(0)}{6} \tag{5.4}$$

dando un sesgo

$$\mathrm{Sesgo} = \mathrm{E}(\hat{f}(x)) - f(x) \approx \frac{1}{6}(h^2 - 3x^2)f''(0) \tag{5.5}$$

e integrando en el intervalo $(-h/2, h/2)$ resulta

$$ISB = \int_{-h/2}^{h/2} \mathrm{Sesgo}^2 dx = \frac{49h^4 f''(0)^2}{2880}h \tag{5.6}$$

expresión similar para el resto de intervalos. Finalmente sumando sobre todos los intervalos y utilizando la aproximación Riemanniana

$$ISB = \sum_k \frac{49}{2880}h^4 f''(kh)h = \frac{49}{2880}h^4 R(f'') + O(h^6) \tag{5.7}$$

Vemos como el sesgo al cuadrado es de orden significativamente menor que el del histograma $O(h^2)$.

El cálculo de la varianza es parecido

$$\mathrm{Var}(\hat{f}(x)) = \left(\frac{1}{2} - \frac{x}{h}\right)^2 \mathrm{Var}\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)^2 \mathrm{Var}\hat{f}_1 + 2\left(\frac{1}{4} - \frac{x^2}{h^2}\right)\mathrm{Cov}(\hat{f}_0, \hat{f}_1) \tag{5.8}$$

siendo

$$\mathrm{Var}(\hat{f}_i) = \frac{np_i(1-p_i)}{(nh)^2} \approx \frac{f(0)(1 - hf(0))}{nh}$$

y

$$\mathrm{Cov}(\hat{f}_0, \hat{f}_1) = \frac{-np_0p_1}{(nh)^2} \approx -\frac{f(0)^2}{n}$$

Substituyendo en (5.8) resulta

$$\mathrm{Var}(\hat{f}(x)) = \left(\frac{2x^2}{nh^3} + \frac{1}{2nh}\right) f(0) - \frac{f(0)^2}{n} + o(n^{-1}) \tag{5.9}$$

e integrando en el intervalo tal y como hicimos con el sesgo

$$IV = \int_{-h/2}^{h/2} \mathrm{Var}\, dx = \left(\frac{2f(0)}{3nh} - \frac{f(0)^2}{n}\right) h \tag{5.10}$$

resultando finalmente

$$IV = \sum_k \left(\frac{2f(kh)}{3nh} - \frac{f(kh)^2}{n}\right) h = \frac{2}{3nh} - \frac{1}{n}R(f) + o(n^{-1}) \tag{5.11}$$

De (5.11) y (5.7) se obtiene el siguiente

Teorema 3 *Sea f'' absolutamente continua y $R(f''') < \infty$. Entonces*

$$AMISE = \frac{2}{3nh} + \frac{49}{2880}h^4R(f'') \tag{5.12}$$

Por tanto

$$h^* = 2\left(\frac{15}{49R(f'')}\right)^{1/5} n^{-1/5} \tag{5.13}$$

$$AMISE^* = \frac{5}{12}\left(\frac{49R(f'')}{15}\right)^{1/5} n^{-4/5} \tag{5.14}$$

Comparando con los resultados obtenidos en el histograma, el AMISE óptimo era de orden $O(n^{-2/3})$, siendo de orden $O(n^{-4/5})$ en el PF. Igualmente el ancho de ventana óptimo es mayor en el FP que en el histograma. Por ejemplo con 800 datos distribuidos normalmente, el ancho de ventana óptimo para el PF es un 50 % mayor que el del histograma.

5.3. Elección del ancho de ventana

De (5.13) el valor óptimo del ancho de ventana para minimizar el AMISE es

$$h^* = 2\left(\frac{15}{49R(f'')}\right)^{1/5} n^{-1/5}$$

5.3.1. Referencia a la distribución Normal

Consideremos $f \sim N(\mu, \sigma)$, entonces

$$R(f'') = \frac{3}{8\sqrt{\pi}\sigma^5}$$

resultando

$$h^* \approx 2,15\sigma n^{-1/5} \tag{5.15}$$

y

$$AMISE^* \approx 0,3870\sigma^{-1}n^{-4/5} \tag{5.16}$$

Substituyendo la desviación típica por su estimación obtenemos

$$\hat{h}_1 = 2,15\hat{s}n^{-1/5} \tag{5.17}$$

o también sustituyendo la estimación más robusta

$$\hat{\sigma} = IQ/1,348$$

5.3.2. Cota superior para el ancho de ventana

Siguiendo un proceso parecido al comentado en los histogramas, entre todas las densidades con varianza σ^2, la más suave (menor R(f")) es

$$f(x) = \frac{35}{96\sigma}\left(1 - \frac{x^2}{9\sigma^2}\right)^3 I_{[-3\sigma,3\sigma]}(x) \Rightarrow R(f'') \geq \frac{35}{243\sigma^5} \tag{5.18}$$

por tanto

$$h \leq \left(\frac{23328}{343}\right)^{1/5} \sigma n^{-1/5} = 2,33\sigma n^{-1/5} \equiv h_{os} \tag{5.19}$$

regla que representa un 108 % de la regla Normal.

Los problemas que presenta el PF sobre el origen de los intervalos, problemas todavía no corregidos, motivan la técnica siguiente, el ASH.

5.4. Problemas

1. Realizar un cuadro comparativo donde suponiendo datos provenientes de una distribución Normal podamos comparar el AMISE óptimo utilizando las técnicas del histograma y del PF. Crear también un cuadro donde para valores fijos del AMISE se nos presente el número de datos necesarios para alcanzarlo con cada técnica.

Capítulo 6

ASH *(Averaged Shifted Histogram)*

6.1. Definición básica

Consiste en promediar varios histogramas desplazados (ASH), el resultado es también un histograma que se puede hacer continua igual que se construyen los polígonos de frecuencias (FP-ASH).

Consideremos una colección de m histogramas $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m$ cada uno con anchura de intervalo h, pero con diferentes orígenes para los intervalos

$$t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \ldots, \frac{(m-1)h}{m}$$

respectivamente. Definimos el estimador ASH como

$$\hat{f}(\cdot) = \hat{f}_{ASH}(\cdot) = \frac{1}{m}\sum_{i=1}^{m} \hat{f}_i(\cdot) \tag{6.1}$$

Destaquemos que el valor del estimador es constante sobre intervalos de longitud $\delta = \frac{h}{m}$.

6.2. Propiedades asintóticas y definición general

Consideremos $B_k = [k\delta, (k+1)\delta]$ y ν_k = número de observaciones en B_k. La altura que presenta la estimación en cada B_k es un promedio de las alturas de los m histogramas desplazados

$$\frac{\nu_{k-m+1} + \cdots + \nu_k}{nh}, \ldots, \frac{\nu_k + \cdots + \nu_{k+m-1}}{nh}$$

por tanto una expresión general equivalente a la presentada en (6.1) será

$$\hat{f}(x;m) = \frac{1}{m}\sum_{i=1-m}^{m-1} \frac{(m-|i|)\nu_{k+i}}{nh} = \frac{1}{nh}\sum_{i=1-m}^{m-1}\left(1 - \frac{|i|}{m}\right)\nu_{k+i} \quad \text{para } x \in B_k \tag{6.2}$$

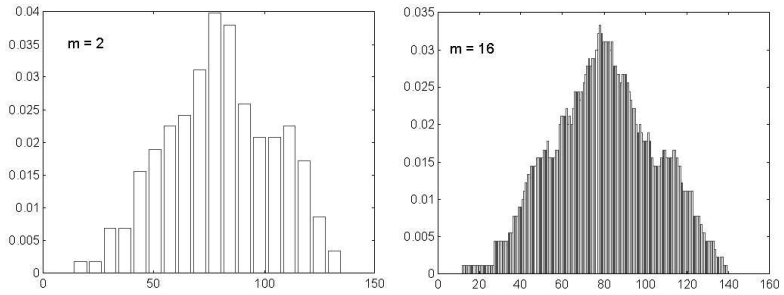

Figura 6.1: ASH para los datos de nevadas en Bufalo, h=13,5 m=2 y m=16 respectivamente.

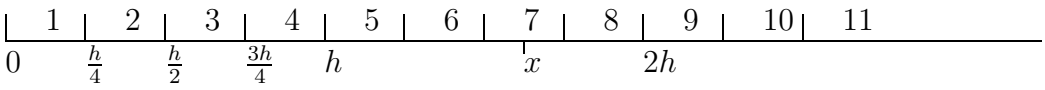

Figura 6.2: Ejemplo de 4 histogramas imbricados (m=4).

Véase la Figura 6.2.

A partir de la figura anterior tenemos que por ejemplo para $x \in B_7$,

$$\hat{f}(x) = \frac{1}{4}\left(\frac{\nu_4+\nu_5+\nu_6+\nu_7}{nh} + \frac{\nu_5+\nu_6+\nu_7+\nu_8}{nh} + \frac{\nu_6+\nu_7+\nu_8+\nu_9}{nh} + \frac{\nu_7+\nu_8+\nu_9+\nu_{10}}{nh}\right) \quad (6.3)$$

La expresión (6.2) puede generalizarse de la forma

$$\hat{f}(x;m) = \frac{1}{nh}\sum_{|i|<m} w_m(i)\nu_{k+i} \quad \text{para } x \in B_k \quad (6.4)$$

donde $w_m(i)$ son una serie de pesos arbitrarios con la única restricción que deben sumar m para que $\int \hat{f}(x;m)dx = 1$. En el caso particular de (6.2) los pesos adoptan la forma de un triángulo isósceles con base (-1,1), pero en general pueden definirse a través de

$$w_m(i) = m \times \frac{K(i/m)}{\sum_{j=1-m}^{m-1} K(j/m)} \quad i = 1-m,\ldots,m-1 \quad (6.5)$$

donde K es una función continua definida en (-1,1).

6.3. Aproximación para $m \to \infty$

Si $m \to \infty$ podemos aislar el efecto de un solo punto x_j sobre la estimación. Si $x \in B_k$ i $x_j \in B_{k+i}$, la influencia de x_j sobre x es proporcional a

$$1 - \frac{|i|}{m} = 1 - \frac{|i|}{m} \cdot \frac{\delta}{\delta} = 1 - \frac{|i|\delta}{h} \approx 1 - \frac{|x - x_j|}{h} \quad \text{si } |x - x_j| < h$$

Si $x_j \notin (x - h, x + h)$ la influencia es 0, (ver Figura 6.2), por tanto

$$\lim_{m\to\infty} \hat{f}(x; m) = \frac{1}{nh} \sum_{j=1}^{n} \left(1 - \frac{|x - x_j|}{h}\right) I_{[-1,1]} \left(\frac{x - x_j}{h}\right) \tag{6.6}$$

y si definimos un peso

$$w(x) = \begin{cases} (1 - |x|) & |x| < 1 \\ 0 & \text{en caso contrario.} \end{cases} \tag{6.7}$$

que corresponde a una densidad en forma de triángulo isósceles, resulta

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w \left(\frac{x - x_i}{h}\right) \tag{6.8}$$

expresión que puede generalizarse con cualquier función peso que represente una densidad y que corresponde a la denominada estimación tipo Núcleo (*Kernel*) que desarrollaremos en la siguiente sección.

6.4. Problemas

1. Demostrar que en (6.4) los pesos $w_m(i)$ deben sumar m para que la densidad estimada integre a 1.

2. Demostrar que si en (6.8) los pesos son nonegativos e integran a 1 la estimación resultant es una verdadera densidad.

Capítulo 7

Naive Estimator (Rosenblatt 1956)

Rosenblatt (1956) propone como estimador

$$\hat{f}(x) = \frac{\sharp\{X_i : X_i \in (x-h, x+h]\}}{2hn} = \frac{F_n(x+h) - F_n(x-h)}{2h} \tag{7.1}$$

basado en que

$$f(x) = \lim_{h\to 0} \frac{1}{2h} P(x-h < X \leq x+h) \tag{7.2}$$

Sabemos que

$$\mathrm{E}F_n(x) = F(x)$$

y por tanto

$$\mathrm{E}\hat{f}(x) = \frac{1}{2h}(F(x+h) - F(x-h))$$

y

$$\lim_{h\to 0} \mathrm{E}\hat{f}(x) = f(x)$$

Además el numerador en (7.1) sigue una distribución $B(n, F(x+h) - F(x-h))$ y se sigue que

$$\mathrm{Var}\hat{f}(x) = \frac{1}{4h^2n^2} n[F(x+h) - F(x-h)][1 - (F(x+h) - F(x-h))] \tag{7.3}$$

por tanto

$$\lim_{h\to 0} \mathrm{Var}\hat{f}(x) = \frac{f(x)}{2hn}$$

Se desprende que la estimación es consistente bajo las condiciones del teorema 2. Puede demostrarse que

$$AMISE = \frac{1}{2hn} + \frac{h^4}{36} R(f'') \tag{7.4}$$

y que dicho valor es minimizado por

$$h^* = \left(\frac{9}{2R(f'')n}\right)^{1/5} \tag{7.5}$$

y el AMISE resultante es

$$AMISE^* = 2^{-4/5}9^{-1/5}\frac{5}{4}(R(f''))^{1/5}n^{-4/5} \tag{7.6}$$

Destaquemos que el estimador definido en (7.1) es discontinuo en $X_i \pm h$ y es constante entre esos valores.

El estimador definido en (7.1) puede generalizarse de la forma

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x - X_i}{h}\right) \tag{7.7}$$

donde

$$w(y) = \begin{cases} 1/2 & y \in [-1, 1) \\ 0 & \text{en caso contrario} \end{cases} \tag{7.8}$$

dado que (7.1) puede también expresarse como

$$\hat{f}(x) = \frac{1}{2hn}\sum_{i=1}^{n} I(x - h < X_i \leq x + h) = \frac{1}{2hn}\sum_{i=1}^{n} I(X_i - h \leq x < X_i + h)$$

La función w puede ser una función más general K función núcleo o función Kernel resultando

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{7.9}$$

Capítulo 8

Estimación tipo Núcleo.

8.1. Definición

Dada la muestra de n observaciones reales $X_1, \ldots, X_n$ definiremos la estimación tipo Núcleo de función núcleo K como

$$\hat{f}_n(x) = \frac{1}{n\,h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{8.1}$$

donde $K(x)$ es una función, denominada función *Kernel* función núcleo o función peso, que satisface ciertas condiciones de regularidad, generalmente es una función de densidad simétrica como por ejemplo la de la distribución normal, y $\{h_n\}$ es una secuencia de constantes positivas conocidas como ancho de ventana, parámetro de suavización o *bandwith*.

En la Tabla 8.1 mostramos algunas de las funciones núcleo univariantes más comunes.

El estimador núcleo puede interpretarse como una suma de protuberancias (del inglés *bump* situadas en las observaciones. La función núcleo K determina la forma de las protuberancias mientras que el parámetro h_n determina su anchura. Al igual que en el histograma h_n también determina la cantidad de suavización de la estimación, siendo el límite cuando h_n tiende a cero una suma de funcionales delta de Dirac en los puntos de las observaciones. También puede interpretarse como una transformación en continua de la función de distribución empírica de acuerdo a la función $K(x)$ que se encarga de redistribuir la masa de probabilidad 1/n en la vecindad de cada punto muestral.

En la Figura 8.1 podemos observar una estimación mostrando los núcleos individuales.

Un inconveniente de la estimación núcleo es que al ser el parámetro de ventana fijo a lo largo de toda la muestra, existe la tendencia a presentarse distorsiones en las colas de la estimación, tal y como muestra la Figura 8.2.

Núcleo	$K(t)$	Rango
Epanechnikov	$\frac{3}{4}(1-t^2)$	$\|t\| < 1$
Gauss	$\frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2}$	$\|t\| < \infty$
Triangular	$1 - \|t\|$	$\|t\| < 1$
Rectangular	$\frac{1}{2}$	$\|t\| < 1$
Biweight	$\frac{15}{16}(1-t^2)^2$	$\|t\| < 1$
Triweight	$\frac{35}{32}(1-t^2)^3$	$\|t\| < 1$
Arco coseno	$\frac{\pi}{4}\cos\frac{\pi}{2}t$	$\|t\| < 1$

Cuadro 8.1: Algunas de las funciones núcleo más comunes.

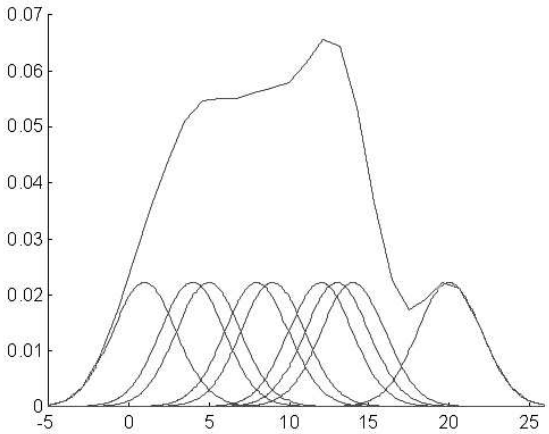

Figura 8.1: Estimación tipo núcleo mostrando los núcleos individuales.

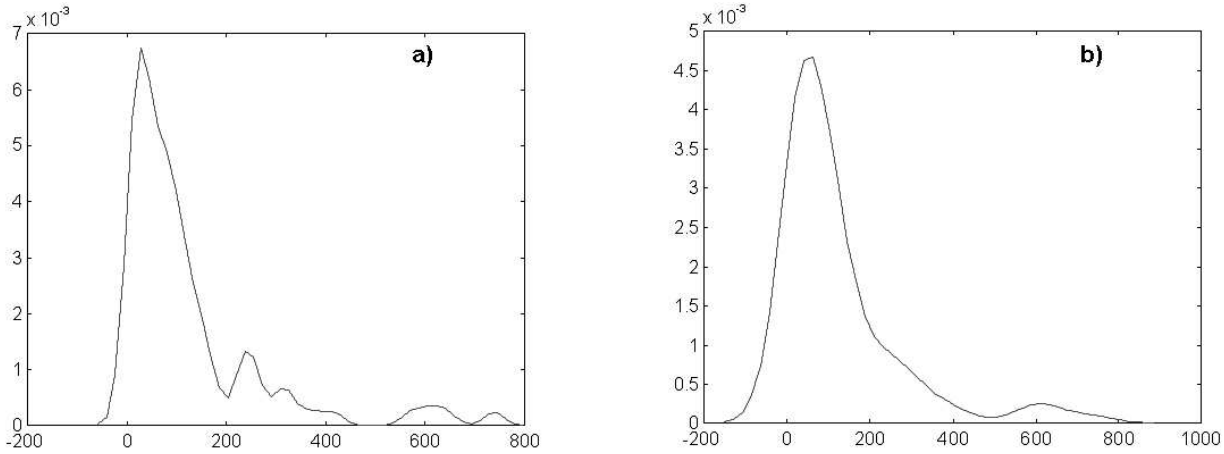

Figura 8.2: Estimación tipo núcleo. a) anchura de ventana 20 y b) 60.

8.2. Propiedades estadísticas

8.2.1. Consistencia

Parzen (1962) basándose en un teorema previo de Bochner (1955), que presentamos a continuación, estudia el sesgo y la consistencia en un punto x para las estimaciones tipo núcleo donde la función núcleo es una función simétrica y acotada que verifica

$$\int_{-\infty}^{\infty} |K(x)|\, dx < \infty \tag{8.2}$$

$$\lim_{x\to\infty} |xK(x)| = 0 \tag{8.3}$$

$$\int_{-\infty}^{\infty} K(x)\, dx = 1 \tag{8.4}$$

Condiciones que son satisfechas por cualquiera de las funciones núcleo presentadas en la Tabla 8.1.

Teorema 4 (Bochner (1955)) *Sea $K(y)$ una función Borel acotada que satisface las condiciones (8.2) y (8.3). Sea $g \in \mathcal{L}^1$. Sea*

$$g_n(x) = \frac{1}{h_n} \int_{-\infty}^{\infty} K(\frac{y}{h})\, g(x-y)\, dy, \tag{8.5}$$

donde $\{h_n\}$ *es una secuencia de constantes positivas que satisfacen* $\lim_{n\to\infty} h_n = 0$. *Entonces si* x *es un punto de continuidad de* g,

$$\lim_{n\to\infty} g_n(x) = g(x) \int_{-\infty}^{\infty} K(y)\, dy. \tag{8.6}$$

Demostración:

Notemos en primer lugar que

$$g_n(x) - g(x)\int_{-\infty}^{\infty} K(y)\, dy \quad = \quad \int_{-\infty}^{\infty} \{g(x-y) - g(x)\} \frac{1}{h_n} K(\frac{y}{h})\, dy$$

Sea ahora $\delta > 0$, y dividamos el dominio de integración en dos regiones, $|y| \le \delta$ y $|y| > \delta$. Entonces

$$\begin{aligned}
& \left| g_n(x) - g(x)\int_{-\infty}^{\infty} K(y)\, dy \right| \le \sup_{|y|\le\delta} |g(x-y) - g(x)| \int_{|z|\le\frac{\delta}{h_n}} |K(z)|\, dz \\
+ & \int_{|y|\ge\delta} \frac{|g(x-y)|}{y} \frac{y}{h_n} K(\frac{y}{h})\, dy + |g(x)| \int_{|y|\ge\delta} \frac{1}{h_n} K(\frac{y}{h})\, dy \\
\le & \sup_{|y|\le\delta} |g(x-y) - g(x)| \int_{-\infty}^{\infty} |K(z)|\, dz \\
+ & \frac{1}{\delta} \sup_{|z|\ge\frac{\delta}{h_n}} |zK(z)| \int_{-\infty}^{\infty} |g(y)|\, dy + |g(x)| \int_{|z|\ge\frac{\delta}{h_n}} |K(z)|\, dz
\end{aligned}$$

Cuando $n \to \infty$, debido a que $h_n \to 0$, el segundo y tercer término tienden a cero, ya que $g \in \mathcal{L}^1$ y $\lim_{y\to\infty} |xK(x)| = 0$. Haciendo entonces que $\delta \to 0$, el primer término tiende a cero debido a que $K \in \mathcal{L}^1$ y a que x es un punto de continuidad de g.

Teniendo ahora en cuenta que

$$\begin{aligned}
\mathrm{E}[\hat{f}_n(x)] &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right)\right] && (8.7)\\
&= \mathrm{E}\left[\frac{1}{h_n} K\left(\frac{x-y}{h}\right)\right] = \int_{-\infty}^{\infty} \frac{1}{h_n} K\left(\frac{x-y}{h}\right) f(y)\, dy && (8.8)
\end{aligned}$$

del teorema anterior se deduce el siguiente

Corolario 1 *La estimación definida en (8.1) es asintóticamente insesgada en todos los puntos* x *en los cuales la función de densidad de probabilidad es continua si las constantes* h_n *satisfacen* $\lim_{n\to\infty} h_n = 0$ *y si la función* $K(y)$ *satisface (8.2),(8.3) y (8.4).*

Teorema 5 *El estimador* $\hat{f}_n(x)$ *definido en (8.1) es consistente, es decir* $\mathrm{MSE}\,[\hat{f}_n(x)] \to 0 \quad \forall x \in \mathbf{R}$ *cuando* $n \to \infty$, *si añadimos la condición adicional de que* $\lim_{n\to\infty} n\, h_n = \infty$.

Demostración:

En efecto, tengamos en cuenta que

$$\mathrm{Var}\left[\hat{f}_n(x)\right] = \frac{1}{n}\,\mathrm{Var}\left[\frac{1}{h_n}\,K\left(\frac{x-y}{h}\right)\right] \tag{8.9}$$

Además

$$\begin{aligned}\frac{1}{n}\,\mathrm{Var}\left[\frac{1}{h_n}\,K\left(\frac{x-y}{h}\right)\right] &\leq \frac{1}{n}\,\mathrm{E}\left[\left(\frac{1}{h_n}\,K\left(\frac{x-y}{h}\right)\right)^2\right] &(8.10)\\ &= \frac{1}{h_n n}\left[\frac{1}{h_n}\int_{-\infty}^{\infty}\left(K\left(\frac{x-y}{h}\right)\right)^2 f(y)\,dy\right] &(8.11)\end{aligned}$$

y por el Teorema 4

$$\frac{1}{h_n}\int_{-\infty}^{\infty}\left(K\left(\frac{x-y}{h}\right)\right)^2 f(y)\,dy \to f(x)\int_{-\infty}^{\infty}K^2(y)\,dy \tag{8.12}$$

ya que $\int_{-\infty}^{\infty}K^2(y)\,dy < \infty$. Es por tanto evidente que

$$\lim_{n\to\infty}\mathrm{Var}[\hat{f}_n(x)] \to 0 \quad \text{si} \quad \lim_{n\to\infty} h_n\,n = \infty \tag{8.13}$$

Finalmente al ser

$$\mathrm{MSE}[\hat{f}_n(x)] = \mathrm{Var}[\hat{f}_n(x)] + \mathrm{sesgo}^2[\hat{f}_n(x)] \tag{8.14}$$

teniendo en cuenta el Corolario 1 el Teorema queda demostrado.

Este resultado ilustra perfectamente el problema básico de la estimación no paramétrica. Una rápida convergencia al cero del parámetro h_n provoca una disminución del sesgo, pero sin embargo la varianza aumentaría de forma considerable. El ancho de ventana ideal debe converger a cero pero a un ritmo más lento que n^{-1}.

Una axiomática más reciente para las funciones núcleo es la propuesta por Nadaraya (1989), donde diremos que una función $K(x)$ pertenece a la clase H_s ($s \geq 2$ es un número par) o bien que es un Kernel de orden s, si satisface las siguientes condiciones de regularidad

$$\begin{aligned}&K(x) = K(-x) &(8.15)\\ &\int_{-\infty}^{\infty}K(x)\,dx = 1 &(8.16)\\ &\sup_{-\infty<x<\infty}|K(x)| < \infty &(8.17)\\ &\int_{-\infty}^{\infty}x^i K(x)\,dx = 0 \quad i = 1,\ldots,s-1 &(8.18)\\ &\int_{-\infty}^{\infty}x^s K(x)\,dx = k_s \neq 0 &(8.19)\\ &\int_{-\infty}^{\infty}x^s|K(x)|\,dx < \infty &(8.20)\end{aligned}$$

Destaquemos que si K es una función de densidad debe verificarse

$$k_2 = \int_{-\infty}^{\infty} x^2 K(x)\, dx \; > 0.$$

Es posible hacer que $k_2 = 0$ si permitimos que K pueda tomar valores negativos.

Si K es una función núcleo simétrica, entonces s debe ser un número par.

Si utilizamos como núcleo una función de densidad simétrica, supuesto bastante común y sobre el que se basa el siguiente apartado, estamos considerando un núcleo de orden 2. En particular corresponden a este modelo las funciones presentadas en la Tabla 8.1.

8.2.2. Minimización del AMISE

La determinación del ancho de ventana se realiza de modo que se minimice algún tipo de error. En general se utiliza como medida de error el error cuadrático medio integrado definido en (3.6) y se minimiza una aproximación al mismo.

Podemos utilizar (8.8) para obtener una expresión aproximada del sesgo. Hagamos en (8.8) el cambio de variable $y = x - h_n t$, obtenemos

$$\mathrm{E}[\hat{f}_n(x)] \;\; = \;\; \int_{-\infty}^{\infty} K(t)\, f(x - h_n t)\, dt \tag{8.21}$$

y haciendo un desarrollo de Taylor en el punto cero

$$f(x - h_n t) = f(x) - f'(x) h_n t + f''(x) \frac{h^2 t^2}{2} + \cdots$$

y sustituyendo en (8.21), obtenemos teniendo en cuenta (8.18)

$$\mathrm{E}[\hat{f}_n(x)] \;\; = \;\; f(x) + \frac{h_n^2 f''(x) k_2}{2} + O(h^4) \tag{8.22}$$

Por consiguiente el sesgo adopta la forma

$$\mathrm{sesgo}\,[\hat{f}_n(x)] \;\; = \;\; \frac{h_n^2 f''(x) k_2}{2} + O(h^4) \tag{8.23}$$

A partir de (8.23) podemos escribir

$$AISB = \frac{1}{4}\, h_n^4\, k_2^2 \int_{-\infty}^{\infty} f''(x)^2\, dx \tag{8.24}$$

y de (8.9) es fácil comprobar que

$$\begin{aligned} \mathrm{Var}[\hat{f}_n(x)] \;\; &= \;\; \frac{1}{n} \int_{-\infty}^{\infty} \frac{1}{h_n^2} \left(K\left(\frac{x-y}{h} \right) \right)^2 f(y)\, dy \; - \; \frac{1}{n} \{ f(x) + \mathrm{sesgo}[\hat{f}_n(x)] \}^2 \\ &\approx \;\; \frac{1}{n h_n} \int_{-\infty}^{\infty} f(x - h_n t) K^2(t)\, dt \; - \; \frac{1}{n} \{ f(x) + O(h_n^2) \}^2 \end{aligned} \tag{8.25}$$

usando la sustitución $y = x - h_n t$ y la aproximación (8.23) para el sesgo. Suponiendo un valor de h_n pequeño y un valor de n grande y expandiendo $f(x - h_n t)$ en serie de Taylor, obtenemos

$$\mathrm{Var}[\hat{f}_n(x)] \approx \frac{1}{nh_n} f(x) \int_{-\infty}^{\infty} K^2(t)\, dt \tag{8.26}$$

Integrando ahora obtenemos

$$AIV = \frac{1}{nh_n} \int_{-\infty}^{\infty} K^2(t)\, dt \tag{8.27}$$

Resultando finalmente a partir de (8.23) y (8.26)

$$AMISE[\hat{f}_n(x)] = \frac{1}{4} h_n^4 k_2^2 \int_{-\infty}^{\infty} f''(x)^2\, dx + \frac{1}{nh_n} \int_{-\infty}^{\infty} K^2(t)\, dt \tag{8.28}$$

Busquemos ahora el valor de h_n que minimiza la expresión anterior, obtenemos

$$h_{opt} = \left\{ \frac{\int_{-\infty}^{\infty} K^2(t)\, dt}{\int_{-\infty}^{\infty} f''(x)^2\, dx} \right\}^{1/5} n^{-1/5}\, k_2^{-2/5} \tag{8.29}$$

Comprobamos como efectivamente la convergencia de h_n a cero es de orden $n^{-1/5}$, menor que n^{-1}, tal y como hemos exigido previamente. Debe hacerse notar la dependencia del valor óptimo respecto a la densidad desconocida que se desea estimar, lo que impide que sea calculable directamente.

Substituyendo (8.29) en (8.28) obtenemos

$$AMISE^* = \frac{5}{4} \{k_2^2 R^4(K)\}^{1/5} \{R(f'')\}^{1/5} n^{-4/5} \tag{8.30}$$

8.2.3. Elección del parámetro de ventana.

Una de las posibilidades a la hora de elegir el parámetro de suavización óptimo es tomar como referencia una distribución standard para obtener el valor de $\int_{-\infty}^{\infty} f''(x)^2\, dx$ en la expresión (8.29). Una de las distribuciones más utilizadas es la distribución normal de media cero y varianza σ^2, resultando entonces

$$\int_{-\infty}^{\infty} f''(x)^2\, dx = \frac{3}{8}\, \pi^{-1/2}\, \sigma^{-5} \approx 0{,}212 \sigma^{-5}. \tag{8.31}$$

Utilizando ahora la función núcleo de Gauss y sustituyendo (46) en (8.29) obtenemos que el ancho de ventana óptimo vendrá dado por

$$h_{opt} = (4\pi)^{-1/10} (\frac{3}{8} \pi^{-1/2})^{-1/5}\, \sigma\, n^{-1/5} = 1{,}06\, \sigma\, n^{-1/5} \tag{8.32}$$

donde σ puede ser substituida por una estimación de la varianza a partir de los datos.

La utilización de (8.32) será adecuada si la población se asemeja en su distribución a la de la normal, sin embargo si trabajamos con poblaciones multimodales se producira una sobresuavización de la estimación, Bowman (1985), Silverman (1986). Una posible modificación del parámetro de suavización es

$$h_n = 1{,}06\, A\, n^{-1/5} \tag{8.33}$$

donde $A = \text{mín}\,(\text{ desviación standard , rango intercuartil}/1{,}349)$, comprobando que se comporta bien trabajando con densidades unimodales y moderadamente bimodales. Silverman también sugiere la reducción del factor 1.06 en (8.33); y propone como nuevo valor del parámetro h_n

$$h_n = 0{,}9\, A\, n^{-1/5}, \tag{8.34}$$

comprobando el autor con diversas simulaciones que el parámetro definido en (8.34) funciona correctamente con un amplio abanico de densidades, teniendo la ventaja adicional de su trivial evaluación. Podemos concluir que para un gran número de datos la elección del parámetro de suavización definido en (8.34) funcionará correctamente, y en otros casos puede ser utilizado como valor inicial para un posterior estudio. Una alternativa razonable es utilizarlo como estimación piloto inicial en posteriores técnicas más refinadas.

8.2.4. Selección de la función núcleo óptima

En (8.30) denominemos

$$C(K) \equiv \{k_2^2 R^4(K)\}^{1/5}. \tag{8.35}$$

En primer lugar veamos que $C(K)$ es invariante frente a transformaciones de la forma

$$K_\delta(\cdot) = \frac{1}{\delta} K(\frac{\cdot}{\delta}) \quad \delta > 0.$$

En efecto

$$\int K_\delta(x) x^2 dx = \frac{1}{\delta} \int k(\frac{x}{\delta}) x^2 dx = \int K(u) \delta^2 u^2 du = \delta^2 k_2,$$

además

$$R(K_\delta) = \int K_\delta^2(x) dx = \frac{1}{\delta^2} \int k^2(\frac{x}{\delta}) dx = \frac{1}{\delta} R(K).$$

Por tanto

$$C(K_\delta) = \{\delta^4 k_2^2 \frac{1}{\delta^4} R^4(K)\}^{1/5} = C(K).$$

El Kernel óptimo es áquel que minimiza

$$\int K^2(x) dx \quad s/t \quad \sigma_K^2 = \sigma^2 = k_2,$$

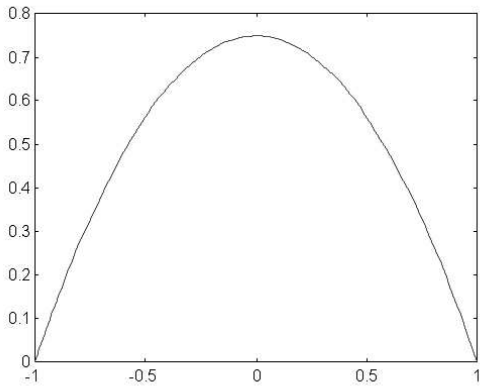

Figura 8.3: Función núcleo de Epanechnikov.

verificándose:

$$K \geq 0 \ , \ \int K(x)dx = 1 \ , \ \int xK(x)dx = 0 \ , \ \int x^2K(x)dx = k_2 \neq 0.$$

Hodges y Lehman (1956) demuestran que la función que minimiza la expresión anterior es la función núcleo de Epanechnikov

$$K_e(t) = \begin{cases} \frac{3}{4}(1-t^2) & |t| \leq 1 \\ 0 & |t| > 1 \end{cases} , \tag{8.36}$$

denominada así por ser Epanechnikov (1969) el primero en utilizarla en el contexto de la estimación no paramétrica. Puede observarse su gráfica en la Figura 8.3

Es inmediato comprobar que $\sigma^2_{K_2} = \frac{1}{5}$ y $R(K_e) = \frac{3}{5}$.

¨Qué sucede si utilizamos una función núcleo K diferente de K_e?. Si realizamos dos estimaciones de una misma densidad f: una con K_e y un tamaño muestral n_1, y otra con otra función núcleo K y un tamaño muestral n_2, los errores son:

$$AMISE^*(K_e) \propto C(K_e)n_1^{-4/5}$$

y

$$AMISE^*(K) \propto C(K)n_2^{-4/5},$$

con la misma constante de proporcionalidad. Para obtener un mismo error

$$AMISE^*(K_e) = AMISE^*(K) \Rightarrow C(K_e)n_1^{-4/5} = C(K)n_2^{-4/5},$$

Núcleo	$K(t)$	Rango	Eficiencia
Epanechnikov	$\frac{3}{4}(1-t^2)$	$\|t\|<1$	1,000
Biweight	$\frac{15}{16}(1-t^2)^2$	$\|t\|<1$	0,994
Triweight	$\frac{35}{32}(1-t^2)^3$	$\|t\|<1$	0,987
Triangular	$1-\|t\|$	$\|t\|<1$	0,986
Gauss	$\frac{1}{\sqrt{2\pi}}e^{-(1/2)t^2}$	$\|t\|<\infty$	0,951
Rectangular	$\frac{1}{2}$	$\|t\|<1$	0,930

Cuadro 8.2: Eficiencias de algunas de las funciones núcleo más comunes.

y resulta

$$\frac{C(K_e)}{C(K)} = \left(\frac{n_2}{n_1}\right)^{-4/5} \Rightarrow \left(\frac{C(K_e)}{C(K)}\right)^{5/4} = \frac{n_1}{n_2},$$

por tanto la cantidad

$$eff(K) = \left(\frac{C(K_e)}{C(K)}\right)^{5/4} \tag{8.37}$$

denominada *eficiencia* de la función núcleo K, representa la razón de tamaños muestrales necesaria para obtener el mismo AMISE dada f usando K_e o usando K.

Por ejemplo para la función núcleo de Epanechnikov $C(K_e) = 0,349086$ mientras que para la función núcleo de Gauss (K_N) y dado que $\sigma_{K_N} = 1$ y que $R(K_N) = \int \frac{1}{2\pi}e^{-x^2}dx = \frac{1}{2\sqrt{\pi}}$ resulta $C(K_N) = 0,363342$, con lo que la eficacia de la función núcleo de Gauss es

$$eff(K_N) = \left(\frac{0,349086}{0,363342}\right)^{5/4} = 0,951198$$

En la Tabla 8.2 mostramos las eficiencias de algunas de las funciones núcleo univariantes más comunes.

De \ A	Gauss	Rectang.	Epan.	Triang.	Biwt.	Triwt.
Gauss	1	1,740	2,214	2,432	2,623	2,978
Rectang.	0,575	1	1,272	1,398	1,507	1,711
Epan.	0,452	0,786	1	1,099	1,185	1,345
Triang.	0,411	0,715	0,910	1	1,078	1,225
Biwt.	0,381	0,663	0,844	0,927	1	1,136
Triwt.	0,336	0,584	0,743	0,817	0,881	1

Cuadro 8.3: Factores de conversión entre funciones núcleo para obtener núcleos equivalentes. De Scott (1992).

8.2.5. Funciones núcleo equivalentes

En ocasiones interesa cambiar el tipo de función núcleo utilizada. La pregunta que nos hacemos es cómo debemos modificar el paramétro de ventana para mantener mínimo el AMISE. Si utilizamos las funciones núcleo K_1 y K_2 con ventanas óptimas h_1 y h_2 respectivamente tenemos que

$$h_1^* = \left\{ \frac{R(K_1)}{R(f'')} \right\}^{1/5} \; n^{-1/5} \, {\sigma_{k_1}^2}^{-2/5}$$

y

$$h_2^* = \left\{ \frac{R(K_2)}{R(f'')} \right\}^{1/5} \; n^{-1/5} \, {\sigma_{k_2}^2}^{-2/5},$$

por tanto

$$\frac{h_1^*}{h_2^*} = \left(\frac{R(K_1)/\sigma_{K_1}^4}{R(K_2)/\sigma_{K_2}^4} \right)^{1/5} = \frac{\sigma_{K_2}}{\sigma_{K_1}} \left(\frac{R(K_1)\sigma_{K_1}}{R(K_2)\sigma_{K_2}} \right)^{1/5}$$

En la Tabla 8.3 mostramos los factores de conversión de las ventanas para pasar de una función núcleo a otra.

Dada la semejanza en las eficiencias una expresión alternativa aproximada es

$$h_2^* \approx \frac{\sigma_{K_1}}{\sigma_{K_2}} h_1^* \tag{8.38}$$

Destaquemos que muchas de las funciones núcleo definidas en la Tabla 8.2 pueden ser consideradas casos particulares de una familia de funciones núcleo con soporte compacto definida por

$$K(x) = k_{rs} \, (1 - |x|^r)^s \;\; I_{[|x| \leq 1]} \tag{8.39}$$

donde

$$k_{rs} = \frac{r}{2Beta(s+1, 1/r)} \quad r > 0 \quad s \geq 0$$

Resultando los casos particulares mostrados en la Tabla 8.4.

Función Núcleo	r	s	k_{rs}
Rectangular		0	1/2
Triangular	1	1	1
Epanechnikov	2	1	3/4
Biweight	2	2	15/16
Triweight	2	3	35/32

Cuadro 8.4: Funciones núcleo como casos particulares.

8.2.6. Reducción del sesgo. Núcleos de orden mayor que 2.

Trabajar con núcleos de orden mayor a 2 permite mejorar el MISE reduciendo la contribución del sesgo. Consideremos

$$\int x^i K(x)dx = 0 \;\; i = 0, \ldots, s-1 \tag{8.40}$$

$$\int x^s K(x)dx = k_s \neq 0, \tag{8.41}$$

con s par. Entonces de (8.21)

$$\mathrm{E}\hat{f}(x) = \int K(t) f(x - ht)dt, \tag{8.42}$$

y considerando el desarrollo de Taylor

$$\mathrm{E}\hat{f}(x) = f(x)\int K - f'(x)h\int tK + \cdots + \frac{f^{(s)}h^s}{s!}\int t^s K dt = f(x) + \frac{f^{(s)}h^s}{s!}k_s,$$

el sesgo al quadrado asintótico resulta

$$IASB = \frac{h^{2s}}{s!^2}k_s^2 R(f^{(s)}). \tag{8.43}$$

La varianza por su parte no se modifica

$$AIV = \frac{R(K)}{nh}. \tag{8.44}$$

Resultando un AMISE igual a

$$AMISE = \frac{R(K)}{nh} + \frac{h^{2s}}{s!^2}k_s^2 R(f^{(s)}). \tag{8.45}$$

Minimizando la expresión anterior como es usual obtenemos el parámetro de ventana óptimo

$$h^* = \left[\frac{s!^2 R(K)}{2sk_s^2 R(f^{(s)})}\right]^{1/(2s+1)} n^{-1/(2s+1)}. \tag{8.46}$$

s	K_s en [-1,1]	AMISE N(0,1)
2	$\frac{3}{4}(1-t^2)$	$0,320\, n^{-4/5}$
4	$\frac{15}{32}(1-t^2)(3-7t^2)$	$0,482\, n^{-8/9}$
6	$\frac{105}{206}(1-t^2)(5-30t^2+33t^4)$	$0,581\, n^{-12/13}$
8	$\frac{315}{4096}(1-t^2)(35-385t^2+1001t^4-715t^6)$	$0,681\, n^{-16/17}$

Cuadro 8.5: Funciones núcleo polinómicas de orden s.

Fijémonos que el AMISE resultante es de orden $n^{-2s/(2s+1)}$ y por tanto en principio es posible aproximar la tasa a n^{-1}, tasa de los estimadores paramétricos, tanto como deseemos.

$$AMISE^* \propto \left[k_s^2 R(K)^{2s} R(f^{(s)})\right]^{1/(2s+1)} n^{-2s/(2s+1)}. \tag{8.47}$$

A partir de Schucany y Sommers (1977) es posible dar un método de construcción de funciones núcleo de orden s. Si K_s es un kernel simétrico de orden s y diferenciable, entonces

$$K_{s+2}(t) = \frac{3}{2}K_s(t) + \frac{1}{2}tK_s'(t) \tag{8.48}$$

es un kernel de orden $s+2$. Por ejemplo si $\phi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2} \Rightarrow \phi'(t) = -t\phi(t)$, y una función núcleo de orden 4 es

$$K_4 = \frac{3}{2}\phi(t) - \frac{1}{2}t^2\phi(t) = \frac{1}{2}(3-t^2)\phi(t). \tag{8.49}$$

Consideramos sólo funciones núcleo de orden par dado que desaparecen todos los momentos de orden impar. En la Tabla 8.5 presentamos algunas de las funciones núcleo polinómicas de orden s más utilizadas junto con el AMISE resultante supuestos datos procedentes de una distribución N(0,1).

Aumentar s presenta algunos problemas:

- Existe una parte negativa que se traslada a la densidad estimada y a mayor s mayor parte negativa. Una posible solución es truncar a la parte positiva y normalizar pero aparecen problemas tales como discontinuidades.
- Las estimaciones aparecen más rugosas para valores más bajos de h.
- Las ventajas no se aprecian hasta que el tamaño muestral es bastante elevado, entonces sí que se pueden observar mejores comportamientos en algunos picos, pero generalmente no se consideran núcleos de orden mayor que 2 o 4.

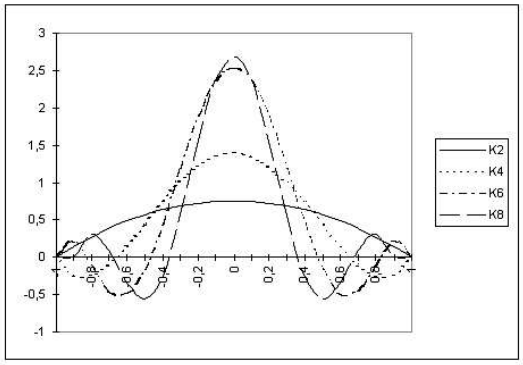

Figura 8.4: Funciones núcleo polinómicas de orden s.

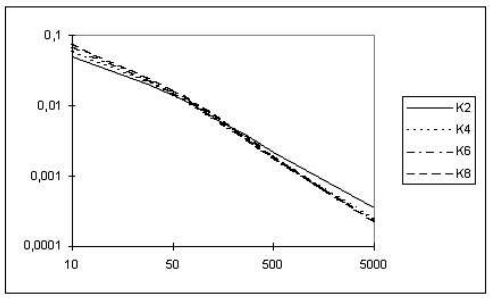

Figura 8.5: AMISE (N(0,1)) de las funciones núcleo polinómicas de orden s.

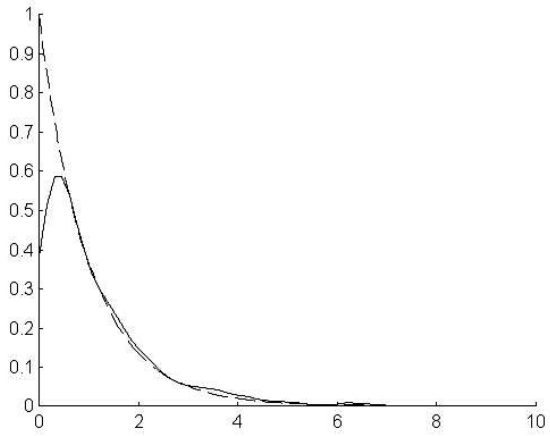

Figura 8.6: Estimación basada en la simulación de una exponencial negativa de esperanza 1. h=0,3.

8.2.7. Dominios acotados

En muchas ocasiones el dominio de definición de una densidad no es toda la recta real sino un subconjunto de la misma. Uno de los ejemplos más claros es la estimación a partir de tiempos o de cualquier variable que sólo pueda tomar valores positivos. La utilización de las funciones núcleo sin realizar ninguna modificación previa conduce a estimaciones inexactas, por ejemplo las infraestimaciones de la densidad que se observan en la Figura 8.6 al estimar una densidad exponencial negativa a partir de una muestra simulada.

Una alternativa en la que podríamos pensar es estimar únicamente en la parte positiva y asignar $\hat{f}(x)$ igual a cero en los valores negativos; renormalizando posteriormente para que la integral fuera 1. Sin embargo este proceso no soluciona la infraestimación que se produce en las cercanias del límite del dominio. Supongamos por ejemplo que $f(x)$ está definida sólo para $x \geq 0$, en (8.8) resulta

$$\mathrm{E}[\hat{f}_n(x)] \;= \int_{-\infty}^{\infty} \tfrac{1}{h_n} K\left(\tfrac{x-y}{h}\right)\, f(y)\, dy = \int_{0}^{\infty} \tfrac{1}{h_n} K\left(\tfrac{x-y}{h}\right)\, f(y)\, dy \tag{8.50}$$

y con la sustitución habitual $y = x - th$ y el desarrollo de Taylor como en (8.21) obtenemos

$$\begin{aligned} \mathrm{E}\hat{f}(x) &= \int_{-\infty}^{x/h} K(t) f(x-th) dt & (8.51)\\ &\approx f(x)\int_{-\infty}^{x/h} K(t)dt - f'(x)h\int_{-\infty}^{x/h} tK(t)dt + f''(x)\frac{h^2}{2}\int_{-\infty}^{x/h} t^2K(t)dt & (8.52)\end{aligned}$$

y la estimación presenta un claro sesgo al ser

$$\lim_{n\to\infty} \mathrm{E}\hat{f}(x) = f(x)\int_{-\infty}^{x/h} K(t)dt < f(x) \quad \text{para } x \in [0,h) \tag{8.53}$$

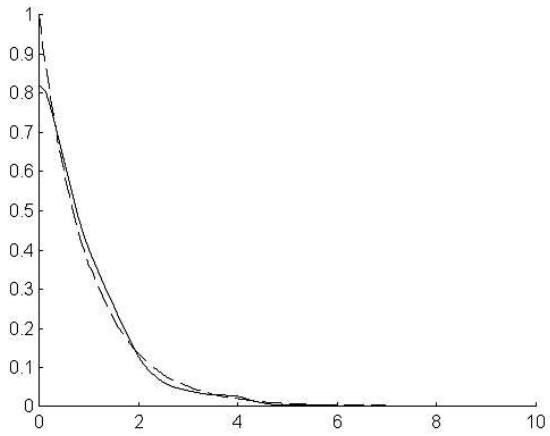

Figura 8.7: Estimación basada en la simulación de una exponencial negativa de esperanza 1 con la reflexión de los datos. h=0,3.

Una técnica que solventa en parte este problema es la *reflexión*. Si los datos son nonegativos y la discontinuidad del dominio se produce en $x = 0$, la densidad se estima de la forma habitual pero la estimación se realiza sobre la muestra ampliada por los correspondientes valores negativos de los datos muestrales $(-x_n, \ldots, -x_1, x_1, \ldots, x_n)$. Si la estimación obtenida sobre esta muestra de tamaño $2n$ es $\hat{f}^*$, una estimación sobre la muestra original viene dada por

$$\hat{f}(x) = \begin{cases} 2\hat{f}^*(x) & \text{para } x \geq 0 \\ 0 & \text{para } x < 0. \end{cases} \tag{8.54}$$

Hay que destacar que el ancho de ventana utilizado debe estar basado en la muestra de tamaño n y no en la de $2n$. En la Figura 8.7 podemos ver el resultado obtenido comparado con la Figura 8.6.

Otra alternativa al método de reflexión es la utilización de funciones núcleo de frontera **"Boundary kernels"**. Los núcleos de frontera son funciones ponderadas que se deben utilizar dentro de la región límite $[0, h)$, es decir para $x = ph \;\; 0 \leq p < 1$. No entramos en detalle en la construcción de tales funciones núcleo pero una de las más utilizadas es

$$B(x) = \frac{[a_2(p) - a_1(p)x]K(x)}{a_0(p)a_2(p) - a_1^2(p)}, \tag{8.55}$$

donde

$$a_l(p) = \int_{-1}^{p} u^l K(u)du.$$

8.3. Selección del ancho de ventana

Siguiendo a Jones, Marron y Sheather (1996a) podemos clasificar las técnicas de selección del ancho de ventana basadas en una muestra en *métodos de primera generación* y *métodos de segunda generación*. La clasificación tiene su origen principal en la superioridad que han mostrado las técnicas desarrolladas recientemente, a partir de 1990 frente a las técnicas de primera generación desarrolladas en su mayoría con anterioridad a 1990.

Entre los métodos de primera generación incluimos:

- Reglas basadas en la distribuciones paramétricas. **Rules of Thumb"**.
- Sobresuavización.
- Reglas de Validación cruzada.

y entre los de segunda:

- Métodos Plug-In.
- Bootstrap suavizado.

8.3.1. Reglas basadas en distribuciones paramétricas.

La idea de este método ha sido comentada en el apartado 8.2.3, su origen se remonta a Deheuvels (1977). La elección de la distribución $N(0, \sigma^2)$ es adecuada desde el momento que es la dispersión o escala más que la posición el factor importante en la elección del paramétro de suavización. En esencia la elección del parámetro de suavización es

$$h_n = 1{,}06\, A\, n^{-1/5} \tag{8.56}$$

donde $A = \text{mín}\,(\text{ desviación standard }, \text{ rango intercuartil}/1{,}349)$, o con la modificación sugerida en Silverman (1986)

$$h_{ROT} = 0{,}9\, A\, n^{-1/5}. \tag{8.57}$$

Destaca la simplicidad del método y su buen comportamiento al menos para distribuciones unimodales, Bowman (1985). Sin embargo si la densidad a estimar no es normal, h_{ROT} no es ni siquira un estimador consistente del parámetro de ventana óptimo.

8.3.2. Sobresuavización

Consiste en resolver el problema variacional

$$\min_f \int_{-\infty}^{\infty} f''(x)^2 dx \quad \text{s/t} \quad \int f = 1 \text{ y } \int x^2 f = 1. \tag{8.58}$$

Podemos ver por ejemplo en Scott (1992) que la solución es

$$f^*(x) = \frac{35}{69{,}984}(9 - x^2)_+^3 \text{ y } R[(f^*)''] = \frac{35}{243} \tag{8.59}$$

Tras un cambio de variable $R[(f^*)''] \geq \frac{35}{243\sigma^5}$, con lo que obtenemos

$$h^* = \left[\frac{R(K)}{n\sigma_K^4 R(f'')}\right]^{1/5} \leq \left[\frac{243\sigma^5 R(K)}{35n\sigma_K^4}\right]^{1/5} \tag{8.60}$$

y la regla finalmente resulta

$$h_{OS} = 3\left[\frac{R(K)}{35\sigma_K^4}\right]^{1/5} \sigma n^{-1/5} \tag{8.61}$$

Por ejemplo para la función núcleo de Gauss $h_{OS} = 1,144\sigma n^{-1/5}$ resultando 1,08 veces mayor que la regla basada en la referencia a la distribución normal.

8.3.3. Reglas de validación cruzada.

Se caracterizan por utilizar la técnica del leave-one-out para minimizar alguna medida de discrepancia entre la densidad y su estimación.

Validación cruzada de mínimos cuadrados. LSCV

Es un método automático para seleccionar el parámetro de ventana. Sugerido por Rudemo (1982) y Bowman (1984). Se basa en la minimización del MISE de la forma siguiente. Dado un estimador $\hat{f}$ de la densidad f, el MISE se expresa

$$MISE\hat{f} = \mathrm{E}\int(\hat{f} - f)^2 = \mathrm{E}\int \hat{f}^2 - 2\mathrm{E}\int \hat{f}f + \mathrm{E}\int f^2. \tag{8.62}$$

El último término no depende de la estimación $\hat{f}$, por tanto la elección de h para minimizar el MISE equivale a la minimización de

$$\Phi(\hat{f}) = \mathrm{E}\left[\int \hat{f}^2 - 2\int \hat{f}f\right]. \tag{8.63}$$

Un estimador de $\int \hat{f}f(x)$ viene dado por

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i), \tag{8.64}$$

donde $\hat{f}_{-i}(x)$ es la densidad estimada a partir de los datos extrayendo de la muestra el dato X_i,

$$\hat{f}_{-i}(x), = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{x - X_j}{h}\right). \tag{8.65}$$

El estimador (8.64) es un estimador insesgado para $\mathrm{E} \int \hat{f} f$, tal y como puede demostrarse,

$$\begin{aligned} \mathrm{E}\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x), &= \mathrm{E}\hat{f}_n(x) = \frac{1}{n-1}\sum_{j \neq i} \mathrm{E}\frac{1}{h} K\left(\frac{X_i - X_j}{h}\right) \\ &= \mathrm{E}\{\frac{1}{h} K\left(\frac{X_1 - X_2}{h}\right)\} = \int\int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(x) f(y) dx dy \\ &= \int \{\int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y) dy\} dx = \int \mathrm{E}\{\hat{f}(x)\} f(x) dx \\ &= \mathrm{E}\{\int \hat{f}(x) f(x) dx\}. \end{aligned} \tag{8.66}$$

Por tanto, un estimador insesgado de

$$\Phi(\hat{f}) = \mathrm{E}\left[\int \hat{f}^2 - 2\int \hat{f} f\right] \tag{8.67}$$

viene dado por

$$\mathrm{LSCV}\,(h) = \int \hat{f}^2(x) dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i). \tag{8.68}$$

El estimador LSCV del parámetro de ventana h_{LSCV} será el valor que minimize la expresión anterior,

$$\hat{h}_{LSCV} = \mathrm{argmin}_h\ \mathrm{LSCV}\ (h) \tag{8.69}$$

Suponemos que el mínimo de (8.68) estara cercano al mínimo de (8.67) y por tanto que el parámetro de ventana obtenido al minimizar (8.68) será una buena elección. Puede demostrarse que la expresión (8.68) es equivalente a

$$\mathrm{LSCV}\,(h) = \frac{R(K)}{nh} + \frac{2}{n^2 h}\sum_{i<j} \gamma(c_{ij}), \tag{8.70}$$

donde

$$\gamma(c) = (K * K)(c) - 2K(c) = \int K(w) K(w+c) dw - 2K(c), \tag{8.71}$$

y

$$c_{ij} = \frac{(X_i - X_j)}{h}. \tag{8.72}$$

Por ejemplo, utilizando la función núcleo de Gauss se obtiene

$$LSCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{n^2 h\sqrt{\pi}}\sum_{i<j}\left(e^{-c_{ij}^2/4} - \sqrt{8}e^{-c_{ij}^2/2}\right). \tag{8.73}$$

Validación cruzada sesgada. BCV

En lugar de trabajar con la expresión exacta del MISE como en (8.62) se trabaja con la expresión aproximada (8.28), que podemos expresar

$$AMISE = \frac{R(K)}{nh} + \frac{h^4 k_2^2 R(f'')}{4}. \tag{8.74}$$

Scott y Terrell (1987) muestran que bajo condiciones de regularidad que implican funciones núcleo con decrecimiento exponencial en las colas, como es el caso de la función núcleo de Gauss, o bien funciones núcleo simétricas con soporte finito en [-1,1] con derivadas de la densidad y de la función núcleo continuas hasta orden 4 y que $K^{(i)}(\pm 1) = 0$ para $0 \leq i \leq 1$, se verifica

$$\mathrm{E}R(\hat{f}'') = R(f'') + \frac{R(K'')}{nh^5} + O(h^2),$$

donde $\frac{R(K'')}{nh^5}$ es asintóticamente constante dado que el parámetro de ventana óptimo es $O(n^{-1/5})$.

Nótese que la función núcleo de Epanechnikov no verifica la condición requerida.

En (8.74) reemplazamos $R(f'')$ por

$$\hat{R(f'')} = R(\hat{f}'') - \frac{R(K'')}{nh^5},$$

y se puede comprobar que resulta

$$BCV(h) = \frac{R(K)}{nh} + \frac{k_2^2}{2n^2 h} \sum_{i<j} \Phi(c_{ij}), \tag{8.75}$$

con c_{ij} igual que en (8.72) y con

$$\Phi(c) = \int K''(w) K''(w+c) dw.$$

Definimos el estimador BCV del parámetro de ventana h_{BCV} como el valor que minimize (8.75)

$$\hat{h}_{BCV} = \mathrm{argmin}_h \ \mathrm{BCV}\ (h) \tag{8.76}$$

Con la función núcleo de Gauss se obtiene

$$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2 h\sqrt{\pi}} \sum_{i<j} \left(c_{ij}^4 - 12c_{ij}^2 + 12\right) e^{-c_{ij}/4}. \tag{8.77}$$

En Hall y Marron (1987) y en Scott y Terrell (1987) se demuestran una serie de convergencias bajo las condiciones de regularidad siguientes:

Condición 1. f''' absolutamente continua; $f^{(iv)}$ integrable; $R(f^{(iv)}(f)^{1/2})$ y $R(f(f^{(iv)})^{1/2})$ finitas.

Condición 2a. $K \geq 0$ simétrico en $[-1,1]$; K' Holder continuo; $k_2 > 0$.

Condición 2b. K'' absolutamente continuo en $(-\infty,\infty)$; K''' continuo en (-1,1);$R(K''') < \infty$.

La función núcleo de Gauss verifica las condiciones anteriores, así como también la función núcleo triweight

$$K(t) = 35/32(1-t^2)^3 I_{[-1,1]}(t),$$

siendo esta última la función núcleo más sencilla que satisface las condiciones 2a y 2 b.

Las convergencias demostradas son:

$$n^{1/10}\left(\frac{\hat{h}_{LSCV}}{h_{MISE}} - 1\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{LSCV}), \tag{8.78}$$

con

$$\begin{aligned} \sigma^2_{LSCV} &= C(f,K)R(\beta) \\ C(f,K) = \frac{2}{25}R(f)\{R(f'')\}^{-1/5}\{k_2\}^{-2/5}\{R(K)\}^{-9/5} & \\ \beta(x) &= \gamma(x) + x\gamma'(x) \end{aligned}$$

con $\gamma(x)$ como en (8.71).

$$n^{1/10}\left(\frac{\hat{h}_{BCV}}{h_{MISE}} - 1\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{BCV}), \tag{8.79}$$

con

$$\begin{aligned} \sigma^2_{BCV} &= C(f,K)R(\beta_{BCV}) \\ \beta_{BCV}(x) &= \gamma_{BCV}(x) + x\gamma'_{BCV}(x) \\ \gamma_{BCV}(x) &= \frac{1}{4}k_2^2(K*K)^{(4)}(x). \end{aligned}$$

Para la función núcleo de Gauss se obtiene

$$\frac{\sigma^2_{LSCV}}{\sigma^2_{BCV}} \approx 15,7 \tag{8.80}$$

La estimación LSCV es más variable que la BCV.

También es posible que LSCV y BCV tengan más de un mínimo local.

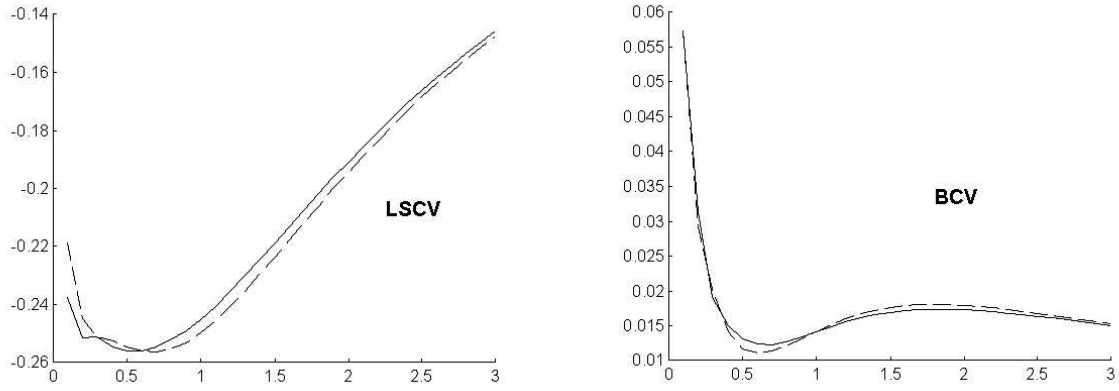

Figura 8.8: LSCV(H) y BCV(h) para dos muestras de tamaño 50 procedentes de una N(0,1) con la función núcleo de Gauss.

En la Figura 8.8 se observa el comportamiento de las estimaciones LSCV y BCV utilizando la función núcleo de Gauss sobre dos muestras simuladas de tamaño 50 de N(0,1). Tal y como puede apreciarse para muchas funciones núcleo, en particular para la normal, se verifica $\lim_{h\to 0^+} BCV(h) = \infty$ y que $\lim_{h\to\infty} BCV(h) = 0$, con $BCV(h) > 0$, para todo $h \in (0, \infty)$, por tanto la función $BCV(h)$ no tiene un mínimo global finito y sí normalmente un mínimo local que determina la estimación del parámetro de ventana.

Validación cruzada pseudo-verosímil

Es una de las técnicas desarrolladas en primer lugar, Habema, Hermans y Van den Broek (1974), pero de resultados no demasidado buenos. Se basa en considerar a h como un parámetro que es estimado por máxima verosimilitud. La función de verosimilitud, dado que la densidad verdadera es desconocida, es estimada por

$$L(h) = \prod_{j=1}^{n} \hat{f}_j(X_j; h),$$

donde $\hat{f}_j$ es la estimación definida sobre la submuestra $X_1, \ldots X_{j-1}, X_{j+1}, \ldots, X_n$. El valor $\hat{h}_{PL}$ es áquel que maximiza $L(h)$.

8.3.4. Métodos Plug-In

Tratan de substituir en la expresión del h_{AMISE} dada en (8.29) el valor $R(f'')$ a través de una muestra piloto. El problema es escoger el parámetro de suavización para esta muestra piloto. Se han propuesto varias aproximaciones.

Sheather y Jones (1991)

$$\begin{aligned} R(f'') &= \int (f'')^2 dx = \int f'' df'(x) = -\int f'''(x) f'(x) dx \\ &= -\int f'''(x) df(x) = +\int f^{(4)}(x) f(x) dx = \psi_4, \end{aligned} \tag{8.81}$$

es decir

$$R(f'') = \psi_4 = \mathrm{E}\{f^{(4)}(x)\}.$$

Un posible estimador es

$$\hat{\psi}_4(g) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}^{(4)}(X_i) = \frac{1}{n^2 g^5} \sum_i \sum_j K^{(4)}\left(\frac{X_j - X_i}{g}\right). \tag{8.82}$$

Definimos

$$\hat{h}_{DPI} = \left\{ \frac{R(K)}{k_2^2 \hat{\psi}_4(g)} \right\}^{1/5} n^{-1/5}. \tag{8.83}$$

Evidentemente $\hat{h}_{DPI}$ depende de g *(pilot bandwith)* que puede ser obtenida de

$$g = \left(\frac{2K^{(4)}(0)}{k_2} \right)^{1/7} R(f''')^{-1/7} n^{-1/7} \tag{8.84}$$

con $R(f''')$ estimado por un proceso análogo obteniendo una función ψ_6, y así sucesivamente hasta que finalmente se estima el término $R(f^{(i)})$ tomando como referencia una distribución paramétrica como puede ser la normal. Generalmente no se realizan más de dos o tres procesos iterativos Jones y Sheather (1991). En Cao et al. (1994) se muestra el buen comportamiento del estimador anterior tomando directamente en (8.84) la referencia a la distribución normal para estimar $R(f''')$.

Hall, Sheather, Jones y Marron (1991)

Trabajan con una mejor aproximación del AMISE, en particular mejoran la aproximación del sesgo, con lo que se obtiene

$$AMISE(h) = \frac{R(K)}{nh} - \frac{R(f)}{n} + \frac{1}{4} h^4 k_2^2 R(f'') - \frac{1}{24} h^6 k_2 k_4 R(f'''). \tag{8.85}$$

Nótese que el segundo término es constante, no depende de h y puede ser ignorado. En el artículo se demuestra que el minimizador de (8.85) viene dado por

$$\hat{h}_{PI} = \left(\frac{J_1}{n} \right)^{1/5} + \left(\frac{J_1}{n} \right)^{3/5} J_2, \tag{8.86}$$

con

$$J_1 = \frac{R(K)}{k_2^2 R_{h_1}(f'')} \qquad \text{y} J_2 = \frac{k_4 R_{h_2}(f''')}{20 k_2 R_{h_1}(f'')}$$

y estimando $R_{h_1}(f'')$ y $R_{h_2}(f''')$ por

$$\hat{R}_{h_1}(f'') = \frac{1}{n(n-1)h_1^5} \sum_{i,j} L^{(4)}\{(X_i - X_j)/h_1\}$$

y

$$\hat{R}_{h_2}(f''') = \frac{-1}{n(n-1)h_2^7} \sum_{i,j} \phi^{(6)}\{(X_i - X_j)/h_2\}.$$

En el artículo los autores sugieren la utilización de

$$L^{(4)}(x) = 135135(-184756x^{10} + 504900x^8 - 491400x^6 + 200200x^4 - 29700x^2 + 756)/16384,$$

con $-1 \leq x \leq 1$ y $\hat{h}_1 = 4,29IQn^{-1/11}$; y la de la función núcleo de Gauss para $\phi^{(6)}$ con $\hat{h}_2 = 0,91IQn^{-1/9}$.

8.3.5. Métodos basados en Bootstrap

La idea básica es estimar el MISE a través de una versión bootstrap de la forma

$$MISE_*(h) = \mathrm{E}_* \int (\hat{f}^*(t;h) - \hat{f}(t;g))^2 dt, \tag{8.87}$$

donde E_* es la esperanza respecto a la muestra bootstrap $X_1^*, \ldots, X_n^*$, g es un paramétro de ventana piloto, $\hat{f}(t;g)$ una estimación de la densidad basada en la muestra original $X_1, \ldots, X_n$ y $\hat{f}^*(t;h)$ una estimación basada en la muestra bootstrap. Escogemos el valor $\hat{h}_{BT}$ que minimiza (8.87). Las diferencias entre las diferentes versiones radican en la elección de g y en la manera de generar la muestra bootstrap.

Destaquemos los trabajos de Taylor (1989) que utilizando $g = h$ llega a

$$MISE_*(h) = \frac{1}{2n^2 h(2\pi)^{1/2}} (\sum_{i,j} exp\left\{-\frac{(X_i - X_j)^2}{8h^2}\right\} \tag{8.88}$$

$$-\frac{4}{3^{1/2}} \sum_{i,j} exp\left\{-\frac{(X_i - X_j)^2}{6h^2}\right\} \tag{8.89}$$

$$+2^{1/2} \sum_{i,j} exp\left\{-\frac{(X_i - X_j)^2}{4h^2}\right\} + n2^{1/2}). \tag{8.90}$$

Otros enfoques pueden encontrarse en Faraway y Jhun (1990), Hall (1990), Cao-Abad (1990).

Se han realizado estudios comparativos por simulación del comportamiento de los diferentes selectores del paramétro de ventana, destaquemos los trabajos de Cao, Cuevas y Manteiga

(1994) y los de Jones, Marron y Sheather (1996a,1996b), Devroye (1997). Las conclusiones obtenidas en los diferentes estudios muestran el buen comportamiento de los estimadores basados en las técnicas Plug-In y Bootstrap frente a los basados en validación cruzada. En cuanto a los estudios teóricos, se ha demostrado que la convergencia de los paramétros de ventana estimados con los métodos Plug-In o Bootstrap es de orden $n^{-5/14}$, mucho más cercana al límite de $n^{-1/2}$, Hall y Marrron (1987), que la de los métodos de validación cruzada $n^{-1/10}$.

Capítulo 9

Estimación de Densidades Multivariantes

9.1. Definición y propiedades básicas

Dada la muestra aleatoria $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de elementos $\mathbf{X}_i \in \Re^d$, definimos la estimación de la densidad por núcleos multivariantes, con función núcleo $K : \Re^d \to \Re$ de la forma

$$\hat{f}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n} K\{\mathbf{H}^{-1}(\mathbf{x} - \mathbf{X}_i)\}. \tag{9.1}$$

donde $\mathbf{H}$ es una matriz simétrica y definida positiva de orden $d \times d$ que será la denominada matriz de anchos de ventana y donde la función núcleo es generalmente una función de densidad multivariante

$$\int_{\Re^d} K(\mathbf{x})\, d\mathbf{x} = 1 \tag{9.2}$$

Las más usuales en $\Re^d$ son:

- Función núcleo multivariante de Gauss (Función de densidad normal multivariante estandar)

$$K_N(\mathbf{x}) = (2\pi)^{-d/2} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})$$

- Función núcleo multivariante de Bartlett-Epanechnikov

$$K_e(\mathbf{x}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-\mathbf{x}^T\mathbf{x}) & \text{si } \mathbf{x}^T\mathbf{x} < 1 \\ 0 & \text{en caso contrario} \end{cases}$$

donde c_d es el volumen de la esfera unidad de dimensión d

$$c_d = \pi^{d/2}/\Gamma((d/2)+1), \tag{9.3}$$

por ejemplo: $c_1 = 2, c_2 = \pi, c_3 = 4\pi/3$, etc.

- Otras funciones útiles para el caso $d = 2$ son:

$$K_2(\mathbf{x}) = \begin{cases} 3\pi^{-1}(1-\mathbf{x}^T\mathbf{x})^2 & \text{si } \mathbf{x}^T\mathbf{x} < 1 \\ 0 & \text{en caso contrario} \end{cases}$$

$$K_3(\mathbf{x}) = \begin{cases} 4\pi^{-1}(1-\mathbf{x}^T\mathbf{x})^3 & \text{si } \mathbf{x}^T\mathbf{x} < 1 \\ 0 & \text{en caso contrario} \end{cases}$$

En la práctica una de las opciones más recomendada es la utilización de

- Producto de funciones núcleo univariantes

$$K(\mathbf{x}) = \prod_{i=1}^{d} K(x_i)$$

Algunas de las condiciones generalmente exigidas a la función núcleo $K(\mathbf{x})$ vienen dadas por las siguientes ecuaciones matriciales

$$\begin{aligned} \int_{\Re^d} K(\mathbf{x})\, d\mathbf{x} &= \mathbf{1} \\ \int_{\Re^d} \mathbf{x}K(\mathbf{x})\, d\mathbf{x} &= \mathbf{0} \\ \int_{\Re^d} \mathbf{x}\mathbf{x}^T K(\mathbf{x})\, d\mathbf{x} &= I_d \end{aligned} \tag{9.4}$$

Si K es una densidad de probabilidad multivariante, las dos últimas ecuaciones anteriores resumen muchas propiedades de las funciones núcleo marginales. La segunda ecuación dice que las medias de las marginales son iguales a cero y la tercera que los kernels marginales son incorrelacionados dos a dos y con varianza unidad.

Volviendo a la matriz $\mathbf{H}$ podemos considerar algunas clases de valores posibles para dicha matriz

$$\mathcal{H}_1 = \{h_1\mathbf{I} : h_1 > 0\}$$

$$\mathcal{H}_2 = \{\text{diag}(h_1, \ldots, h_d) : h_1, \ldots, h_d > 0\}$$

o en el caso bivariante (d=2)

$$\mathcal{H}_3 = \left\{ \begin{pmatrix} h_1 & h_{12} \\ h_{12} & h_2 \end{pmatrix} : h_1, h_2 > 0, h_{12}^2 < h_1 h_2 \right\}.$$

Nótemos que $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3$ y que en el caso bivariante cada clase representa estimadores con uno, dos o tres parámetros de suavización independientes.

Es fácil observar que utilizando la función núcleo de Gauss

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{x}) = (2\pi)^{-1}|\mathbf{H}|^{-1} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{H}^{-2}\mathbf{x}\right), \tag{9.5}$$

que es la densidad de una distribución normal multivariante con vector de medias $\mathbf{0}$ y matriz de covarianzas $\mathbf{H}^2$. La pertenencia a $\mathcal{H}_1$ significa que la masa de la función núcleo será esférica, a $\mathcal{H}_2$ significa que será elíptica con con los ejes ortogonales y en el caso $\mathcal{H}_3$ elíptica con los ejes en cualquier orientación.

Bajo la axiomática anterior y con la parametrización

$$\mathbf{H} = h \cdot A \tag{9.6}$$

donde A es una matriz $d \times d$ con $|A| = 1$ y $h > 0$, en Scott (1992) a través de la forma multidimensional del desarrollo de Taylor pero siguiendo el mismo esquema que en el caso univariante, se muestra que para una estimación como la definida en (9.1) el error cuadrático medio integrado asintótico toma la forma

$$AMISE = \frac{R(K)}{nh^d} + \frac{1}{4}h^4 \int_{\Re^d} [tr\{AA^T\nabla^2 f(\mathbf{x})\}]^2 d\mathbf{x}, \tag{9.7}$$

donde $R(K) = \int_{\Re^d} K(\mathbf{x})^2 d\mathbf{x}$ y $\nabla^2 f(\mathbf{x})$ es $[\partial^2 f/(\partial x_i \partial x_j)]$. El primer sumando corresponde a la AIV y el segundo al IASB.

Bajo la parametrización anterior se tiene que si por ejemplo $\mathbf{H} \in \mathcal{H}_2$

$$H = \begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_d \end{pmatrix}; \quad \text{entonces} \quad H = h \cdot \begin{pmatrix} h_1/h & & 0 \\ = & \ddots & \\ 0 & & h_d/h \end{pmatrix},$$

donde $h = (h_1 h_2 \cdots h_d)^{1/d}$.

9.2. Selección del parámetro de suavización

9.2.1. Referencia a la distribución Normal

La elección óptima de la matriz de anchos de ventana será aquella que minimza el AMISE. Silverman (1986) presenta algunos resultados para el parámetro de suavización en el caso $\mathbf{H} \in \mathcal{H}_1$, es decir $\mathbf{H} = h\mathbf{I}$, se obtiene

$$AMISE = \frac{1}{nh^d} R(K) + \frac{1}{4}h^4 k_2^2 \int [\nabla^2 f(\mathbf{x})\}]^2 d\mathbf{x}, \tag{9.8}$$

Función núcleo	Dimensión	$A(K)$
Mult. Gaussian	2	1
Mult. Gaussian	d	$\{4/(d+2)\}^{1/(d+4)}$
Mult. Epanechnikov	2	2.40
Mult. Epanechnikov	d	$\{8c_d^{-1}(d+4)(2\sqrt{\pi})^d\}^{1/(d+4)}$
K_2	2	2,78
K_3	2	3,12

Cuadro 9.1: Valor de la constante $A(K)$ para diversas funciones núcleo multivariantes

y se obtiene un parámetro óptimo

$$h^* = \left\{ \frac{dR(K)}{k_2^2 \int [\nabla^2 f(\mathbf{x})\}]^2 d\mathbf{x} n} \right\}^{1/(d+4)} \tag{9.9}$$

versión multivariante de la forma obtenida en el caso univariante. Una posibilidad es considerar los datos procedentes de una distribución normal multivariante de varianza unidad, se obtiene un valor óptimo para el ancho de ventana que minimiza el AMISE

$$h^* = A(K)\, n^{-1/(d+4)}$$

donde la constante $A(K)$ depende de la función núcleo utilizada según se muestra en la tabla 9.1.

Para unos datos cualesquiera con matriz de covarianzas S, tomaremos como ancho de ventana óptimo

$$h^* = \sigma \cdot h$$

donde $\sigma^2 = d^{-1} \sum_i s_{ii}$.

Scott (1992) en forma análoga propone para datos normales con las variables independientes y la función núcleo normal multivariante

$$h_i^* = \left(\frac{4}{d+2} \right)^{1/(d+4)} \sigma_i n^{-1/(d+4)}. \tag{9.10}$$

Aquí considerando diferente parámetro de suavización para cada dimensión. Dado que la constante en (9.10) varía entre 0,924 y 1,059 una forma sencilla de asignar un parámetro de suavización, según propone Scott, es

$$\hat{h}_i^* = \hat{\sigma}_i n^{-1/(d+4)} \tag{9.11}$$

En Wand (1992) se muestra que para datos normales bivariantes con coeficiente de correlación ρ y utilizando la función núcleo de Gauss para $\mathbf{H} \in \mathcal{H}_3$ el valor que minimiza el AMISE es

$$\mathbf{H} = \mathbf{\Sigma}^{1/2}\, n^{-1/6} \tag{9.12}$$

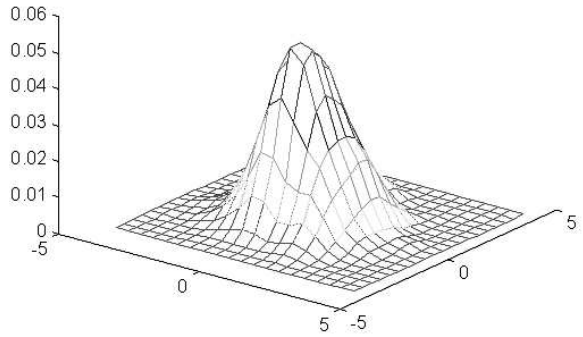

Figura 9.1: Estimación tipo núcleo multivariante.

Para otras funciones núcleo, parámetros que resulten en funciones núcleo equivalente pueden obtenerse dividiendo por la desviación standard de la función núcleo tal y como realizábamos en el caso univariante.

Scott (1992) también propone para el caso bidimensional y utilizando el producto de funciones núcleo univariantes para el caso de datos normales bivariantes y funciones núcleo de Gauss

$$\hat{h}_i = \sigma_i(1-\rho^2)^{5/12}(1+\rho^2/2)^{-1/6}n^{-1/6} \quad i = 1, 2.$$

No existen demasiados estudios al respecto pero pueden adaptarse algunos de los métodos de selección del parámetro de suavización vistos en el caso univariante.

En la Figura 9.1 presentamos una estimación de una densidad bivariante con una muestra de tamaño 200 simulada de una normal standard bivariante, utilizando la función núcleo de Gauss bivariante y un parámetro de ventana único de $h = 0{,}4135$.

9.3. Consideraciones sobre el tamaño muestral

Supongamos que la verdadera densidad es la normal multivariante standard y que la función núcleo es también la normal. Supongamos que deseamos estimar la densidad en el punto $\mathbf{0}$ y que el parámetro de ventana ha sido elegido de forma que minimiza el error cuadrático medio en este punto. Presentamos en la tabla 9.2 el tamaño muestral requerido para que el MSE sea menor a 0,1 en función de la dimensión.

Son evidentes de la contemplación de la tabla los problemas con los que nos encontraremos cuando realizemos estimaciones de la densidad con un elevado número de variables.

Dimensión	Tamaño muestral
1	4
2	19
3	67
4	223
5	768
6	2.790
7	10.700
8	43.700
9	187.000
10	842.000

Cuadro 9.2: Tamaño muestral requerido para asegurar un MSE menor que 0,1 en el punto cero cuando estimamos una normal standard con la función núcleo de Gauss y parámetro de suavización óptimo.

Capítulo 10

Estimación por núcleos adaptables

10.1. Introducción

La idea básica que se encierra en estos métodos es considerar que el parámetro de suavización no tiene que ser fijo sino que puede variar para los diferentes datos muestrales según la densidad de observaciones presentes en un entorno de los mismos. Zonas con baja densidad de observaciones, por ejemplo en densidades con largas colas, permiten un paramétro de suavización mayor que al mismo tiempo evite distorsiones en las estimaciones resultantes.

Fix y Hodges (1951) en su trabajo pionero centrado en los problemas del análisis discriminante, propusieron el método del vecino más próximo (*nearest neighbor estimator*). En un punto fijo $\mathbf{x}$ y para un entero fijo k, sea $D_k(\mathbf{x})$ la distancia euclídea de $\mathbf{x}$ a su k-ésimo vecino más próximo entre $\mathbf{X}_1, \ldots, \mathbf{X}_n$, y sea $V_k(\mathbf{x}) = c_d[D_k(\mathbf{x})]^d$ el volumen de la esfera d-dimensional de radio $D_k(\mathbf{x})$, donde c_d es el volumen de la esfera unidad d-dimensional tal y como se define en (9.3). El estimador de la densidad del k-ésimo vecino más próximo (k-NN) se define por

$$\hat{f}(x) = \frac{k/n}{V_k(\mathbf{x})}. \tag{10.1}$$

Loftsgaarden and Quesenberry (1965) demostraron que el estimador anterior es consistente si $k = k_n \to \infty$ y $k_n/n \to 0$ cuando $n \to \infty$. El estimador (10.1) puede ser redefinido como un estimador tipo núcleo

$$\hat{f}(x) = \frac{1}{n[D_k(\mathbf{x})]^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{D_k(\mathbf{x})}\right), \tag{10.2}$$

donde el parámetro de suavización es k y donde la función núcleo es la rectangular. Sin embargo estudiadas las propiedades del estimador anterior resulta discontinuo y con integral infinita y por tanto con problemas tanto teóricos como prácticos para su utilización.

Un intento de superar los problemas del estimador anterior es la definición del estimador

de núcleo variable (*variable kernel estimator*), definido de la siguiente forma

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H_{ik}^{d}}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{H_{ik}}\right) \tag{10.3}$$

donde la ventana variable $H_{ik} = hD_k(\mathbf{X}_i)$ no depende de $\mathbf{x}$ como en (10.2), h es un parámetro de suavización, y k controla el comportamiento local de H_{ik}. El estimador (10.3) sí que es una verdadera densidad bajo las condiciones usuales de regularidad de las funciones núcleo y también está demostrada su consistencia, Wagner (1975), Devroye (1985).

10.2. Estimador por núcleos adaptables

10.2.1. Definición

El método de estimación por núcleos adaptables es un proceso que consta de dos fases. Una estimación inicial se utiliza para tener una idea de la densidad y posteriormente esta estimación se utiliza para definir los diferentes parámetros de suavización que corresponden a cada observación y construir la estimación final.

El algoritmo de construcción es el siguiente:

1. Construir una estimación piloto $\tilde{f}_h^0$ con ancho de ventana fijo h satisfaciendo $\tilde{f}_h^0(\mathbf{X}_i) > 0 \ \forall i$.

2. Definimos los anchos de ventana locales como

   $$\lambda_i = (\tilde{f}_h^0(\mathbf{X}_i)/g)^{-\alpha}$$

   donde g es la media geométrica de $\tilde{f}_h^0(\mathbf{X}_i)$:

   $$\log g = n^{-1}\sum \log \tilde{f}_h^0(\mathbf{X}_i)$$

   y α es el parámetro de sensibilidad, un número que satisfaga $0 \leq \alpha \leq 1$.

3. Definimos la estimación de núcleo adaptable como

   $$\hat{f}_h(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} h^{-d}\lambda_i^{-d}K\{h^{-1}\lambda_i^{-1}(\mathbf{x}-\mathbf{X}_i)\} \tag{10.4}$$

   donde K es la función núcleo con las condiciones habituales y h el parámetro de suavización.

En los diferentes estudios realizados se ha comprobado una cierta independencia de los resultados obtenidos respecto de la estimación piloto de la densidad, por tanto cualquier estimación razonable parece aceptable, Silverman (1986). Una alternativa es utilizar una estimación tipo núcleo con ancho de ventana fijo obtenido por ejemplo tomando como referencia una distribución standard. Este parámetro fijo h obtenido puede ser el mismo que se incluye posteriormente en la estimación final (10.4), con la única desventaja de requerir una recalculación de la densidad piloto si se desea cambiar h, pero con la ventaja de implicar una suavización general de los datos del mismo orden en la estimación piloto o en la final. En la estimación piloto no se necesita ninguna condición particular, por lo que es aconsejable trabajar con funciones núcleo simples como puede ser la de Bartlett-Epanechnikov.

Hacer que los anchos de ventana locales dependan de una potencia de la densidad piloto permite dotar al método de una gran flexibilidad, dado que si por ejemplo $\alpha = 0$ el método se reduce a una estimación con parámetro fijo igual a h. α es el denominado parámetro de sensibilidad de la estimación.

10.2.2. Elección del parámetro de sensibilidad

Una primera elección puede ser $\alpha = 1/d$. La razón de esta elección es que, supuesto un valor pequeño para h, el número esperado de observaciones en una esfera de radio $hf(\mathbf{x})^{-1/d}$ centrada en $\mathbf{x}$ es aproximadamente igual a $f(\mathbf{x}) \times$ (volumen de la esfera) $= c_d h^d$, por tanto el número de observaciones afectadas por la función núcleo es aproximadamente el mismo en cualquier punto de la densidad independientemente del punto concreto.

Sin embargo una de las elecciones que ha demostrado mejor comportamiento en la práctica, Abramson (1982), Silverman (1986), es $\alpha = 1/2$. La razón de este buen comportamiento puede explicarse al calcular el sesgo de la estimación en un punto, Silverman (1986) muestra que en el caso univariante y tomando $\lambda_i = f(X_i)^{-1/2}$ se verifica

$$\mathrm{E}\hat{f}(t) - f(t) \approx \frac{h^4}{24 f(t)} A(t) \int y^4 K(y) dy + o(h^4),$$

donde

$$\begin{aligned} A(t) = &- \frac{f^{(4)}(t)}{f(t)} + \frac{8 f'''(t) f'(t)}{f(t)^2} \\ &+ \frac{6 f''(t)^2}{f(t)^2} - \frac{36 f''(t) f'(t)^2}{f(t)^3} + \frac{24 f'(t)^4}{f(t)^4} \end{aligned} \tag{10.5}$$

El sesgo es de orden $O(h^4)$, el mismo orden que se obtiene con un estimador con ventana fija pero con una función núcleo de orden 4, es decir $\int t^2 K(t) dt = 0$, pero al utilizar la estimación con núcleo adaptable utilizamos una función núcleo simétrica y no negativa y por tanto la estimación resultante es siempre no negativa y en consecuencia una verdadera densidad.

10.3. Aplicación al Análisis Discriminante

Dado un conjunto de poblaciones $E_1, \ldots, E_k$ y un individuo problema representado por un vector de observaciones $\mathbf{x}$ podemos plantear el problema de asignar el individuo problema a una de las poblaciones anteriores del modo siguiente. Supongamos conocidas las probabilidades a priori de que el individuo pertenezca a cada una de las k poblaciones $P(E_i)$ y también supongamos que conocemos o podemos estimar a partir de un conjunto de muestras control clasificadas las probabilidades o verosimilitudes del individuo problema condicionadas a cada una de las poblaciones $L_{E_i}(\mathbf{x}) = f_{E_i}(\mathbf{x})$. Bajo estos supuestos, la asignación del individuo problema a una de las poblaciones se efectuará segun la regla

$$\mathbf{x} \in E_i \Longleftrightarrow P(E_i/\mathbf{x}) = \text{máx}\{P(E_1/\mathbf{x}), \ldots, P(E_k/\mathbf{x})\}, \tag{10.6}$$

donde las probabilidades condicionadas de las diferentes poblaciones se obtienen a través de la regla de Bayes

$$P(E_i/\mathbf{x}) = \frac{f_{E_i}(\mathbf{x})P(E_i)}{\sum_{j=1}^{k} f_{E_j}(\mathbf{x})P(E_j)}$$

Si desconocemos y no podemos hacer suposiciones sobre las funciones de densidad condicionadas la forma de estimarlas es utilizando técnicas de estimación no paramétrica de la densidad.

La selección de la técnica de estimación así como de los parámetros de suavización puede realizarse por alguno de los métodos discutidos con anterioridad. Dado que el objetivo fundamental es la correcta clasificación de los individuos, una alternativa a la elección del parámetro de suavización h es buscar aquel valor que minimiza el error de clasificación de los individuos control a través de un leave-one-out. Puede utilizarse el mismo parámetro para todas las poblaciones o quizá pudiera ser más apropiado utilizar un parámetro distinto para cada una.

Presentamos a continuación diversos resultados obtenidos aplicando algunos métodos clásicos de discriminación como son los discriminadores lineal y cuadrático de Fisher, y el método MDP (Villarroya, Ríos, Oller (1995)). Los resultados obtenidos se comparan con los que se obtienen utilizando una estimación no paramétrica de la densidad con la técnica de los núcleos adaptables utilizando las funciones núcleo de Gauss o Epanechnikov multivariantes, parámetro de sensibilidad 0,5 ó 1 y un parámetro de suavización inicial h seleccionado de forma que se minimize el error de clasificación por el método del leave-one-out. El parámetro de suavización se expresa como un porcentaje del valor ótpimo h_{opt} suponiendo una distribución normal multivariante de las observaciones.

10.3.1. Aplicación a diversos ejemplos clásicos y pruebas de simulación

Para la comparación el método MDP, se han considerado los mismos ejemplos y simulaciones que los realizados por los autores del método.

Método	Prob. error de clasif. (Leave-one-out)
LDF	0.0676
QDF	0.0946
$\text{MDP}_{\lambda=1}$	0.0541

Función núcleo	Parámetro de sensibilidad	Ventana óptima	Prob. error (Leave-one-out)
Mult. Gaussian	0.5	70 % - 100 % h_{opt}	0.0541
Mult. Gaussian	1.0	80 % - 100 % h_{opt}	0.0405
Mult. Epanechnikov	0.5	80 % h_{opt}	0.0541
Mult. Epanechnikov	1.0	70 % i 100 % h_{opt}	0.0541

Cuadro 10.1: Lubischew (1962). Discriminación entre tres especies de *Chaetocnema*

En las pruebas de simulación se han realizado 10000 simulaciones consistentes en: dos poblaciones control, ambas de tamaño muestral 25, y un individuo problema con probabilidad 0.5 de pertenecer a cada una. Los resultados se resumen en las tablas 10.1 a 10.6

10.3.2. Generalización para datos discretos y mixtos

Consideremos el caso de datos binarios multivariantes donde cada observación toma los valores 0 ó 1. La distribución de un vector multivariante binario de longitud k viene dada por las probabilidades de cada uno de los 2^k posibles resultados. Sea B^k el espacio $\{0,1\}^k$ de psobiles observaciones multivariantes binarias. Dados dos vectores $\mathbf{x}$ y $\mathbf{y}$ en B^k sea $d(\mathbf{x},\mathbf{y})$ el número de desacuerdos en los correspondientes elementos de $\mathbf{x}$ e $\mathbf{y}$; se verifica

$$d(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}). \tag{10.7}$$

Para cualquier λ tal que $\frac{1}{2} \leq \lambda \leq 1$, definamos la función núcleo K como

$$K(\mathbf{y}|\mathbf{x},\lambda) = \lambda^{k-d(x,y)}(1-\lambda)^{d(x,y)}. \tag{10.8}$$

Que satisface

$$\sum_y K(\mathbf{y}|\mathbf{x},\lambda) = 1 \quad \forall \mathbf{x}\,\text{y}\,\lambda.$$

Dada una muestra $\mathbf{X}_1,\ldots,\mathbf{X}_n$ de observaciones procedentes de una distribución p en B^k, la estimación núcleo de p es

$$\hat{p}(\mathbf{y}) = n^{-1}\sum_i K(\mathbf{y}|\mathbf{X}_i,\lambda). \tag{10.9}$$

Método	Prob. error de clasif. (Leave-one-out)
LDF	0.0769
QDF	0.1026
$\text{MDP}_{\lambda=1}$	0.0769

Función núcleo	Parámetro de sensibilidad	Ventana óptima	Prob. error (Leave-one-out)
Mult. Gaussian	0.5	110 % - 120 % h_{opt}	0.0769
Mult. Gaussian	1.0	100 % h_{opt} i endavant	0.1026
Mult. Epanechnikov	0.5	100 %- 110 % h_{opt}	0.1026
Mult. Epanechnikov	1.0	90 %- 110 % h_{opt}	0.1026

Cuadro 10.2: Lubischew (1962). Discriminación entre dos especies de *Haltica*

Método	Prob. error de clasif. (Leave-one-out)
LDF	0.2258
QDF	0.1935
$\text{MDP}_{\lambda=1}$	0.1613

Función núcleo	Parámetro de sensibilidad	Ventana óptima	Prob. error (Leave-one-out)
Mult. Gaussian	0.5	70 % - 110 % h_{opt}	0.1290
Mult. Gaussian	1.0	60 % - 110 % h_{opt}	0.1290
Mult. Epanechnikov	0.5	90 % h_{opt}	0.1290
Mult. Epanechnikov	1.0	70 %- 90 % h_{opt}	0.1290

Cuadro 10.3: Huang y Li (1991). Discriminación entre dos grupos de mujeres: normales y con enfermedad coronaria.

Método	Prob. error de clasif. (Leave-one-out)
LDF	0.2449
QDF	0.2245
$MDP_{\lambda=20}$	0.1020

Función núcleo	Parámetro de sensibilidad	Ventana óptima	Prob. error (Leave-one-out)
Mult. Gaussian	0.5	30 % h_{opt}	0.1429
Mult. Gaussian	1.0	20 % h_{opt}	0.1429

Cuadro 10.4: Hand (1981). Discriminación entre dos tipos de usuarios del centro de computación de la Universidad de Londres.

Método	% clasificación erronea (Leave-one-out)
LDF	49.33
QDF	45.92
$MDP_{\lambda=1}$	45.57
$MDP_{\lambda=10}$	38.33

Función nucleo	Parámetro de sensibilidad	Ventana óptima	% error (Leave-one-out)	% error con h_{opt}
Mult. Gaussian	0.5	0.6 h_{opt}	38.03	39.25

Cuadro 10.5: Primera simulación: Población 1 - Normal bivariante (0,I) ; Población 2 - Corona circular (R=2 , r=1)

Método	% clasificación erronea (Leave-one-out)
LDF	46.69
QDF	4.53
$\text{MDP}_{\lambda=1}$	34.28
$\text{MDP}_{\lambda=10}$	0.23

Función nucleo	Parámetro de sensibilidad	Ventana óptima	% error (Leave-one-out)	% error con h_{opt}
Mult. Gaussian	0.5	2.0 h_{opt}	0.16	0.98

Cuadro 10.6: Segunda simulación: Población 1 - Normal trivariante (0,I) ; Población 2 - Dos uniformes trivariantes en (-4,-3) y (3,4)

El parámetro λ controla la cantidad de suavización. Cuando $\lambda = 1$ todo el peso del núcleo está concentrado en $\mathbf{y} = \mathbf{x}$, y $\hat{p}(\mathbf{y})$ es la proporción de datos para los cuales $\mathbf{X}_i = \mathbf{y}$. Por otro lado cuando $\lambda = 1/2$, $K(\mathbf{y}|\mathbf{X}_i, \lambda)$ proporciona el mismo peso $(1/2)^k$ a todo $\mathbf{y}$ en B^k y por tanto la estimación es una uniforme discreta sobre B^k.

Para la elección automática del parámetro λ puede utilizarse el método de la validación cruzada máximo verosímil. La función que debe maximizarse es

$$\sum_i \log \hat{p}_{-1}(\mathbf{X}_i). \tag{10.10}$$

La regla que permite asignar una observación $\mathbf{x}$ a una población es idéntica a la presentada en (10.6) con $f_{E_i}(\mathbf{x})$ substituido por $\hat{p}_{E_i}(\mathbf{x})$. Entre los trabajos aplicados que utilizan el método comentado destaquemos Anderson et al. (1972) con dos grupos de individuos uno de ellos con *Keratoconjunctivitis sicca* (KCS) y otro normal y un vector de 10 características binarias. Calculando λ por validación cruzada máximo verosímil dado un grupo de 41 nuevos pacientes todos ellos fueron diagnosticados correctamente.

Si los datos tienen k_1 componentes binarias y k_2 componentes continuas, es posible aplicar la técnica discriminante comentada con la elección de una función núcleo adecuada. Una posible función núcleo es

$$K(\mathbf{y}|\mathbf{x}, \lambda, h) = \lambda^{k_1 - d_1(\mathbf{x},\mathbf{y})}(1-\lambda)^{d_1(\mathbf{x},\mathbf{y})} h^{-k_2} \phi\{h^{-1} d_2(\mathbf{x}, \mathbf{y})\}(2\pi)^{\frac{1}{2}(1-k_2)}, \tag{10.11}$$

donde d_1 es la distancia entre las componentes binarias definida en (10.7), d_2 es la distancia euclídea entre las componentes continuas, ϕ es la función de densidad normal, y λ y h son parámetros de suavización. Si $\mathcal{S}$ es el espacio de posibles observaciones, entonces una

estimación de la densidad dada la muestra $\mathbf{X}_i, \ldots, \mathbf{X}_n$ es

$$\hat{f}(\mathbf{y}) = \frac{1}{n} \sum_i K(\mathbf{y}|\mathbf{x}, \lambda, h) \quad para\ \mathbf{y} \in \mathcal{S}. \tag{10.12}$$

Capítulo 11

Otros métodos de estimación no paramétrica

11.1. Estimación por series ortogonales

Un enfoque diferente al problema de la estimación no paramétrica de la densidad es el proporcionado por el método de las series ortogonales, propuesto en primer lugar por Cencov (1962) y desarrollado posteriormente por varios autores, Schwartz (1967), Kronmal y Tarter (1968), Walter (1977) y un largo etcétera.

En su versión más actual puede formularse de la siguiente manera. Supongamos que queremos estimar una determinada densidad $f(x)$ que suponemos de cuadrado integrable. El método supone que podemos representar a dicha función por un desarrollo en serie de la forma

$$f(x) = \sum_{i=-\infty}^{\infty} a_i \, \varphi_i(x) \quad x \in \Omega \tag{11.1}$$

donde $\{\varphi_i\}$ es un sistema ortonormal completo de funciones reales definidas sobre un conjunto Ω de la recta real, es decir satisfaciendo

$$\int_{\Omega} \varphi_i(x) \, \varphi_j(x) \, dx = \delta_{ij}$$

donde δ_{ij} es la delta de Kronecker, y a_i son los coeficientes de la serie y que vienen definidos por $a_i = \mathrm{E}[\varphi_i(x)]$. Definidos de esta forma es fácil comprobar que los coeficientes a_i minimizan la expresión

$$\begin{aligned} R(a) &= \| \, f(x) - \sum_{i=-\infty}^{\infty} a_i \, \varphi_i(x) \, \|^2 = \\ &= \int f^2(x) dx - 2 \sum_{i=-\infty}^{\infty} a_i \int f(x) \varphi_i(x) + \sum_{i=-\infty}^{\infty} a_i^2 \end{aligned} \tag{11.2}$$

Derivando parcialmente respecto a_i obtenemos

$$\frac{\partial R(a)}{\partial a_i} = -2 \int f(x)\varphi_i(x)dx + 2a_i$$

con lo que efectivamente

$$a_i = \int f(x)\varphi_i(x)dx = \mathrm{E}[\varphi_i(x)] \tag{11.3}$$

minimiza la expresión $R(a)$.

En el caso general de que la única información sobre la función de densidad $f(x)$ provenga de una muestra aleatoria $x_1, \ldots, x_n$, una estimación insesgada de los parámetros a_i viene dada por

$$\hat{a}_i = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(X_j) \tag{11.4}$$

y por tanto resultaría la estimación

$$\hat{f}_n(x) = \sum_{i=-\infty}^{\infty} \hat{a}_i \, \varphi_i(x), \tag{11.5}$$

estimación que sería de poca utilidad si no efectuáramos un proceso de suavización, que en su forma más general consiste en la introducción de una secuencia de pesos b_i simétricos ($b_{-i} = b_i$) con $0 < b_i < 1$, que satisfagan $b_i \to 0$ cuando $i \to \infty$, y definir la nueva estimación

$$\hat{f}_n(x) = \sum_{i=-\infty}^{\infty} b_i \, \hat{a}_i \, \varphi_i(x) \tag{11.6}$$

La velocidad de convergencia a cero determinará la cantidad de suavización introducida. La elección $b_i = 1$ para $-k \leq i \leq k$ y $b_i = 0$ en caso contrario, conduce a una estimación en forma de suma parcial del desarrollo anterior

$$\hat{f}_n(x) = \sum_{i=-k}^{k} \hat{a}_i \, \varphi_i(x). \tag{11.7}$$

Al efectuar el truncamiento no podemos garantizar que en todos los casos las estimaciones verifiquen $\hat{f}_n(x) > 0$ para todo x o que $\int \hat{f}_n(x)dx \equiv 1$, tan solo se cumplirá para elecciones particulares del sistema ortonormal y de la secuencia de pesos.

Una vez escogido un sistema ortonormal de funciones, la bondad y la suavización de la estimación obtenida dependerá evidentemente del número de términos que intervengan en el desarrollo. Se han propuesto algunas reglas para la elección del número óptimo de términos

pero sin embargo ninguna de ellas carece de inconvenientes, vease por ejemplo, Kronmal y Tarter (1968), Hart (1985).

La elección del sistema de funciones también tiene una gran repercusión sobre la calidad de la estimación. Si, como es habitual, no disponemos de ningún conocimiento previo de la forma de la densidad, una de las razones para la elección del sistema puede ser la simplicidad de su implementación. Entre los sistemas ortonormales más utilizados destacan el sistema trigonométrico de Fourier y el sistema ortonormal de Hermite, aunque sin olvidar tampoco los sistemas ortonormales de Laguerre y Legendre.

11.2. Máxima verosimilitud penalizada.

Uno de los métodos de estimación de parámetros más populares es el método de la máxima verosimilitud. Es un resultado conocido la no existencia en general de un máximo para la función de verosimilitud si el subespacio $F \subset \mathcal{L}^1$ al cual pertenece la función de densidad es de dimensión infinita. La verosimilitud sería maximizada por una combinación de funcionales delta de Dirac en los puntos de la muestra

$$\hat{f}_n(x) \to \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \tag{11.8}$$

donde δ es la función delta de Dirac, definida por

$$\begin{aligned} \int_{-\infty}^{\infty} \delta(y)\, dy &= 1 \\ \delta(y) &= 0 \qquad \text{si } y \neq 0. \end{aligned} \tag{11.9}$$

El método de la máxima verosimilitud penalizada se basa en modificar la verosimilitud con un término que cuantifique de alguna manera la rugosidad"de la curva estimada, fue considerado en primer lugar por Good y Gaskins (1971). Sea $R(g)$ un funcional que cuantifica la rugosidad de g al que denominaremos función penalizadora. Definiremos el logaritmo de la verosimilitud penalizado por

$$l_\alpha(g) = \sum_{i=1}^{n} \ln g(X_i) \; - \; \alpha R(g) \tag{11.10}$$

donde α es un parámetro de suavización positivo.

La función de densidad estimada $\hat{f}$ se denomina estimación máximo verosímil penalizada si maximiza $l_\alpha(g)$ sobre la clase de todas las funciones que verifican $\int_{-\infty}^{\infty} g = 1$, $g(x) \geq 0$ para todo x y $R(g) < \infty$. El parámetro α controla la cantidad de suavización.

Good y Gaskins en su trabajo proponen trabajar con la raiz cuadrada de la densidad siendo $\gamma = \sqrt{f}$, proponiendo como funciones penalizadoras

$$R(f) = \int \gamma'^2 \tag{11.11}$$

equivalente a

$$4R(f) = \int \frac{f'^2}{f} \tag{11.12}$$

Dicha función penaliza la pendiente de las estimaciones. Otra función penalizadora propuesta también por Good y Gaskins es la función

$$R(f) = \int \gamma''^2 \tag{11.13}$$

que tomará valores altos si γ posee una gran curvatura local, y el valor cero si γ es una linea recta.

La ventaja de trabajar con γ en lugar de con f es que la restricción $f(x) \geq 0$ se satisface automáticamente si γ es real; además la restricción $\int f = 1$ se sustituye por $\int \gamma^2 = 1$. Good y Gaskins (1971,1980) proponen utilizar un desarrollo de γ en forma de serie de funciones ortonormales siguiendo un desarrollo análogo al de la sección anterior

$$\gamma(x) = \sum_{k=0}^{m} \gamma_k \varphi_k(x) \tag{11.14}$$

La estimación máximo verosímil penalizada se obtiene ahora buscando los coeficientes γ_k que maximizan (11.11) y sustituyendo en (11.14) se forma finalmente la estimación $\hat{f}_n(x) = \hat{\gamma}(x)^2$.

11.3. Secuencias delta.

Muchos de los métodos descritos hasta ahora son casos particulares de la siguiente clase general. Sea $\delta_\lambda(x, y)$ $(x, y \in \mathbf{R})$, una función acotada con un parámetro de suavización $\lambda > 0$. La secuencia $\{\delta_\lambda(x, y)\}$ se llama una secuencia delta sobre $\mathbf{R}$ si $\int_{-\infty}^{\infty} \delta_\lambda(x, y)\varphi(y)\, dy \to \varphi(x)$ cuando $\lambda \to \infty$ para toda función φ infinitamente diferenciable sobre $\mathbf{R}$. Cualquier estimador que pueda ser escrito en la forma

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \delta_\lambda(x, X_j), \qquad x \in \mathbf{R} \tag{11.15}$$

es llamado un estimador por secuencias delta. Los histogramas, los estimadores tipo núcleo y los estimadores por series ortogonales pueden expresarse de la forma (11.15). En algunos casos (histogramas y series ortogonales), λ tomará valores enteros, mientras que en otros (núcleos) tomará valores reales.

Estimador	Secuencia delta
Histogramas	$\delta_m(x, X_j) = \sum_{i=1}^{m} (t_{i+1} - t_i)^{-1} I_{T_i}(x) I_{T_i}(X_j)$
Núcleos	$\delta_h(x, X_j) = \frac{1}{h} K\left(\frac{x - X_j}{h}\right)$
Series ortogonales	$\delta_r(x, X_j) = \sum_{k=-r}^{r} \varphi_k(x)\, \varphi_k(X_j)$

Cuadro 11.1: Equivalencia con las secuencias delta de algunos métodos de estimación no paramétrica

Bibliografía

1. **Abramson, I.S.** (1982). .ºn bandwidth variation in kernel estimates - a square root law."*Ann. Statist, 10, 1217-1223.*

2. **Anderson, J.A., Whaley, K., Williamson, J. and Buchanan, W.W.** *(1972). .ª statistical aid to the diagnosis of Keratoconjunctivitis sicca."Quart. J. Med., 41, 175-189.*

3. **Beran, R.** *(1977). "Minimum Hellinger Distance Estimates for Parametric Models."The Annals of Statistics , 5 , 3 445-463.*

4. **Birgé, L.** *(1985). "Non-Asymptotic Minimax Risk for Hellinger Balls."Probability and Mathematical Statistics , 5 , 1 21-29.*

5. **Birgé, L.** *(1986). .ºn Estimating a Densitiy Using Hellinger Distance and Some Other Strange Facts."Probab. Theory and Related Fields , 71 , 271-291.*

6. **Bochner, S.** *(1955). Harmonic analysis and the Theory of Probability. Univ. of California Press.*

7. **Bonan, S., Lubinsky, D.S. and Nevai, P.** *(1987). .ºrthogonal polynomials and their derivatives,II."SIAM J. Math. Anal. , 18 , 4 1163-1176.*

8. **Boneva, L.I., Kendall, D. and Stefanov, I.** *(1971) "Spline transformations: Three New Diagnostic Aids for the Statistical data-analyst."Journal of the Royal Statistical Society. Series B., 33, 1-70.*

9. **Bowman, A.W.** *(1984). .ªn Alternative Method of Cross-Validation for the Smoothing of Density Estimates."Biometrika, 71, 353-360.*

10. **Bowman, A.W.** *(1985). .ª comparative study of some kernel-based nonparametric density estimators."J. Statist. Comput. Simul., 21, 313-327.*

11. **Bowman, A.W., Hall, P. and Titterington, D.M.** *(1984). Çross-validation in nonparametric estimation of probabilities and probability."Biometrika , 71 , 2 341-351.*

12. **Brunk, H.D.** *(1978). Ünivariate density estimation by orthogonal series. "Biometrika, 65(3): 521-528.*

13. **Brunk, H.D. and Pierce, D.** *(1974). "Estimation of discrete multivariate densities for computer-aided differential."Biometrika , 61 , 3 493-499.*

14. **Buckland, S.T.** *(1985). "Perpendicular distance models for line transect sampling "Biometrics, 41, 177-195.*

15. **Cao, R., Cuevas, A. and Manteiga, W.G.** *(1994). .A comparative study of several smoothing methods in density estimation."Comp. Statis. and Data Analysis , 17 , 2 153-176.*

16. **Cao-Abad, R.** *(1990). Aplicaciones y nuevos rsultados del método bootstrap en la estimación no paramétrica de curvas. Tesis Doctoral, Universidad de Santiago de Compostela.*

17. **Cencov, N.N.** *(1962). "Evaluation of an Unknown Distribution Density from Observations."Soviet Mathematics, 3, 1559-1562.*

18. **Crain, B.R.** *(1976). "More on estimation of distributions using orthogonal expansions. "JASA, 71, 741-745.*

19. **Crain, B.R.** *(1976) "More on Estimation of Distributions Using Orthogonal Expansions."Jour. Amer. Statist. Assoc. , 71 , 355 741-745.*

20. **Cristobal Cristobal, J.A., Faraldo Roca, P. and Gonz lez Manteiga, W.** *(1987). .A class of linear regression parameter estimators constructed by nonparametric estimation."The Annals of Statistics, 15(2), 603-609.*

21. **Csörgo, M. and Révész, P.** *(1986). .A Nearest Neighbour-Estimator for the Score Function."Probab. Theory and Related Fields , 71 , 293-305.*

22. **Cutler, A. and Cordero-Braña, O.I.** *(1996). "Minimum Hellinger distance estimation for finite mixture models."JASA , 91 , 436 1716-1723.*

23. **Cwik, J. and Koronacke, J.** *(1997). .A combined adaptive-mixtures/plug-in estimator of multivariate probability densities."Computational Statistics and Data Analysis, 26, 2, 199-218.*

24. **Cwik, J. and Mielniczuk, J.** *(1995). "Nonparametric rank discrimination method."Comp. Statist. and Data Analysis , 19 , 1 59-74.*

25. **Deheuvels, P.** *(1977). "Estimation nonparametrique de la densité par histogrammes generelisés."Rev. Statist. Appl., 35, 5-42.*

26. **Devroye, L.** *(1985). .A note on the L_1 consistency of variable kernel estimates."The Annals of Statistics, 13, 1041-1049.*

27. **Devroye, L.** *(1997). Üniversal smoothing factor selection in density estimation: theory and practice."Test, 6,2, 223-282.*

28. **Devroye, L. and Gyorfi, L.** *(1985). Nonparametric Density Estimation. The L^1 view. New York. John Wiley.*

29. **Epanechnikov, V.A.** *(1969). "Non-parametric Estimation of a Multivariate Probability Density. "Theory of Probability and its Applications, 14, 153-158.*

30. **Faraway, J. and Jhun, M.** *(1990). "Bootstrap choice of bandwidth for density estimation."J. Amer. Statist. Assoc., 85, 1119-1122.*

31. **Fellner, W.H.** *(1974). "Heuristic estimation of probability densities "Biometrika, 61(3), 485-492.*

32. **Fisher, N.I., Mammen, E. and Marron, J.S.** *(1994). "Testing for multimodality."Comp. Statist. and Data Analysis , 18 , 5 499-512.*

33. **Fix, E. and Hodges, J.L.** *(1951). "Discriminatory analysis, nonparametric estimation: consistency properties."Report No 4, Project no 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.*

34. **Freedman, D. and Diaconis, P.** *(1981). .On the Histogram as a Density Estimator: L_2 Theory.". Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 57, 453-476.*

35. **Friedman, J.H., Stuetzle, W. and Schroeder, A.** *(1984). "Projection Pursuit Density Estimation. "JASA, 79(387), 599-608.*

36. **Gajek, L.** *(1986). .On Improving Density Estimators Which are not Bona Fide Functions."The Annals of Statistics, 14, 1612-1618.*

37. **Gasser, T., Kneip, A. and Köhler, W.** *(1991). .A Flexible and Fast Method for Automatic Smoothing."J. Amer. Statist. Assoc., 86, 643-652.*

38. **Gasser, T., Müller, H.G. and Mammitzsch, V.** *(1985). "Kernels for Nonparametric Curve Estimation."J.R. Statist. Soc. B. , 47 , 2 238-252.*

39. **Glivenko, V.I.** *(1934). Çourse in Probability Theory."Moscow (en ruso).*

40. **González-Manteiga, W., Cao, R. and Marron, J.S.** *(1996). "Bootstrap Selection of the Smoothing Parameter in Nonparametric Hazard Rate."JASA , 91 , 435 1130-1140.*

41. **Good, I.J.** *(1971). "Non-parametric roughness penalty for probability densities. "Nature Physical Science, 229, 29-30.*

42. **Good, I.J. and Gaskins, R.A.** *(1971). "Nonparametric roughness penalties for probability densities. "Biometrika, 58(2), 255-277.*

43. **Good, I.J. and Gaskins, R.A.** *(1980). "Density estimation and Bump-Hunting by the penalized likelihood method exemplified by scattering and meteorite data."JASA, 75(369), 42-73.*

44. **Green, P.J.** *(1987). "Penalized Likelihood for General Semi-parametric Regression Models. "International Statistical Review, 55, 245-259.*

45. **Grund, B., Hall, P. and Marron, J.S.** *(1994). "Loss and Risk in Smoothing Parameter Selection."Journal of Nonparametric Statistics, 4, 107-132.*

46. **Habbema, J.D.F., Hermans, J. and van der Broek, K.** *(1974). .ª stepwise discrimination program using density estimation."En Bruckman, G. (ed.), Compstat 1974. Viena. Physica Verlag, pp. 100-110.*

47. **Hall, P.** *(1990). Üsing the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems."Journal of Multivariate Analysis, 32, 177-203.*

48. **Hall, P. and Hannan, E.J.** *(1988). .ºn stochastic complexity and nonparametric density estimation."Biometrika , 75 , 4 705-714.*

49. **Hall, P. and Marron, J.S.** *(1987). .ºn the Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator."Ann. Statist., 15, 163-181.*

50. **Hall, P. and Wand, M.P.** *(1988). "Minimizing L_1 Distance in Nonparametric Density Estimation."Journal of Multivariate Analysis, 26, 59-88.*

51. **Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S.** *(1991). .ºn optimal data-based bandwith selection in kernel density estimation."Biometrika , 78 , 2 263-269.*

52. **Hand, D.J.** *(1981). Discrimination and Classification. John Wiley and Sons, New York.*

53. **Hart, J.D.** *(1985). .ºn the choice of Truncation Point in Fourier Series Density Estimation."Journal of Statistical Computation and Simulation, 21, 95-116.*

54. **Hjort, N.L.** *(1986). .ºn Frequency Polygons and Averaged Shifted Histograms in Higher Dimensions."Technical report 22, Stanford University.*

55. **Hodges, J.L. and Lehmann, E.L.** *(1956). "The efficiency of some nonparametric competitors of the t-test."Ann. Math. Statist., 27, 324-335.*

56. **Huang, X. and Li, B.** *(1991). .A new discriminant technique: Bayes-Fisher discrimination. Biometrics, 47, 741-744.*

57. **Izenman, A.J.** *(1991). Recent Developments in Nonparametric Density Estimation. "JASA, 86(413), 205-224.*

58. **Jones, M.C. and Sheather, S.J.** *(1991). Üsing non-stochastic terms to advantage in kernel-based estimation of inegrated squared density derivatives. Statistics and Probability letters, 11, 511-514.*

59. **Jones, M.C., Marron, J.S. and Sheather S.J.** *(1996a). .A brief Survey of Bandwith Selection for Density Estimation."JASA , 91 , 433 401-407.*

60. **Jones, M.C., Marron, J.S. and Sheather, S.J.** *(1996b). "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation."Computational Statistics , 11 , 337-381.*

61. **Koronacki, J. and Luboinska, U.** *(1994). "Estimating the density of a functional of several random variables."Comp. Statist. and Data Analysis , 18 , 3 317-330.*

62. **Kronmal, R. and Tarter, M.** *(1968). "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods."Journal of the American Statistical Association, 63, 925-952.*

63. **Kullback, S. and Leibler, R.A.** *(1951). .On Information and Sufficiency."Ann. Math. Statist. , 22 , 79-86.*

64. **Liang, W. and Krishnaiah, P.R.** *(1985). "Multi-stage Nonparametric Estimation of Density Function using Orthonormal Systems."J. of Multivariate Analysis, 17, 228-241.*

65. **Loftsgaarden D.O. and Quesenberry, C.P.** *(1965). .A nonparametric estimate of a multivariate density function. "The Annals of Mathematical Statistics, 36, 1049-1051.*

66. **Lubischev, A.** *(1962). .On the use of discriminant functions in taxonomy."Biometrics, 18, 455-477.*

67. **Marchette, D.J. and Wegman, E.J.** *(1997). "The Filtered Mode Tree."J. of Computational and Graphical Stat. , 6 , 2 143-159.*

68. **Marchete, D., Preibe, C.E., Rogers, G.W. and Solka, J.L.** *(1996). "Filtered Kernel Density Estimation."Computational Statistics, 11, 95-112.*

69. **Marron, J.S. and Tsybakov, A.B.** *(1995). "Visual Error Criteria for Qualitative Smoothing."JASA , 90 , 430 499-507.*

70. ***Miñarro, A. and Oller, J.M.*** *(1992). On a class of probability density functions and their information metric. Sankhyà, Series A , 55 , 2 214-225.*

71. **Móricz, F.** *(1984). .ªpproximation theorems for double orthogonal series."Jour. Approximation Theory , 42 , 107-137.*

72. **Nadaraya, E.A.** *(1965). .ºn non-parametric estimates of density functions and regression curves. "Theory of Probability and its applications, 10, 186-190.*

73. **Nadaraya, E.A.** *(1989). Nonparametric Estimation of Probability Densities and Regression Curve. Kluwer Academic Publishers, Dordrecht, Germany.*

74. **Ott, J. and Kronmal, R.A.** *(1976). "Some classification procedures for multivariate binary data using orthogonal."Jour. Amer. Statist. Assoc. , 71 , 354 391-399.*

75. **Park, B.U. and Marron, J.S.** *(1990). Çomparison of Data-Driven Bandwith Selectors."JASA, vol. 85, No. 409, 66-72.*

76. **Parzen, E.** *(1962). .ºn estimation of a probability density function and mode. "Ann. of Math. Stat. 33, 1065-1076.*

77. **Rosenblatt, M.** *(1956). Remarks on some nonparametric estimatees of a density function. "The Annals of Mathematical Statistics., 27, 832-837.*

78. **Rosenblatt, M.** *(1971). Çurve Estimates. "The Annals of Statistics, 42(6), 1815-1842.*

79. **Rudemo, M.** *(1982). "Empirical Choice of Histograms and Kernel Density Estimators."Scandinavian Journal of Statistics, 9, 65-78.*

80. **Sain, S.R. and Scott, D.W.** *(1996). .ºn locally adaptive density estimation."JASA , 91 , 436 1525-1534.*

81. **Schwartz, S.C.** *(1967). "Estimation of Probability Density by an Orthogonal Series. "Ann. of Math. Statist., 38, 1261-1265.*

82. **Scott, D.W.** *(1992). Multivariate Density Estimation. John wiley and Sons, New York.*

83. **Scott, D.W. and Terrell, G.R.** *(1987). "Biased and Unbiased Cross-Validation in Density Estimation.". J. Amer. Statist. Assoc., 82, 1131-1146.*

84. **Scott, D.W. and Wand, M.P.** *(1991). "Feasibility of Multivariate Density Estimates.". Biometrika, 78, 197-206.*

85. **Silverman, B.W.** *(1978). Çhoosing the window width when estimating a density."Biometrika , 65 , 1 1-11.*

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

87. **Simonoff, J.S.** *(1995). "The anchor position of histograms and frequency polygons: quantitative and qualitative smoothing."Comm. Statist. Simulation Comput., 24, 691-710.*

88. **Smith, E.P. and van Bell, G.** *(1984). "Nonparametric Estimation of Species Richness."Biometrics, 40,119-129.*

89. **Susarla, V. and Walter, G.** *(1981). "Estimation of a multivariate density function using delta sequences."The Annals of Statistics , 9 , 2 347-355.*

90. **Swanepoel, J.W.H.** *(1986). .ºn the construction of Nonparametric Density Function Estimators using the bootstrap."Commun.Statist.-Theor.Meth., 15(5), 1399-1415.*

91. **Tapia, R.A. and Thompson J.R.** *(1978). Nonparametric Probability Density Estimation. Johns Hopkins University Press. Baltimore.*

92. **Tarter, M. and Kronmal, R.** *(1970). .ºn multivariate density estimates based on orthogonal expansions. "The Annals of Mathematical Statistics, 41(2), 718-722.*

93. **Tarter, M.E., Freeman, W. and Hopkins, A.** *(1986). .ª FORTRAN implementation of univariate Fourier series density estimation."Commun. Statist. - Simula. 15(3), 855-870.*

94. **Taylor, C.C.** *(1989). "Bootstrap choice of the smoothing parameter in kernel density estimation."Biometrika , 76 , 4 705-712.*

95. **Terrell, G.R.** *(1990). "The maximal smoothing principle in density estimation."J. Amer. Statist. Assoc., 85, 470-477.*

96. **Tou, J. and González, R.C.** *(1974). Pattern Recognition Principles. Addison Wesley Pu. Co., London, England.*

97. **Turnbull, B.W. and Mitchell, T.J.** *(1984). "Nonparametric Estimation of the Distribution of Time to Onset for Specific Diseases in Survival/Sacrifice Experiments."Biometrics 40, 41-50.*

98. **Villarroya, A., R¡os, M. and Oller, J.M.** *(1995). "Discriminant Analysis Algorithm Based on a Distance Function and on a Bayesian Decision."Biometrics, 51, 908-919.*

99. **Wagner, T.J.** *(1975). "Nonparametric Estimates of Probability Densities."IEEE Transactions on Information Theory, 21, 438-440.*

100. **Walter, G.** *(1977). "Properties of Hermite Series Estimation of Probability Density."Ann. of Statistics, 5(6), 1258-1264.*

101. **Walter, G. and Blum, J.** *(1979). "Probability density estimation usign delta sequences."The Annals of Statistics, 7(2), 328-340.*

102. **Wand, M.P.** *(1992). "Error Analysis for General Multivariate Kernel Estimators."Journal of Nonparametric Statistics, 2, 1-15.*

103. **Wand, M.P. and Devroye, L.** *(1993). "How easy is a given density to estimate?."Compu. Statist. and Data Analysis , 16 , 311-323.*

104. **Watson, G.S.** *(1969). "Density Estimation by Orthogonal Series."The Annals of Mathematical Statistics, 40(4), 1496-1498.*

105. **Watson, G.S. and Leadbetter, M.R.** *(1963). .°n the estimation of the Probability Density."Ann. of Math. Statist., 34, 480-491.*