8 Estimación de densidades

Disponemos de una muestra de datos: $X_1, X_2, \ldots, X_n$ que proviene de una distribución P desconocida, cuya función de densidad f es continua.
Objetivo: Estimar f de manera no paramétrica (sin hacer hipótesis paramétricas sobre f).

Si suponemos que f pertenece a la familia de las densidades normales, entonces el estimador natural de f sería la densidad de la normal con media $\overline{X}_n$ y varianza S_n^2 :

$$\widehat{f}_n(x) = \frac{1}{S_n\sqrt{2\pi}} e^{-\frac{1}{2}\left(x-\overline{X}_n\right)^2 / S_n^2}$$

Sin embargo, en esta sección no haremos ninguna hipótesis sobre la forma de f y dejaremos "los datos hablar, en la medida de lo posible, por sí mismos".

8.1 El estimador de núcleos

Sea $K(x)$ una función de densidad de una distribución con media 0 y varianza 1 (por ejemplo la normal estándar). El estimador de núcleos de f se define por:

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

donde $h > 0$ es el parámetro de suavizado o ancho de banda. Se puede ver el estimador $\widehat{f}_n$ como una suma de "pequeñas montañas". Cada pequeña montaña esta centrada en una observación X_i y tiene una superficie de $1/n$.

- Si h es demasiado pequeño, las montañas estarán muy separadas y observaremos muchos picos.
- Si h es demasiado grande, observaremos una sola montaña plana.
- Un valor intermedio de h debería dar el mejor resultado.

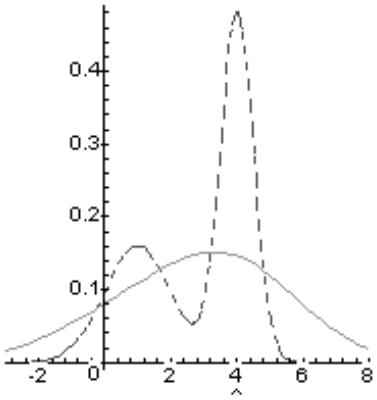
El criterio el más utilizado para evaluar el error de estimación de $\widehat{f}_n$, es el Error Cuadrático Medio:

$$\begin{aligned} \text{EQM}\left(\widehat{f}_n\right) &= \int \text{E}\left(\widehat{f}_n(x) - f(x)\right)^2 dx \\ &= \int \text{var}\left(\widehat{f}_n(x)\right) dx + \int \underbrace{\text{E}^2\left(\widehat{f}_n(x) - f(x)\right)}_{\text{Sesgo de } \widehat{f}_n(x)} dx \end{aligned}$$

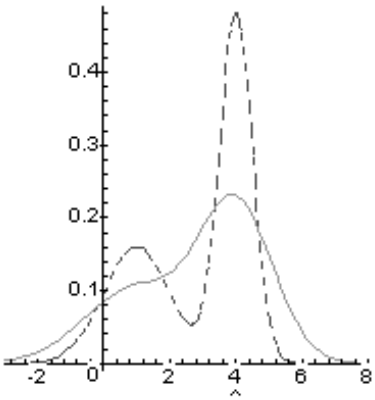
- El termino de varianza es del orden de $\frac{1}{nh}$
- El termino de Sesgo es del orden h^4

Por tanto, el valor de h optimó tiene que ser del orden de $n^{-1/5}$ (cuando $n \to \infty$) lo que conduce a un EQM del orden de $n^{-4/5}$.

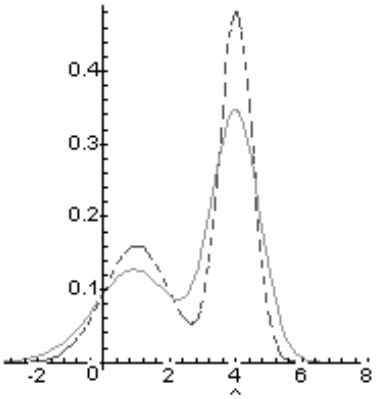
Ejemplo 15 *Disponemos de una muestra de* $n = 1000$ *realizaciones de la distribución* $0.4N(1,1) + 0.6N(4,0.5)$. *En los gráficos a continuación, aparece la verdadera función de densidad* f *(linea discontinua) versus su estimador* $\widehat{f}_n$ *para distintos valores del parámetro de suavizado* h.



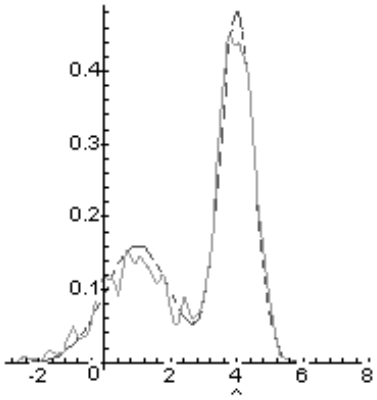
parámetro de suavizado: h=2



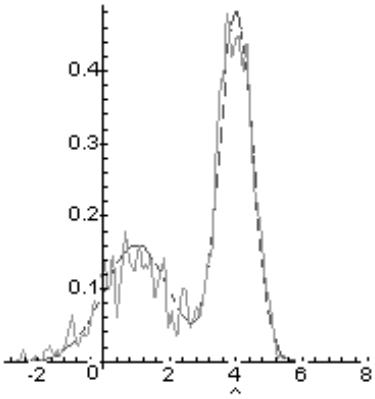
parámetro de suavizado h=1



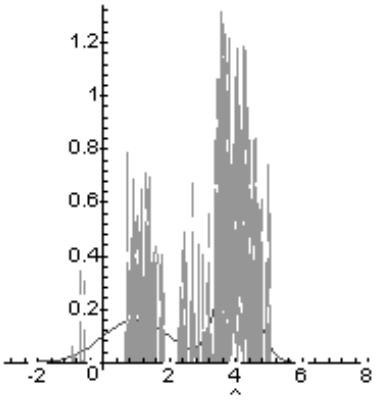
parámetro de suavizado h=0.5



parámetro de suavizado h=0.1



parámetro de suavizado h=0.05



parámetro de suavizado h=0.001