2.4. Escogiendo el ancho de banda

Nota 2.11

La principal característica del ancho de banda

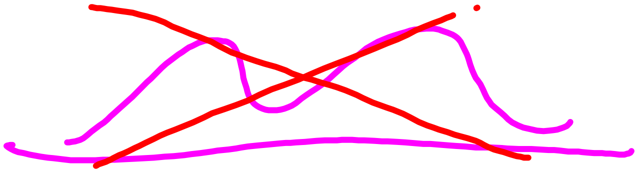
$$h_{opt} = \left(\frac{||K||_2^2}{||f''||_2^2 \left(\mu_2(K)\right)^2 n} \right)^{1/5} = O\left(n^{-1/5}\right).$$

f es desconocid

ES QUE ¡NO FUNCIONA!

Veremos dos métodos para determinar un h que funcione:

- Referencia normal.
- $\boxed{\text{Validación cruzada.}}$

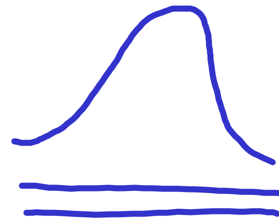


2.4.1. Referencia normal

Cuidado 2.12

Este método es más efectivo si se conoce que la verdadera distribución es bastante suave, unimodal y simétrica.

Más adelante veremos otro método para densidades más generales.



Asuma que f es normal distribuida y se utiliza un kernel K gausiano. Entonces se tiene que

$f \sim N(\mu, \sigma^2)$

$1.06\hat{\sigma}\, n^{-1/5}$

Fácil

$K \sim N(0,1)$

$$\hat{h}_{rn} = \left(\frac{\|K\|_2^2}{\|f''\|_2^2 \left(\mu_2(K)\right)^2 n} \right)^{1/5} = O\left(n^{-1/5}\right)$$
$$= 1{,}06\hat{\sigma} n^{-1/5}.$$

donde

$$\boxed{\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}^2)}}$$

TAREA

Pregunta 2.13

Pruebe que la ecuación anterior es verdadera. Es decir, calcule $\|K\|_2^2$, $\|f''\|_2^2$ y $\mu_2(K)$

Nota 2.14

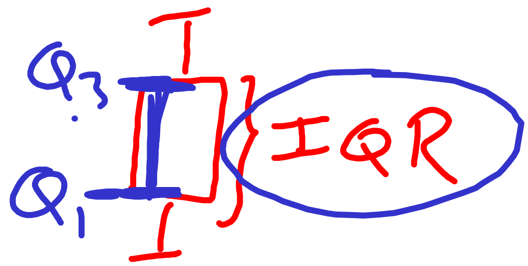
Un problema con $\hat{h}_{rn} = 1{,}06\hat{\sigma}n^{-1/5}$ es su sensibilidad a los valores extremos.

Ejemplo 2.15

La varianza empírica de 1, 2, 3, 4, 5, es 2.5.

La varianza empírica de 1, 2, 3, 4, 5, 99, es 1538.

El rango intercuantil IQR se define como



$$\mathrm{IQR} = Q_3^X - Q_1^X$$

donde Q_1^X y Q_3^X son el primer y tercer de un conjunto de datos $X_1, \ldots, X_n$.

Con el supuesto que $X \sim \mathcal{N}(\mu, \sigma^2)$ entonces $Z = \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Entonces,

$$\begin{aligned}
\mathrm{IQR} &= Q_3^X - Q_1^X \\
&= \left(\cancel{\mu} + \sigma Q_3^Z\right) + \left(\cancel{\mu} + \sigma Q_1^Z\right) \\
&= \sigma\left(Q_3^Z - Q_1^Z\right) \\
&\approx \sigma\left(0{,}67 - (0{,}67)\right) \\
&= 1{,}34\sigma.
\end{aligned}$$

Por lo tanto $\hat{\sigma} = \dfrac{\cancel{R}\ \mathrm{IQR}}{1{,}34}$

$R = IQR$

Podemos sustituir la varianza empírica de la fórmula inicial y tenemos

$$\hat{h}_{rn} = 1{,}06 \frac{R}{1{,}34} n^{-\frac{1}{5}} \approx \underline{0{,}79 \hat{R} n^{-\frac{1}{5}}}$$

Combinando ambos estimadores, podemos obtener,

$$\hat{h}_{rn} = 1{,}06 \min \underbrace{\left\{ \frac{R}{1{,}34}, \hat{\sigma} \right\}} n^{-\frac{1}{5}}$$

2.4.2. Validación Cruzada

Defina el *error cuadrático integrado* como

$$\begin{aligned}\mathrm{ISE}(\hat{f}_h) &= \int \left(\hat{f}_h(x) - f(x)\right)^2 dx \\ &= \int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx + \boxed{\int f^2(x)dx.}\end{aligned}$$

Nota 2.16

El MISE es el valor esperado del ISE.

$E(\mathrm{ISE}(\hat{f}_h)) = \mathrm{MISE}$

Nuestro objetivo es minimizar el ISE con respecto a h.

Primero note que $\int f^2(x)dx$ NO DEPENDE de h. Podemos minimizar la expresión

$$\mathrm{ISE}(\hat{f}_h) - \int f^2(x)dx = \int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx$$

Vamos a resolver esto en dos pasos partes

$E[X] = \int x f(x)dx \to \bar{X} = \frac{1}{n}\sum X_i$

$E[\hat{f}_h(x)] = \int \hat{f}_h(x) f(x) dx$

Integral $\int \hat{f}_h(x) f(x) dx$

El término $\int \hat{f}_h(x) f(x) dx$ es el valor esperado de $\mathrm{E}\left[\hat{f}(X)\right]$. Su estimador es

$i=1$

$j = 1, \dots, n$

$$\widehat{\mathrm{E}\left[\hat{f}(X)\right]} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_h(X_i) = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{X_j - X_i}{h}\right).$$

$i=2$

$j = 1, 2, \dots, n$

Cuidado 2.17

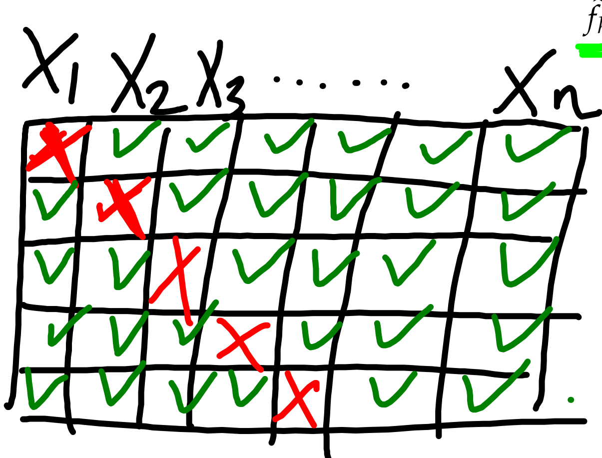
El problema con esta expresión es que las observaciones que se usan para estimar la esperanza son las misma que se usan para estimar $\hat{f}_h(x)$ (Se utilizan doble).

La solución es remover la $i^{\text{ésima}}$ observación de $\hat{f}_h$ para cada i. Redefiniendo el estimador anterior tenemos $\int \hat{f}_h(x) f(x) dx$ como

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

where

$$\hat{f}_{h,-i}(x) = \frac{1}{(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n} K\left(\frac{x - X_i}{h}\right) \frac{1}{h}$$



$j \neq i$

Integral $\int \hat{f}_h^2(x)dx$

Siguiendo con el término $\int \hat{f}_h^2(x)dx$ note que este se puede reescribir como

$$
\begin{aligned}
\int \hat{f}_h^2(x)dx &= \int \left(\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right)^2 dx \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{i=1}^{n}\int K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{i=1}^{n}\int K(u)\,K\left(\frac{X_i-X_j}{h}-u\right)du \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{i=1}^{n} K*K\left(\frac{X_i-X_j}{h}\right).
\end{aligned}
$$

donde $K*K$ es la convolución de K consigo misma.

$$K*K(s) = \int K(u)\cdot K(s-u)\,du.$$

Finalmente tenemos la función,

NO DEPENDE DE f.

$$\mathrm{CV}(h) = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} K * K\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_h(X_i - X_j).$$

Nota 2.18

Note que $\mathrm{CV}(h)$ no depende de f o sus derivadas.

Nota 2.19

Para efectos prácticos es mejor utilizar el criterio

$$CV(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_h(X_i - X_j)$$

y luego calcular numéricamente la integral.