Estimación No Paramétrica de la Densidad

Ana M. Bianco & Paula M. Spano

Introducción al Aprendizaje Estadístico

Estimadores de núcleos: Selección de ventana

En el gráfico anterior se muestra que la elección de la ventana es crucial.

- Una ventana h pequeña dará un estimador muy rugoso, con muchos picos y difícil de interpretar
- una ventana h grande sobresuaviza al estimador de la densidad y enmascara estructuras de los datos.

Sesgo y Varianza de $\widehat{f}(x)$

Comprobemos en el caso del estimador de Parzen que para que todo funcione bien deben pasar dos cosas:

- $h \longrightarrow 0$
- $nh \longrightarrow \infty$

$$\widehat{f}(x) = \frac{1}{2h\,n} \sum_{i=1}^{n} \mathcal{I}_{(x-h,x+h)}(X_i)$$

Sesgo y Varianza de $\widehat{f}(x)$

$$\widehat{f}(x) = \frac{1}{2h\,n}\sum_{i=1}^{n}\mathcal{I}_{(x-h,x+h)}(X_i) \quad \text{con} \sum_{i=1}^{n}\mathcal{I}_{(x-h,x+h)}(X_i) \sim Bi(n,p)$$

donde $p = \mathbb{P}(x-h < X < x+h) = F(x+h) - F(x-h) = p_{x,h}$.

Sesgo y Varianza de $\widehat{f}(x)$

$$\widehat{f}(x) = \frac{1}{2h\,n}\sum_{i=1}^{n}\mathcal{I}_{(x-h,x+h)}(X_i) \quad \text{con} \sum_{i=1}^{n}\mathcal{I}_{(x-h,x+h)}(X_i) \sim Bi(n,p)$$

donde $p = \mathbb{P}(x-h < X < x+h) = F(x+h) - F(x-h) = p_{x,h}$.
Luego, usando las propiedades de esperanza y varianza:

$$\mathbb{E}[\hat{f}(x)] = \frac{F(x+h) - F(x-h)}{2h},$$
$$\mathbb{V}\mathrm{ar}[\hat{f}(x)] = \frac{F(x+h) - F(x-h)}{4nh^2} - \frac{(F(x+h) - F(x-h))^2}{4nh^2}.$$

Sesgo y Varianza de $\widehat{f}(x)$

$$\widehat{f}(x) = \frac{1}{2h\,n}\sum_{i=1}^{n}\mathcal{I}_{(x-h,x+h)}(X_i) \quad \text{con} \sum_{i=1}^{n}\mathcal{I}_{(x-h,x+h)}(X_i) \sim Bi(n,p)$$

donde $p = \mathbb{P}(x-h < X < x+h) = F(x+h) - F(x-h) = p_{x,h}$.
Luego, usando las propiedades de esperanza y varianza:

$$\mathbb{E}[\hat{f}(x)] = \frac{F(x+h) - F(x-h)}{2h},$$
$$\mathbb{V}\mathrm{ar}[\hat{f}(x)] = \frac{F(x+h) - F(x-h)}{4nh^2} - \frac{(F(x+h) - F(x-h))^2}{4nh^2}.$$

Si

- $h \to 0$, entonces $\mathbb{E}[\hat{f}(x)] \to f(x)$ y $\mathbb{V}\mathrm{ar}[\hat{f}(x)] \approx \frac{f(x)}{2nh} - \frac{f(x)^2}{n} \to \infty$
- $nh \to \infty$, $n \longrightarrow \infty$ $h \longrightarrow 0$, entonces sesgo y varianza se reducen.

Sesgo y Varianza de $\widehat{f}(x)$

Comprobamos en el caso fácil que para que todo funcione bien deben pasar dos cosas:

- $h \longrightarrow 0$
- $nh \longrightarrow \infty$

Sesgo y Varianza de $\widehat{f}(x)$

Comprobamos en el caso fácil que para que todo funcione bien deben pasar dos cosas:

- $h \longrightarrow 0$
- $nh \longrightarrow \infty$
- Se puede probar que bajo condiciones generales del núcleo K

$$\begin{aligned} \mathbb{B}[\widehat{f}(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\ \mathbb{V}\mathrm{ar}[\widehat{f}(x)] &\approx \frac{1}{nh} C_2(K)\, f(x) \end{aligned}$$

Sesgo y Varianza de $\widehat{f}(x)$

Comprobamos en el caso fácil que para que todo funcione bien deben pasar dos cosas:

- $h \longrightarrow 0$
- $nh \longrightarrow \infty$
- Se puede probar que bajo condiciones generales del núcleo K

$$\begin{aligned} \mathbb{B}[\widehat{f}(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\ \mathbb{V}\mathrm{ar}[\widehat{f}(x)] &\approx \frac{1}{nh} C_2(K)\, f(x) \end{aligned}$$

- El sesgo es proporcional a $h^2 \Rightarrow$ elijamos h pequeña

Sesgo y Varianza de $\widehat{f}(x)$

Comprobamos en el caso fácil que para que todo funcione bien deben pasar dos cosas:

- $h \longrightarrow 0$
- $nh \longrightarrow \infty$
- Se puede probar que bajo condiciones generales del núcleo K

$$\begin{aligned} \mathbb{B}[\widehat{f}(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\ \mathbb{V}\mathrm{ar}[\widehat{f}(x)] &\approx \frac{1}{nh} C_2(K)\, f(x) \end{aligned}$$

- El sesgo es proporcional a $h^2 \Rightarrow$ elijamos h pequeña
- El sesgo depende de $f''(x)$ que mide la curvatura de f en x

Sesgo y Varianza de $\widehat{f}(x)$

Comprobamos en el caso fácil que para que todo funcione bien deben pasar dos cosas:

- $h \longrightarrow 0$
- $nh \longrightarrow \infty$
- Se puede probar que bajo condiciones generales del núcleo K

$$\begin{aligned} \mathbb{B}[\widehat{f}(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\ \mathbb{V}\mathrm{ar}[\widehat{f}(x)] &\approx \frac{1}{nh} C_2(K)\, f(x) \end{aligned}$$

- La varianza dismimuye a medida que nh crece

Sesgo y Varianza de $\widehat{f}(x)$

Comprobamos en el caso fácil que para que todo funcione bien deben pasar dos cosas:

- $h \longrightarrow 0$
- $nh \longrightarrow \infty$
- Se puede probar que bajo condiciones generales del núcleo K

$$\begin{aligned} \mathbb{B}[\widehat{f}(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\ \mathbb{V}\mathrm{ar}[\widehat{f}(x)] &\approx \frac{1}{nh} C_2(K)\, f(x) \end{aligned}$$

- La varianza dismimuye a medida que nh crece
- Para disminuir la varianza necesitamos h o n grandes.

Error Cuadrático Medio de $\widehat{f}(x)$

Compromiso Sesgo-Varianza

Tenemos que

$$\mathsf{MSE}[\widehat{f}(x)] = \mathsf{Sesgo}^2[\widehat{f}(x)] + \mathbb{V}\mathrm{ar}[\widehat{f}(x)]$$

Por otro lado:

$$\begin{aligned}
\mathbb{B}[\widehat{f}(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\
\mathbb{V}\mathrm{ar}[\widehat{f}(x)] &\approx \frac{1}{nh} C_2(K)\, f(x)
\end{aligned}$$

Error Cuadrático Medio de $\widehat{f}(x)$

Compromiso Sesgo-Varianza

Tenemos que

$$\mathsf{MSE}[\widehat{f}(x)] = \mathsf{Sesgo}^2[\widehat{f}(x)] + \mathbb{V}\mathrm{ar}[\widehat{f}(x)]$$

Por otro lado:

$$\begin{aligned}
\mathbb{B}[\widehat{f}(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\
\mathbb{V}\mathrm{ar}[\widehat{f}(x)] &\approx \frac{1}{nh} C_2(K)\, f(x)
\end{aligned}$$

$$\Rightarrow \mathsf{MSE}[\widehat{f}(x)] \approx \frac{h^4}{4} C_1^2(K)\, (f''(x))^2 + \frac{1}{nh} C_2(K)\, f(x)$$

Si $h \longrightarrow 0$ y $nh \longrightarrow \infty$ $\widehat{f}(x)$ es un estimador consistente de $f(x)$.

$\widehat{f}_h(x)$

Por lo que vimos, el efecto de la ventana sobre la estimación de la densidad puede ser crucial, por lo tanto haremos explícita esta dependencia, denotando

$$\widehat{f}_h(x)$$

al estimador basado en una ventana h.

Error Cuadrático Medio de $\widehat{f}_h(x)$

Compromiso Sesgo-Varianza

Tenemos que

$$\mathsf{MSE}[\widehat{f}_h(x)] = \mathsf{Sesgo}^2[\widehat{f}_h(x)] + \mathbb{V}\mathrm{ar}[\widehat{f}_h(x)]$$

Por otro lado:

$$\begin{aligned} \mathbb{B}[\widehat{f}_h(x)] &\approx \frac{h^2}{2} C_1(K)\, f''(x) \\ \mathbb{V}\mathrm{ar}[\widehat{f}_h(x)] &\approx \frac{1}{nh} C_2(K)\, f(x) \end{aligned}$$

$$\Rightarrow \mathsf{MSE}[\widehat{f}_h(x)] \approx \frac{h^4}{4} C_1^2(K)\, (f''(x))^2 + \frac{1}{nh} C_2(K)\, f(x)$$

Si $h \to 0$ y $nh \to \infty$ $\widehat{f}_h(x)$ es un estimador consistente de $f(x)$.

Error Cuadrático Integrado y Medio Integrado de $\widehat{f}_h$

Como medida global: Error Cuadrático Integrado

$$\mathsf{ISE}[\widehat{f}_h] = \int \left(\widehat{f}_h(x) - f(x)\right)^2 \, \mathrm{d}x$$

Como medida global: Error Cuadrático Medio Integrado

$$\mathsf{MISE}[\widehat{f}_h] = \int \mathsf{MSE}[\widehat{f}_h(x)] \, dx$$

Error Cuadrático Integrado y Medio Integrado de $\widehat{f}_h$

Como medida global: Error Cuadrático Integrado

$$\mathsf{ISE}[\widehat{f}_h] = \int \left(\widehat{f}_h(x) - f(x)\right)^2 \, \mathrm{d}x$$

Como medida global: Error Cuadrático Medio Integrado

$$\mathsf{MISE}[\widehat{f}_h] = \int \mathsf{MSE}[\widehat{f}_h(x)] \, dx$$

Asintóticamente: Error Cuadrático Medio Integrado Asintótico

$$\mathsf{AMISE}[\widehat{f}_h(x)] \;\; = \;\; \frac{h^4}{4} C_1^2(K) \int (f''(x))^2 dx + \frac{1}{nh} C_2(K)$$

Error Cuadrático Integrado y Medio Integrado de $\widehat{f}_h$

Como medida global: Error Cuadrático Integrado

$$\mathsf{ISE}[\widehat{f}_h] = \int \left(\widehat{f}_h(x) - f(x)\right)^2 \mathrm{d}x$$

Como medida global: Error Cuadrático Medio Integrado

$$\mathsf{MISE}[\widehat{f}_h] = \int \mathsf{MSE}[\widehat{f}_h(x)]\, dx$$

Asintóticamente: Error Cuadrático Medio Integrado Asintótico

$$\mathsf{AMISE}[\widehat{f}_h(x)] \quad = \quad \frac{h^4}{4} C_1^2(K) \int (f''(x))^2 dx + \frac{1}{nh} C_2(K)$$

Si llamamos $\|g\|_2^2 = \int g(t)^2 dt$

$$\mathsf{AMISE}[\widehat{f}_h(x)] \quad = \quad \frac{h^4}{4} C_1^2(K) \|f''\|_2^2 + \frac{1}{nh} \|K\|_2^2$$

Error Cuadrático Medio Integrado de $\widehat{f}_h$

$$\mathsf{AMISE}[\widehat{f}_h(x)] \quad = \quad \frac{h^4}{4} C_1^2(K) \|f''\|_2^2 + \frac{1}{nh} \|K\|_2^2$$

- Notemos que MISE es una función de la ventana h.

- Derivando a AMISE respecto de h e igualando a 0, obtenemos que el valor de h que minimiza el AMISE es

$$h = \left\{ \frac{\|K\|_2^2}{C_1^2(K) \|f''\|_2^2 \, n} \right\}^{\frac{1}{5}}$$

Error Cuadrático Medio Integrado de $\widehat{f}_h$

$$\text{AMISE}[\widehat{f}_h(x)] \;=\; \frac{h^4}{4}C_1^2(K)\|f''\|_2^2 + \frac{1}{nh}\|K\|_2^2$$

- Notemos que MISE es una función de la ventana h.

- Derivando a AMISE respecto de h e igualando a 0, obtenemos que el valor de h que minimiza el AMISE es

$$h = \left\{\frac{\|K\|_2^2}{C_1^2(K)\|f''\|_2^2\, n}\right\}^{\frac{1}{5}} = algo * \left(\frac{1}{n}\right)^{\frac{1}{5}} \sim n^{-\frac{1}{5}}$$

Selección de h: Regla de Silverman

El AMISE$[\widehat{f}_h(x)]$ se minimiza en

$$h = \left\{ \frac{\|K\|_2^2}{C_1^2(K)\|f''\|_2^2 \, n} \right\}^{\frac{1}{5}}$$

• Silverman propone reemplazar a $\|f''\|_2^2$ por su valor cuando f es normal:

$$\|f''\|_2^2 = \sigma^{-5} \frac{3}{8\sqrt{\pi}} \approx 0.212 \, \sigma^{-5}$$

y a σ por un estimador $\widehat{\sigma}$

Selección de h: Regla de Silverman

El AMISE$[\widehat{f}_h(x)]$ se minimiza en

$$h = \left\{ \frac{\|K\|_2^2}{C_1^2(K)\|f''\|_2^2\, n} \right\}^{\frac{1}{5}}$$

• Silverman propone reemplazar a $\|f''\|_2^2$ por su valor cuando f es normal:

$$\|f''\|_2^2 = \sigma^{-5}\frac{3}{8\sqrt{\pi}} \approx 0.212\,\sigma^{-5}$$

y a σ por un estimador $\widehat{\sigma}$

$$h_{Sil} = \left(\frac{4\widehat{\sigma}^5}{3\,n}\right)^{\frac{1}{5}} \approx 1.06\,\widehat{\sigma}\,n^{-\frac{1}{5}}$$

- Si f es normal, la ventana h_{Sil} es óptima.
- Si f no es normal, h_{Sil} dará una ventana no muy alejada de la óptima cuando la distribución no es muy diferente a la normal.

Selección de h: Regla de Silverman

$$h_{Sil} = \left(\frac{4\widehat{\sigma}^5}{3\,n}\right)^{\frac{1}{5}} \approx 1.06\,\widehat{\sigma}\,n^{-\frac{1}{5}}$$

- σ puede estimarse por S (sd(datos) en R).

  o

- σ puede estimarse por la distancia intercuartil IQR (IQR(datos) en R). Para que coincida con σ bajo la distribución normal debe dividirse por 1.349
- La ventana óptima de acuerdo a la regla de Silverman resulta:

$$h_{Sil} = 1.06\ \min\left(S, \frac{IQR}{1.349}\right)\ n^{-\frac{1}{5}}$$

Volviendo un poco para atrás...

$$
\begin{aligned}
\mathrm{ISE}[\widehat{f}_h(\cdot)] &= \int \left(\widehat{f}_h(x) - f(x)\right)^2 \,\mathrm{d}x \\
&= \int \left(\widehat{f}_h(x)\right)^2 \,\mathrm{d}x - 2\int \widehat{f}_h(x) f(x)\,\mathrm{d}x \\
&\quad + \int \left(f(x)\right)^2 \,\mathrm{d}x.
\end{aligned}
$$

Como el último término no involucra a h, podemos minimizar

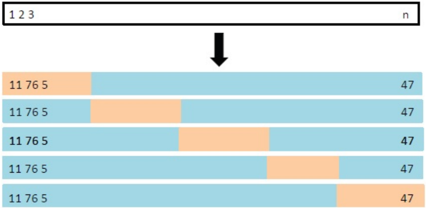
$$
A = \int \left(\widehat{f}_h(x)\right)^2 \,\mathrm{d}x - 2\int \widehat{f}_h(x) f(x)\,\mathrm{d}x
$$

Convalidación Cruzada: Splitting the data

Para evitar sobreajuste usamos *Splitting the data*:

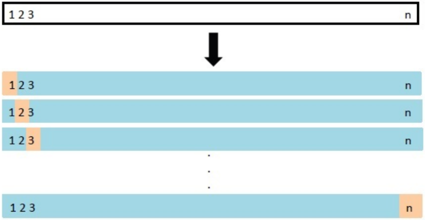
(Fuente: James, Witten, Hastie y Tibshirani, 2013)

Convalidación Cruzada: Splitting the data



Convalidación Cruzada

Para evitar sobreajuste usamos *leave-one-out-Cross-Validation*:



Convalidación Cruzada

Bowman (1984) propone estimar a A por

$$\frac{1}{n}\sum_{i=1}^{n}\int\left(\widehat{f}_h^{(-i)}(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_h^{(-i)}(x_i) \qquad (1)$$

siendo $\widehat{f}_h^{(-i)}(x_i)$ la densidad estimada en el punto x_i sin utilizar al punto x_i:

$$\widehat{f}_h^{(-i)}(x_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x_i - x_j}{h}\right)$$

Una versión más simple que se usa en general es:

$$\mathrm{LSCV}(h) = \int\left(\widehat{f}_h(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_h^{(-i)}(x_i) \qquad (2)$$

Convalidación Cruzada

$$\mathrm{LSCV}(h) = \int \left(\widehat{f}_h(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_h^{(-i)}(x_i) \tag{3}$$

En general, buscamos el mínimo sobre una grilla $h_1, \ldots, h_q$ y luego, eventualmente, se refina.