Estimación de densidad

Teoría de predicción

- Tenemos Y una variable aleatoria cuya distribución desconocemos.
- Queremos usar otras variables para tratar de explicar o predecir a Y.
- Necesitamos un conjunto de posibles predictores $\mathcal{P}$ y una medida para decidir cual de todos ellos ajusta mejor a Y

Medida: $ECM(Y, \widehat{Y}) = E((Y - \widehat{Y})^2)$ donde $\widehat{Y} \in \mathcal{P}$

Mejor predictor constante: $\mathcal{P} = \{\widehat{Y} : \widehat{Y} \text{ es constate}\}$
Buscamos $a_0 = \arg \min_{a \in \mathbb{R}} E(Y - a)^2$ entonces $a_0 = E(Y)$ luego
$\widehat{Y} = E(Y)$

Mejor predictor lineal: $\mathcal{P} = \{\widehat{Y} : \widehat{Y} = aX + b\}$

Buscamos $(a_0, b_0) = \arg \min_{a,b \in I\!R} E(Y - aX - b)^2$ entonces

$\widehat{Y} = \frac{\text{cov}(X,Y)}{V(X)}(X - E(X)) + E(Y)$

Mejor predictor basado en X:

$\mathcal{P} = \{\widehat{Y} : \widehat{Y} = t(X) \text{ medible } V(t(X)) < \infty\}$

Buscamos $\arg \min_{t \in \mathcal{P}} E(Y - t(X))^2$ entonces $\widehat{Y} = E(Y|X)$

Problema: si no tenemos la distribución de Y entonces $E(Y)$? $\text{cov}(X, Y)$? $E(Y|X)$?

La idea: vía una muestra aleatoria estimar los objetos que desconocemos.

Que es una muestra aleatoria???

Es una sucesión de variables independientes $X_1, \ldots, X_n$ con idéntica distribución que X.

Como estimar $F_X(t)$ a partir de $X_1, \ldots, X_n$?

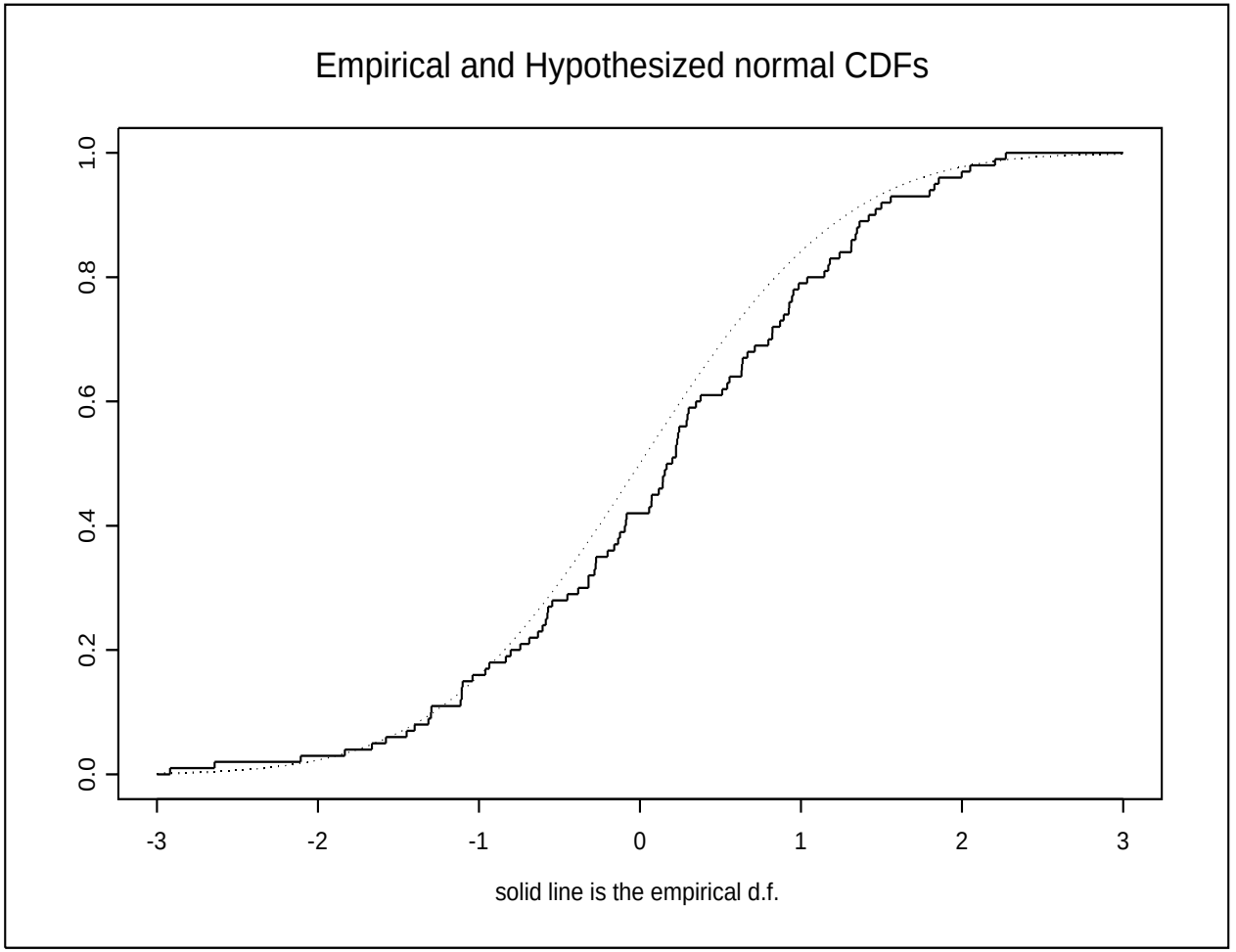
$F(t) = P(X \leq t)$ entonces la primera propuesta es la función de distribución empírica

$$\widehat{F}_n(t) = \frac{\#\{X_i \leq t\}}{n} = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i).$$

- $E(\widehat{F}_n(t)) = F(t)$
- $\widehat{F}_n(t) \to F(t)$

Entonces si queremos estimar $E(X) = \int x \, dF(x)$ en lugar de F ponemos $\widehat{F}_n$

$\widehat{E}(X) = \int x \, d\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n$


Empirical and Hypothesized normal CDFs
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
solid line is the empirical d.f.

Que hacer si tenemos mas información de X?

Por ejemplo X una v.a. continua $\widehat{F}_n$ no es derivable mientras que F si.

Se puede mejorar la estimación?

Hay dos enfoques diferentes.

Estadística paramétrica:

Acá se supone que F pertenece a una familia de distribución conocidad, pero que no conocemos los paramétros.

Por ejemplo $X \sim \mathcal{N}(\mu_0, \sigma_0^2)$ buscamos estimar μ_0 y σ_0

Estadística no paramétrica:

Acá no se hace ningún supuesto sobre la densidad.

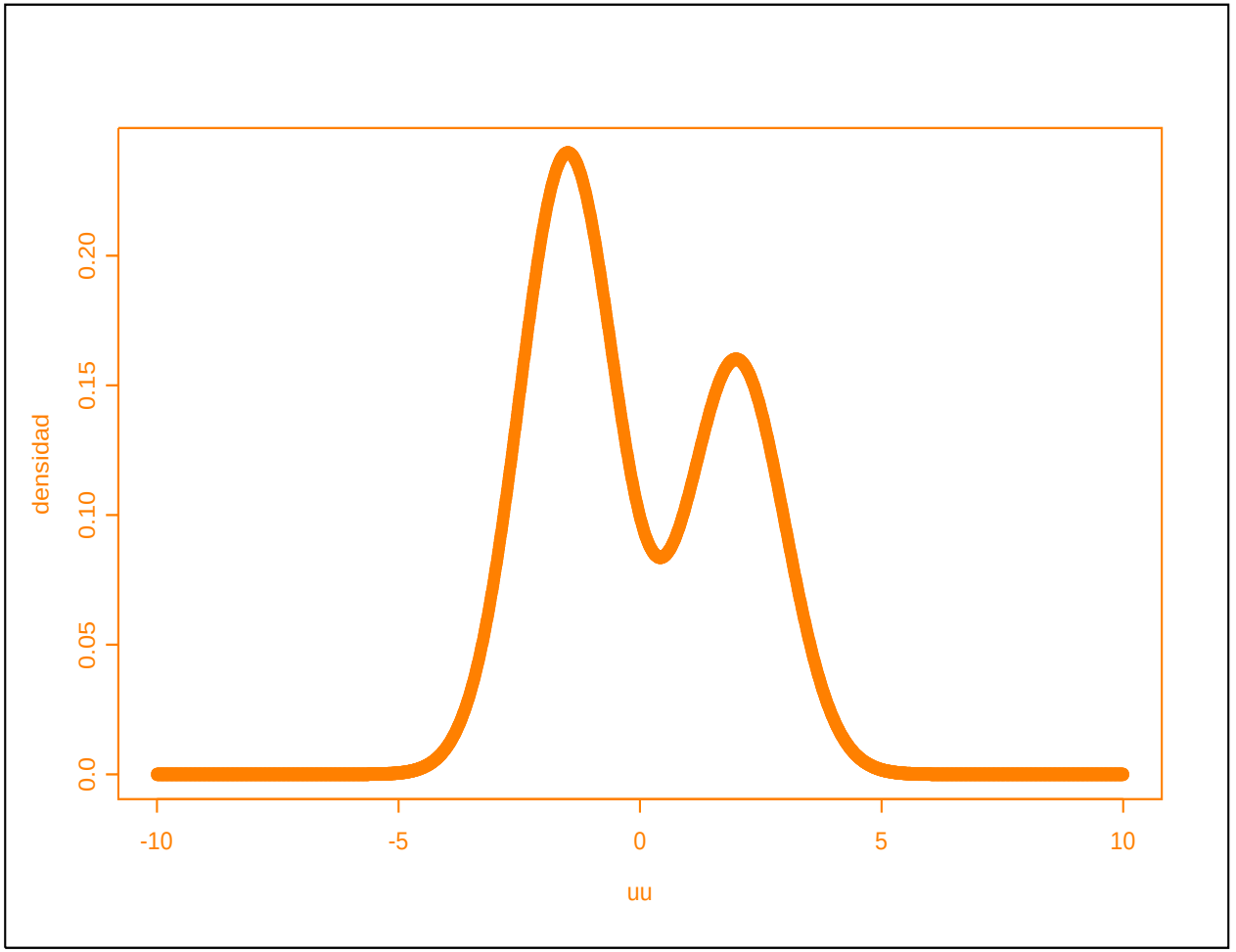
Estimación no paramétrica de la densidad

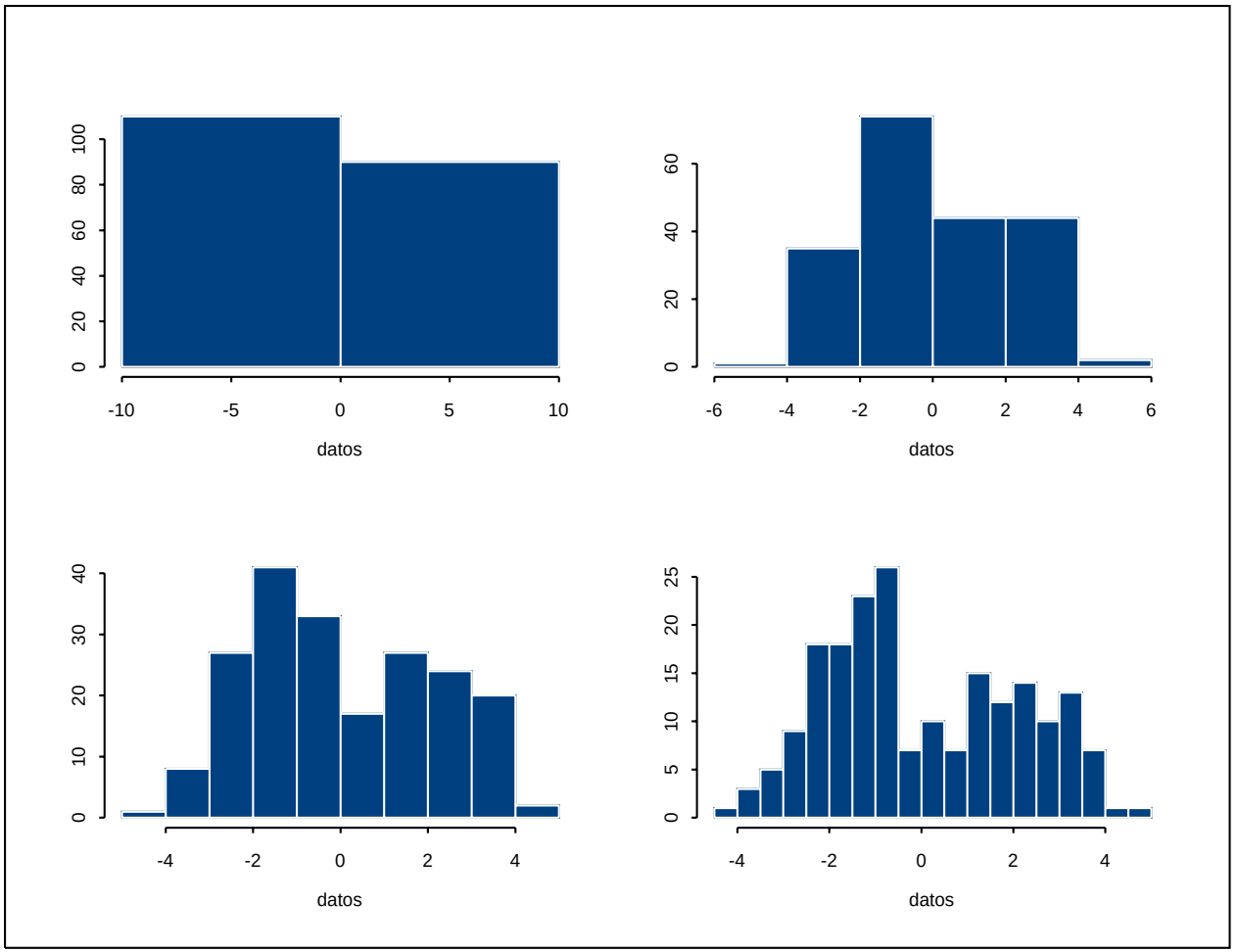
Rosenblatt(1956)

Histograma

Sea A_j una partición de $\mathbb{R} = \cup A_j$ entonces para cada $x \in A_j$ se define $\widehat{f}(x) = \frac{\#\{X_i : X_i \in A_j\}}{n|A_j|}$.

Ejemplo: Sea $X \sim f(x) = 0.6\mathcal{N}(-1.5, 1) + 0.4\mathcal{N}(2, 1)$


densidad
0.0
0.05
0.10
0.15
0.20
-10
-5
0
5
10
uu


0
20
40
60
80
100
-10
-5
0
5
10
datos
0
20
40
60
-6
-4
-2
0
2
4
6
datos
0
10
20
30
40
-4
-2
0
2
4
datos
0
5
10
15
20
25
-4
-2
0
2
4
datos

Estimador de Núcleos

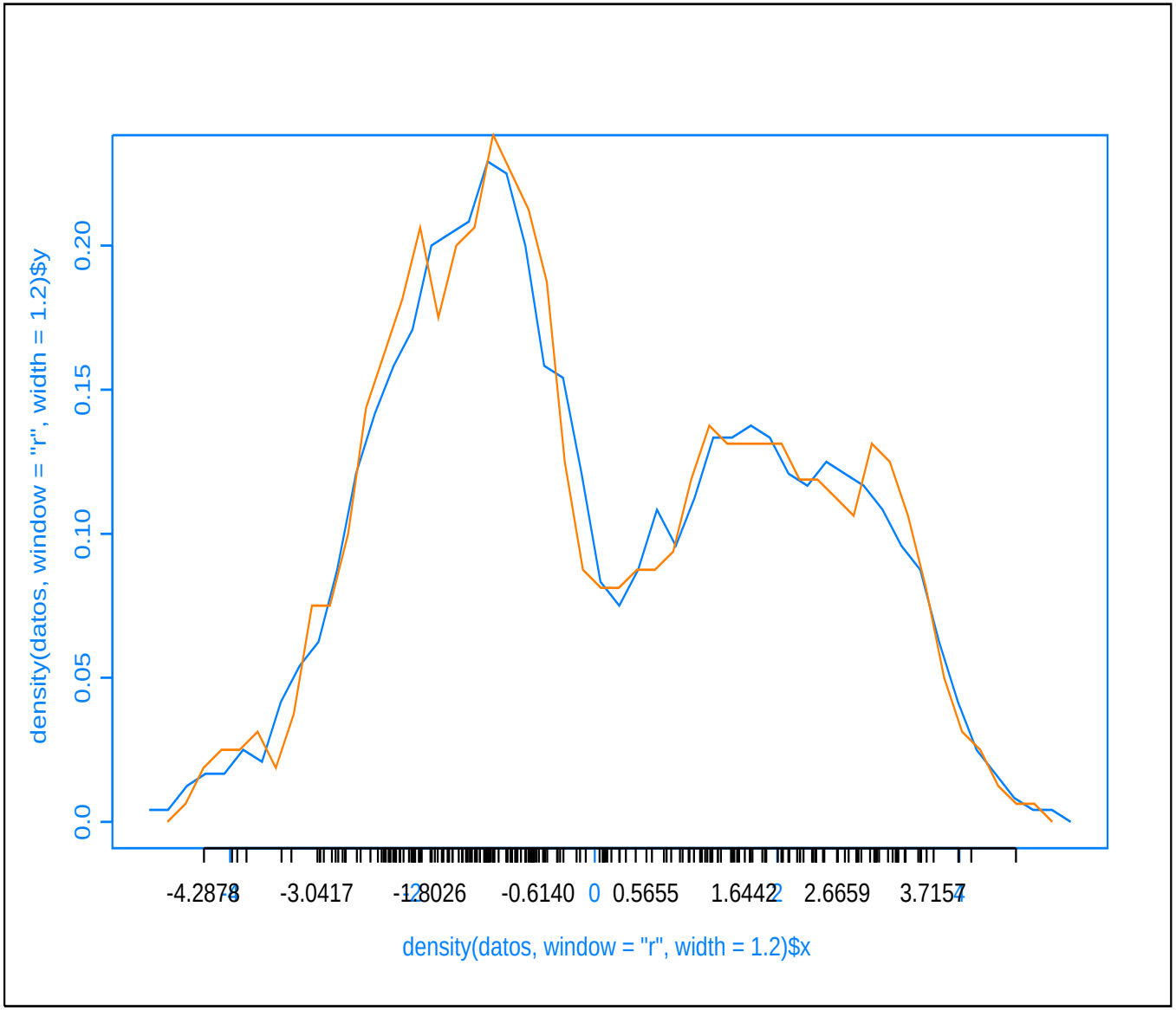
$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

como $F(x+h) - F(x-h) = P(X \in (x-h, x+h])$ entonces un estimador podría ser $\widehat{F}_n(x+h) - \widehat{F}_n(x-h) = \frac{\#\{X_i : X_i \in (x-h, x+h]\}}{n}$

Para un h chico la propuesta es

$$\widehat{f}_n(x) = \frac{\#\{X_i : X_i \in (x-h, x+h]\}}{n2h}$$

si $K(x) = \frac{1}{2} I_{(-1,1]}(x)$ entonces $\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$


density(datos, window = "r", width = 1.2)$y
0.0
0.05
0.10
0.15
0.20
-4.2878
-3.0417
-1.8026
-0.6140
0
0.5655
1.6442
2.6659
3.7157
density(datos, window = "r", width = 1.2)$x

El estimador sigue sin tener la suavidad que queremos. Podemos reemplazar el núcleo K por otro núcleo que cumpla: $\int K = 1$ y $K \geq 0$ y considerar nuevamente

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

Hay dos cosas para elegir:

K núcleo $K \geq 0$ y $\int K = 1$.

h ventana

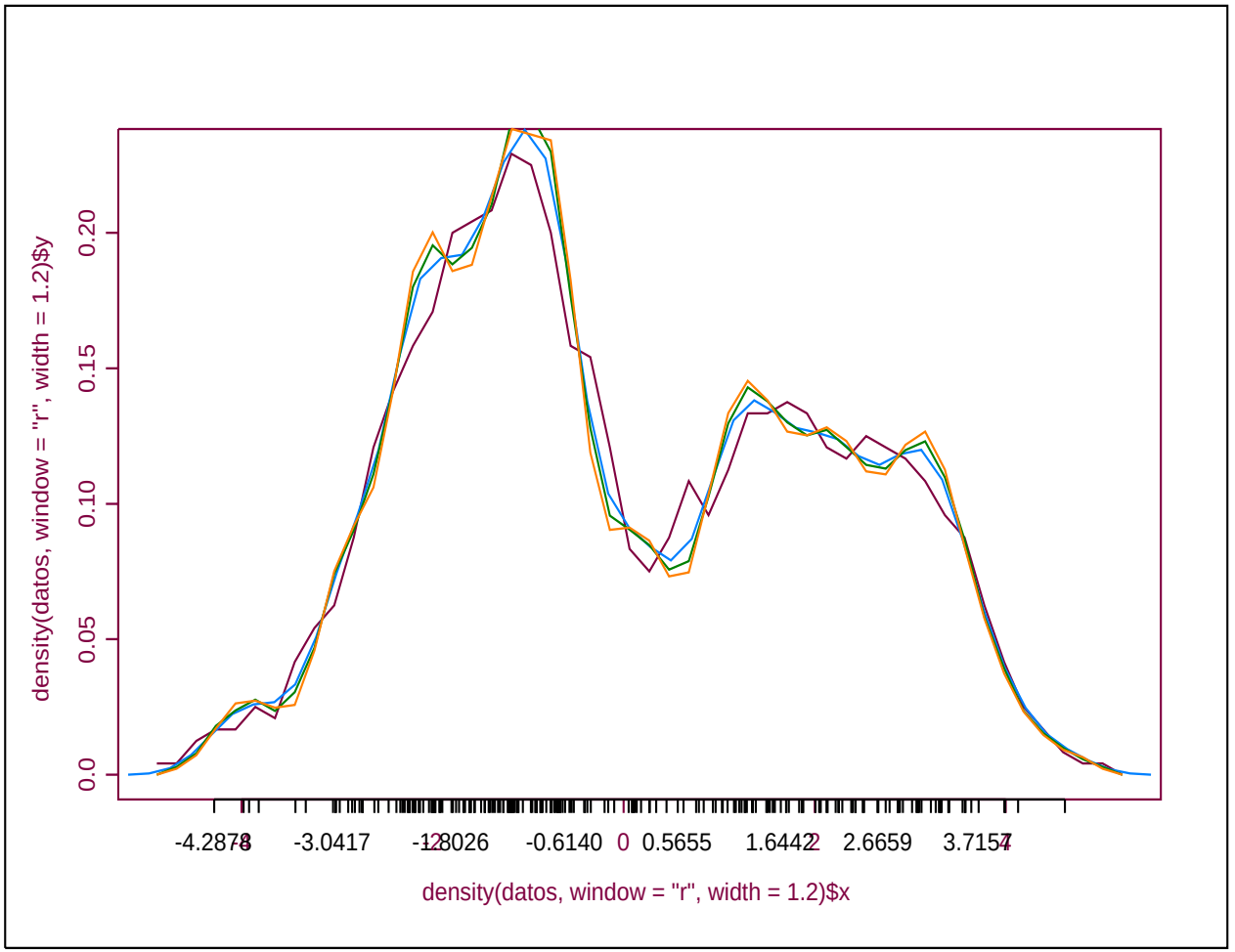


Figure 1: estimación con diferentes núcleos

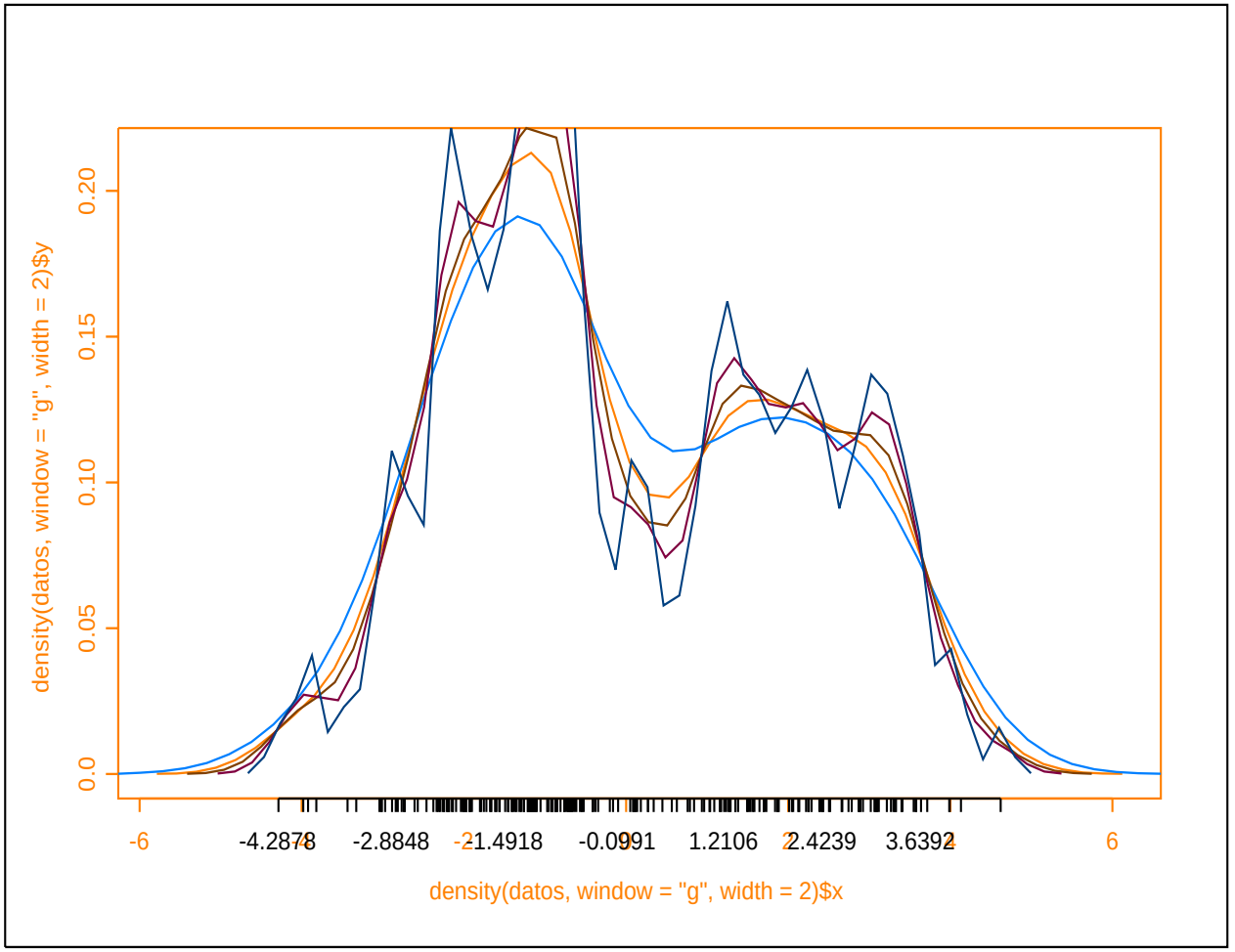


Figure 2: estimación con diferentes ventanas (3,2,1.5,1,0.5)

Propiedades:

- $E(\widehat{f}_n(t)) = f * K_h(t) \to f(t)$ si $h \to 0$ $\qquad K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$
- $V(\widehat{f}_n(t)) \to 0$ si $h \to 0$ y $nh \to \infty$

Otras propuestas de estimación de densidad:

- vecinos mas cercanos
- series ortogonales
- máxima verosimilitud penalizada

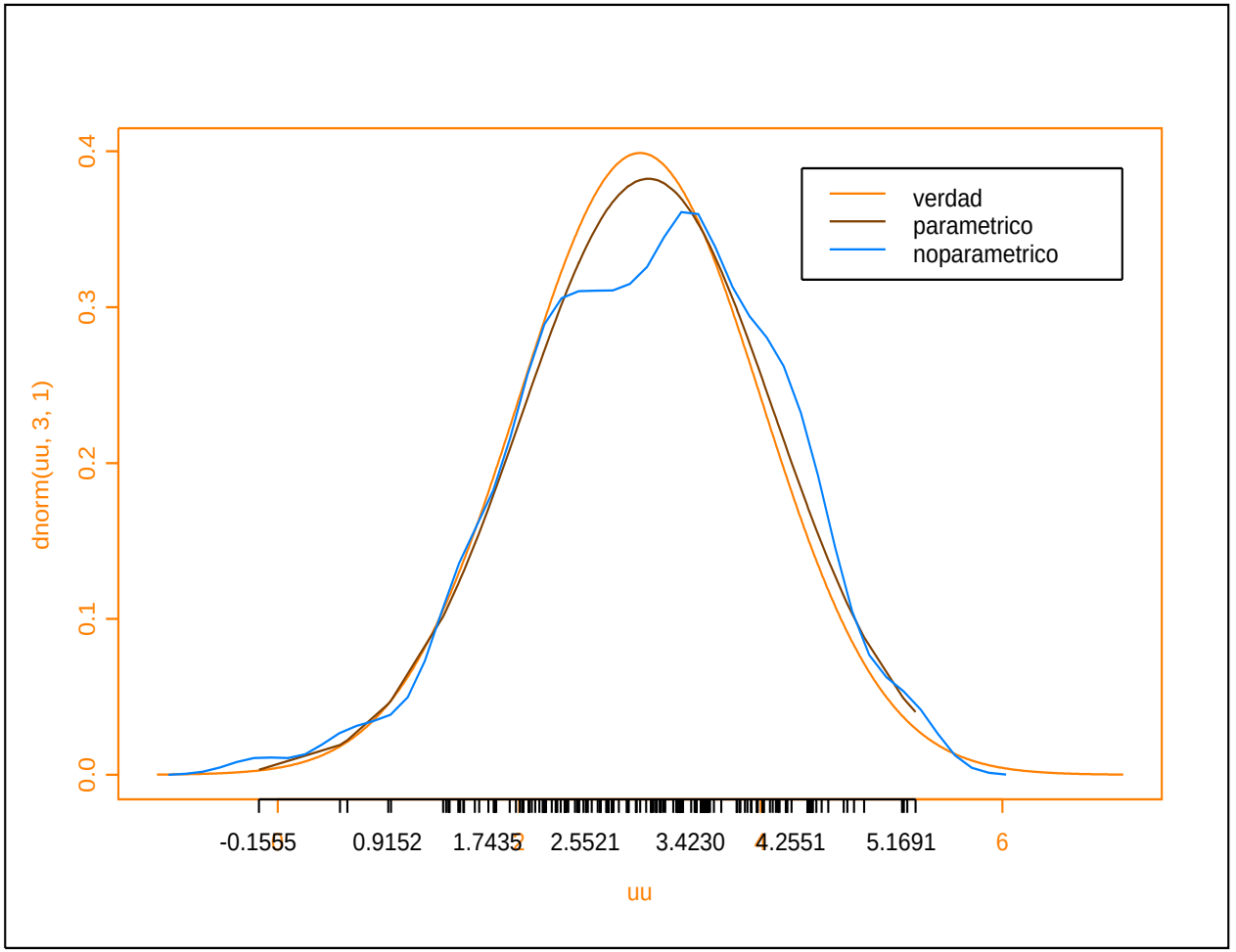


Figure 3: Normal(3,1) estimadores (3.0684,1.0431)

Modelos de Regresión

Sea Y una variable aleatoria cuya distribución desconocemos y queremos explicarla mediante X. Una relación que nos puede interesar es por ejemplo $E(Y|X = x)$.

Según la estructura que asumimos para esta relación surgen diversos modelos

- Modelo de regresión paramétrica $E(Y|X = x) = \beta_0 x + a_0$
- Modelo de regresión no paramétrica $E(Y|X = x) = m_0(x)$

tanto β_0 como m_0 son desconocidos y el objetivo es estimarlos.

Regresión no paramétrica

La "heurística" para estimar $m_0(x) = E(Y|X = x)$.

$E(Y|X = x) = \int y \frac{f_{XY}(x,y)}{f_X(x)} dy$ entonces un estimador posible sería

$$\widehat{E(Y|X = x)} = \int y \frac{\widehat{f}_{n,XY}(x,y)}{\widehat{f}_{n,X}(x)} dy$$

donde

$$\begin{aligned} \widehat{f}_n(x) &= \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) \\ \widehat{f}_{n,XY}(x,y) &= \frac{1}{nh^2} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) K(\frac{y - Y_i}{h}) \end{aligned}$$

$$\widehat{E(Y|X=x)} = \frac{\frac{1}{nh^2}\int y \sum_{i=1}^{n} K(\frac{x-X_i}{h})K(\frac{y-Y_i}{h})dy}{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})}$$

es fácil ver que si K es simétrico

$$\widehat{E(Y|X=x)} = \sum_{i=1}^{n} \frac{K(\frac{x-X_i}{h})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})} Y_i$$

el estimador obviamente depende de h y K, heredando las mismas propiedades y desventajas que el estimador de densidad

Ejemplos

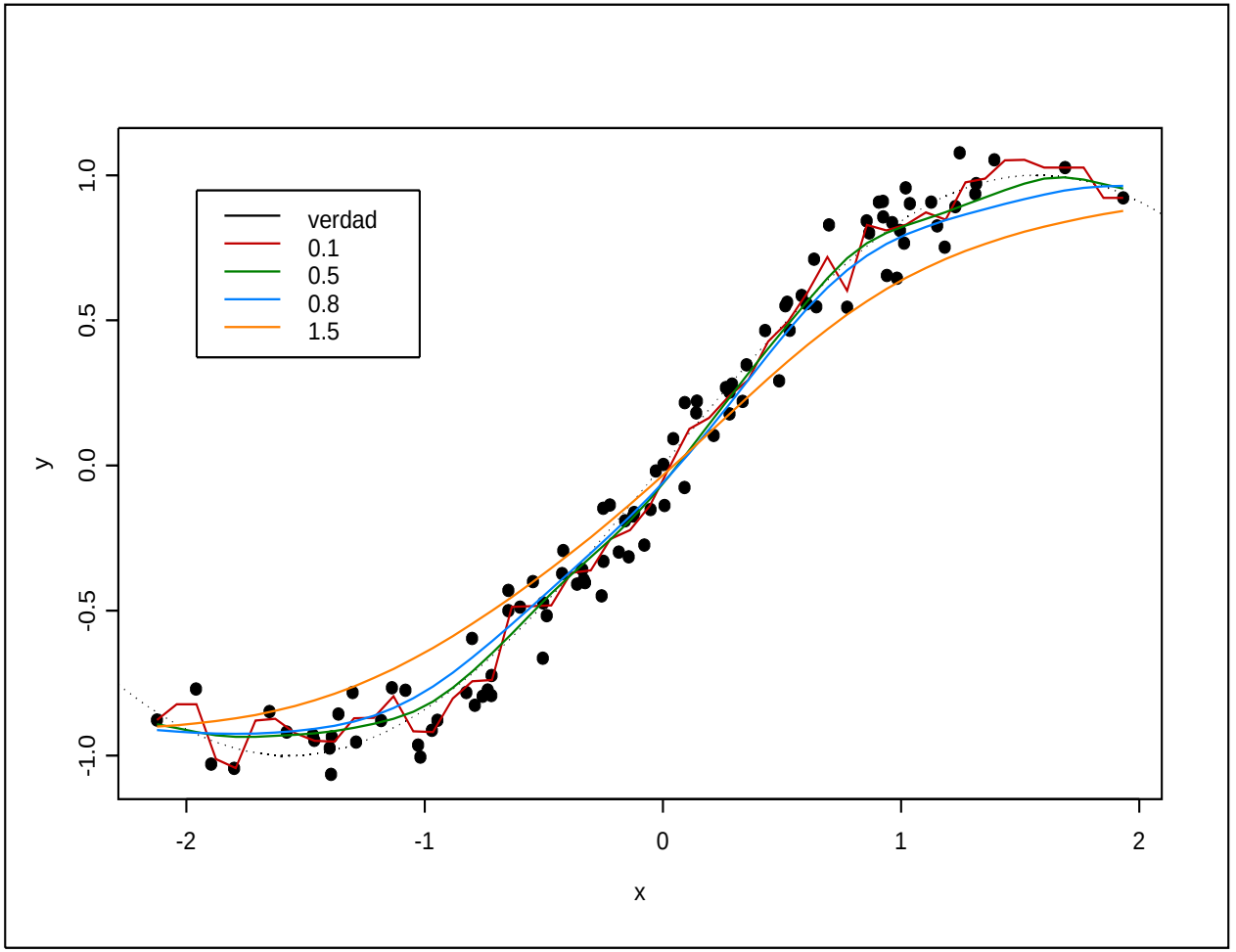
Gráfico 1:

- $Y = \text{sen}(X) + \varepsilon$
- $X \sim \mathcal{N}(0, 1)$ y $\varepsilon \sim \mathcal{N}(0, 0.1)$ ambas independientes.
- $E(Y|X) = \text{sen}(X)$

Gráfico 2:

- $Y = 4X + 1 + \varepsilon$
- $X \sim \mathcal{N}(0, 1)$ y $\varepsilon \sim \mathcal{N}(0, 0.9)$ ambas independientes.
- $E(Y|X) = 4X + 1$

en todos los caso genere una muestra de tamao 100.


verdad
0.1
0.5
0.8
1.5
y
x
1.0
0.5
0.0
-0.5
-1.0
-2
-1
0
1
2

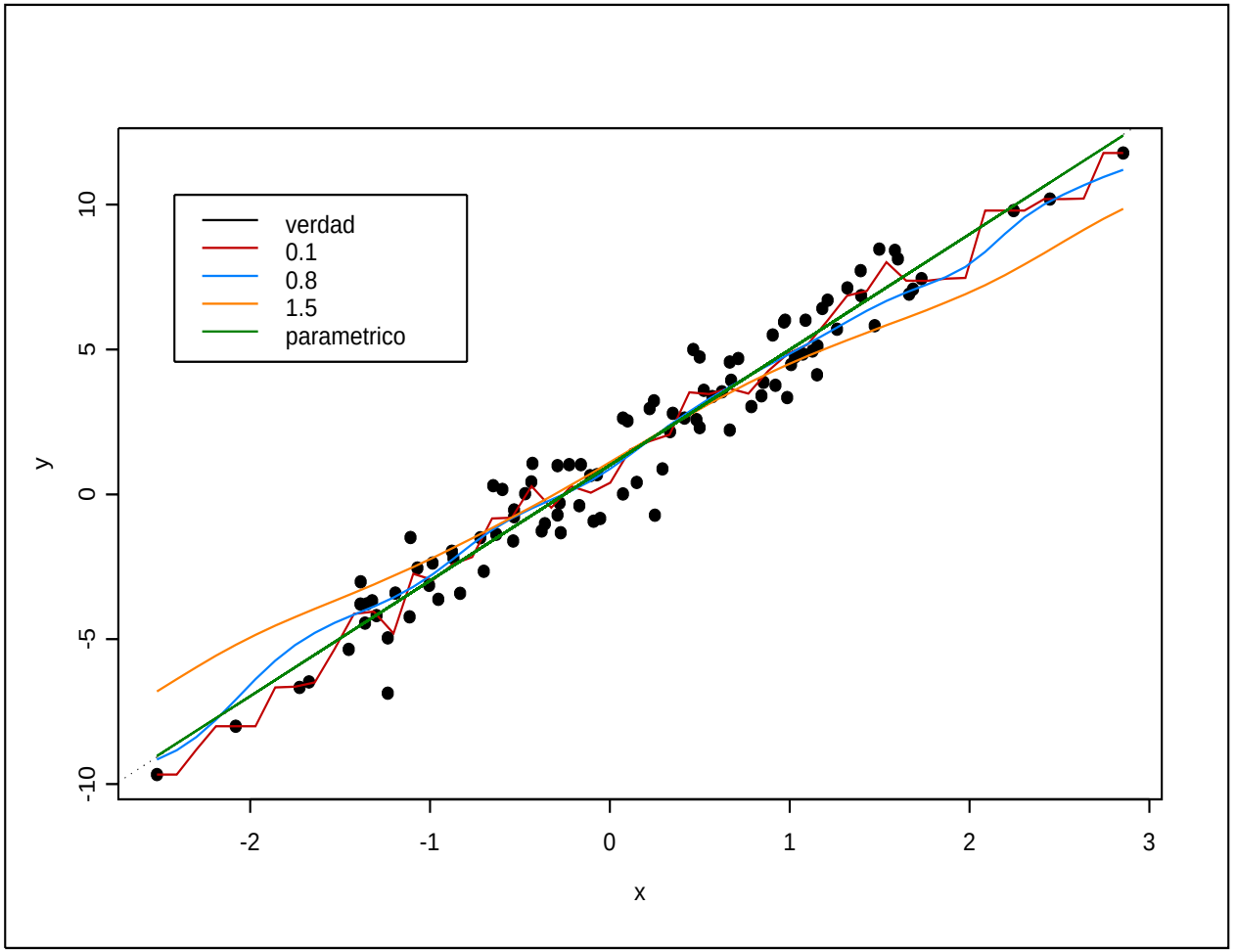


Figure 4: parámetros estimados $(1.005294, 3.986668)$

Estimación robusta de modelos de regresión parcialmente lineales generalizados

- Modelo de regresión parcialmente lineales

$E(Y|X = x, T = t) = m_0(t) + \beta_0 x$

donde $m_0(t)$ es una función desconocida a estimar y β_0 es un parámetro desconocido a estimar.

- Modelo de regresión parcialmente lineales generalizados

$E(Y|X = x, T = t) = H(m_0(t) + \beta_0 x)$

donde $m_0(t)$ desconocida, β_0 desconocida y H un función conocida.