Estimación de densidades basada en núcleos: algunos elementos básicos.

Isabel Cañette

Seminario de Reconocimiento de Patrones.
Seminario de Probabilidad y Estadística.
Diciembre, 2002

Introducción.

Decimos que una variable aleatoria X tiene densidad f, si

$$P\{a < X < b\} = \int_a^b f(x)\,dx \;\forall a < b.$$

Estimar una densidad es reconstruir dicha función f a partir de un conjunto de variables $X_1, \ldots X_n$ con la misma distribución que X.

- Estimación paramétrica: Si conocemos la distribución a la cuál pertenece f, (por ejemplo, gaussiana), entonces basta estimar los parámetros para tener una estimación de f.

- Estimación no paramétrica: Se usa cuando no queremos asumir hipótesis sobre la distribución de nuestra muestra.

Antecedentes y motivación.

Histogramas.
El estimador de densidades más sencillo (y más antiguo) es el histograma. Dado un punto x_0 y un ancho de intervalo h, consideramos los intervalos: $\{[x_0 + mh, x_0 + (m+1)h], m \in \mathbb{Z}\}$, y luego, dado un punto x,

$$\widehat{f}(x) = \frac{1}{nh}\#\{i : X_i \text{ está en el mismo intervalo que } x\}$$

Estimador "naive".
Por definición de densidad, tenemos que:

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

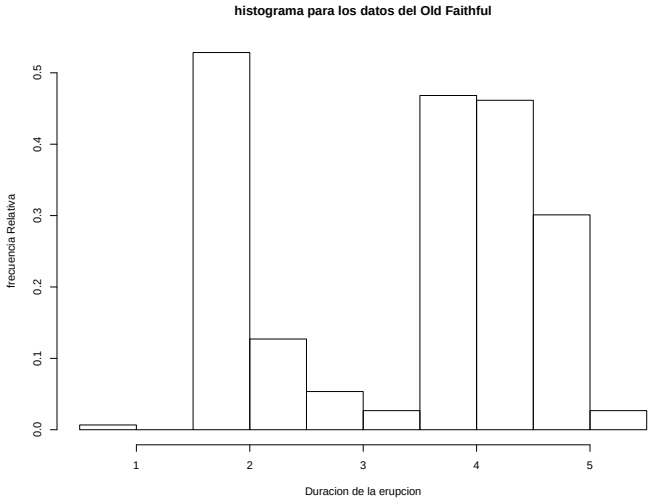
Estimando $P\{x - h < x < x + h\}$ con :

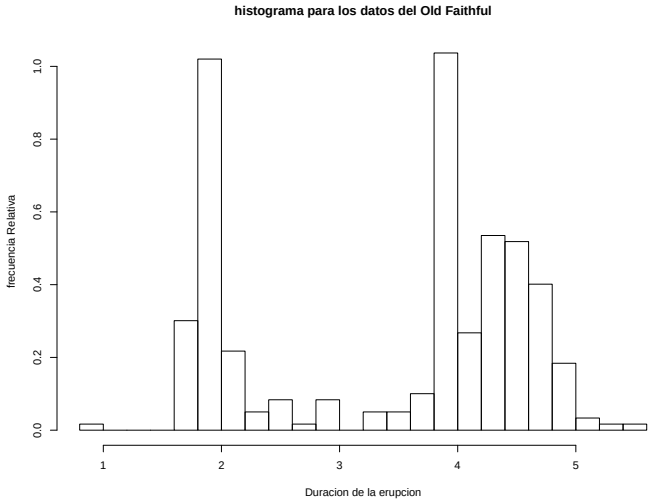
$$\frac{\text{No. de } X_i \text{ en } [x-h, x+h]}{n},$$

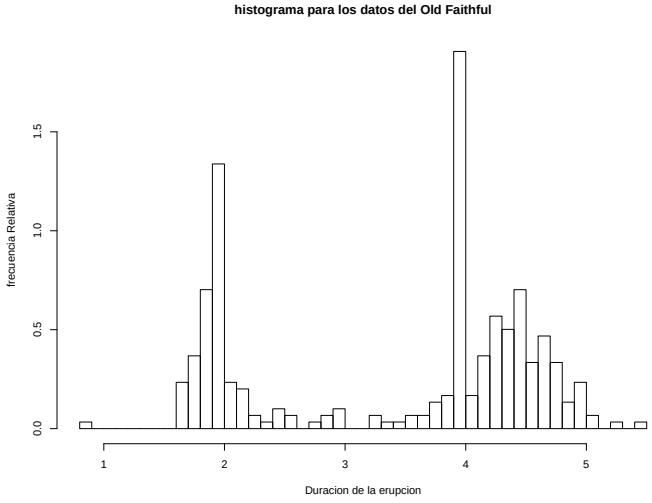
obtenemos:

$$\widehat{f}(x) = \frac{1}{2nh}\{\text{No. de } X_i \text{ en } [x-h, x+h]\}.$$

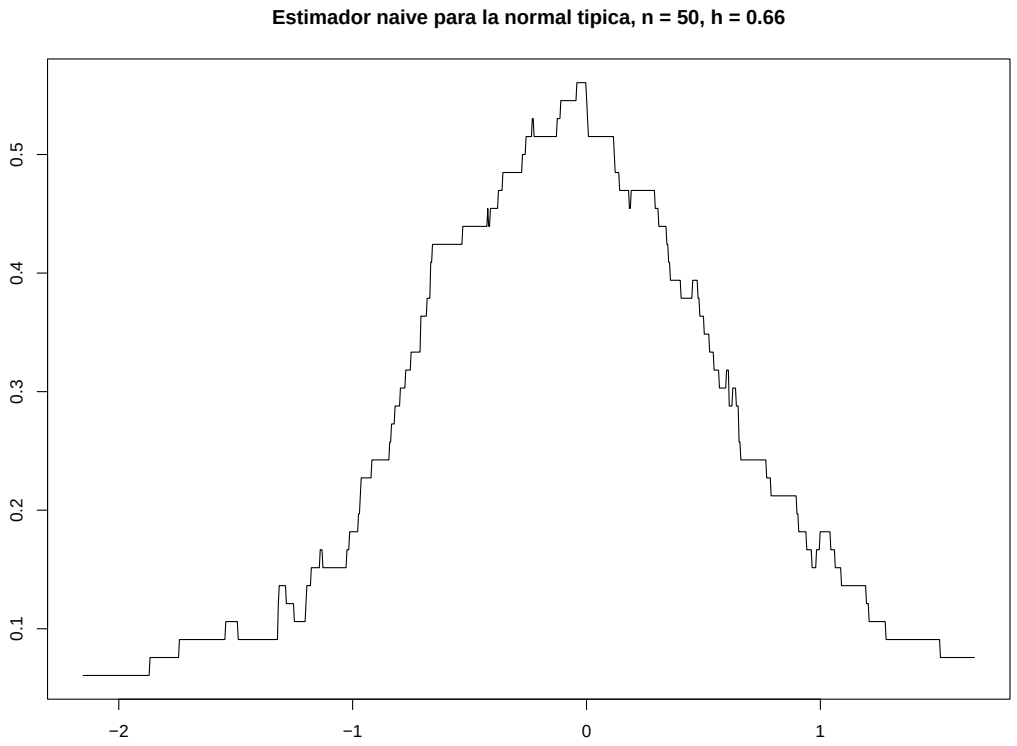
Histogramas de la duración de las erupciones del geyser Old Faithful, con 10, 20 y 35 intervalos respectivamente.

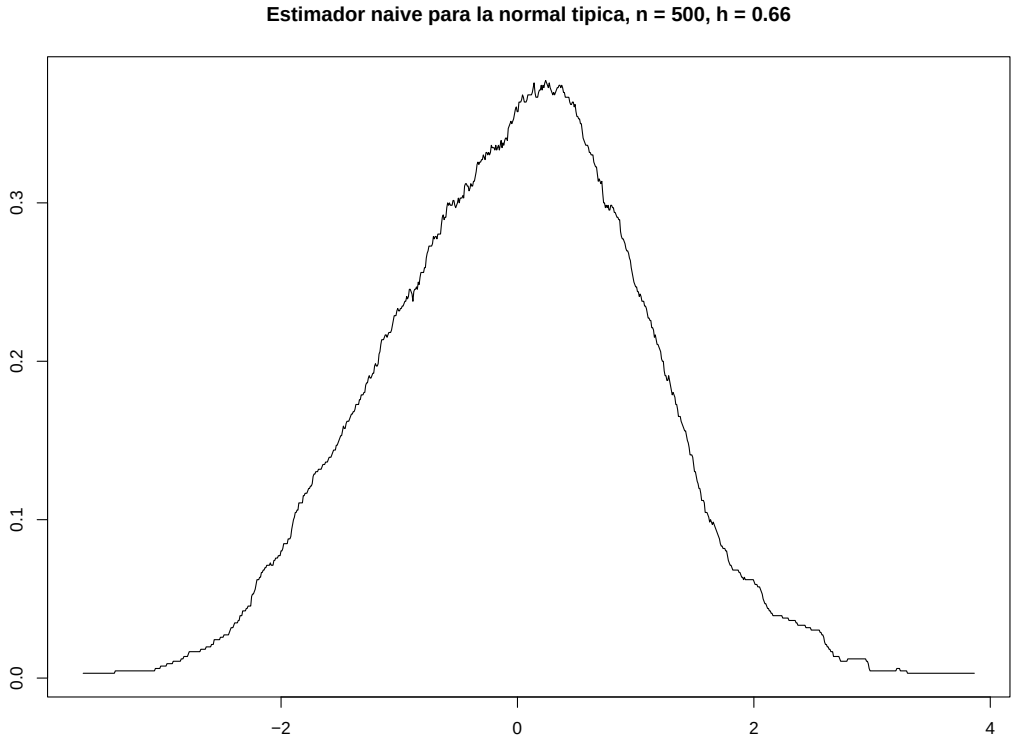






Estimaciones 'naive' para datos con distribución normal típica, para 50 y 500 datos respectivamente.





El estimador "naive" se puede expresar de la forma siguiente: sea la función w definida como:

$$w(x) = \begin{cases} 1/2 & \text{si } |x| < 1 \\ 0 & \text{si no} \end{cases} \tag{1}$$

Entonces,

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

Estimadores basados en núcleos (Akaike (54), Rosenblatt(56)).
Podemos generalizar la definición anterior (del estimador "naive") sustituyendo w por una función K tal que $\int K = 1$, es decir:

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

$\widehat{f}$ es una función de densidad, y hereda la regularidad de K.

ECM, ECMI, sesgo y varianza.

Para cada x, consideramos el error cuadrático medio de nuestro estimador en x.

$$ECM(x) = E\Big(\widehat{f}(x) - f(x)\Big)^2 = Var(\widehat{f}(x)) + \Big(\underbrace{E(\widehat{f}(x) - f(x))}_{B(\widehat{f}(x))}\Big)^2.$$

Una medida global del error está dada por el error cuadrático medio integrado, es decir:

$$ECMI = \int Var(\widehat{f}(x))\,dx + \int B^2(\widehat{f}(x))\,dx.$$

Si f tiene derivada segunda continua y acotada, y el núcleo K es simétrico y con 2do. momento finito, entonces:

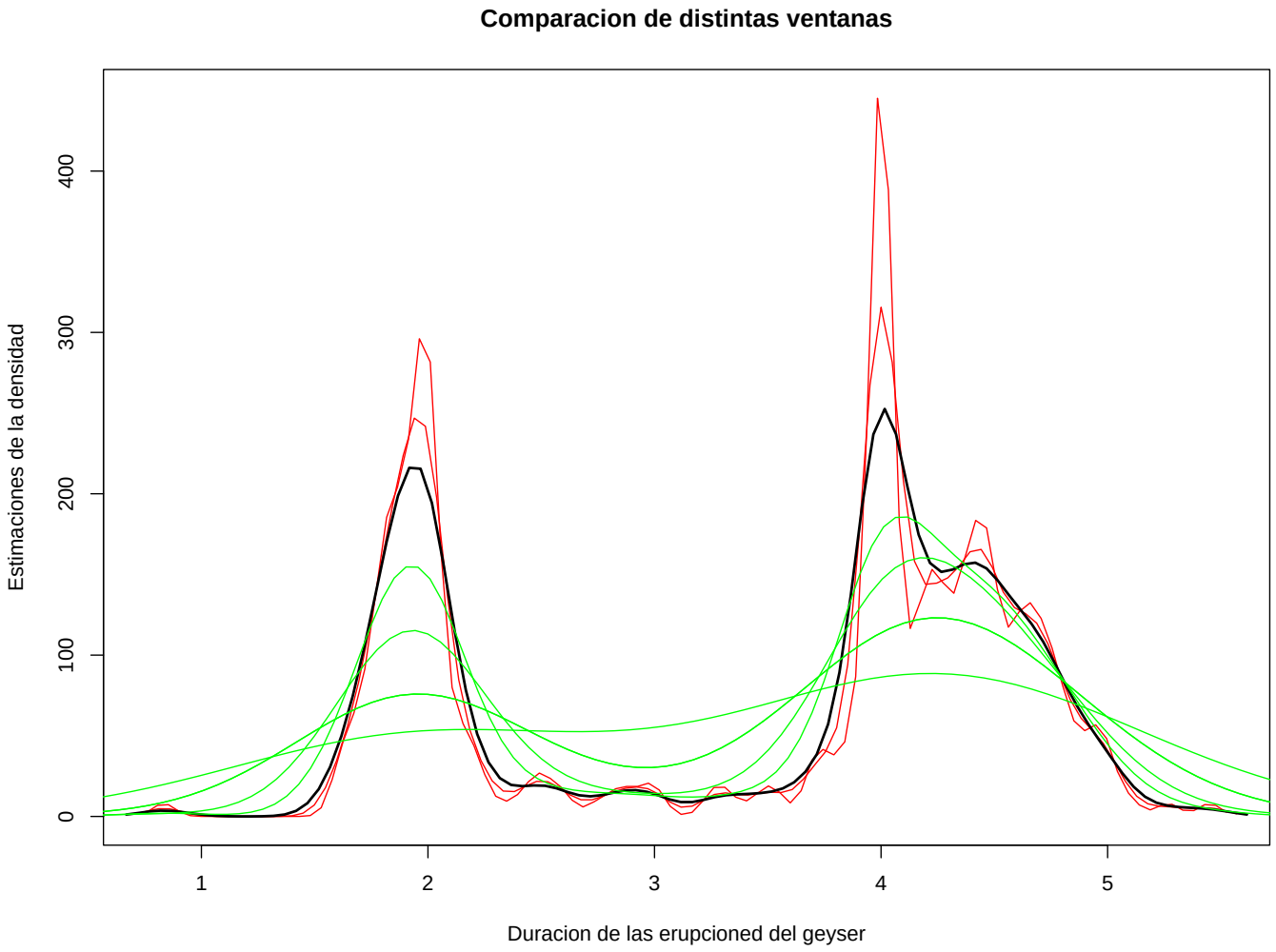
$$B(\widehat{f}(x)) \approx \frac{1}{2}h^2 \underbrace{\int u^2 K(u)\,du}_{k_2}\ f''(x),$$

$$Var(\widehat{f}(x)) \approx \frac{1}{nh} f(x) \int K^2(u)\,du.$$

De donde:

$$ECMI \approx \frac{1}{nh}\int K^2(u)\,du + \frac{1}{4}h^4 k_2^2 \int (f''(x))^2\,dx, \tag{2}$$

Observación: Hay un compromiso entre el sesgo y la varianza para h. Cuando h es muy pequeña el sesgo disminuye, pero aumenta la varianza.



La ventana óptima.

Minimizando (2) en h, obtenemos:

$$h_{opt} = k_2^{-2/5}\Big(\int K^2(t)\,dt\Big)^{1/5}\Big(\int (f''(x))^2\,dx\Big)^{-1/5} n^{-1/5}. \tag{3}$$

Observaciones:

- Esta h depende de f''
- De todos modos, (3) nos da elementos para buscar una h.

Métodos de selección de ventanas

Referencia a una distribución conocida (Silverman's "rule of thumb").

Sea ϕ la densidad normal típica. Si f fuese normal con varianza σ, tendríamos:

$$\int \Big(f''(x)\Big)^2\,dx = \sigma^{-5}\int \Big(\phi''(x)\Big)^2\,dx = \frac{3}{8}\pi^{-1/2}\sigma^{-5}.$$

Si además K es el núcleo gaussiano, entoces $k_2 = 1$, y $\int K^2 = (4\pi)^{-1/2}$, por lo que:

$$h_{opt} = \underbrace{k_2^{-2/5}}_{1} \underbrace{\left(K^2(t)\,dt\right)^{1/5}}_{(4\pi)^{-1/10}} \underbrace{\left(\int \left(f''(x)\right)^2 dx\right)^{-1/5}}_{(\frac{3}{8}\pi^{-1/2}\sigma^{-5})^{-1/5}} n^{-1/5} = \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-1/5}. \quad (4)$$

Sustituyendo σ por un estimador (por ejemplo el desvío estandar de la muestra), en (4), obtenemos un estimador de h_{opt}.

Métodos de sustitución. ("Plug-in")
Los métodos de sustitución de Park y Marron, y de Sheater y Jones, utilizan estimadores basados en núcleos para obtener una estimación de $\int f''$.

Métodos de validación cruzada. El método de validación cruzada basado en mínimos cuadrados (Rudemo(82) y Bowman(84)) minimiza el ECMI:

$$ECMI = E\int(\widehat{f} - f)^2 = \underbrace{E\int \widehat{f}^2 - 2E\int \widehat{f} f}_{R(f)} + \int f^2.$$

Basta minimizar $R(f)$, y un estimador insesgado de este valor es:

$$CV(h) = \int \widehat{f}^2 - \frac{2}{n}\sum_i \widehat{f}_{-i}(X_i), \qquad (5)$$

Donde $\widehat{f}_{-i}(x)$ es la estimación en x que se obtiene utilizando todos los datos menos el $i-$ésimo.
Esto se debe a que, si:

$$K_h(x) = \frac{1}{h}K(x/h)$$

entonces:

$$E\int \widehat{f}f = E\Big[\int \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)f(x)\,dx\Big] = E(K_h(X - Y))$$

donde X e Y son independientes, ambas con densidad f. Luego, un estimador de $E\widehat{f}f$ será un estimador de $E(K_h(X - Y))$, por ejemplo,

$$\frac{1}{n(n-1)}\sum\sum_{i\neq j} K_h(X_i - X_j),$$

que es el término que restamos (multiplicado por 2) en (5). Por otra parte, un estimador de $E\int \widehat{f}^2$ es precisamente $\int \widehat{f}^2$.

Se busca h que minimice $CV(h)$. Si el núcleo es simétrico, entonces:

$$CV(h) = \frac{1}{n^2}\sum_i\sum_j K_h^{(2)}(X_i - X_j) - \frac{2}{n^2}\sum\sum_{i\neq j} K_h(X_i - X_j), \tag{6}$$

con

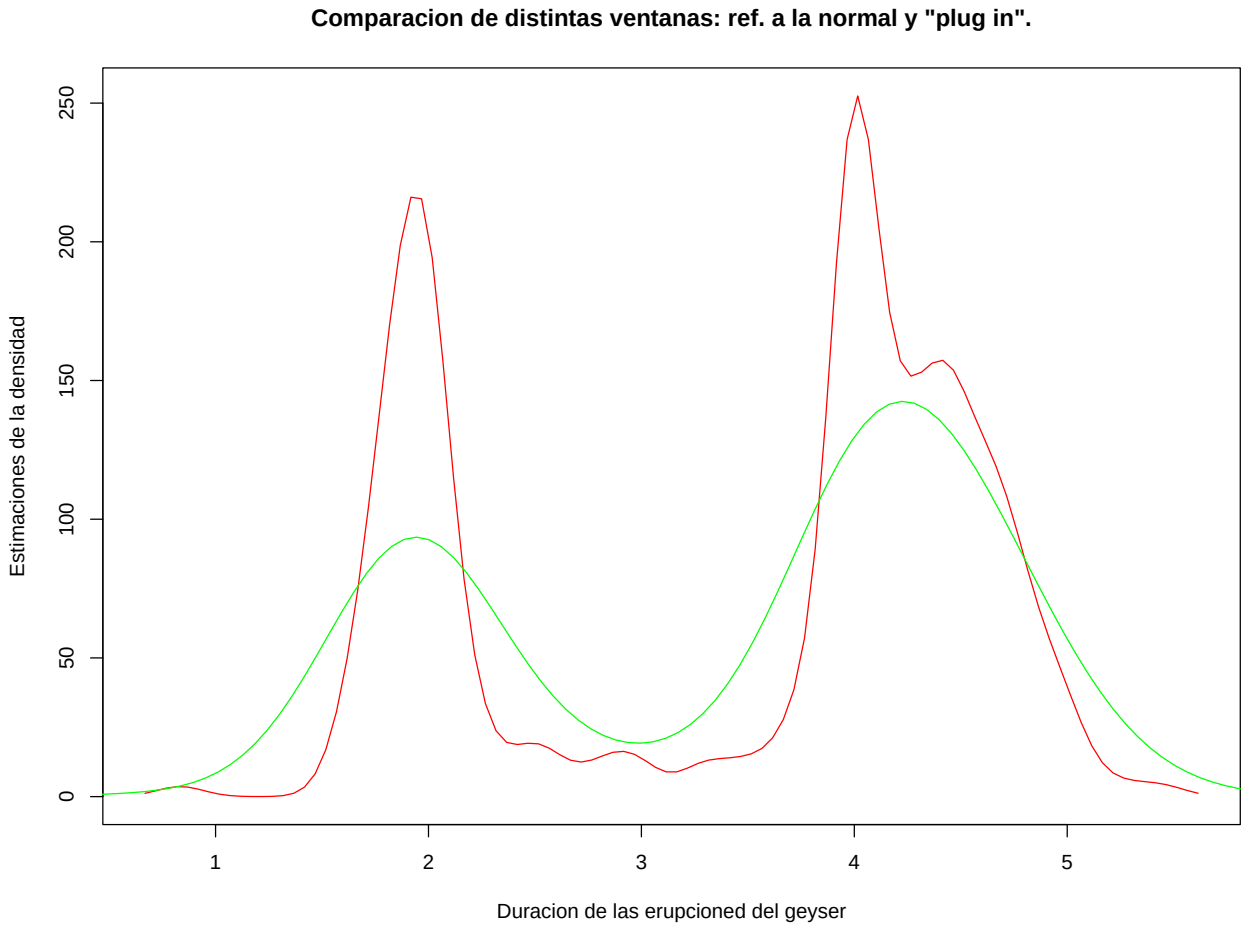
$$K_h^{(2)}(x) = \int K_h(x-y)K_h(y)\,dy = (K_h * K_h)(x).$$

El método de validación cruzada por máxima verosimilitud maximiza

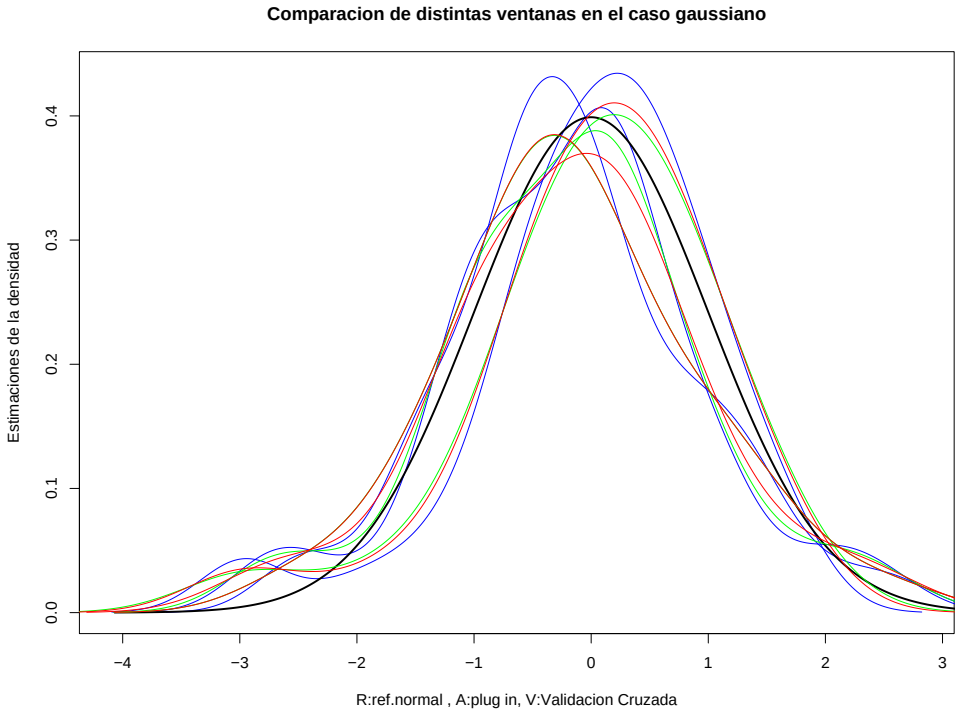
$$CV_2(h) = \frac{1}{n}\sum_{i=1}^{n} log\widehat{f}_{-i}(X_i)$$

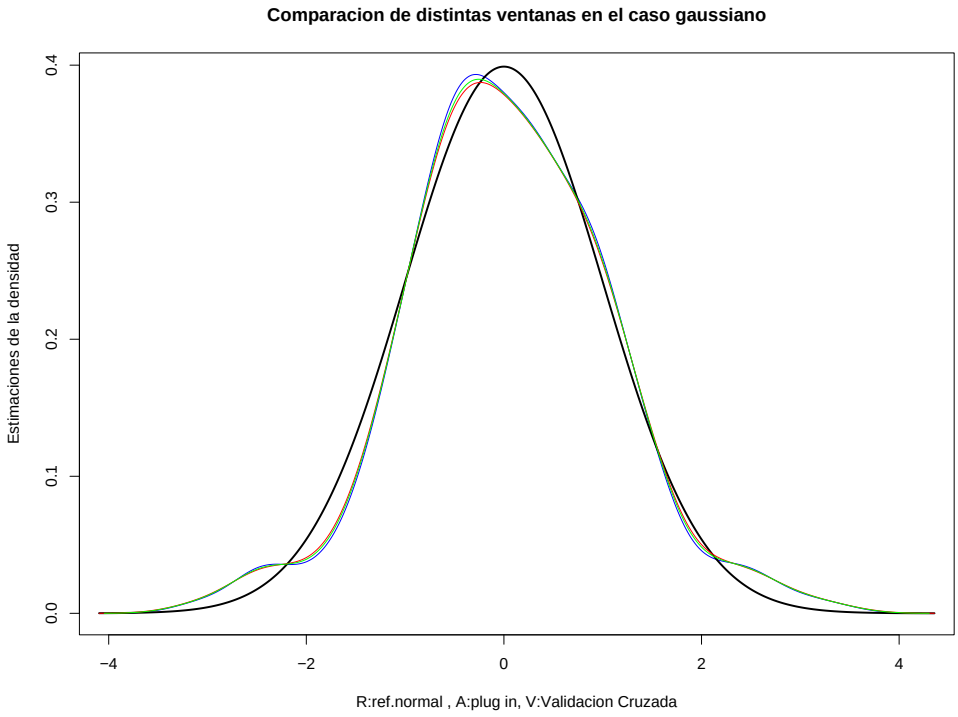
(este método es muy sensible a la presencia de outliers).

Comparación de la estimación obtenida con la ventana referida a la distribución normal, y la ventana obtenida por sustitución.



Comparación de estimaciones obtenidas con distintas ventanas para la distribución normal, para n = 50 y n = 500.





Aplicación del método para d≥2.

Veremos ahora el caso de datos multivariados. Sea d la dimensión del espacio. Sea $K : \mathbb{R}^d \to \mathbb{R}$, con $\int_{\mathbb{R}^d} K(x)\,dx = 1$. Nuestros estimadores serán de la forma:

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

En general se utilizan funciones K radialmente simétricas, de la forma

$$K(x) = k(x^t x) = k(||x||); \tag{7}$$

por ejemplo, para el núcleo gaussiano, $k(u) = (2\pi)^{-d/2} exp((-1/2)u)$, y el estimador es:

$$\widehat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} (2\pi)^{-d/2} \exp\Big[-\frac{1}{2}\Big(\frac{x - X_i}{h}\Big)^t \Big(\frac{x - X_i}{h}\Big)\Big] \tag{8}$$

Datos asimétricos.

El problema de usar núcleos simétricos con una sola ventana, es que el núcleo tendrá la misma escala en todas las dimensiones.

En el caso de datos asimétricos, podemos proceder de la forma siguiente: Sea S un estimador de la matriz de covarianza de X, por ejemplo, la covarianza muestral. Tomamos el núcleo $K(x) = (1/\sqrt{det(S)})k(x^t S^{-1} x)$, en lugar de (7). Esto es equivalente a reescalar los datos. Por ejemplo, en el caso Gaussiano queda:

$$\widehat{f}(x) = \frac{(det(S))^{-1/2}}{nh^d} \sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}} e^{-\left[\frac{1}{2}\left(\frac{x-X_i}{h}\right)^t S^{-1}\left(\frac{x-X_i}{h}\right)\right]} \tag{9}$$

Sesgo, varianza y parámetros de suavizado

Podemos generalizar la mayoría de las técnicas desarrolladas para 1 dimensión. Sea K una densidad simétrica, y sean:

$$\alpha = \int t^2 K(t)\, dt, \;\; \beta = \int K^2\, dt.$$

Entonces,

$$B(x) \approx \frac{1}{2} h^2 \alpha \nabla^2 f(x)$$

$$Var(\widehat{f}(x)) \approx \frac{1}{nh^d}\beta f(x).$$

Minimizando el ECMI en h, obtenemos:

$$h_{opt} = \underbrace{\left(d\beta\alpha^{-2}\left(\int(\nabla^2 f)^2\right)^{-1}\right)^{\frac{1}{d+4}}}_{A(f,K)} n^{-\frac{1}{d+4}} \qquad (10)$$

Al igual que en dimensión 1, podemos, dado un núcleo K, hallar $A(f, K)$ en el caso en que $f = \phi$, siendo ϕ la distribución normal estandar en dimensión d. Utilizando que:

$$\int(\nabla^2\phi)^2 = (2\sqrt{\pi})^{-d}(\frac{1}{2}d + \frac{1}{2}d^2) \qquad (11)$$

Basta sustituir (11), y los α y β correspondientes al núcleo K en (10) para obtener $A(\phi, K)$. En particular, si $K = \phi$, $A(\phi, \phi) = 1$.

Si estamos utilizando el estimador (9), entonces podemos usar directamente la ventana (10). Si estamos utilizando el estimador (8) con el núcleo

simétrico, en lugar de h_{opt} utilizaremos σh_{opt}, donde σ^2 es un estimador de la varianza marginal promedio.

Otras técnicas de estimación de la ventana

El método de validación cruzada se traslada al caso en $\mathbb{R}^d$; vale la fórmula (6), donde

$$K_h(x) = \frac{1}{h^d} K(x/h).$$

El método de sustitución también se traslada a $\mathbb{R}^d$. Algunos paquetes utilizan este método para estimar por separado distintas ventanas para las distintas dimensiones.

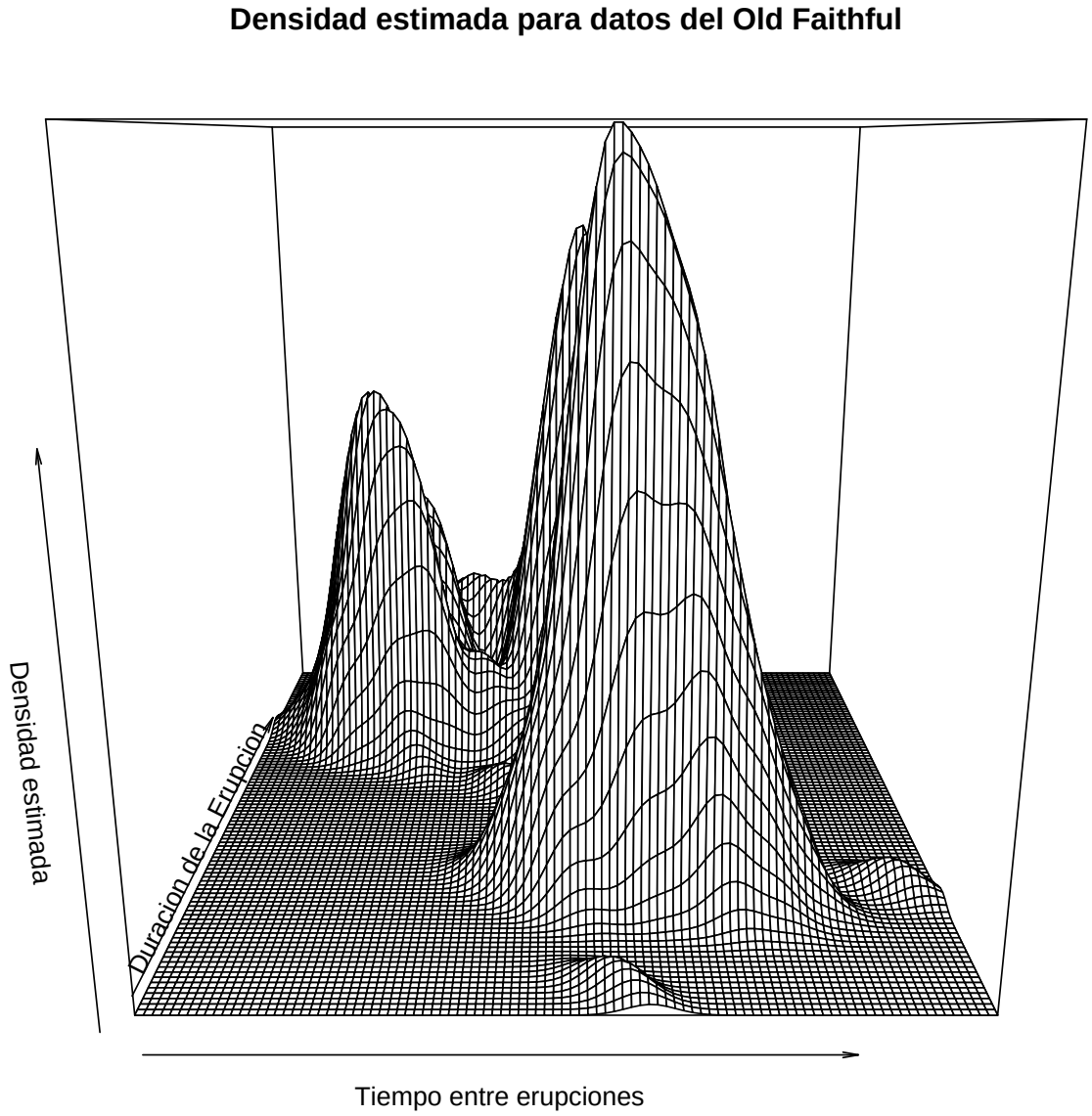
El problema de la dimensión.

Si observamos las fórmulas de la varianza y la ventana óptima, tenemos que: $Var\widehat{f} \approx \frac{1}{nh^d}$, y $h \approx n^{-\frac{1}{d+4}}$, de donde:

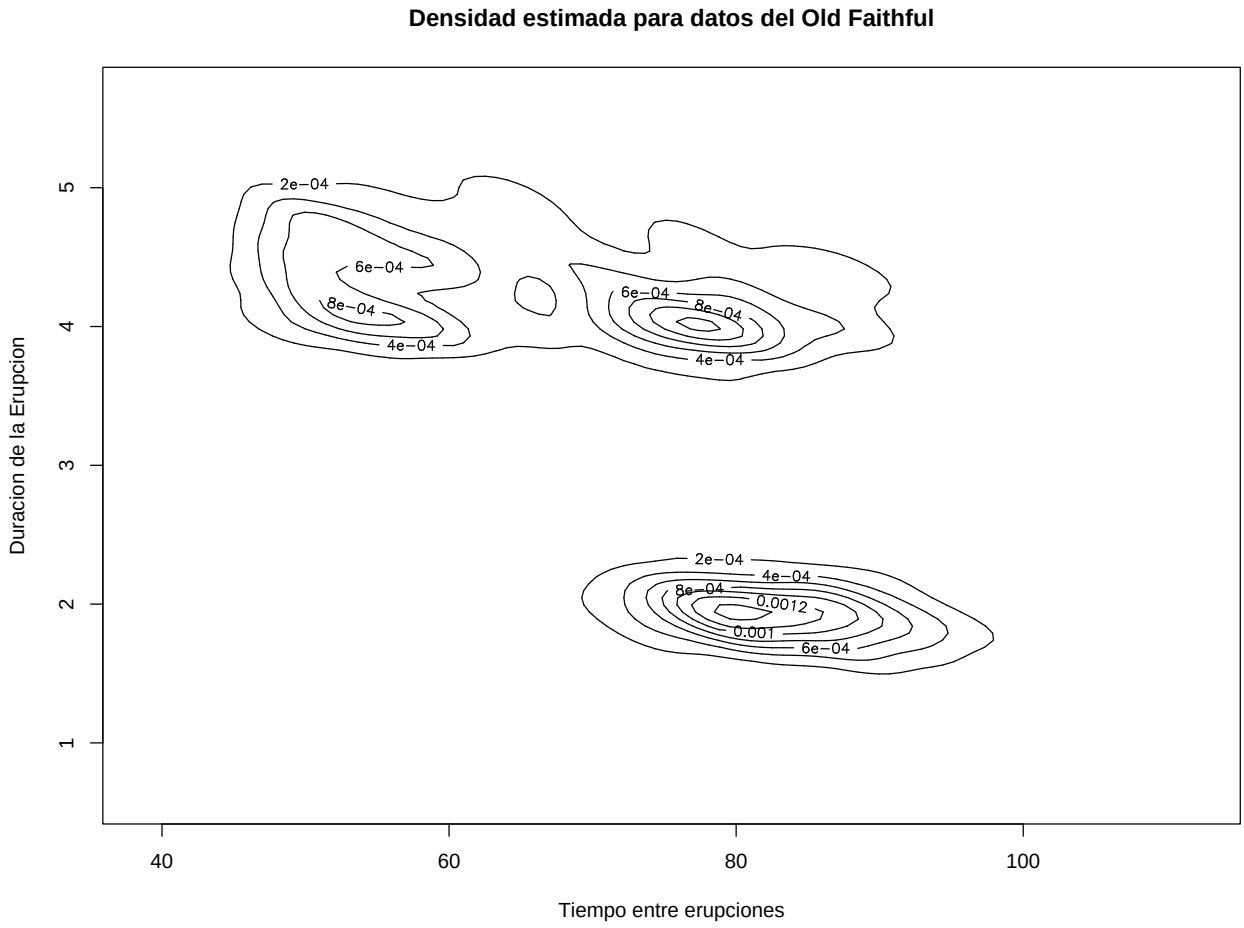
$$Var\widehat{f} \approx \frac{1}{n^{1-\frac{d}{d+4}}} = n^{-\frac{4}{d+4}}.$$

Es decir, la varianza del estimador tiende a cero como $\frac{1}{\sqrt[d]{n}}$, lo que hace necesario un n de orden exponencial en d para mantener una varianza dada.

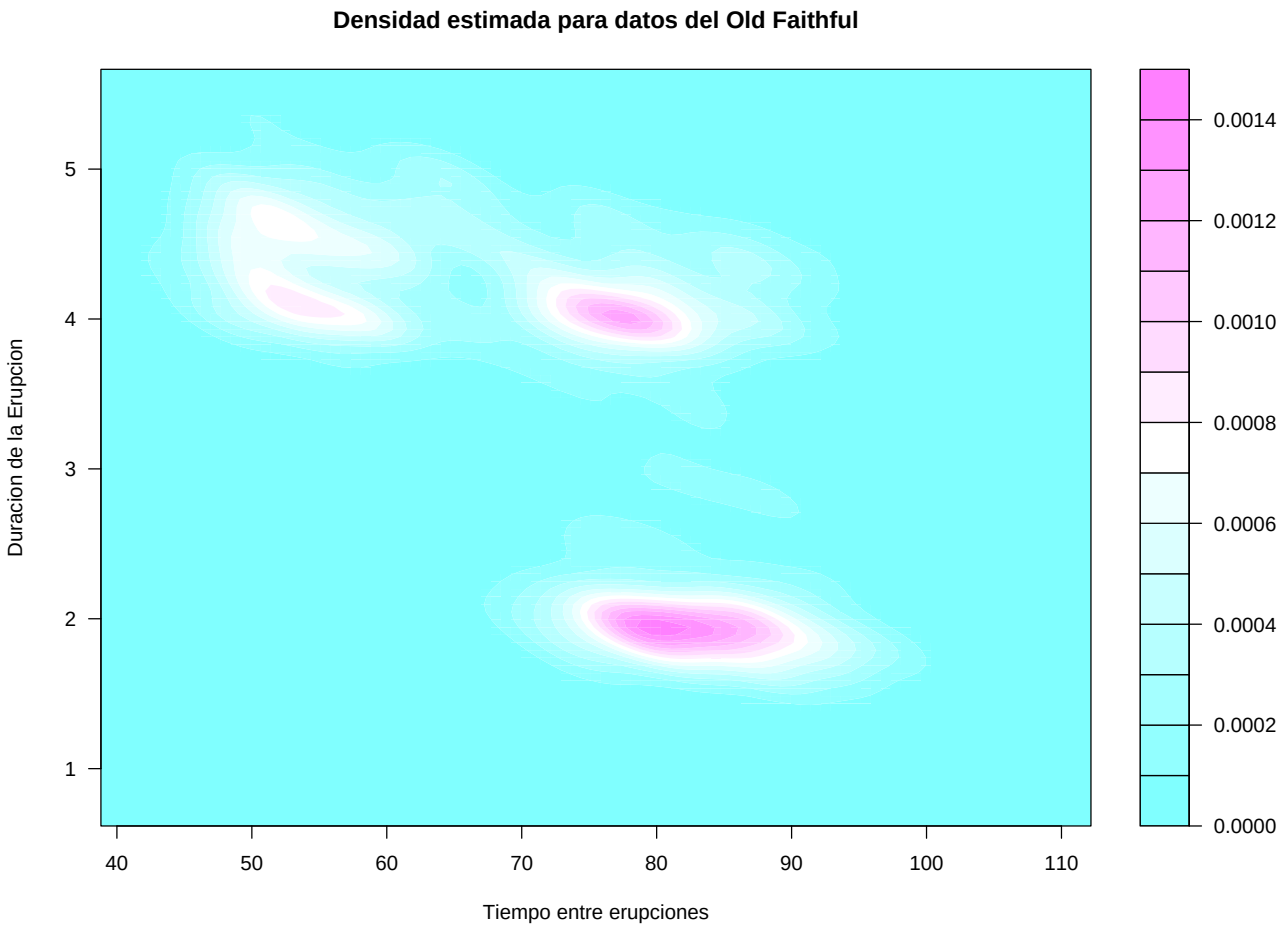
Estimación de la densidad conjunta para los datos del geyser.



Estimación de la densidad conjunta para los datos del geyser.



Estimación de la densidad conjunta para los datos del geyser.



Algunas referencias y sitios de interés.

Libros:

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

- Wand, M. P. and Jones, M. C. (1995).*Kernel Smoothing*. Chapman and Hall, London.

- Bowman, A.W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations.* Oxford University Press, Oxford.

Sitios de interés:

- CRAN.
  Sitio de desarrollo del lenguaje R, (versión libre S-Plus) Tiene módulos para estimación de densidades en 1,2 y 3 dimensiones, selección de ventanas, etc. Aparte de esto, hay muchísimos módulos para aplicar técnicas estadísticas.
  http://www.r-project.org/

- Octave home page.
  Sitio de desarrollo del lenguaje Octave (versión libre de Matlab). (Si se obtiene algún mensaje de error en un código escrito para Matlab, lo más probable es que se arregle transponiendo una matriz).
  http://www.octave.org/

- Sitio Personal de Steve Marron.
  Tiene código escrito para hacer estimación de densidades en Matlab. También tiene preprints de interés.

  http://www.unc.edu/depts/
  statistics/faculty/marron.html

- Wölfan Härdle. *Applied Nonparametric Regression*
  Libro completo que utiliza las técnicas de suavizado para el caso de regresión.

  http://www.quantlet.de/mdstat/
  scripts/anr/html/