Corrección de sesgo de borde

Lucia Coudet - Daniel Czarnievicz

Octubre de 2018

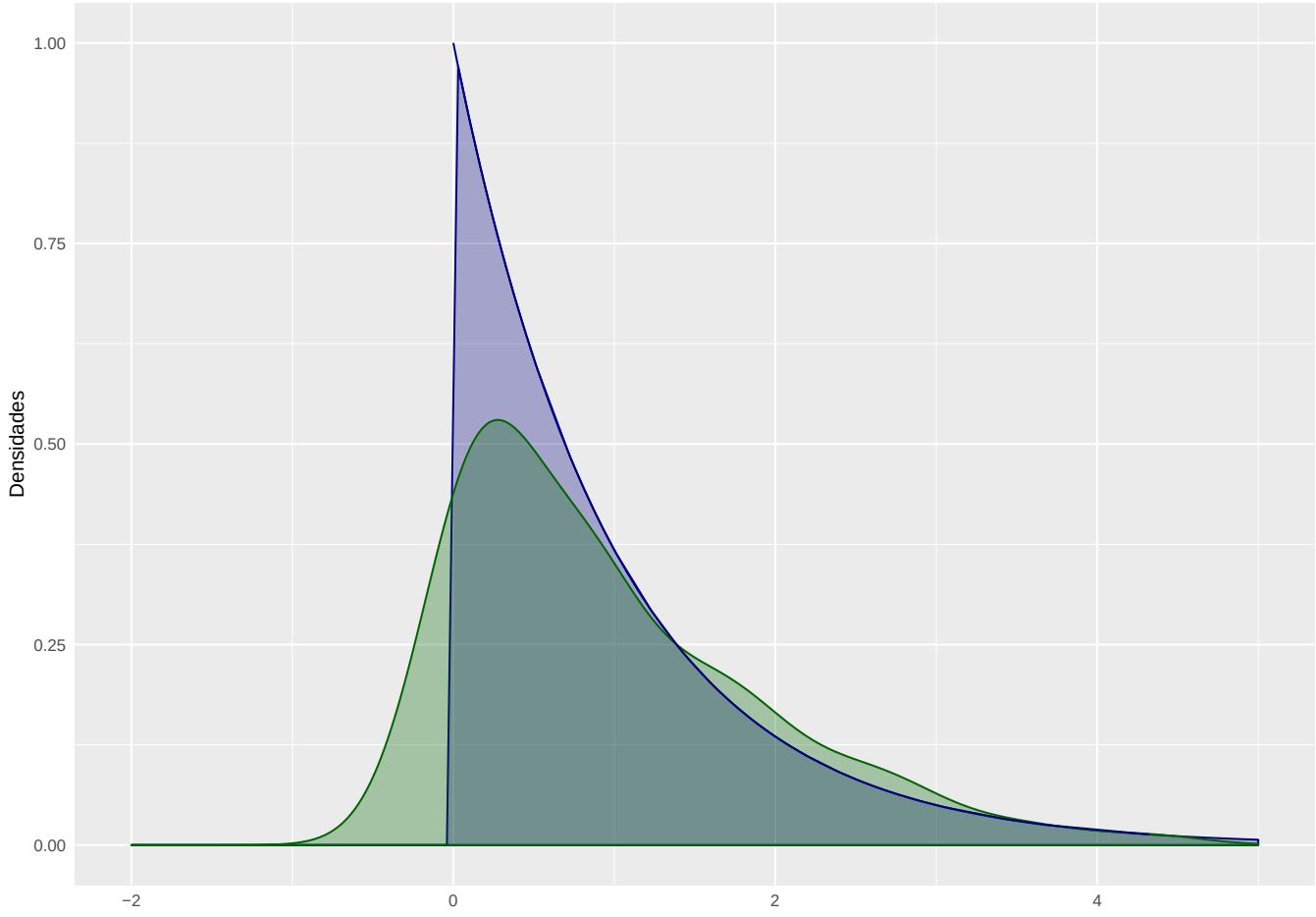
Motivación

Motivación

Supongamos que necesitamos estimar la densidad, mediante métodos kernel, de una muestra proveniente de una distribución exponencial $X \sim Exp(\theta)$ con esperanza θ^{-1}.

Motivación

```
## Warning: Calling `as_tibble()` on a vector is discourage
## This warning is displayed once per session.
```



Motivación

Como puede observarse en la `Figura 1` el kernel gaussiano no estima de forma desable la densidad en y cerca del borde $x = 0$.

Esto se debe a que la estimación kernel supone que se cumplen ciertas **condiciones de suavidad** sobre toda la recta real, lo cual no se cumple para el caso de la densidad exponencial, cuya derivada primera contiene un punto de discontinuidad en $x = 0$.

Sesgo de borde

Sesgo de borde

Supongamos que f tiene dos derivadas continuas en todo su recorrido y que $n \to \infty$, $h \to 0$ y $nh \to \infty$. Entonces,

En el interior del soporte:

$$E(\hat{f}(x)) \approx f(x) + \frac{1}{2}h^2 \int u^2 K(u)du \; f''(x) = f(x) + O(h^2)$$

En y cerca del borde:

$$E(\hat{f}(x)) \approx a_0(p)f(x) - ha_1(p)f'(x) + \frac{1}{2}h^2 a_2(p)f''(x)$$

donde $x = ph$, $a_i(p) = \int_{-\infty}^{p} u^i K(u)du$, $i = 1, \ldots, 3$.

Sesgo de borde

Sesgo en el interior del soporte

Lejos del borde, lo cual significa que $x \geq h$, no hay superposición de los kernels que contribuyen a la estimación de la densidad con el borde mismo, por lo tanto, la expresión usual para la media asintótica aplica. Si suponemos que f tiene deriavas primera y segunda contínuas en todo el soporte, y que $n \to \infty$, $h = h(n) \to 0$, y $nh \to \infty$, entonces:

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

Por lo tanto:

$$E(\hat{f}(x)) = E\left(\frac{1}{n} \sum_{i=1}^{n} K_h(x - y)\right) = \frac{n}{n} E(K_h(x-y)) = E(K_h(x-X_i))$$

Sesgo de borde

Dado que el kernel es una función de la variable aleatoria y:

$$E(K_h(x-y)) = \int K_h(x-y)f(y)dy = \int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy$$

Cambio de variable: $\frac{x-y}{h} = u$

$$\int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy = \int K(u)f(x-uh)du$$

Sesgo de borde

Desarrollo de taylor de orden 2 en $f(x-uh)$:

$$f(x-uh) = f(x) + f^{'}(x)(x-uh-x) + f^{''}(x)\frac{(x-uh-x)^2}{2!} + o(h^2) =$$

$$= f(x) - uhf^{'}(x) + \frac{(hu)^2}{2}f^{''}(x) + o(h^2)$$

Por lo tanto:

$$\int K(u)f(x-uh)du = \int K(u)\left[f(x) - uhf^{'}(x) + \frac{(uh)^2}{2!}f^{''}(x) + o(h^2)\right]du =$$

Sesgo de borde

Aplicando distributiva

$$= f(x)\int K(u)du - f'(x)h\int uK(u)du + \frac{h^2}{2}f''(x)\int u^2K(u)du +$$

$$+ o(h^2)\int k(u)du$$

Si se cumple que:

- $\int K(u)du = 1$
- $\int uK(u)du = 0$

$$E(\hat{f}(x)) = f(x) + \frac{h^2}{2}f''(x)\int u^2K(u)du + o(h^2)$$

Sesgo de borde

Entonces tenemos que:

$$sesgo(\hat{f}(x)) = f(x) + \frac{h^2}{2} f''(x) \int u^2 K(u) du + o(h^2) - f(x)$$

$$\boxed{sesgo(\hat{f}(x)) \approx \frac{h^2}{2} f''(x) \int u^2 K(u) du}$$

Sesgo de borde

Sesgo en y cerca del borde

En y cerca del borde, el problema viene dado porque se estima masa de probabilidad fuera del mismo. Es decir, se da una pérdida de masa de probabilidad.

Suponiendo (sin pérdida de generalidad) que el soporte de f es $[0, \infty)$ y tomando $x = ph$, donde $0 < p < 1$ (observe que si $p > 1$ se está en un punto lejos del borde por lo cual se está en el interior del soporte), entonces hay que prestar atención a los límites de la integral:

$$E(\hat{f}(x)) = \int_{-\infty}^{p} K(z) f(x - hz) dz$$

Sesgo de borde

Aplicando desarrollo de Taylor para $f(x - hz)$, ahora $\int_{-\infty}^{p} K(z)dz \neq 1$ por lo que la $f(x)$ queda ponderada por un término $\neq 1$. No se logra entonces una estimación consistente en los puntos cerca y en el borde.

$$E(\hat{f}(x)) = f(x) \int_{-\infty}^{p} K(u)du + o(1)$$

Sesgo de borde

Si aplicamos desarrollo de orden 2 obtenemos la expresión presentada anteriormente:

$$E(\hat{f}(x)) = \int_{-\infty}^{p} K(u)\left[f(x) - f'(x)(uh) + f''(x)\frac{(uh)^2}{2!} + o(h^2)\right]du =$$

$$\approx f(x)\underbrace{\int_{-\infty}^{p} K(u)du}_{a_0(p)} + f'(x)h\underbrace{\int_{-\infty}^{p} uK(u)du}_{a_1(p)} + \frac{h^2}{2}f''(x)\underbrace{\int_{-\infty}^{p} u^2K(u)du}_{a_2(p)}$$

Sesgo de borde

Por lo tanto:

$$E(\hat{f}(x)) \approx a_0(p)f(x) - ha_1(p)f'(x) + \frac{1}{2}h^2 a_2(p)f''(x)$$

Siendo el sesgo:

$$\boxed{sesgo(\hat{f}(x)) \approx a_0(p)f(x) - ha_1(p)f'(x) + \frac{1}{2}h^2 a_2(p)f''(x) - f(x)}$$

Corrección del sesgo de borde

Como puede observarse en la expresión aproximada para la esperanza, en y cerca del borde existe una pérdida de masa de probabilidad más allá del mismo.

Una primera opción para corregir este problema es simplemente truncar la estimación al intervalo $[0, \infty)$, lo cual resulta inapropiado ya que no corrige el problema del sesgo, incluso si se trunca y se renormaliza $\hat{f}$ para que integre 1.

Existen varios métodos que tratan de corregirlo, algunos de ellos son:

- Mirroring
- Jackknife generalizado
- Linear multiples, local linear regression, and local linear density estimation, entre otros.

Mirroring

Mirroring

El método mirroring se basa en reubicar la masa de probabilidad perdida (más allá del borde) *reflejando* la misma en el borde. Esto es, utilizar:

$$\hat{f}_R(x) = \hat{f}(x) + \hat{f}(-x)$$

o equivalentemente reemplazar el kernel $K_h(x - X_i)$ por

$$K_h(x - X_i) + K_h(-x - X_i)$$

De esta forma se recupera la consistencia del estimador:

$$E\left[\hat{f}_R(x)\right] \approx f(x) - h2[a_1(p) + p(1 - a_o(p))]f^{'}(p)$$

Mirroring

Velocidad de convergencia

El sesgo anterior es una $O(h)$. Existen alternativas que permiten obtener un sesgo $O(h^2)$ tanto cerca del borde como al interior.

Una alternativa es usar el método *jackknifing* generalizado.

Generalized jackknifing

Generalized jackknifing

Metodología

Tomar una combinación lineal de K (el kernel) y una función L, relacionada a K, de tal forma que el kernel resultante tenga las siguientes propiedades:

- $a_0(p) = 1$
- $a_1(p) = 0$

Sean $c_l(p) = \int_{-\infty}^{p} u^l L(u) du$, Jones (1993) demuestra que la combinación lineal

$$\frac{c_1(p)K(x) - a_1(p)L(x)}{c_1(p)a_0(p) - a_1(p)c_0(p)}$$

tiene sesgo $O(h^2)$

Generalized jackknifing

En particular, Rice (1984) propone utilizar $L(x) = cK(cx)$ lo cual implica trabajar con una combinación lineal que utiliza un solo kernel, K, y dos anchos de banda, h y ch, con $0 < c < 1$.

De este modo, se obtiene el siguiente kernel de borde:

$$K_c(x) = \frac{\big[a_1(pc) - a_1(c)\big]K(x) - a_1(p)c^2K(cx)}{\big[a_1(pc) - a_1(c)\big]a_0(p) - a_1(p)c\big[a_0(pc) + a_0(c) - 1\big]}$$

El cual constituye una familia de kernels según el valor de c considerado.

Generalized jackknifing

En el caso particular en que se combinen las funciones $K(x)$ y $xK(x)$ el kernel resultante es:

$$K_L(x) = \frac{a_2(p) - a_1(p)x}{a_0(p)a_2(p) - a_1^2(p)} K(x)$$

lo cual implica asignarle un sistema de pesos al kernel selccionado.

Otra opción es considerar la derivada primera del kernel seleccionado, lo cual implica imponer condiciones de suavidad. Se obtiene el siguiente kernel:

$$K_D(x) = \frac{a_1^{'}(p)K(x) - a_1(p)K^{'}(x)}{a_1^{'}(p)a_0(p) - a_1(p)a_o^{'}(p)}$$

Generalized jackknifing - Mirroring

Dado que el método mirroring alcanzan un sesgo de orden h, es posible combinarlo con el método de *jackknifing* para alcanzar un sesgo de orden h^2 en todo el recorrido.

Esto se logra combinando $K(x)$ y $K(2p-x)$ de forma tal que la estimación de la función de densidad utiliza el siguiente kernel:

$$K_{R1}(x) = \frac{2p\big[1 - a_0(p) + a_1(p)\big]K(x) - a_1(p)K(2p-x)}{\big[2p(1-a_0(p)) + a_1(p)\big]a_0(p) - a_1(p)(1-a_0(p))}$$

Sesgo de borde

Jones (1993) demuestra que el sesgo es $\frac{1}{2}h^2 f''(x)B(p)$ donde:

$$B(p) = \frac{c_1(p)a_2(p) - a_1(p)c_2(p)}{c_1(p)a_0(p) - a_1(p)c_0(p)}$$

el cual es una $O(h^2)$, propiedad que se deseaba alcanzar.

Ejemplo en R

Implementación de K_L

```
library(bde)
kernel <- jonesCorrectionMuller94BoundaryKernel(
   dataPoints = datos, mu = 2, # mu: biweight kernel
   lower.limit = 0, upper.limit = 5)
kernel <- ggplot_build(
   gplot(kernel, show = FALSE,
         includePoints = FALSE))$data[[1]]
as_tibble(kernel) %>%
   dplyr::select(x, y) %>%
   ggplot() +
   geom_area(aes(x), stat="function", fun=dexp,
             alpha=0.3, color="navy", fill="navy") +
   geom_area(aes(x, y), color="darkgreen",
             fill="darkgreen", alpha=0.3) +
   labs(x=NULL, y="Densidades") +
   theme(axis.ticks=element_blank())
```

Implementación de K_L

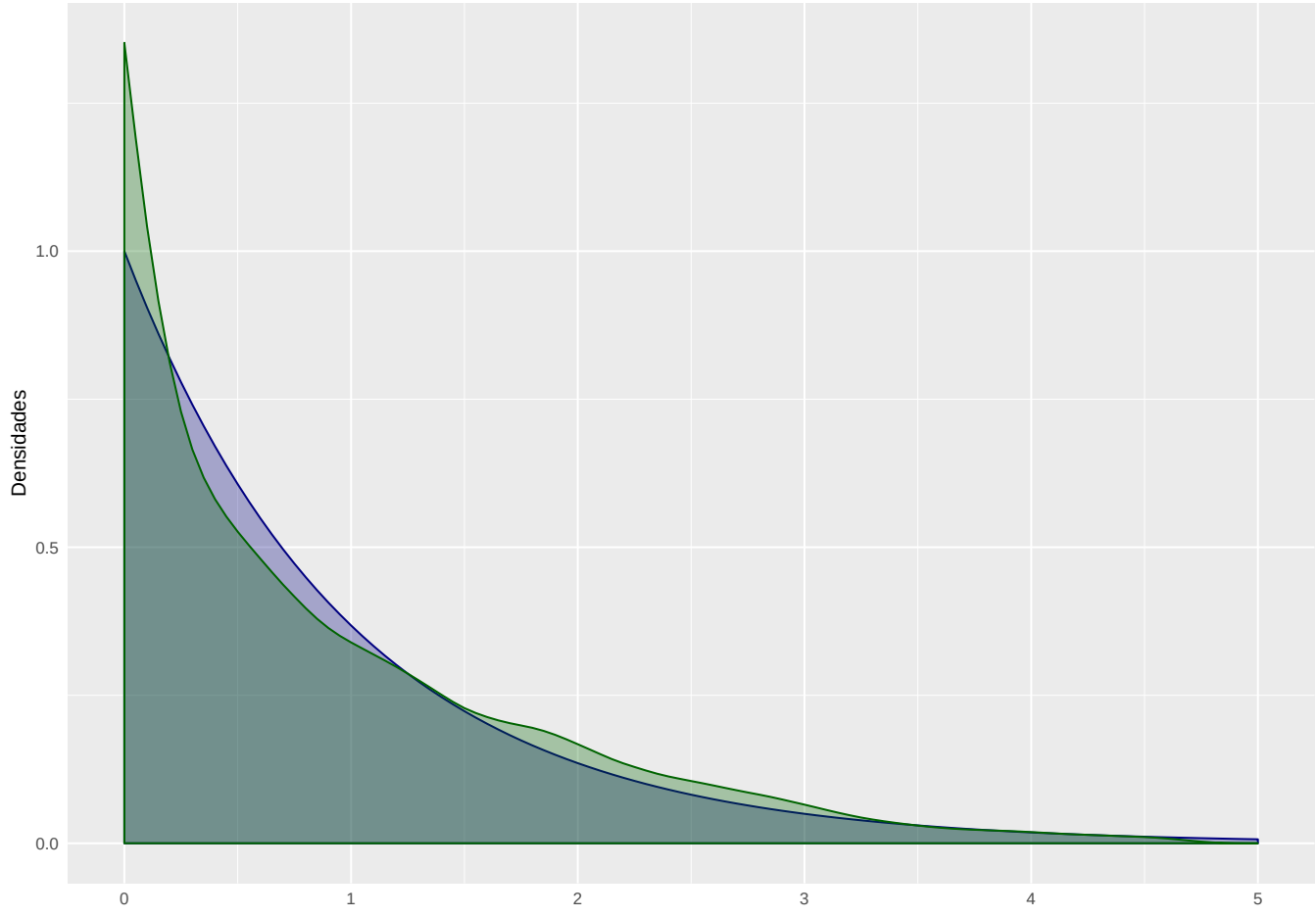


Figure 2: Densidad exponencial (violeta) y estimación utilizando K_L (verde).

Referencias

Referencias

Ivanka, Horová, Kolacek Jan, and Zelinka Jiri. 2012. *Kernel Smoothing in Matlab: Theory and Practice of Kernel Smoothing*. World scientific.

Jones, M Chris. 1993. "Simple Boundary Correction for Kernel Density Estimation." *Statistics and Computing* 3 (3). Springer: 135–46.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rice, John. 1984. "Boundary Modification for Kernel Regression." *Communications in Statistics-Theory and Methods* 13 (7). Taylor & Francis: 893–900.

Santafe, Guzman, Borja Calvo, Aritz Perez, and Jose A. Lozano. 2015. *Bde: Bounded Density Estimation*. https://CRAN.R-project.org/package=bde.