Entrega 4 - Estimación de densidades

Coudet & Czarnievicz

Diciembre 2018

El código para esta entrega puede encontrarse haciendo click aquí.

Ejercicio 1

Se define el estimador de histograma de la siguiente forma:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathrm{I}_{\{x_i \in B_k\}}$$

siendo B_k el k-ésimo bin.

El error global de tipo L_2 se define como:

$$L_2 = \int \left(\hat{f}_n(x) - f(x)\right)^2 dx$$

A partir de la definición de error global L_2 se obtiene la siguiente expresión para el AMISE (Asymptotic Integrated Mean Squared Error):

$$AMISE(h) = \frac{1}{nh} + \frac{1}{12} h^2 R(f')$$

De lo anterior se obtiene el ancho de banda óptimo:

$$h^* = \left(\frac{6}{R(f')}\right)^{1/3} n^{-1/3}$$

Las *reglas de refencia normal* de plantean utilizar la distribución normal de media μ y varianza σ^2 como aproximaciones para el término $R(f')$. Estas son:

- Scott (1979): $\hat{h} = 3.5\hat{\sigma}\, n^{-1/3}$ donde $\hat{sigma}$ es la varianza muestral.
- Freedman & Diaconis (1981): $\hat{h} = 2\, IQR\, n^{-1/3}$ donde IQR es el Rango Interquartilico.

Por lo tanto, dada la densidad de una distribución normal:

$$f(x) = (2\pi\sigma^2)^{-1/2} \ \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

Derivamos la misma:

$$\frac{\partial f(x)}{\partial x} = -(2\pi\sigma^2)^{-1/2} \ \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \ \left(\frac{x-\mu}{\sigma^2}\right)$$

Calculamos el cuadrado de su valor absoluto:

$$|f'(x)|^2 = \left[f'(x)\right]^2 = \left(-(2\pi\sigma^2)^{-1/2} \ \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \ \left(\frac{x-\mu}{\sigma^2}\right)\right)^2 =$$

$$= \frac{1}{2\pi\sigma^4}\left(\frac{x-\mu}{\sigma}\right)^2 \ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

Por lo que obtenemos la siguiente expresión para el término $R(f\text{t})$:

$$R(f') = \int_{-\infty}^{+\infty} |f'(x)|^2 dx = \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma^4}\left(\frac{x-\mu}{\sigma}\right)^2 \ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx$$

$$= \frac{1}{2\pi\sigma^4} \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{\sigma}\right)^2 \ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx$$

Realizando el siguiente cambio de variable: $\frac{x-\mu}{\sigma} = z \Rightarrow dx = \sigma dz$, e integrando por partes:

$$= \frac{1}{2\pi\sigma^3} \int_{-\infty}^{+\infty} z^2 \ \exp\left\{-z^2\right\} dz = \frac{1}{2\pi\sigma^3}\left[\underbrace{z\left(-\frac{1}{2}\exp\left\{-z^2\right\}\right)}_{0}\Bigg|_{-\infty}^{+\infty} + \frac{1}{2}\underbrace{\int_{-\infty}^{+\infty} z \ \exp\left\{-z^2\right\}}_{\sqrt{\pi}}\right] = \frac{1}{2\pi\sigma^3}\frac{\sqrt{\pi}}{2}$$

Obtenemos la expresión para $R(f'):

$$\boxed{R(f') = \frac{1}{4\sqrt{\pi}\sigma^3}}$$

Ejercicio 2

Dado un kernel K (o función núcleo) y un ancho de banda $h > 0$, el estimador de densidad kernel o estimador de densidad por núcleos está definido como:

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

En lo que sigue usaremos las siguientes funciones núcleo:

- Gaussian: $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}u^2\right\}$ cpm $u \in \mathbb{R}$.
- Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)$ con $|u| < 1$.
- Rectangular (uniform): $K(u) = \frac{1}{2}$ con $|u| < 1$.
- Triangular: $K(u) = (1 - |u|)$ con $|u| < 1$.
- Biweight (quadratic): $K(u) = \frac{15}{16}(1 - u^2)^2$ con $|u| < 1$.
- Cosine: $K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)$

Y los siguientes criterior de selección para el ancho de ventana h:

- Scott (normal refernce rule) $h_n = 1.06\hat{\sigma}n^{-\frac{1}{5}}$ donde $\hat{sigma} = \min\left\{s, \frac{IQR}{1.34}\right\}$ con IQR el rango intercuartílico.
- Silverman: $h_n = 0.9\hat{\sigma}n^{-\frac{1}{5}}$ donde $\hat{sigma} = \min\left\{S^2, IQR\right\}$.
- Unbiased cross validation (UCV): Encuentra el ancho de banda óptimo h_* mediante validación cruzada (leave-one-out). Para ello utiliza como función de riesgo el error cuadrático integrado $ISE(\hat{f}(x))$, con lo cual busca minimizar la siguiente función objetivo:
$$\hat{J}(h) = \int(\hat{f}_n(x))^2 dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{(-i)}(X_i)$$
donde $\hat{f}_{(-i)}(X_i)$ es el estimador de la densidad luego de remover la i-ésima observación. $\hat{J}(h)$ es conocido como el score de cross validation o el riesgo estimado.
- Biased cross validation (BCV): A diferencia de UCV encuentra h_* que minimiza el $AMISE(\hat{f}(x))$ en lugar del que minimiza el $ISE(\hat{f}(x))$. Reemplazando $R(f'')$ por $\hat{R}(f'') - \frac{1}{nh^5}R(K'')$ en la expression del $AMISE$. La función objetivo es entonces:
$$BCV(h) = \frac{1}{nh}R(K) + \frac{h^4\mu_2^2(k)}{4}\left[R(\hat{f}'') - \frac{1}{nh^5}R(K'')\right]$$

- Sheather & Jones: Implementa las mejoras propuestas por Sheather and Jones (1991) al estimador propuesto por Park and Marron (1990). El método se basa en elegit el ancho de banda que minimiza el error cuadrático medio integrado (MISE).

a)

```
data("geyser", package="locfit")
tibble(x = density(geyser, bw = 0.4, kernel = "gaussian")$x,
       Gaussian = density(geyser, bw = 0.4, kernel = "gaussian")$y,
       Epanechnikov = density(geyser, bw = 0.4, kernel = "epanechnikov")$y,
       Rectangular = density(geyser, bw = 0.4, kernel = "rectangular")$y,
       Triangular = density(geyser, bw = 0.4, kernel = "triangular")$y,
       Biweight = density(geyser, bw = 0.4, kernel = "biweight")$y,
       Cosine = density(geyser, bw = 0.4, kernel = "cosine")$y) %>%
  gather(key=key, value=value, -x) %>%
  ggplot() +
  geom_line(aes(x=x, y=value, color=key), show.legend=FALSE) +
  facet_wrap(~key) +
  labs(x="\nEruption duration", y="Estimated density\n") +
  ggthemes::theme_economist() +
  theme(axis.title=element_text(face="bold", size=12))
```

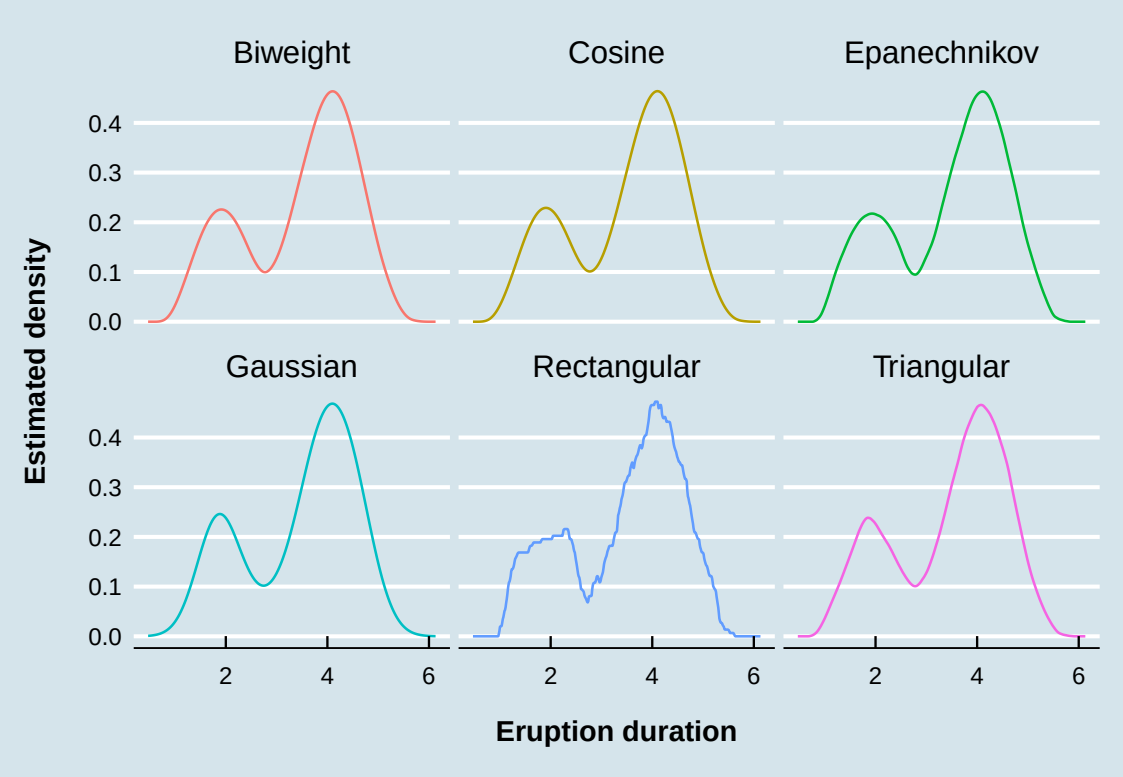


Figure 1: Kernel estimation of eruption duration

En la `Figura 1` es posible apreciar que todos los estimadores núcleo estiman una densidad bimodal. Las diferencias obtenidas surgen del propio método subyacente en cada uno de ellos:

- El núcleo *Epanechnikov* estima con forma tendiente a cúpulas.
- El núcleo *Rectangular* es notoriamente más rugoso.
- El núcleo *Triangular* estima con forma tendiente a triángulos.
- Los núcleos *Biweight*, *Cosine*, y *Gaussian* son los que logran una estimación más suave.

b)

```
data("geyser", package="locfit")
x <- tibble(x_005 = density(geyser, bw = 0.05, kernel = "gaussian")$x,
            x_01 = density(geyser, bw = 0.1, kernel = "gaussian")$x,
            x_02 = density(geyser, bw = 0.2, kernel = "gaussian")$x,
            x_04 = density(geyser, bw = 0.4, kernel = "gaussian")$x,
            x_06 = density(geyser, bw = 0.6, kernel = "gaussian")$x,
            x_1 = density(geyser, bw = 1, kernel = "gaussian")$x) %>%
   gather(key=keyx, value=x)
y <- tibble(h_0.05 = density(geyser, bw = 0.05, kernel = "gaussian")$y,
            h_0.1 = density(geyser, bw = 0.1, kernel = "gaussian")$y,
            h_0.2 = density(geyser, bw = 0.2, kernel = "gaussian")$y,
            h_0.4 = density(geyser, bw = 0.4, kernel = "gaussian")$y,
            h_0.6 = density(geyser, bw = 0.6, kernel = "gaussian")$y,
            h_1 = density(geyser, bw = 1, kernel = "gaussian")$y) %>%
   gather(key=keyy, value=y)
as_tibble(cbind(x,y))%>%
   mutate(labels = sub("_", " = ", keyy)) %>%
   select(-starts_with("key")) %>%
   ggplot() +
   geom_line(aes(x=x, y=y, color=labels), show.legend = FALSE) +
   facet_wrap(~ labels) +
   labs(x="\nEruption duration", y="Estimated density\n") +
   ggthemes::theme_economist() +
   theme(axis.title=element_text(face="bold", size=12))
```

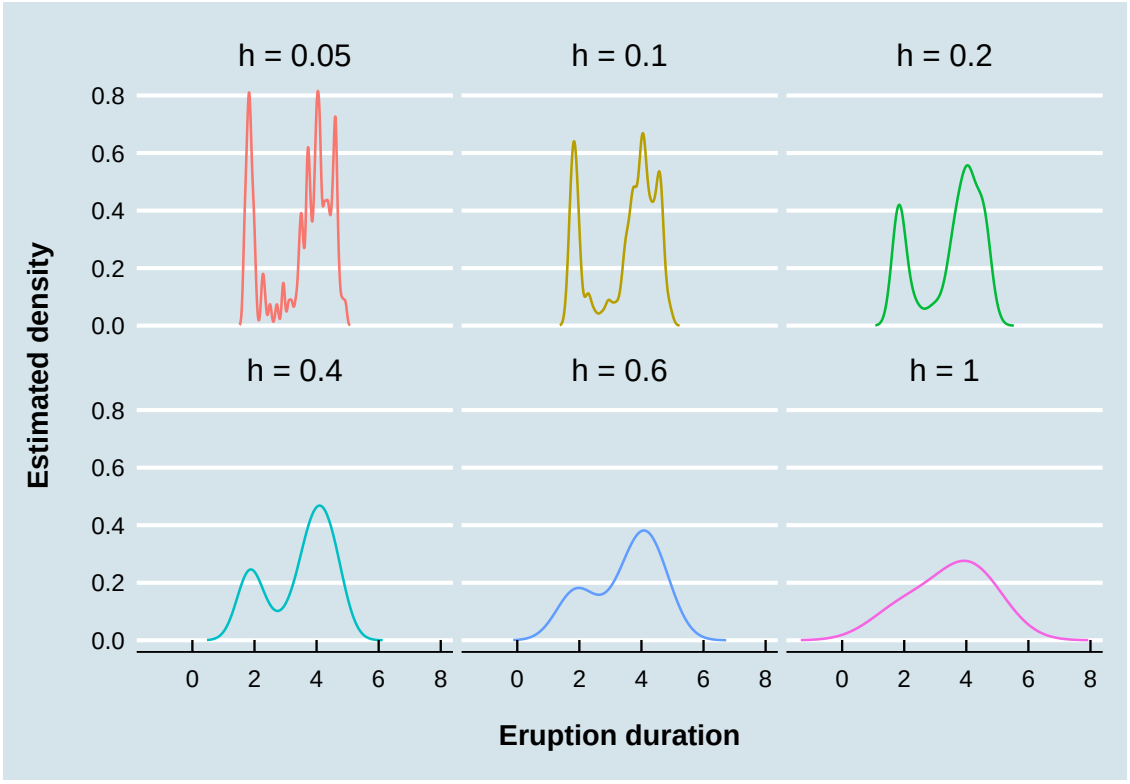


Figure 2: Kernel estimation of eruption duration for different values of h (Gaussian kernel)

Se observa claramente la transición de *undersmoothing* a *oversmoothing* a medida que aumenta el valor de h. Las estimaciones de densidades que utilizan anchos de banda más pequeños son notoriamente más rugosas, y aparecen nuevos modos respecto al gráfico anterior. En particular, si se utiliza $h = 1$ la distribución se vuelve unimodal.

c)

```
data("geyser", package="locfit")
x <- tibble(x_nrd0 = density(geyser, bw = 'nrd0', kernel = "gaussian")$x,
            x_nrd = density(geyser, bw = 'nrd', kernel = "gaussian")$x,
            x_ucv = density(geyser, bw = 'ucv', kernel = "gaussian")$x,
            x_bcv = density(geyser, bw = 'bcv', kernel = "gaussian")$x,
            x_SJ = density(geyser, bw = 'SJ', kernel = "gaussian")$x) %>%
   gather(key=keyx, value=x)
y <- tibble(Silverman = density(geyser, bw = 'nrd0', kernel = "gaussian")$y,
            Scott = density(geyser, bw = 'nrd', kernel = "gaussian")$y,
            Unbiased_CV = density(geyser, bw = 'ucv', kernel = "gaussian")$y,
            Biased_CV = density(geyser, bw = 'bcv', kernel = "gaussian")$y,
            Sheather_Jones = density(geyser, bw = 'SJ', kernel = "gaussian")$y) %>%
   gather(key=keyy, value=y)
```

```
as_tibble(cbind(x,y))%>%
   mutate(labels = sub("_", " ", keyy)) %>%
   select(-starts_with("key")) %>%
   ggplot() +
   geom_line(aes(x=x, y=y, color=labels), show.legend = FALSE) +
   facet_wrap(~ labels) +
   labs(x="\nEruption duration", y="Estimated density\n") +
   ggthemes::theme_economist() +
   theme(axis.title=element_text(face="bold", size=12))
```

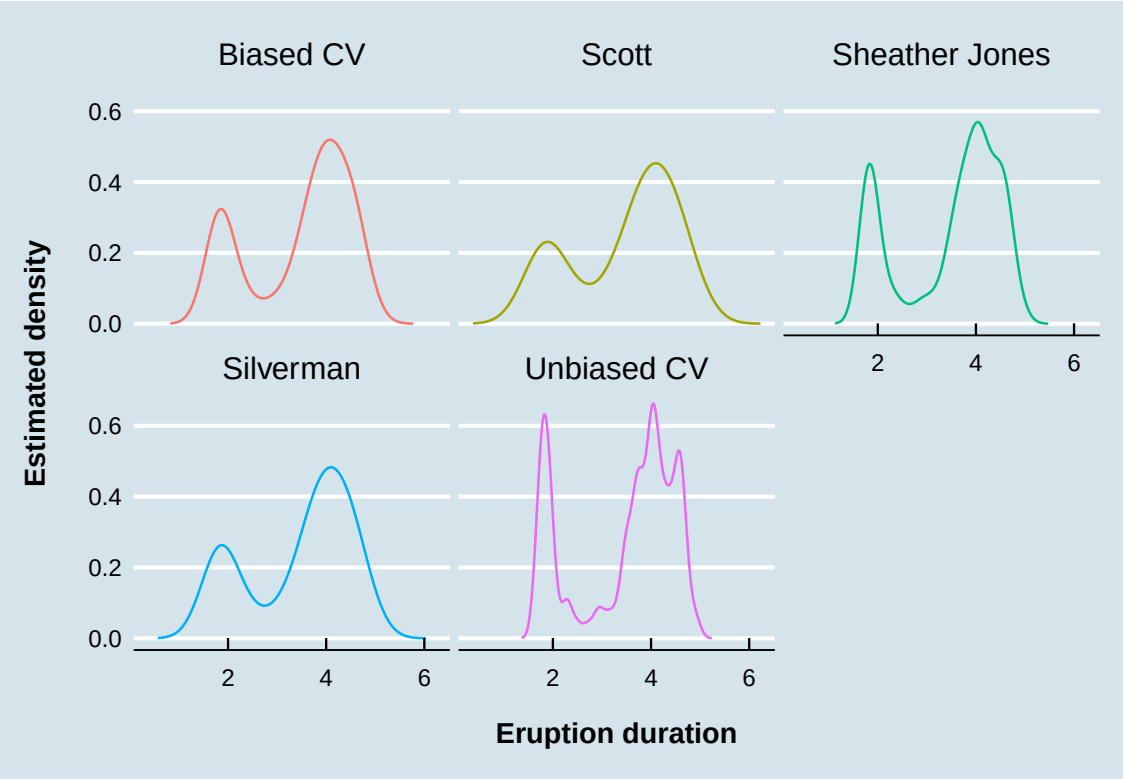


Figure 3: Kernel estimation of time eruptions using the different criteria for the selection of h (Gaussian kernel)

Ejercicio 3

Sea la siguiente mixtura normal (distribución *Bart Simpson*):

$$F(x) = \frac{1}{2}\Phi(x,\, 0,\, 1) + \frac{1}{10}\sum_{i=0}^{4}\Phi(x,\, \frac{i}{2} - 1,\, \frac{1}{10})\, \mathrm{I}_{(x\in\mathbb{R})}$$

```
n <- 1000
x <- rnorm(n)
Densidad <- 0.5 * dnorm(x) + (1/10) * (dnorm(x, mean=-1, sd=1/10) +
                                       dnorm(x, mean=-1/2, sd=1/10) +
                                       dnorm(x, mean=0, sd=1/10) +
                                       dnorm(x, mean=1/2, sd=1/10) +
                                       dnorm(x, mean=1, sd=1/10))
Distribucion <- 0.5 * pnorm(x) + (1/10) * (pnorm(x, mean=-1, sd=1/10) +
                                           pnorm(x, mean=-1/2, sd=1/10) +
                                           pnorm(x, mean=0, sd=1/10) +
                                           pnorm(x, mean=1/2, sd=1/10) +
                                           pnorm(x, mean=1, sd=1/10))
as_tibble(cbind(x, Densidad, Distribucion)) %>%
   gather(key=key, value=value, -x) %>%
   ggplot() +
   geom_line(aes(x, value, color=key), show.legend=FALSE) +
   facet_wrap(~key, scales="free", ncol=1) +
   labs(x=NULL, y=NULL) +
   ggthemes::theme_economist()
```

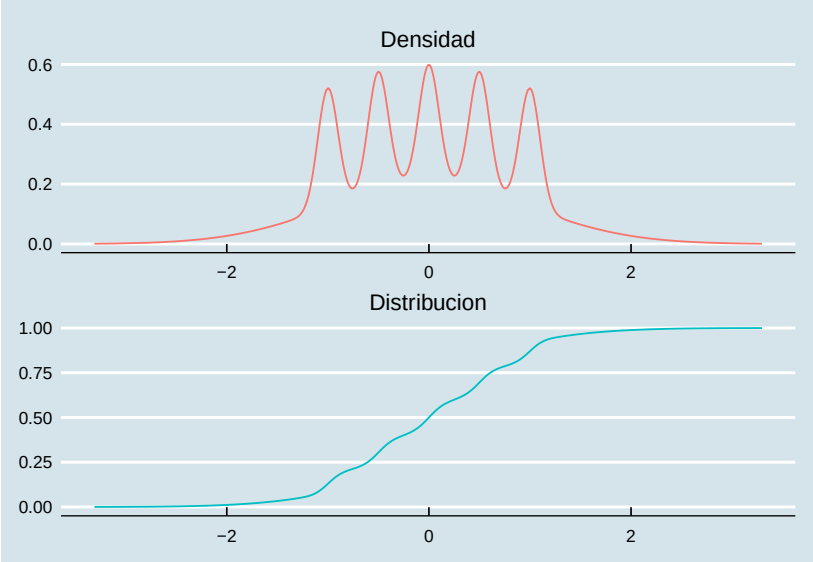


Figure 4: Distribución Bart Simpson

a)

```
n <- 1000
set.seed(58456)
y <- runif(n)
dens_bart <- function(x) {0.5 * pnorm(x) + (1/10) * (pnorm(x, mean=-1, sd=1/10) +
                                                     pnorm(x, mean=-1/2, sd=1/10) +
                                                     pnorm(x, mean=0, sd=1/10) +
                                                     pnorm(x, mean=1/2, sd=1/10) +
                                                     pnorm(x, mean=1, sd=1/10))}
sim_bart <- inverse(dens_bart)
datos <- NULL
for (i in 1:length(y)) {
   datos <- c(datos, sim_bart(y[i]))
}
x <- tibble(x_gaus_nrd = density(datos, bw = 'nrd', kernel = "gaussian")$x,
            x_gaus_ucv = stats::density(datos, bw = 'ucv', kernel = "gaussian")$x,
            x_gaus_SJ = stats::density(datos, bw = 'SJ', kernel = "gaussian")$x,
            x_gaus_0.03 = stats::density(datos, bw = 0.03, kernel = "gaussian")$x,
            x_rect_nrd = stats::density(datos, bw = 'nrd', kernel = "rectangular")$x,
            x_rect_ucv = stats::density(datos, bw = 'ucv', kernel = "rectangular")$x,
            x_rect_SJ = stats::density(datos, bw = 'SJ', kernel = "rectangular")$x,
            x_rect_0.03 = stats::density(datos, bw = 0.03, kernel = "rectangular")$x,
            x_bi_nrd = stats::density(datos, bw = 'nrd', kernel = "biweight")$x,
            x_bi_ucv = stats::density(datos, bw = 'ucv', kernel = "biweight")$x,
            x_bi_SJ = stats::density(datos, bw = 'SJ', kernel = "biweight")$x,
            x_bi_0.03 = stats::density(datos, bw = 0.03, kernel = "biweight")$x) %>%
```

```
   gather(key=keyx, value=x)
```

```
## Warning in bw.ucv(x): minimum occurred at one end of the range

## Warning in bw.ucv(x): minimum occurred at one end of the range

## Warning in bw.ucv(x): minimum occurred at one end of the range
```

```
y <- tibble(y_gaus_nrd = stats::density(datos, bw = 'nrd', kernel = "gaussian")$y,
            y_gaus_ucv = stats::density(datos, bw = 'ucv', kernel = "gaussian")$y,
            y_gaus_SJ = stats::density(datos, bw = 'SJ', kernel = "gaussian")$y,
            y_gaus_0.03 = stats::density(datos, bw = 0.03, kernel = "gaussian")$y,
            y_rect_nrd = stats::density(datos, bw = 'nrd', kernel = "rectangular")$y,
            y_rect_ucv = stats::density(datos, bw = 'ucv', kernel = "rectangular")$y,
            y_rect_SJ = stats::density(datos, bw = 'SJ', kernel = "rectangular")$y,
            y_rect_0.03 = stats::density(datos, bw = 0.03, kernel = "rectangular")$y,
            y_bi_nrd = stats::density(datos, bw = 'nrd', kernel = "biweight")$y,
            y_bi_ucv = stats::density(datos, bw = 'ucv', kernel = "biweight")$y,
            y_bi_SJ = stats::density(datos, bw = 'SJ', kernel = "biweight")$y,
            y_bi_0.03 = stats::density(datos, bw = 0.03, kernel = "biweight")$y) %>%
   gather(key=keyy, value=y)
```

```
## Warning in bw.ucv(x): minimum occurred at one end of the range

## Warning in bw.ucv(x): minimum occurred at one end of the range

## Warning in bw.ucv(x): minimum occurred at one end of the range
```

```
as_tibble(cbind(x,y)) %>%
   mutate(
      kernel = if_else(grepl("_gaus_", keyx), "Gaussian",
               if_else(grepl("_rect_", keyx), "Rectangular", "BiWeight")),
      regla = if_else(grepl("_nrd", keyx), "Scott",
              if_else(grepl("_ucv", keyx), "Unbiased CV",
              if_else(grepl("_SJ", keyx), "Sheather Jones", "h = 0.03")))
      ) %>%
   select(-starts_with("key")) %>%
   ggplot() +
   geom_line(aes(x=x, y=y, color=kernel), show.legend = FALSE) +
   facet_grid(vars(kernel), vars(regla)) +
   labs(x=NULL, y="Densidad estimada\n") +
   xlim(c(-3,3)) +
```

```
  ggthemes::theme_economist() +
  theme(axis.title=element_text(face="bold", size=12))
```

```
## Warning: Removed 954 rows containing missing values (geom_path).
```

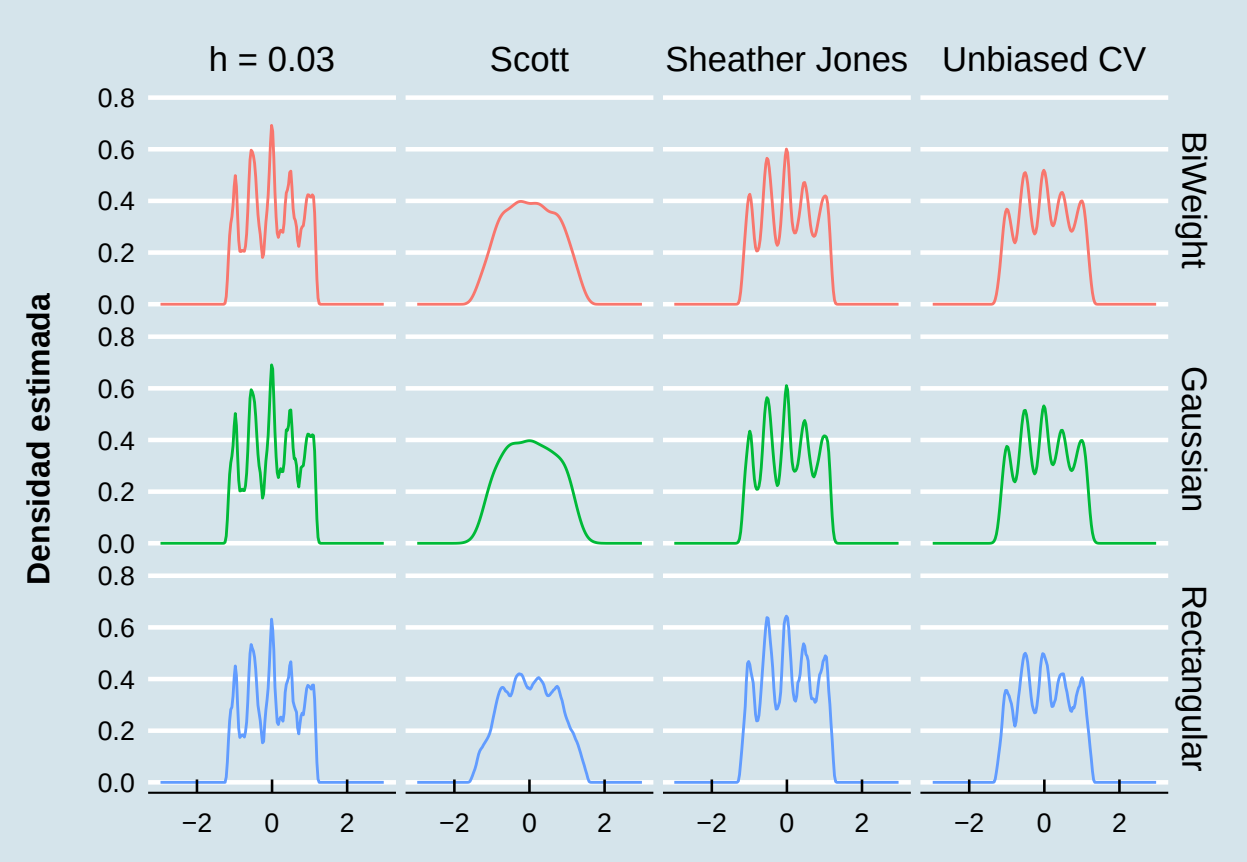


Figure 5: Estimaciones Kernel para datos provenientes de la distribución Bart Simpson, utilizando distintos criterios de selección de h y distintos Kernels

En la `Figura 5` se aprecia como la estimación obtenida es relativamente insensible a la elección del Kernel, pero muy sensible a la elección del ancho de banda h: igual elección de ancho de banda resulta en estimaciones similares independientemente del Kernel seleccionado.

Ejercicio 4

Dada la estimación de densidad por núcleo:

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

Se deine el error Cuadrático Medio Integrado (MISE) como:

$$MISE_{\hat{f}}(\hat{f}_h(x)) = \int\limits_{Rec(X)} MSE_{\hat{f}}(\hat{f}_h(x))dx = \int\limits_{Rec(X)} \left[V_{\hat{f}}(\hat{f}_h(x)) + B^2_{\hat{f}}(\hat{f}_h(x))\right] dx =$$

$$= \int\limits_{Rec(X)} V_{\hat{f}}(\hat{f}_h(x))dx + \int\limits_{Rec(X)} B^2_{\hat{f}}(\hat{f}_h(x))dx =$$

$$= \int\limits_{Rec(X)} V_{\hat{f}}(\hat{f}_h(x))dx + \int\limits_{Rec(X)} \left[E_{\hat{f}}(\hat{f}_h(x)) - f(x)\right]^2 dx$$

Comenzando con la esperanza:

$$E_{\hat{f}}(\hat{f}(x)) = E_X\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right] = \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left[\frac{1}{h}K\left(\frac{x - X_i}{h}\right)\right] = E_Y\left[\frac{1}{h}K\left(\frac{x - Y}{h}\right)\right] =$$

$$= \frac{1}{h}\int\limits_{Rec(Y)} K\left(\frac{x - Y}{h}\right) f_Y(y)dy =$$

Realizando el siguiente cambio de variable:

$$\begin{array}{rcl} Z = \dfrac{x - Y}{h} & \Rightarrow & Y = x - Zh \\ \dfrac{dz}{dy} = -\dfrac{1}{h} & \Rightarrow & dy = -dzh \\ y \to -\infty & \Rightarrow & z \to +\infty \\ y \to +\infty & \Rightarrow & z \to -\infty \end{array}$$

$$E_{\hat{f}}(\hat{f}(x)) = -\int\limits_{+\infty}^{-\infty} K(z)f(x - zh)dz = \int\limits_{-\infty}^{+\infty} K(z)f(x - zh)dz = \int\limits_{Rec(Z)} K(z)f(x - zh)dz$$

Utilizando un desarrollo de Taylor alrededor del punto x:

$$f(x-zh) \approx f(x) + (x - zh - x)f'(x) + \frac{1}{2}(x - zh - x)^2 f''(x) = f(x) - zhf'(x) + \frac{1}{2}(-zh)^2 f''(x)$$

Obtenemos que:

$$E_{\hat{f}}(\hat{f}(x)) = \int\limits_{Rec(Z)} K(z)f(x-zh)dz \approx \int\limits_{Rec(Z)} K(z)\left[f(x) - zhf'(x) + \frac{1}{2}(-zh)^2 f''(x)\right]dz =$$

$$= \int\limits_{Rec(Z)} K(z)f(x)dz - h\int\limits_{Rec(Z)} K(z)zf'(x)dz + \frac{1}{2}h^2 \int\limits_{Rec(Z)} K(z)z^2 f''(x)dz$$

Si se cumple que:

- $\int\limits_{Rec(Z)} K(z)dz = 1$
- $\int\limits_{Rec(Z)} zK(z)dz = 0$

Entonces,

$$E_{\hat{f}}(\hat{f}(x)) \approx f(x)\int\limits_{Rec(Z)} K(z)dx - hf'(x)\int\limits_{Rec(Z)} zK(z)dz + \frac{1}{2}h^2 f''(x)\int\limits_{Rec(Z)} z^2 K(z)dz =$$

$$= f(x) + \frac{1}{2}h^2 f''(x)\int\limits_{Rec(Z)} z^2 K(z)dz$$

Por lo cual el sesgo cuadrático será:

$$B^2(\hat{f}(x)) \approx \left[f(x) + \frac{1}{2}h^2 f''(x)\int\limits_{Rec(Z)} z^2 K(z)dz - f(x)\right]^2 =$$

$$= \left[\frac{1}{2}h^2 f''(x)\int\limits_{Rec(Z)} z^2 K(z)dz\right]^2 = \frac{1}{4}h^4[f''(x)]^2\left[\int\limits_{Rec(Z)} z^2 K(z)dz\right]^2 =$$

$$= \frac{1}{4}h^4[f''(x)]^2\mu_2^2(K)$$

Siendo el sesgo integrado asintótico:

$$\int\limits_{Rec(X)} B^2(\hat{f}(x))dx \approx \int\limits_{Rec(X)} \frac{1}{4}h^4[f''(x)]^2\mu_2^2(K)dx =$$

$$= \frac{1}{4}h^4\mu_2^2(K)\int\limits_{Rec(X)} [f''(x)]^2 dx = \frac{1}{4}h^4\mu_2^2(K)R(f''(x))$$

$$\boxed{AISB(\hat{f}(x)) = \frac{1}{4}h^4\mu_2^2(K)R(f''(x))}$$

Luego la varianza será:

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

$$= \frac{1}{(nh)^2}\sum_{i=1}^{n} V_{X_i}\left[K\left(\frac{x-X_i}{h}\right)\right] = \frac{1}{nh^2}V_Y\left[K\left(\frac{x-Y}{h}\right)\right] =$$

$$= \frac{1}{h}\left[\frac{1}{n}E_Y\left(\frac{1}{h}K^2\left(\frac{x-Y}{h}\right)\right) - \frac{1}{n}E_Y^2\left(\frac{1}{h}K\left(\frac{x-Y}{h}\right)\right)\right]$$

Trabajando primero con la esperanza de K^2.

$$\frac{1}{n}E_Y\left(\frac{1}{h}K^2\left(\frac{x-Y}{h}\right)\right) = \frac{1}{n}\int\limits_{Rec(Y)} \frac{1}{h}K^2\left(\frac{x-Y}{h}\right) f_Y(y)dy$$

Realizando el mismo cambio de variable que para la esperanza obtenemos que:

$$\frac{1}{n}\int\limits_{Rec(Y)} \frac{1}{h}K^2\left(\frac{x-Y}{h}\right) f_Y(y)dy = \frac{1}{n}\int\limits_{Rec(Z)} K^2(z)f(x-zh)dz$$

Utilizando un desarrollo de Taylor de orden 1 obtenemos que:

$$\frac{1}{n}\int\limits_{Rec(Z)} K^2(z)f(x-zh)dz \approx \frac{1}{n}\int\limits_{Rec(Z)} K^2(z)f(x)dz - \frac{1}{n}\int\limits_{Rec(Z)} K^2(z)zhf'(x)dz =$$

$$= \frac{1}{n}f(x)R(K) - \frac{1}{n}hf'(x)\int\limits_{Rec(Z)} zK^2(z)dz = \frac{1}{n}f(x)R(K) - O(n^{-1}) \approx \frac{1}{n}f(x)R(K)$$

En lo que respecta al segundo término,

$$\frac{1}{n}E_Y^2\left(\frac{1}{h}K\left(\frac{x-Y}{h}\right)\right) = \frac{1}{n}\left[f(x) + B(\hat{f}(x))\right]^2 =$$

$$= \frac{1}{n}\left[f(x) + o(h^2)\right]^2 \approx \frac{1}{n}[f(x)]^2 = O(n^{-1})$$

Por lo tanto, la varianza integrada asintótica es:

$$AIV(\hat{f}(x)) = \frac{1}{h} \int\limits_{Rec(X)} \frac{1}{n} f(x) R(K) dx = \frac{1}{nh} R(K) \int\limits_{Rec(X)} f(x) dx = \frac{1}{nh} R(K)$$

$$AIV(\hat{f}(x)) = \frac{1}{nh} R(K)$$

Entonces,

$$AMISE(\hat{f}(x)) = AIV(\hat{f}(x)) + AISB(\hat{f}(x))$$

$$\boxed{AMISE(\hat{f}(x)) = \frac{1}{nh} R(K) + \frac{1}{4} h^4 \mu_2^2(K) R(f''(x))}$$

b)

A partir de la expresión encontrada en la parte anterior, es posible encontrar el ancho de banda óptimo h^* derivando dicha expresión con respecto de h e igualando a 0:

$$\frac{\partial\, AMISE(\hat{f}(x))}{\partial h} = (-1)\frac{1}{n} h^{-2} R(K) + \frac{1}{4} 4 h^3 \mu_2^2(K) R(f''(x)) = 0 \Rightarrow$$

$$\Rightarrow \frac{1}{n} h^{-2} R(K) = h^3 \mu_2^2(K) R(f''(x)) \Rightarrow h^5 = \frac{R(K)}{n \mu_2^2(K) R(f'')} \Rightarrow$$

$$\Rightarrow \boxed{h^* = \left(\frac{R(K)}{\mu_2^2(K) R(f'')} \right)^{1/5} n^{-1/5}}$$

Calculando la derivada segunda se comprueba que es un mínimo:

$$\frac{\partial^2\, AMISE(\hat{f}(x))}{\partial^2 h} = 2\frac{1}{n} h^{-3} R(K) + 3 h^2 \mu_2^2(K) R(f''(x)) > 0$$

dado que $h > 0$, $R(K) > 0$, $\mu_2^2(K) > 0$, $R(f'') > 0$. Por lo tanto, estamos en presencia de un mínimo.

c)

$$f'' = \frac{\partial^2 f}{\partial^2 x} = \frac{\partial f'}{\partial x} = \frac{\partial}{\partial x}\left(-(2\pi\sigma^2)^{-1/2}\ \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\ \left(\frac{x-\mu}{\sigma^2}\right)\right) =$$

$$= -(2\pi\sigma^2)^{-1/2}\left[\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\ \left(-\frac{2}{2\sigma}\left(\frac{x-\mu}{\sigma}\right)\right)\left(\frac{x-\mu}{\sigma^2}\right) + \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\ \frac{1}{\sigma^2}\right] =$$

$$= -(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\left[-\left(\frac{x-\mu}{\sigma^2}\right)^2 + \frac{1}{\sigma^2}\right]$$

Entonces,

$$R(f'') = \int [f'']^2\, dx = \int \left[-(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\left[-\left(\frac{x-\mu}{\sigma^2}\right)^2 + \frac{1}{\sigma^2}\right]\right]^2 dx$$

Desarrollando el cuadrado dentro de la integral:

$$R(f'') = \int \left[(2\pi\sigma^2)^{-1}\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\}\left(\frac{\sigma^2-(y-\mu)^2}{\sigma^4}\right)^2\right] dx =$$

$$= \int \left[(2\pi\sigma^2)^{-1}\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\}\left(\frac{\sigma^4 - 2\sigma^2(y-\mu)^2 + (y-\mu)^4}{\sigma^8}\right)\right] dx =$$

$$= (2\pi\sigma^2)^{-1}\sigma^{-4}\int\left[\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\}\right] dx -$$

$$-\frac{2}{\sigma^6}(2\pi)^{-1}\int\left[\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\}\left(\frac{y-\mu}{\sigma}\right)^2\right] dx +$$

$$+\frac{1}{\sigma^4}(2\pi\sigma^2)^{-1}\int\left[\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\}\left(\frac{y-\mu}{\sigma}\right)^4\right] dx =$$

$$= (2\pi\sigma^2)^{-1}\sigma^{-4}\sigma\underbrace{\int\left[\exp\left\{-z^2\right\}\right] dz}_{\sqrt{\pi}} -$$

$$-\frac{2}{\sigma^6}(2\pi)^{-1}\sigma\underbrace{\int\left[\exp\left\{-z^2\right\} z^2\right] dz}_{\frac{\sqrt{\pi}}{2}} +$$

$$+\frac{1}{\sigma^4}(2\pi\sigma^2)^{-1}\sigma \underbrace{\int \left[\exp\left\{-z^2\right\} z^4\right] dz}_{\frac{3}{4}\sqrt{\pi}} =$$

$$= (2\pi)^{-1}\sigma^{-5}\sqrt{\pi} - \frac{2}{\sigma^5}(2\pi)^{-1}\frac{\sqrt{\pi}}{2} + \frac{1}{\sigma^5}(2\pi)^{-1}\frac{3}{4}\sqrt{\pi} =$$

$$\frac{1}{2\sqrt{\pi}\sigma^5} - \frac{1}{2\sqrt{\pi}\sigma^5} + \frac{3}{8\sqrt{\pi}\sigma^5} = \frac{3}{8\sqrt{\pi}\sigma^5}$$

Por otra parte:

$$R(K) = \int (K(z))^2 dz = \int \left((2\pi)^{-1/2} e^{-\frac{1}{2}z^2}\right)^2 dz = \int (2\pi)^{-1} e^{-z^2} dz = (2\pi)^{-1} \underbrace{\int e^{-z^2} dz}_{\sqrt{\pi}} = \frac{1}{2\sqrt{\pi}}$$

Mientras que

$$\mu_2^2(K) = \int z^2 K(z) dz = \int z^2 (2\pi)^{-1/2} e^{1\frac{1}{2}z^2} dz = E(Z^2) = V(Z) + E^2(Z) = 1$$

dado que $Z \sim N(0,1)$

Por lo tanto:

$$h^* = \left(\frac{\frac{1}{2\sqrt{\pi}}}{\frac{3}{8\sqrt{\pi}\sigma^5}}\right)^{1/5} n^{-1/5} = \left(\frac{8\sqrt{\pi}}{6\sqrt{\pi}}\right)^{1/5} \sigma n^{-1/5} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} \approx 1.06\, \sigma\, n^{-1/5}$$

Ejercicio 5

```
no_heart <- c(
   195, 348, 237, 174, 205, 158, 201, 171, 190, 085, 180, 082, 193, 210, 170,
   090, 150, 167, 200, 154, 228, 119, 169, 086, 178, 166, 251, 211, 234, 143,
   222, 284, 116, 087, 157, 134, 194, 121, 130, 064, 206, 099, 158, 087, 167,
   177, 217, 114, 234, 116, 190, 132, 178, 157, 265, 073, 219, 098, 266, 486,
   190, 108, 156, 126, 187, 109, 149, 146, 147, 095, 155, 048, 207, 195, 238,
   172, 168, 071, 210, 091, 208, 139, 160, 116, 243, 101, 209, 097, 221, 156,
   178, 116, 289, 120, 201, 072, 168, 100, 162, 227, 207, 160)
no_heart <- as_tibble(matrix(no_heart, ncol=2, byrow=TRUE)) %>%
   rename(pl_chol = V1, pl_trig = V2) %>%
```

```
   mutate(condition = "No heart")
narrow <- c(
   184, 145, 263, 142, 185, 115, 271, 128, 173, 056, 230, 304, 222, 151, 215,
   168, 233, 340, 212, 171, 221, 140, 239, 097, 168, 131, 231, 145, 221, 432,
   131, 137, 211, 124, 232, 258, 313, 256, 240, 221, 176, 166, 210, 092, 251,
   189, 175, 148, 185, 256, 184, 222, 198, 149, 198, 333, 208, 112, 284, 245,
   231, 181, 171, 165, 258, 210, 164, 076, 230, 492, 197, 087, 216, 112, 230,
   090, 265, 156, 197, 158, 230, 146, 233, 142, 250, 118, 243, 050, 175, 489,
   200, 068, 240, 196, 185, 116, 213, 130, 180, 080, 208, 220, 386, 162, 236,
   152, 230, 162, 188, 220, 200, 101, 212, 130, 193, 188, 230, 158, 169, 112,
   181, 104, 189, 084, 180, 202, 297, 232, 232, 328, 150, 426, 239, 154, 178,
   100, 242, 144, 323, 196, 168, 208, 197, 291, 417, 198, 172, 140, 240, 441,
   191, 115, 217, 327, 208, 262, 220, 075, 191, 115, 119, 084, 171, 170, 179,
   126, 208, 149, 180, 102, 254, 153, 191, 136, 176, 217, 283, 424, 253, 222,
   220, 172, 268, 154, 248, 312, 245, 120, 171, 108, 239, 092, 196, 141, 247,
   137, 219, 454, 159, 125, 200, 152, 233, 127, 232, 131, 189, 135, 237, 400,
   319, 418, 171, 078, 194, 183, 244, 108, 236, 148, 260, 144, 254, 170, 250,
   161, 196, 130, 298, 143, 306, 408, 175, 153, 251, 117, 256, 271, 285, 930,
   184, 255, 228, 142, 171, 120, 229, 242, 195, 137, 214, 223, 221, 268, 204,
   150, 276, 199, 165, 121, 211, 091, 264, 259, 245, 446, 227, 146, 197, 265,
   196, 103, 193, 170, 211, 122, 185, 120, 157, 059, 224, 124, 209, 082, 223,
   080, 278, 152, 251, 152, 140, 164, 197, 101, 172, 106, 174, 117, 192, 101,
   221, 179, 283, 199, 178, 109, 185, 168, 181, 119, 191, 233, 185, 130, 206,
   133, 210, 217, 226, 072, 219, 267, 215, 325, 228, 130, 245, 257, 186, 273,
   242, 085, 201, 297, 239, 137, 179, 126, 218, 123, 279, 317, 234, 135, 264,
   269, 237, 088, 162, 091, 245, 166, 191, 090, 207, 316, 248, 142, 139, 173,
   246, 087, 247, 091, 193, 290, 332, 250, 194, 116, 195, 363, 243, 112, 271,
   089, 197, 347, 242, 179, 175, 246, 138, 091, 244, 177, 206, 201, 191, 149,
   223, 154, 172, 207, 190, 120, 144, 125, 194, 125, 105, 036, 201, 092, 193,
   259, 262, 088, 211, 304, 178, 084, 331, 134, 235, 144, 267, 199, 227, 202,
   243, 126, 261, 174, 185, 100, 171, 090, 222, 229, 231, 161, 258, 328, 211,
   306, 249, 256, 209, 089, 177, 133, 165, 151, 299, 093, 274, 323, 219, 163,
   233, 101, 220, 153, 348, 154, 194, 400, 230, 137, 250, 160, 173, 300, 260,
   127, 258, 151, 131, 061, 168, 091, 208, 077, 287, 209, 308, 260, 227, 172,
   168, 126, 178, 101, 164, 080, 151, 073, 165, 155, 249, 146, 258, 145, 194,
   196, 140, 099, 187, 390, 171, 135, 221, 156, 294, 135, 167, 080, 208, 201,
   208, 148, 185, 231, 159, 082, 222, 108, 266, 164, 217, 227, 249, 200, 218,
   207, 245, 322, 242, 180, 262, 169, 169, 158, 204, 084, 184, 182, 206, 148,
   198, 124, 242, 248, 189, 176, 260, 098, 199, 153, 207, 150, 206, 107, 210,
   095, 229, 296, 232, 583, 267, 192, 228, 149, 187, 115, 304, 149, 140, 102,
   209, 376, 198, 105, 270, 110, 188, 148, 160, 125, 218, 096, 257, 402, 259,
   240, 139, 054, 213, 261, 178, 125, 172, 146, 198, 103, 222, 348, 238, 156,
   273, 146, 131, 096, 233, 141, 269, 084, 170, 284, 149, 237, 194, 272, 142,
   111, 218, 567, 194, 278, 252, 233, 184, 184, 203, 170, 239, 038, 232, 161,
```

```
   225, 240, 280, 218, 185, 110, 163, 156, 216, 101)
narrow <- as_tibble(matrix(narrow, ncol=2, byrow=TRUE)) %>%
   rename(pl_chol = V1, pl_trig = V2) %>%
   mutate(condition = "Narrow")
datos <- dplyr::bind_rows(no_heart, narrow)
```

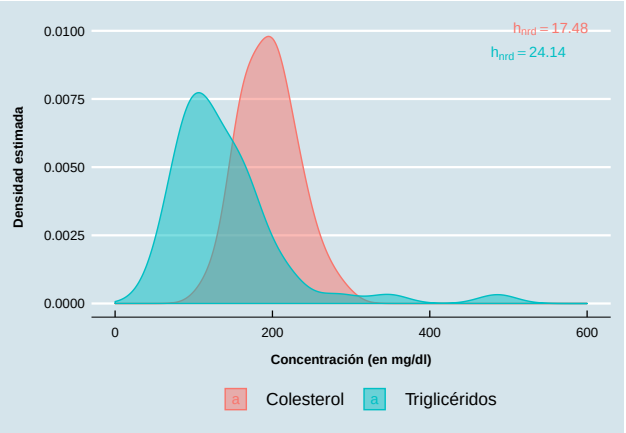
```
# Condition: no evidence of heart disease
metodos <- c("nrd", "nrd0", "ucv", "bcv", "SJ")
variables <- names(datos)[-3]
bws_noheart <- matrix(NA, ncol=length(metodos), nrow=length(variables),
                      dimnames=list(variables, metodos))
for (i in variables) {
   for (j in metodos) {
      bws_noheart[i, j] <- density(as.numeric(as.matrix(no_heart[,i])), bw=j)$bw
   }
}
knitr::kable(bws_noheart, digits=2,
             col.names=c("Scott", "Silverman", "Unbiased CV", "Biased CV",
                         "Sheather Jones"),
             caption="Bandwidth estimation using different methods for
             plasma triglycerids and plasma cholesterol, males with no
             evidence of heart disease")
```

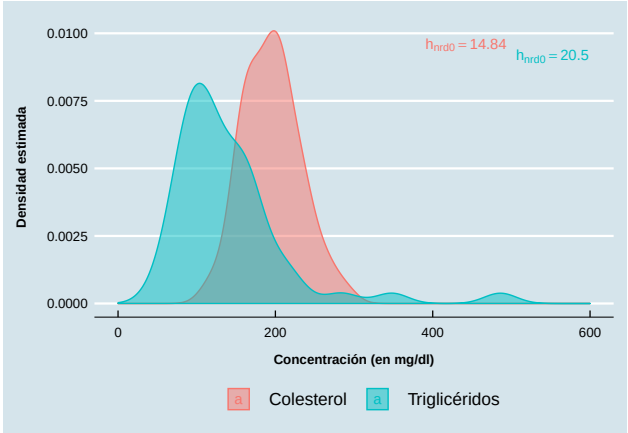
Table 1: Bandwidth estimation using different methods for plasma triglycerids and plasma cholesterol, males with no evidence of heart disease

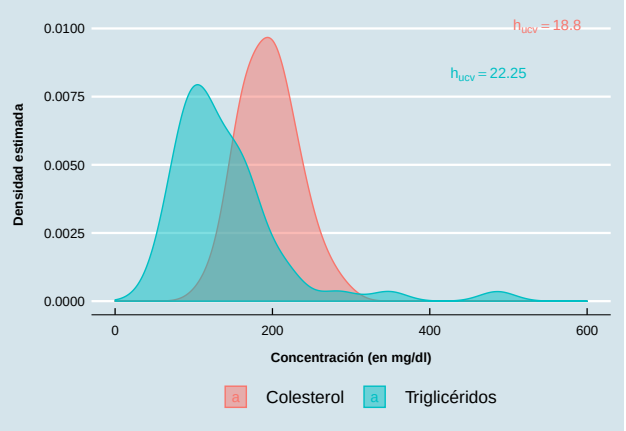
	Scott	Silverman	Unbiased CV	Biased CV	Sheather Jones
pl_chol	17.48	14.84	18.80	18.79	19.11
pl_trig	24.14	20.50	22.25	29.01	19.47

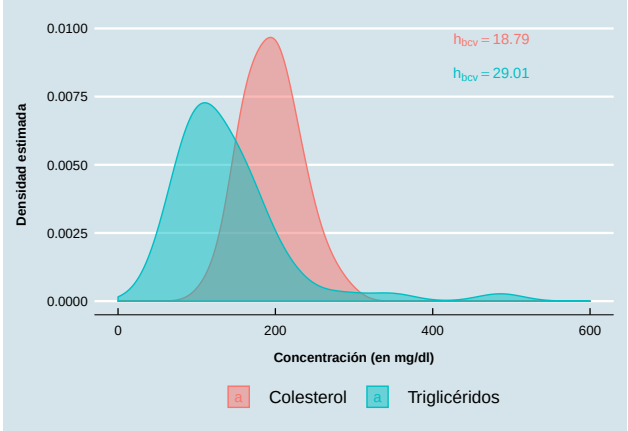
```
# Generamos las etiquetas para los gráficos
eti <- as_tibble(bws_noheart) %>%
   mutate(variable = c("Colesterol", "Triglicéridos")) %>%
   gather(value="h", key="metodo", -variable) %>%
   mutate(etiqueta = paste0("h[", metodo, "] == ", round(h, 2)))
make_my_plots <- function(m){
   no_heart %>%
   dplyr::select(-condition) %>%
   gather(key="variable", value=value) %>%
   mutate(variable = if_else(variable == "pl_chol", "Colesterol", "Triglicéridos")) %>%
   ggplot() +
   geom_density(aes(value, fill = variable, color = variable),
                kernel = "gaussian", bw = m, alpha = 0.5) +
   geom_text_repel(data=filter(eti, metodo == m),
                   aes(x=c(500,500), y=c(0.01, 0.00875), label=etiqueta, color=variable),
                   parse=TRUE) +
   xlim(c(0,600)) +
   labs(x="\nConcentración (en mg/dl)", y="Densidad estimada\n") +
```

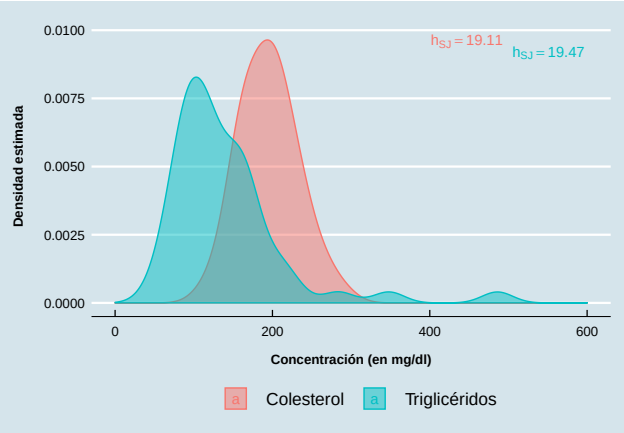
```
   ggthemes::theme_economist() +
   theme(legend.position="bottom",
         legend.title=element_blank(),
         legend.spacing.x=unit(0.5, "cm"),
         axis.title=element_text(face="bold"))
}
plot_list <- lapply(metodos, make_my_plots)
for (i in 1:length(plot_list)) {
   print(plot_list[[i]])
}
```











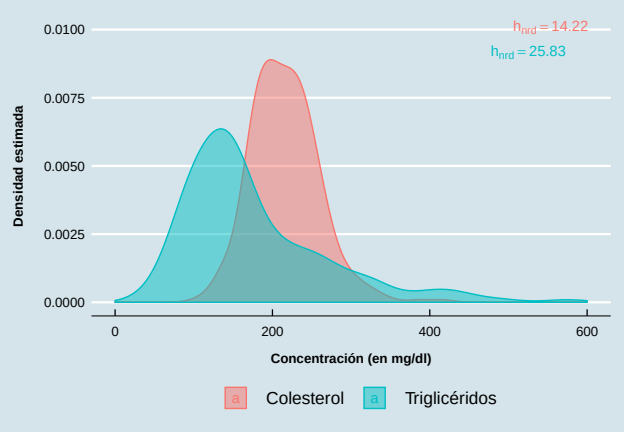

```
# Condition: narrowing of the arteries
metodos <- c("nrd", "nrd0", "ucv", "bcv", "SJ")
variables <- names(datos)[-3]
bws_narrow <- matrix(NA, ncol=length(metodos), nrow=length(variables),
                     dimnames=list(variables, metodos))
for (i in variables) {
   for (j in metodos) {
      bws_narrow[i, j] <- density(as.numeric(as.matrix(narrow[,i])), bw=j)$bw
   }
}
knitr::kable(bws_narrow, digits=2,
             col.names=c("Scott", "Silverman", "Unbiased CV", "Biased CV",
                         "Sheather Jones"),
             caption="Bandwidth estimation using different methods for
             plasma triglycerids and plasma cholesterol, males with
             narrowing of arteries")
```

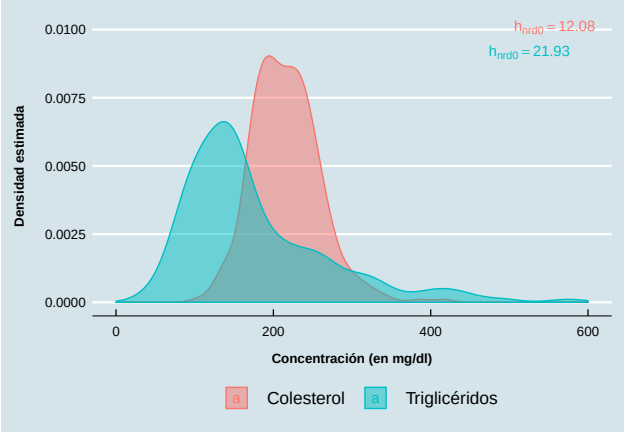
Table 2: Bandwidth estimation using different methods for plasma triglycerids and plasma cholesterol, males with narrowing of arteries

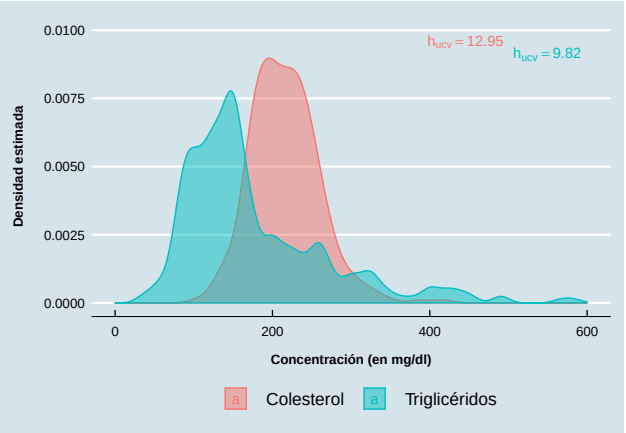
	Scott	Silverman	Unbiased CV	Biased CV	Sheather Jones
pl_chol	14.22	12.08	12.95	14.62	12.06
pl_trig	25.83	21.93	9.82	20.06	14.66

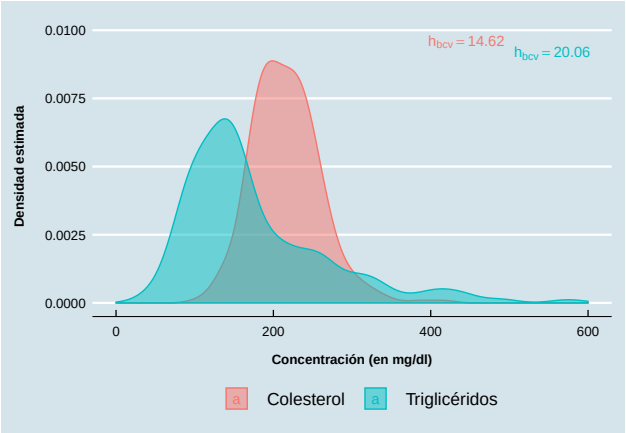
```
# Generamos las etiquetas para los gráficos
eti <- as_tibble(bws_narrow) %>%
   mutate(variable = c("Colesterol", "Triglicéridos")) %>%
   gather(value="h", key="metodo", -variable) %>%
   mutate(etiqueta = paste0("h[", metodo, "] == ", round(h, 2)))
```

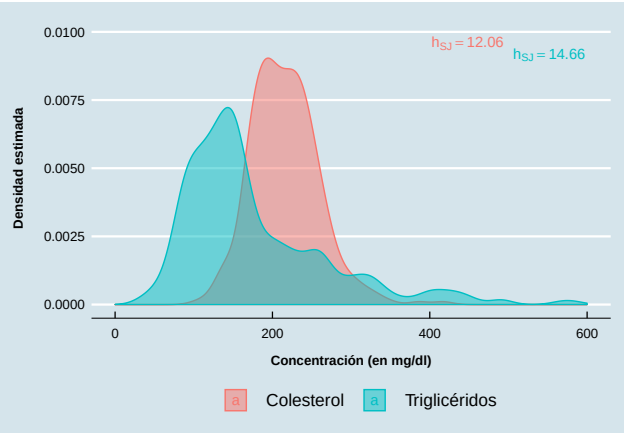
```
make_my_plots <- function(m){
   narrow %>%
   dplyr::select(-condition) %>%
   gather(key="variable", value=value) %>%
   mutate(variable = if_else(variable == "pl_chol", "Colesterol", "Triglicéridos")) %>%
   ggplot() +
   geom_density(aes(value, fill = variable, color = variable),
                kernel = "gaussian", bw = m, alpha = 0.5) +
   geom_text_repel(data=filter(eti, metodo == m),
                   aes(x=c(500L,500L), y=c(0.01, 0.00875), label=etiqueta, color=variabl
                   parse=TRUE) +
   xlim(c(0,600)) +
   labs(x="\nConcentración (en mg/dl)", y="Densidad estimada\n") +
   ggthemes::theme_economist() +
   theme(legend.position="bottom",
         legend.title=element_blank(),
         legend.spacing.x=unit(0.5, "cm"),
         axis.title=element_text(face="bold"))
}
plot_list <- lapply(metodos, make_my_plots)
for (i in 1:length(plot_list)) {
   print(plot_list[[i]])
}
```









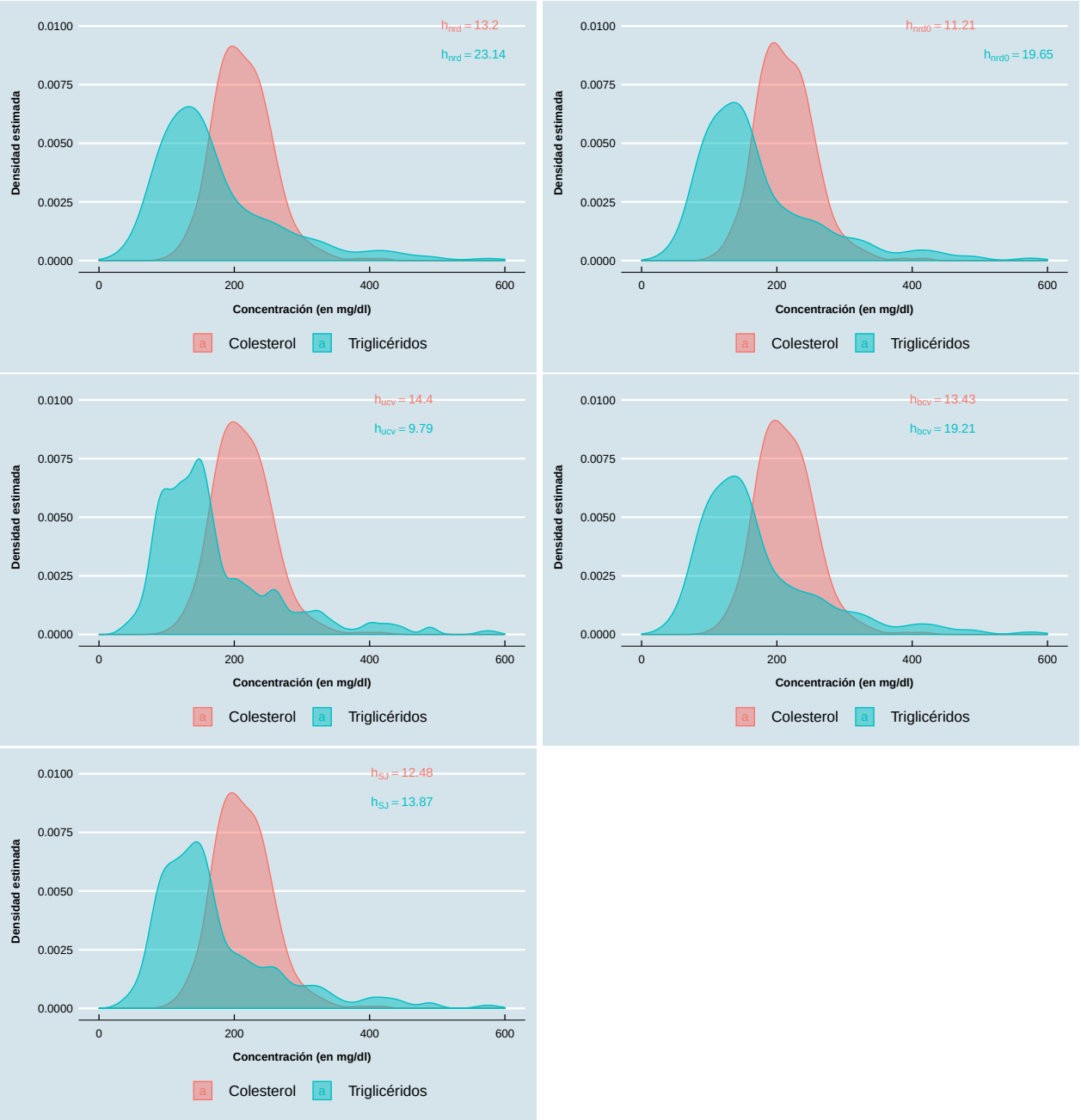



```
# Condition: both (complete sample)
metodos <- c("nrd", "nrd0", "ucv", "bcv", "SJ")
variables <- names(datos)[-3]
bws <- matrix(NA, ncol=length(metodos), nrow=length(variables),
              dimnames=list(variables, metodos))
for (i in variables) {
   for (j in metodos) {
      bws[i, j] <- density(as.numeric(as.matrix(datos[,i])), bw=j)$bw
   }
}
knitr::kable(bws, digits=2,
             col.names=c("Scott", "Silverman", "Unbiased CV", "Biased CV",
                         "Sheather Jones"),
             caption="Bandwidth estimation using different methods for
             plasma triglycerids and plasma cholesterol, complete sample")
```

Table 3: Bandwidth estimation using different methods for plasma triglycerids and plasma cholesterol, complete sample

	Scott	Silverman	Unbiased CV	Biased CV	Sheather Jones
pl_chol	13.20	11.21	14.40	13.43	12.48
pl_trig	23.14	19.65	9.79	19.21	13.87

```
# Generamos las etiquetas para los gráficos
eti <- as_tibble(bws) %>%
   mutate(variable = c("Colesterol", "Triglicéridos")) %>%
   gather(value="h", key="metodo", -variable) %>%
   mutate(etiqueta = paste0("h[", metodo, "] == ", round(h, 2)))
make_my_plots <- function(m){
   datos %>%
   dplyr::select(-condition) %>%
   gather(key="variable", value=value) %>%
   mutate(variable = if_else(variable == "pl_chol", "Colesterol", "Triglicéridos")) %>%
   ggplot() +
   geom_density(aes(value, fill = variable, color = variable),
                kernel = "gaussian", bw = m, alpha = 0.5) +
   geom_text_repel(data=filter(eti, metodo == m),
                   aes(x=c(500, 500), y=c(0.01, 0.00875), label=etiqueta, color=variable
                   parse=TRUE, direction="x") +
   xlim(c(0,600)) +
   labs(x="\nConcentración (en mg/dl)", y="Densidad estimada\n") +
   ggthemes::theme_economist() +
   theme(legend.position="bottom",
         legend.title=element_blank(),
         legend.spacing.x=unit(0.5, "cm"),
         axis.title=element_text(face="bold"))
}
plot_list <- lapply(metodos, make_my_plots)
for (i in 1:length(plot_list)) {
   print(plot_list[[i]])
}
```



References

Loader, Catherine. 2013. *Locfit: Local Regression, Likelihood and Density Estimation.* https://CRAN.R-project.org/package=locfit.

Park, Byeong U, and James S Marron. 1990. "Comparison of Data-Driven Bandwidth Selectors." *Journal of the American Statistical Association* 85 (409). Taylor & Francis

Group: 66–72.

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

Scott, David W. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Sheather, Simon J, and Michael C Jones. 1991. "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation." *Journal of the Royal Statistical Society: Series B (Methodological)* 53 (3). Wiley Online Library: 683–90.

Wasserman, Larry. 2007. *All of Nonparametric Statistics*. Springer, New York.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.