STAT 210
Applied Statistics and Data Analysis
Density Estimation

Joaquin Ortega

Fall 2020

Introduction[1]

[1] Partly based on **Notes for Nonparametric Statistics**, Eduardo García, https://bookdown.org/egarpor/NP-UC3M/

Introduction

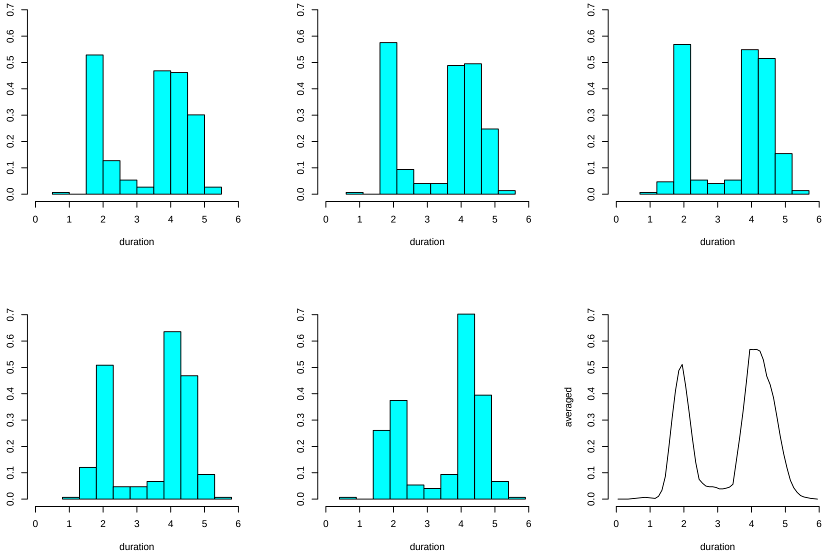
Histograms can be unreliable as estimators of a density.

They depend on two parameters, the bandwidth and the anchor.

These are frequently chosen arbitrarily and can produce very different graphs.

In the next slide, we give an example using the Old Faithful data, from the book by Venables and Ripley.

Introduction



Introduction

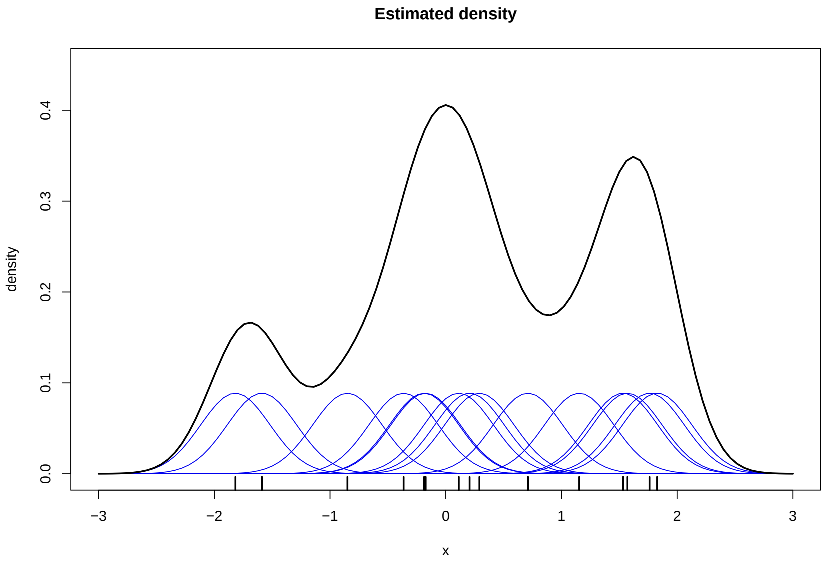
We will consider two instances of nonparametric function estimation:

- Density estimation
- Nonparametric regression

The idea for density estimation was proposed by M. Rosemblat and E. Parzen in 1956.

Instead of considering each datum as a point, they proposed using smooth functions centered on the datum, obtained from a kernel.

Introduction



Estimation kernels

Estimation kernels

At each sample point x_i, we place a function $K_h(x - x_i)$ where K is an **estimation kernel**.

To estimate the value of the density at each point we average the values of these functions at that point:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

where

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

and the estimator of the density becomes

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

Estimation kernels

Estimation kernels K have the following characteristics

- kernels are non-negative,
- symmetric with respect to the origin and
- integrate to 1.

Since estimation kernels are probability densities, kernel estimators are also densities: It is immediate that $\hat{f}_h(x)$ is nonnegative and

$$\begin{aligned}\int \hat{f}_h(x)\,dx &= \frac{1}{nh}\sum_{i=1}^{n}\int K\Big(\frac{x-x_i}{h}\Big)\,dx\\ &= \frac{1}{nh}\sum_{i=1}^{n}\int hK(x)\,dx = 1\end{aligned}$$

Estimation kernels

The estimated density inherits the regularity of the kernel: If K is d times differentiable, $\hat{f}_h(x)$ is also d times differentiable

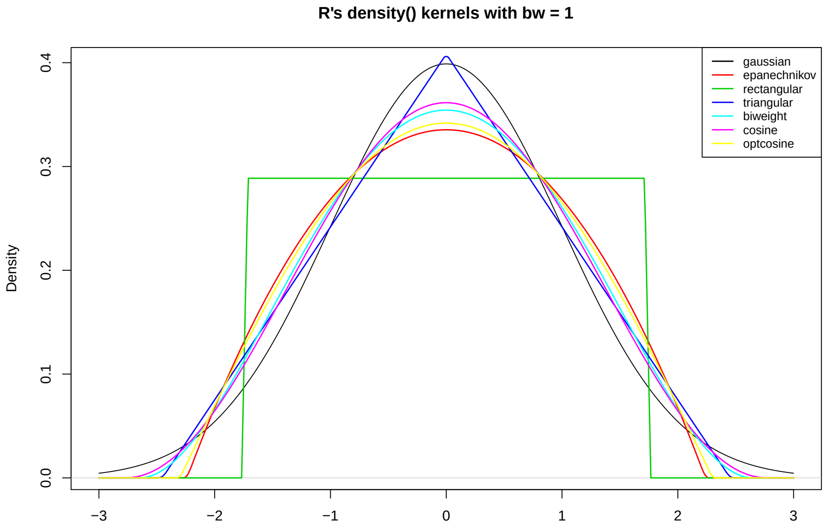
Kernel density estimators depend only on the **bandwidth** h (and the choice of the kernel)

A small bandwidth will give a rough estimate, while a large bandwidth will give a smooth estimate.

Estimation kernels

Nombre	$\mathbf{K}(u)$
Uniform	$(1/2)\mathbf{1}(\lvert u\rvert \leq 1)$
Triangular	$(1-\lvert u\rvert)\mathbf{1}(\lvert u\rvert \leq 1)$
Epanechnikov	$(3/4)(1-u^2)\mathbf{1}(\lvert u\rvert \leq 1)$
Biweight	$(15/16)(1-u^2)^2\mathbf{1}(\lvert u\rvert \leq 1)$
Gaussian	$(1/\sqrt{2\pi})\exp(-u^2/2)$
Cosine	$(\pi/4)\cos(\pi u/2)\mathbf{1}(\lvert u\rvert \leq 1)$

Estimation kernels

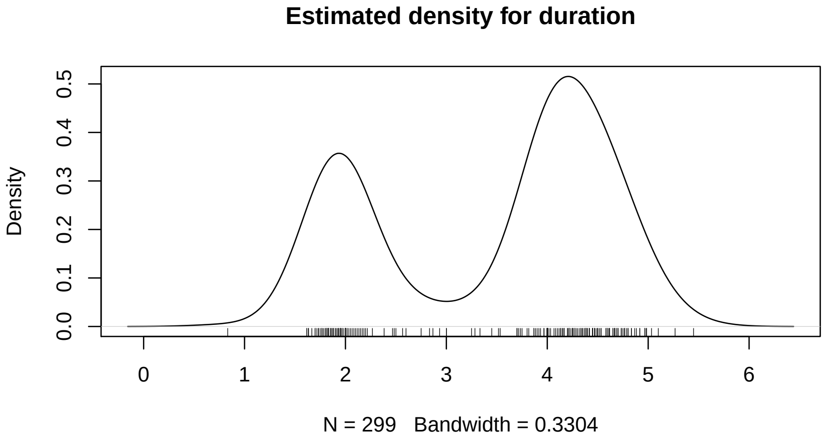


Estimation kernels

R has the function `density` for density estimation.

This function does not produce a graph. It has to be combined with `plot()`.

```
plot(density(duration), main = 'Estimated density for duration')
rug(duration)
```

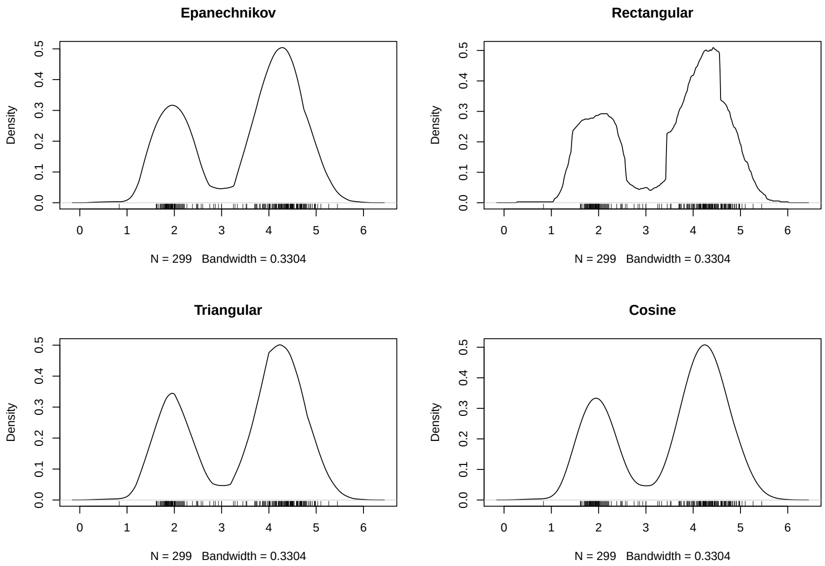


Estimation kernels

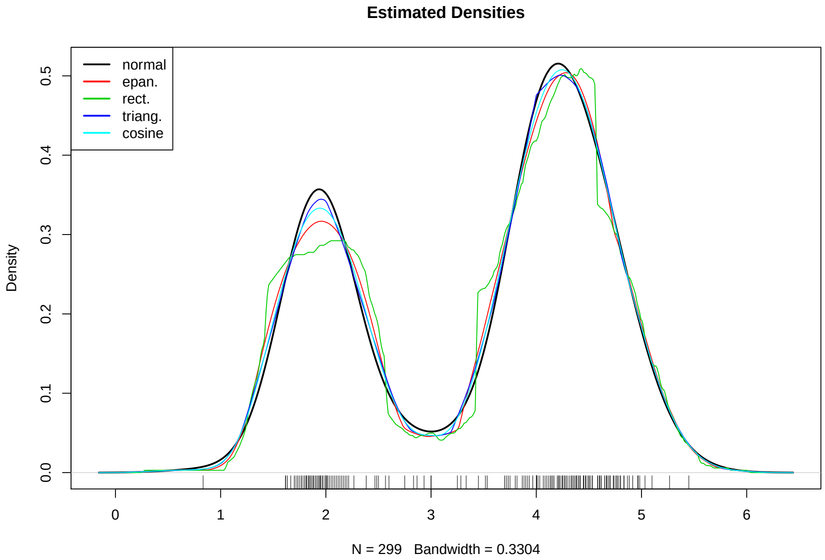
The default kernel is `gaussian`, which is the kernel used in the previous graphic.

Other kernels are used in the next figure.

Estimation kernels



Estimation kernels



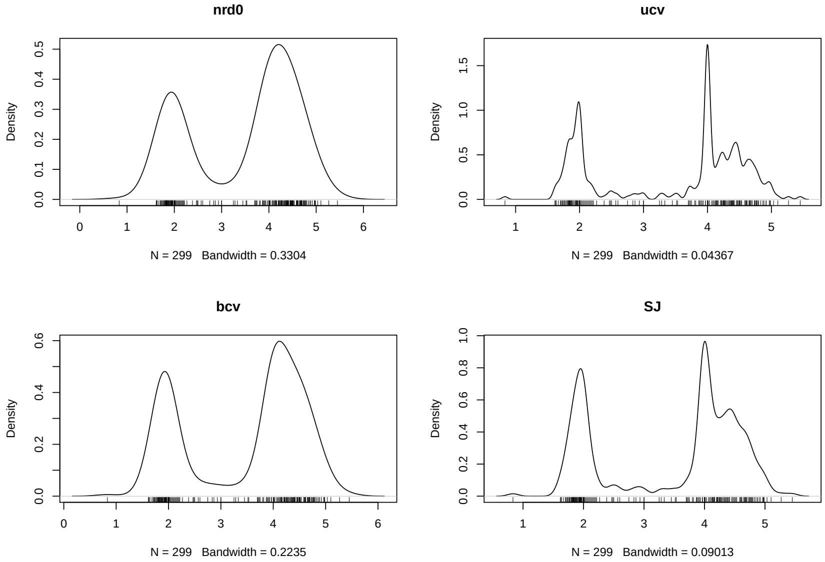
Estimation kernels

There are two parameters to control the bandwidth, `bw` and `adjust`.

`bw` is the bandwidth for the smoothing kernel. 'The kernels are scaled such that this is the standard deviation of the smoothing kernel'. Not necessarily a good choice.

Its value can be a number or a method for calculating the bandwidth (see the help for `density`).

Estimation kernels

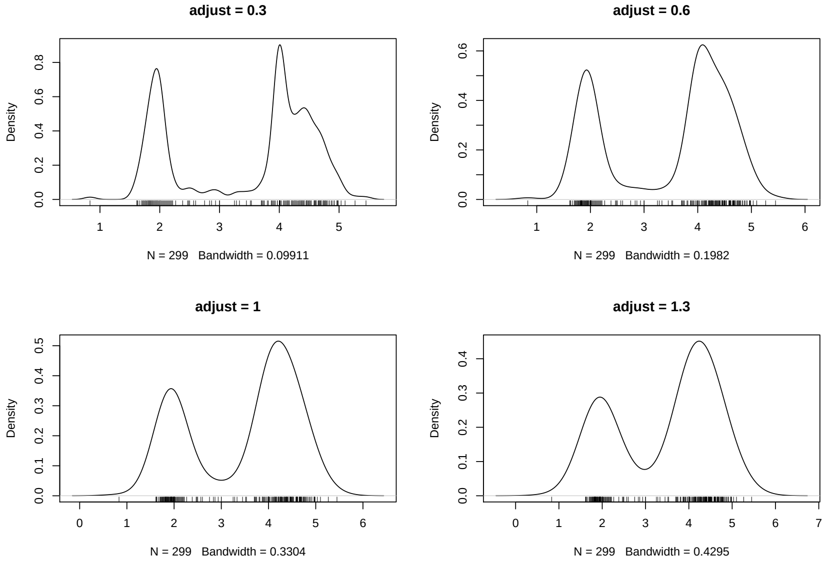


Estimation kernels

The bandwidth can also be controlled using the `adjust` parameter, which by default is set to 1.

This parameter is multiplied by the calculated bandwidth and is easier to handle than the bandwidth.

Estimation kernels



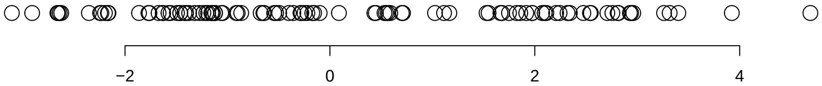
Estimation kernels

We now look at a simulated example. We consider a population with distribution given by a mixture of normal densities. 60% of the population come from a $N(-1, 1)$ distribution, while the remaining 40% come from a $N(2, 1)$.

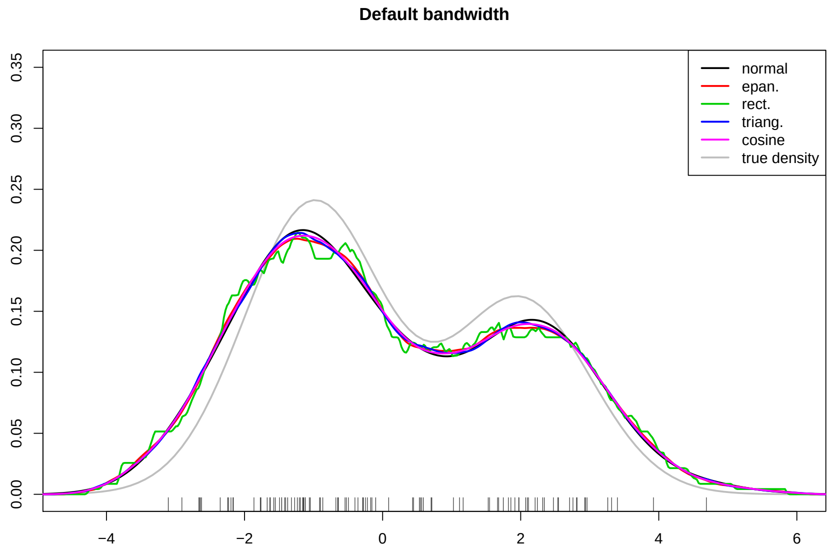
We draw a sample of size 100 and use it to estimate the density for the population distribution.

Estimation kernels

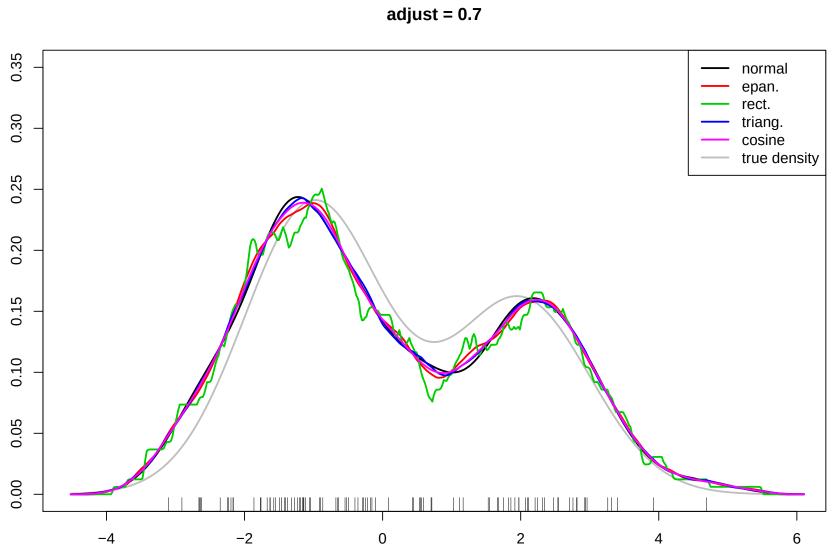
```
set.seed(569921); data.1 <- runif(100) <= 0.6
data.dens <- data.1*rnorm(100,mean=-1)+ (1-data.1)*rnorm(100,mean=2)
plot(data.dens, rep(0,100),ylab='',xlab='', cex = 2,
     axes = FALSE)
axis(1)
```



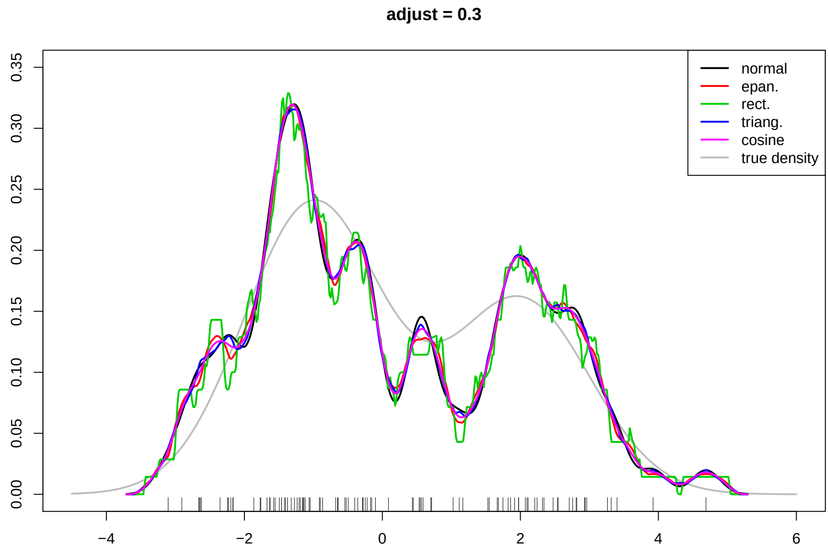
Estimation kernels



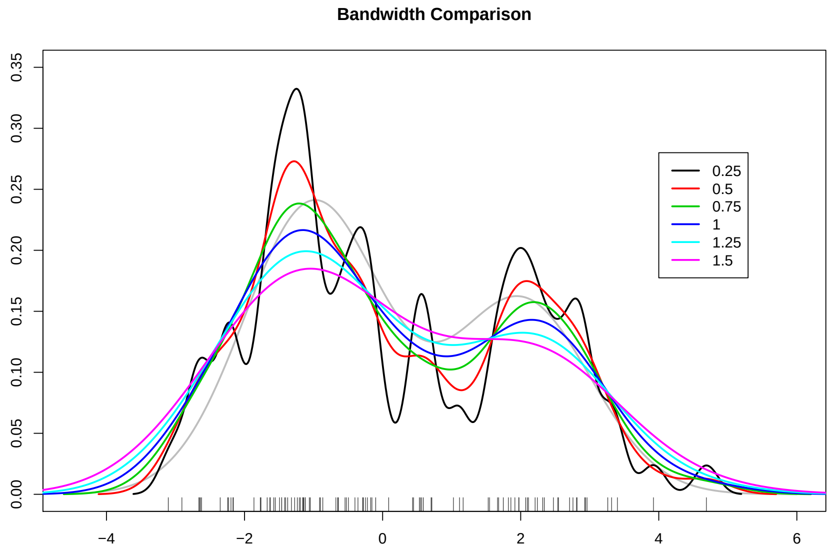
Estimation kernels



Estimation kernels



Estimation kernels



Estimation kernels

https://bookdown.org/egarpor/NP-UC3M/kde-i-kde.html

https://ec2-35-177-34-200.eu-west-2.compute.amazonaws.com/kde/

Bandwidth selection

Bandwidth selection

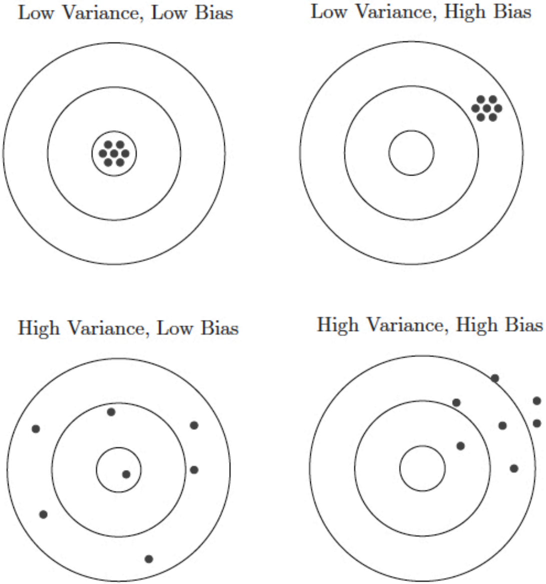
We want to choose a bandwidth that minimizes the difference between the estimated density and the true one.

The mean square error of $\hat{f}(x)$ is a function of x and is given by

$$\begin{aligned} MSE(\hat{f}(x)) &= E[(\hat{f}(x) - f(x))^2] \\ &= E[(\hat{f}(x) \pm E[\hat{f}(x)] - f(x))^2] \\ &= (E[\hat{f}(x)] - f(x))^2 + E[(\hat{f}(x) - E[\hat{f}(x)])^2] \\ &= (\text{Bias}(\hat{f}(x)))^2 + \text{Var}(\hat{f}(x)) \end{aligned}$$

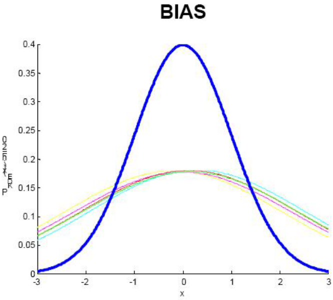
Bias represents systematic error while variance represents the random error present in the estimation process.

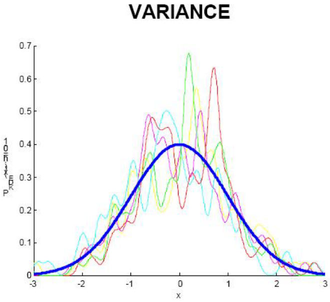
Bandwidth selection



Estimation kernels: bandwidth selection

- If we choose a large bandwidth we have small variance but large bias.
- if we choose a small bandwidth we have small bias but large variance.





Bandwidth selection

A global measure of error can be obtained integrating over all values of x, this is known as the mean integrated square error (MISE)

$$
\begin{aligned}
MISE(\hat{f}) &= E\Big[\int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 \, dx\Big] \\
&= \int_{-\infty}^{\infty} MSE(\hat{f}(x)) \, dx \\
&= \int_{-\infty}^{\infty} (\mathrm{Bias}(\hat{f}(x)))^2 \, dx + \int_{-\infty}^{\infty} \mathrm{Var}(\hat{f}(x)) \, dx
\end{aligned}
$$

and it can be shown under certain mild conditions that

$$\mathrm{Bias}(\hat{f}(x)) = \frac{1}{2} h^2 f''(x) \int z^2 K(z) \, dz + o(h^2)$$

and

$$\mathrm{Var}(\hat{f}(x)) = \frac{R(K)}{nh} f(x) + o((nh)^{-1})$$

Bandwidth selection

Assuming that the true distribution is Gaussian and using a Gaussian kernel, Silverman showed that the (asymptotic) optimal bandwidth is

$$h_{opt} = 1.05 s n^{-1/5}$$

where s is the empirical standard deviation.

Silverman's empirical rule for bandwidth selection:

$$h_{opt} = 0.9 A n^{-1/5}$$

where $A = \min\{s, IQR/1.34\}$.

Bandwidth selection

https://bookdown.org/egarpor/NP-UC3M/kde-i-bwd.html

https://ec2-35-177-34-200.eu-west-2.compute.amazonaws.com/kde-bwd/

The `ks` Package

Estimation kernels: `kde`

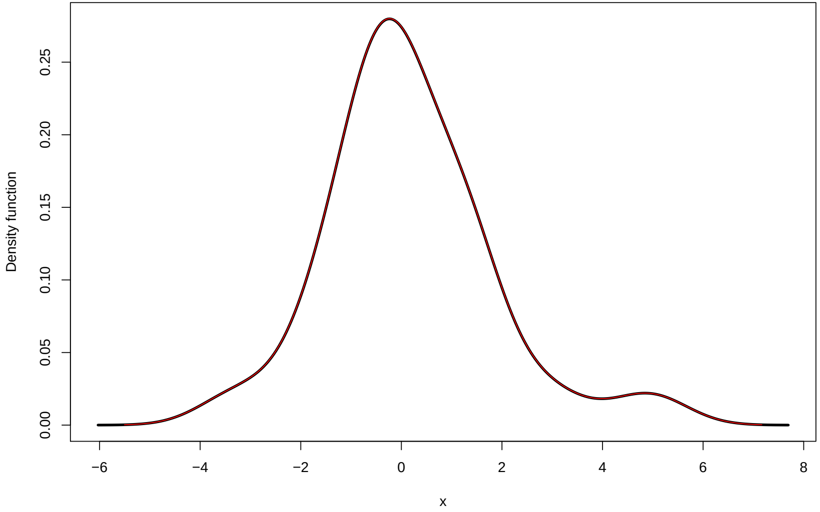
An alternative for density estimation that has certain advantages is the package `ks` that extends to the multivariate case and allows evaluating the estimated density at arbitrary points.

The main function is `kde` that gives the same output as the default configuration for `density`.

The only limitation is that `kde` only considers normal kernels.

Estimation kernels: kde

```
library(ks)
plot(kde <- kde(x = samp_t, h = bw), lwd = 3) # ?ks::plot.kde for options
lines(density(x = samp_t, bw = bw), col = 2)
```



Estimation kernels: kde

Evaluating the estimated density at specific points

```
dens1 <- kde(x = samp_t, h = bw, eval.points = -2:2)
dens1$eval.points
```

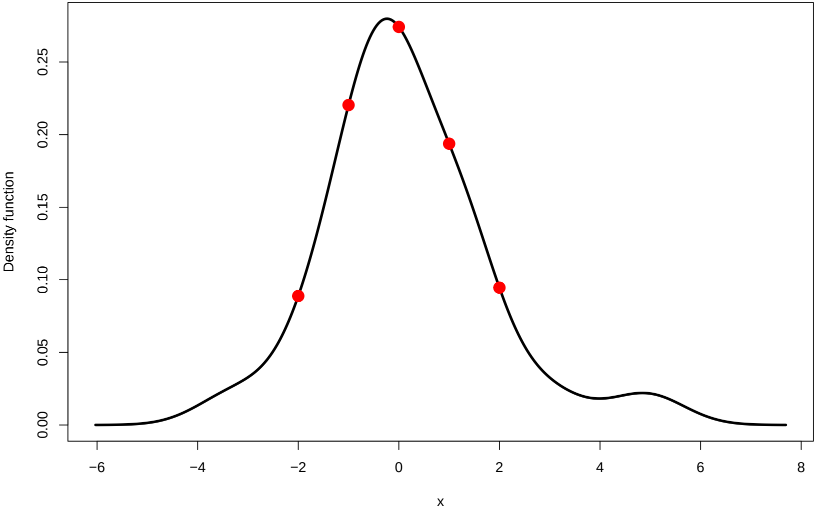
```
## [1] -2 -1  0  1  2
```

```
round(dens1$estimate,4)
```

```
## [1] 0.0888 0.2204 0.2742 0.1937 0.0946
```

Estimation kernels: ‘kde

```
plot(kde <- kde(x = samp_t, h = bw), lwd = 3)
points(dens1$eval.points,dens1$estimate, pch=19, col='red',
```



Estimation kernels: `kde`

Sampling from an estimated density: Use `rkde`.

```
round(rkde(n = 5, fhat = dens1),4)
```

```
## [1] -0.6155  2.5214 -1.3567 -0.2321 -1.0299
```