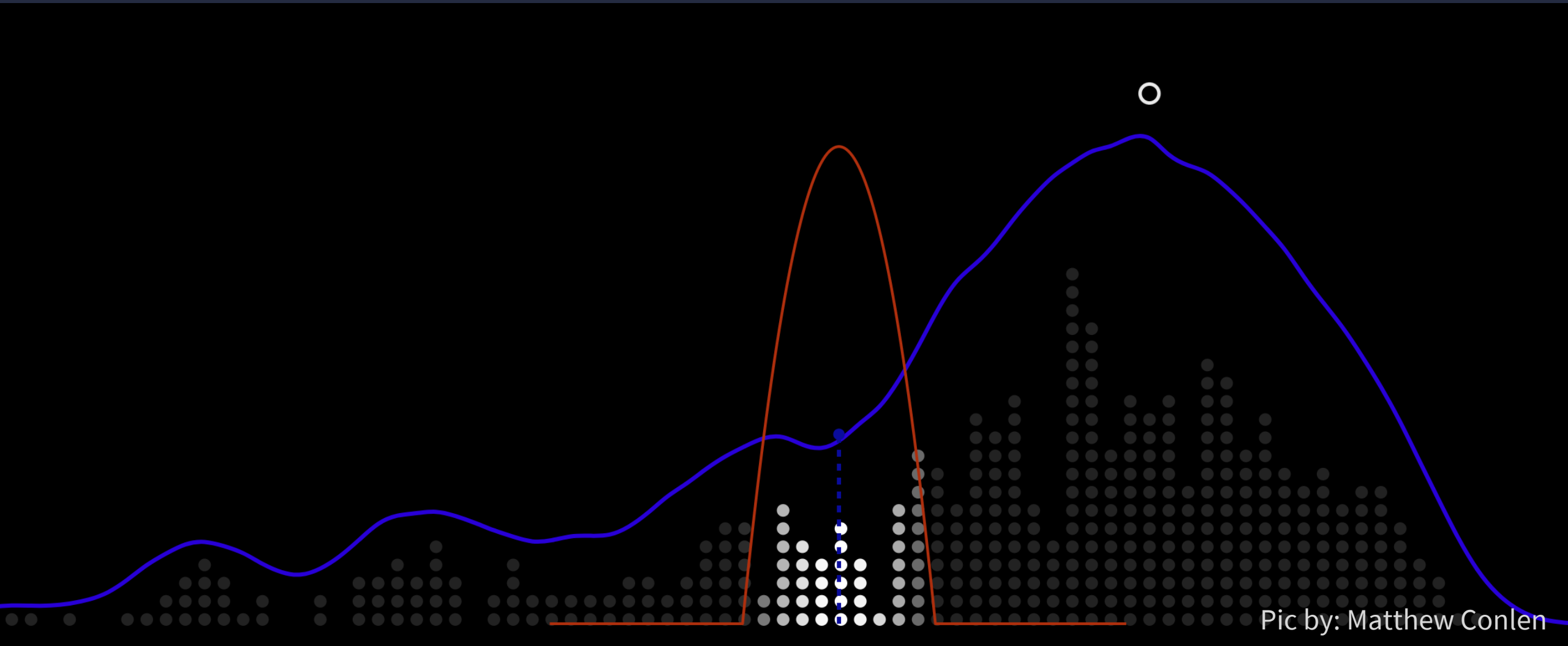
Kernel Density Estimation

https://github.com/ST541-Fall2020/Tingyu_project_KDE

Tingyu Zhu



Pic by: Matthew Conlen

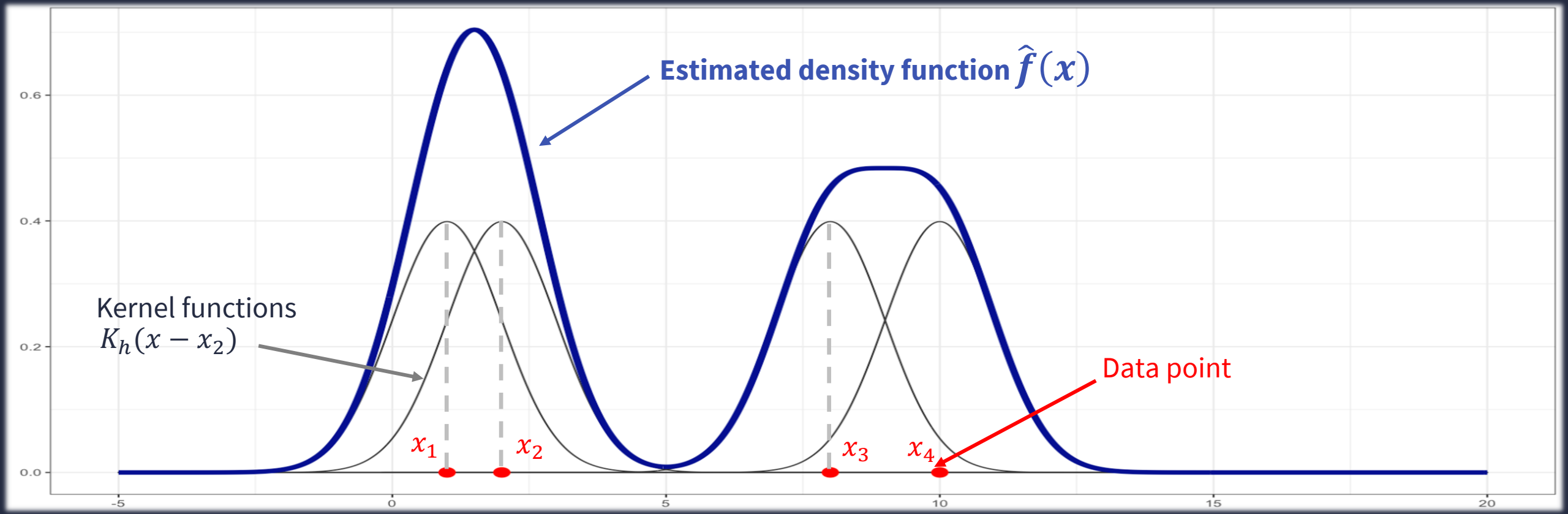
Kernel density estimator

Let $\{x_1, x_2, \dots, x_n\}$ be i.i.d. samples ~ unknown $f(x)$.We want to estimate the shape of $f(x)$.

The kernel density estimator is given by:

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

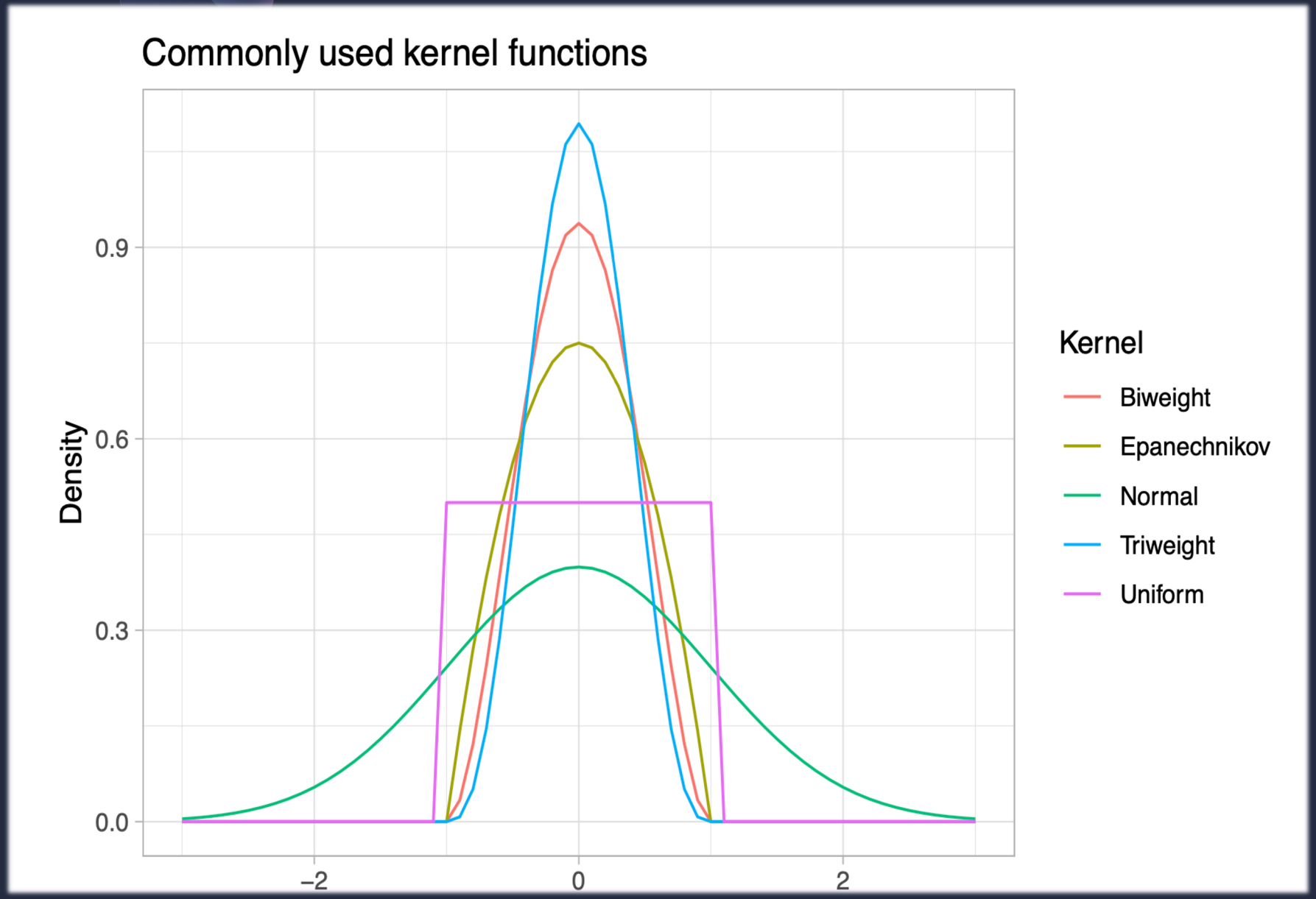
- *K:* the ***kernel*** function, K(.)>0, symmetry
- *h > 0:* ***bandwidth***. *A* smoothing parameter



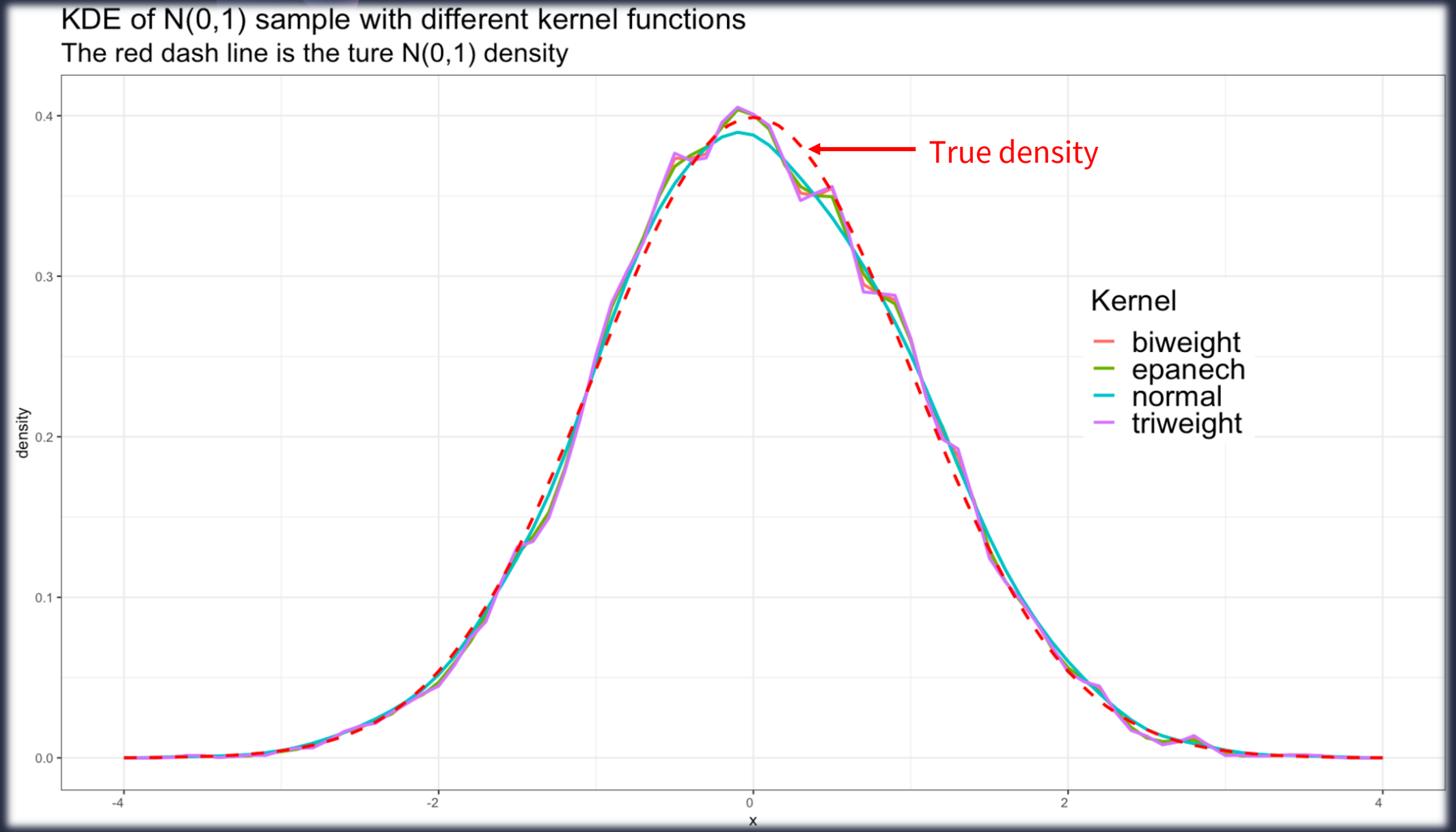
My task:

1. The effect of different choices of **kernel function** and **bandwidth**
2. Apply the KDE to a real dataset

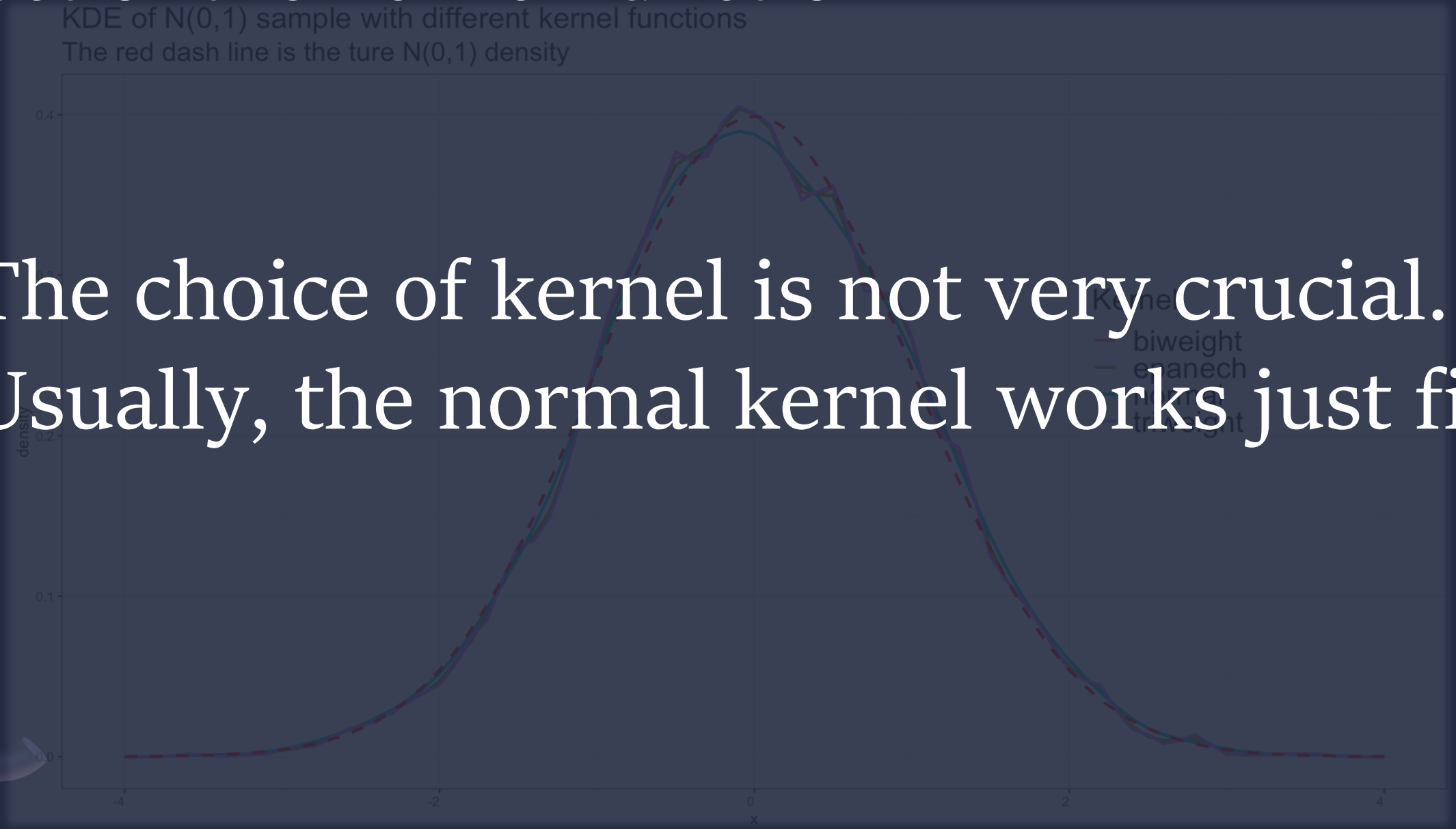
Effect of the kernel function



Effect of the kernel function

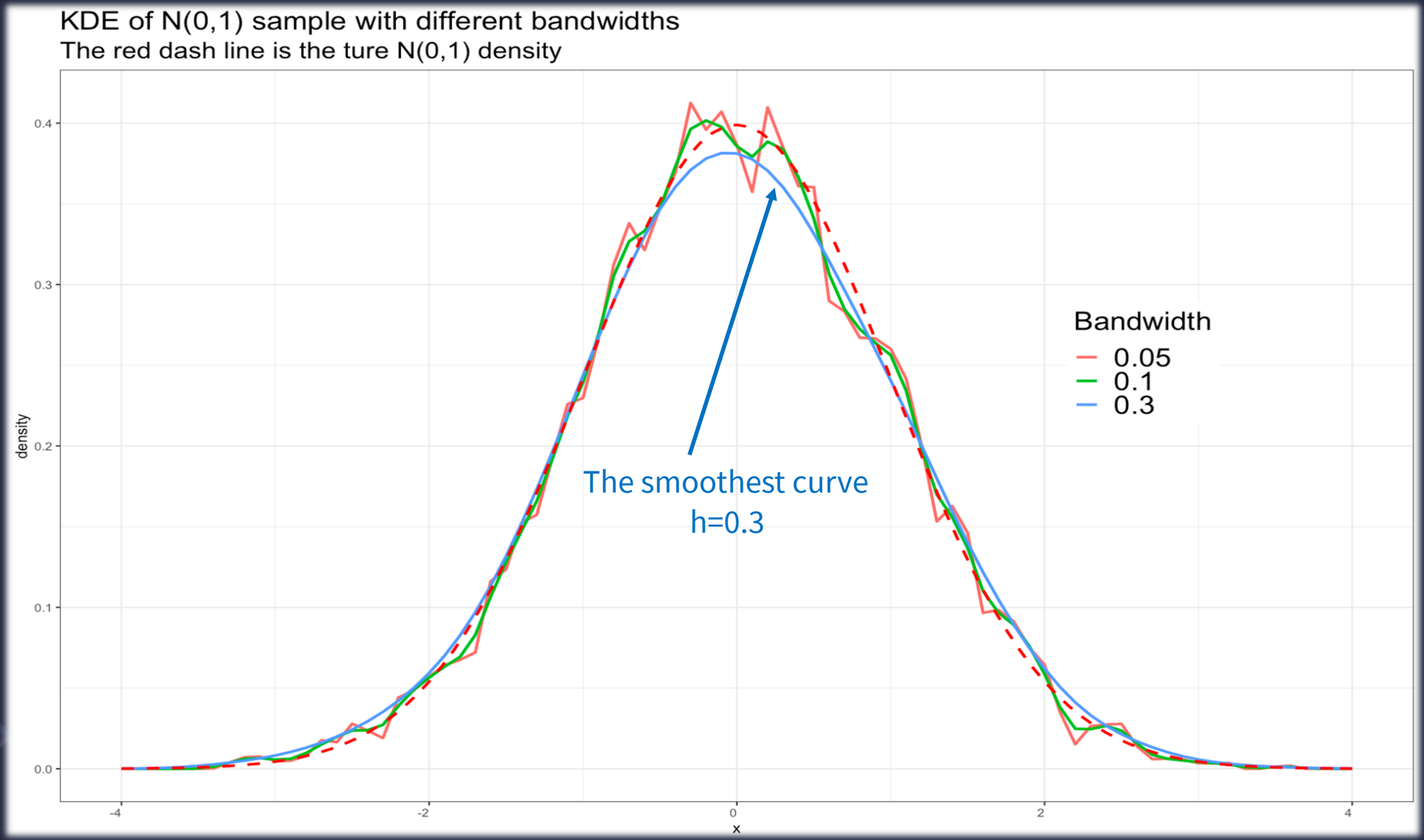


Effect of the kernel function



The choice of kernel is not very crucial.
Usually, the normal kernel works just fine.

Effect of the bandwidth

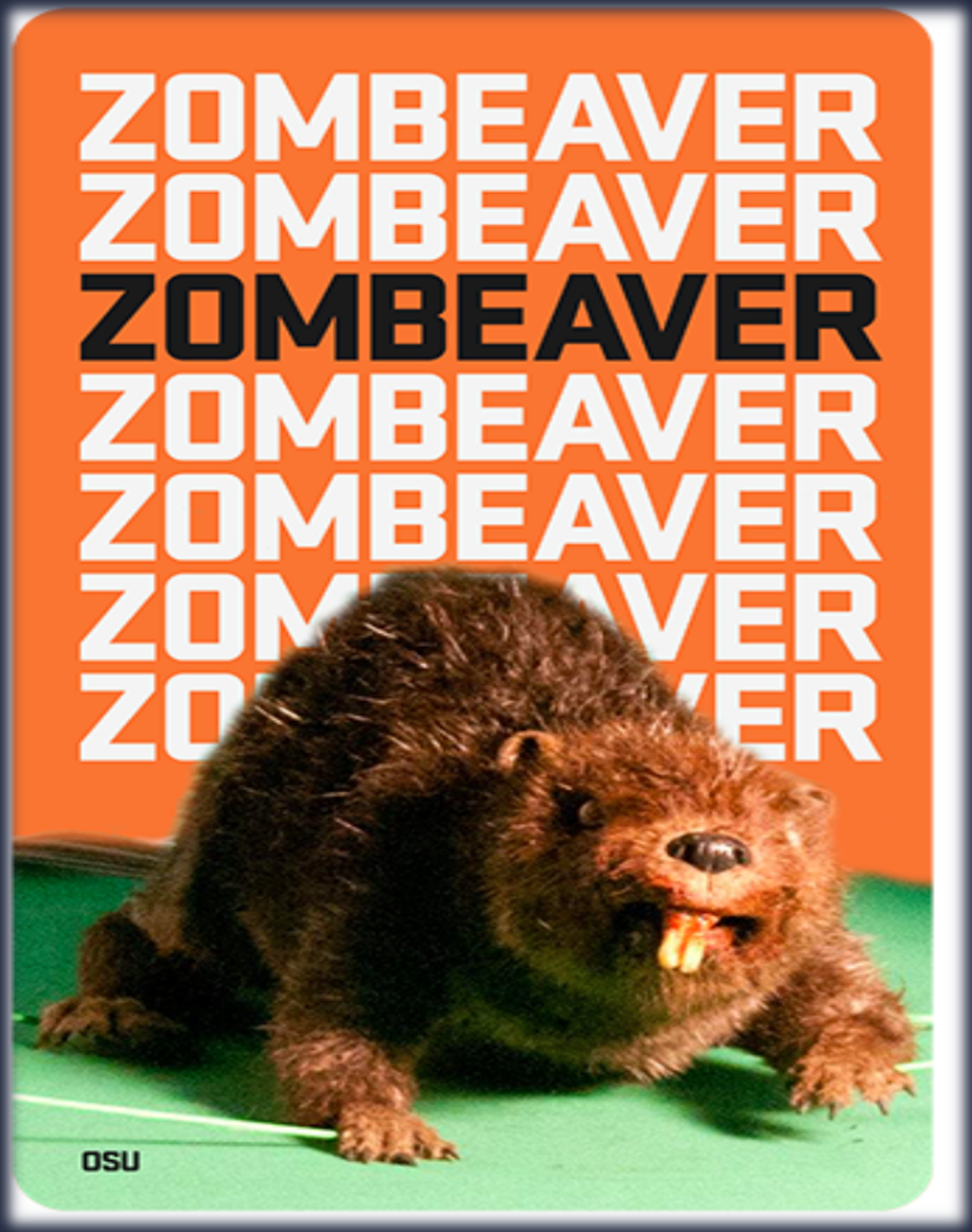


Effect of the bandwidth

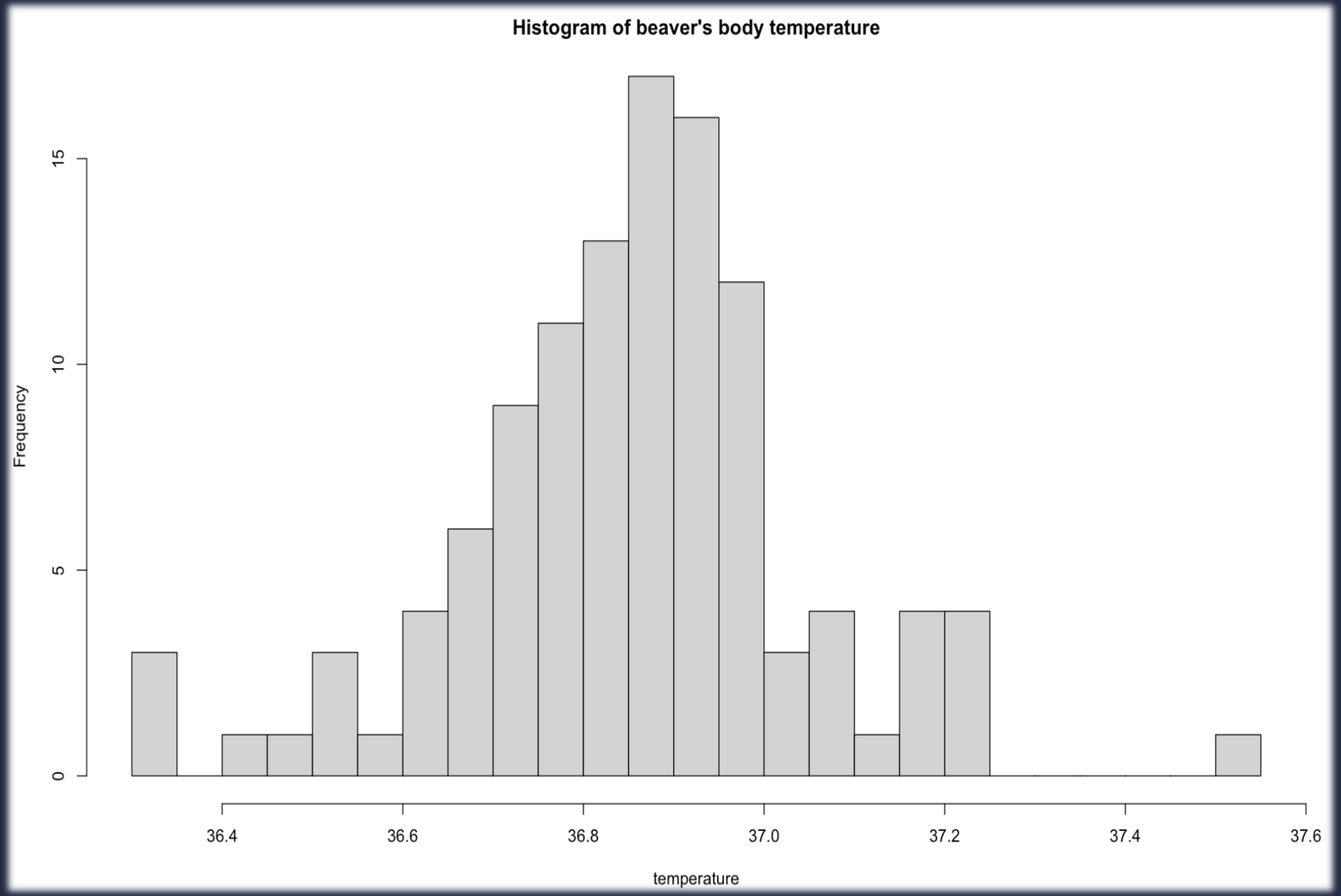
The choice of bandwidth is very important!

Different ways have been developed to select a proper bandwidth.

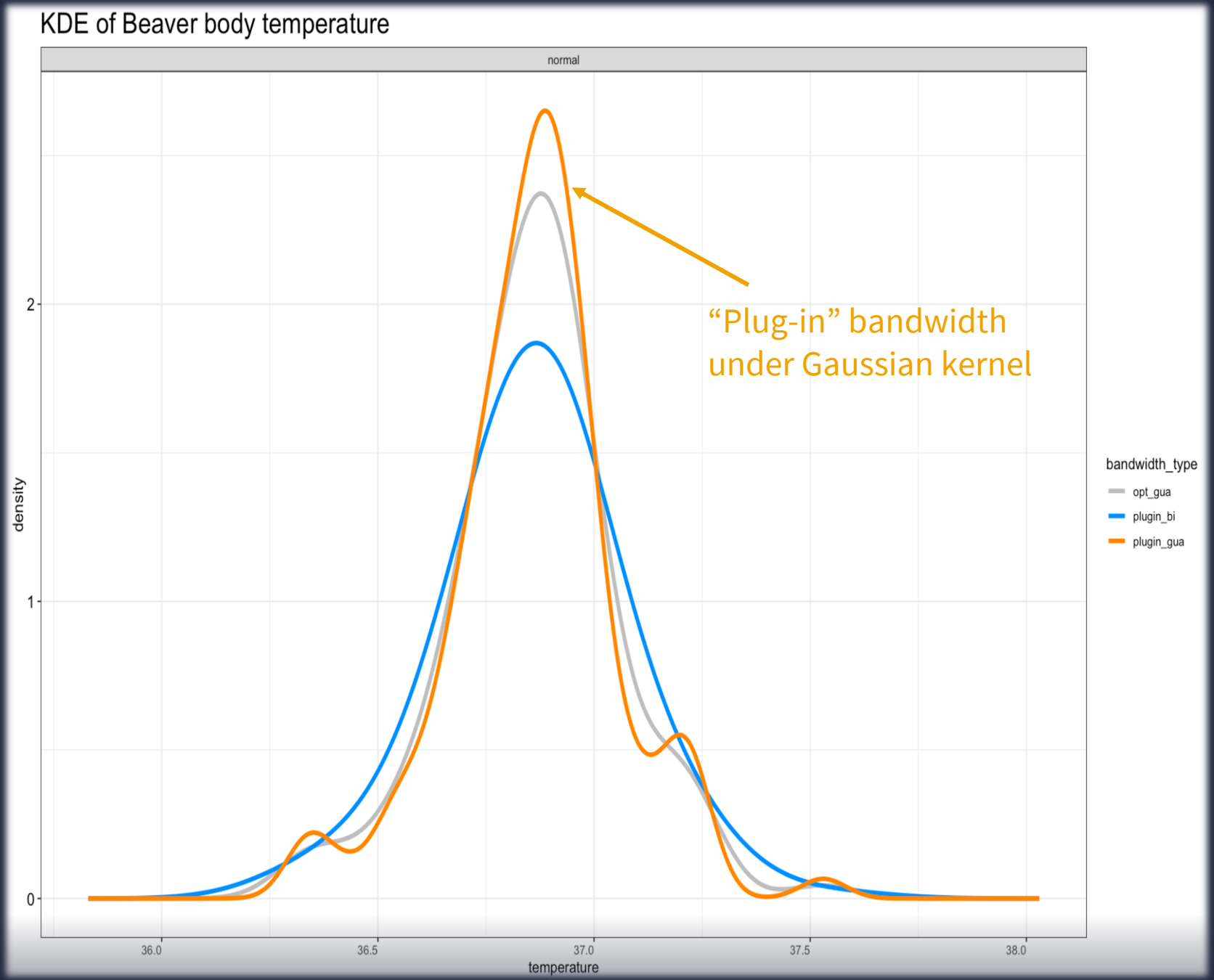
Application



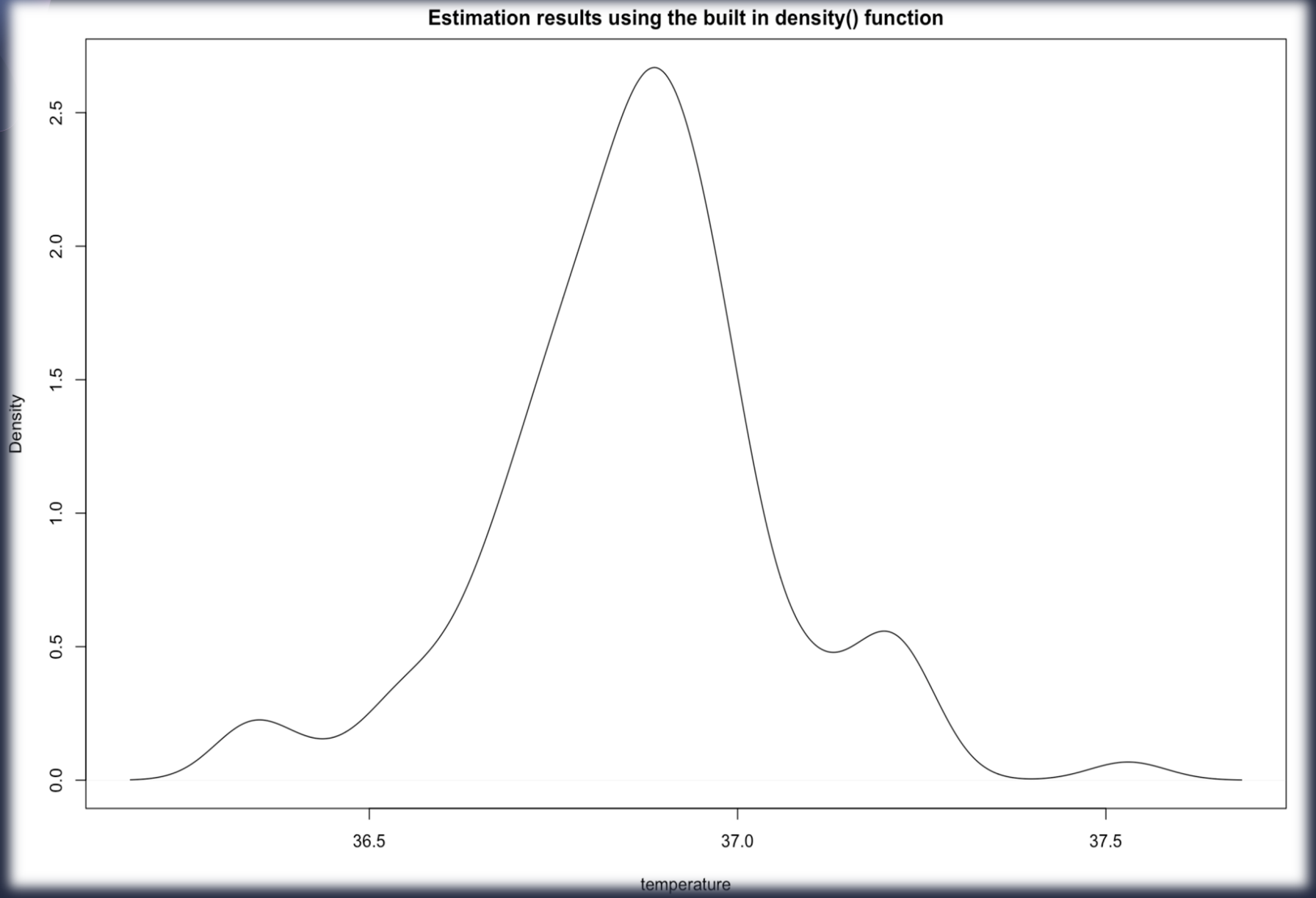
Histogram for Beaver's body temperature dataset in R



KDE for Beaver's body temperature



Estimated density using the built-in “density()” function in R

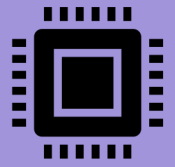


Reference:

- 1. https://mathisonian.github.io/kde/
- 2. https://en.wikipedia.org/wiki/Kernel_density_estimation
- 3. Fan, J. and Yao, Q., 2008. Nonlinear time series: nonparametric and parametric methods. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-69395-8

What else?

Simulations results of KDE with combinations of different size of samples, choices of kernels and bandwidths.



Evidence showing that to get a good KDE, we need to find the optimal bandwidth under the corresponding kernel function. (Choice of kernel function matters to some extent) .

“Hand-made” KDE functions, codes and plots!