Option 7: What is kernel density estimation? (KDE)? What are the benefits of KDE? What do the results tell you?

What is kernel density estimation?

Kernel density estimation (KDE) is an algorithm that takes a sample and finds an appropriately smooth PDF that fits the data.

Kernel density estimators belong to a class of estimators called *non-parametric* density estimators. In comparison to parametric estimators where the estimator has a fixed functional form (structure) and the parameters of this function are the only information we need to store, Non-parametric estimators have no fixed structure and depend upon all the data points to reach an estimate.

To understand kernel estimators, we first need to understand histograms whose disadvantages provides the motivation for kernel estimators. When we construct a histogram, we need to consider the width of the bins (equal sub-intervals in which the whole data interval is divided) and the end points of the bins (where each of the bins start). As a result, the problems with histograms are that they are *not smooth, depend on the width of the bins and the end points of the bins.* We can alleviate these problems by using *kernel density estimators*.

What are the benefits of KDE?

To remove the dependence on the end points of the bins, kernel estimators center a kernel function at each data point. If we use a smooth kernel function for our building block, then we will have a smooth density estimate. This way we have eliminated two of the problems

associated with histograms. The problem of bin-width still remains which is tackled using a technique discussed later on.

More formally, Kernel estimators smooth out the contribution of each observed data point over a local neighborhood of that data point. The contribution of data point x(i) to the estimate at some point x depends on how apart x(i) and x are. The extent of this contribution is dependent upon the shape of the kernel function adopted and the width (bandwidth) accorded to it. If we denote the kernel function as *K* and its bandwidth by *h*, the estimated density at any point x is

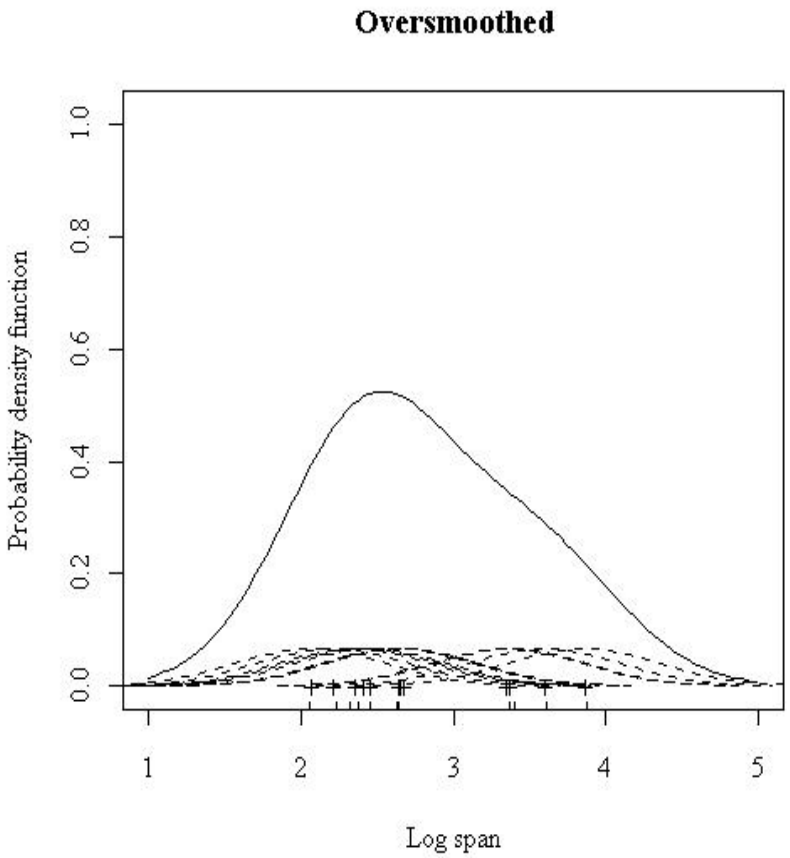
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right)$$

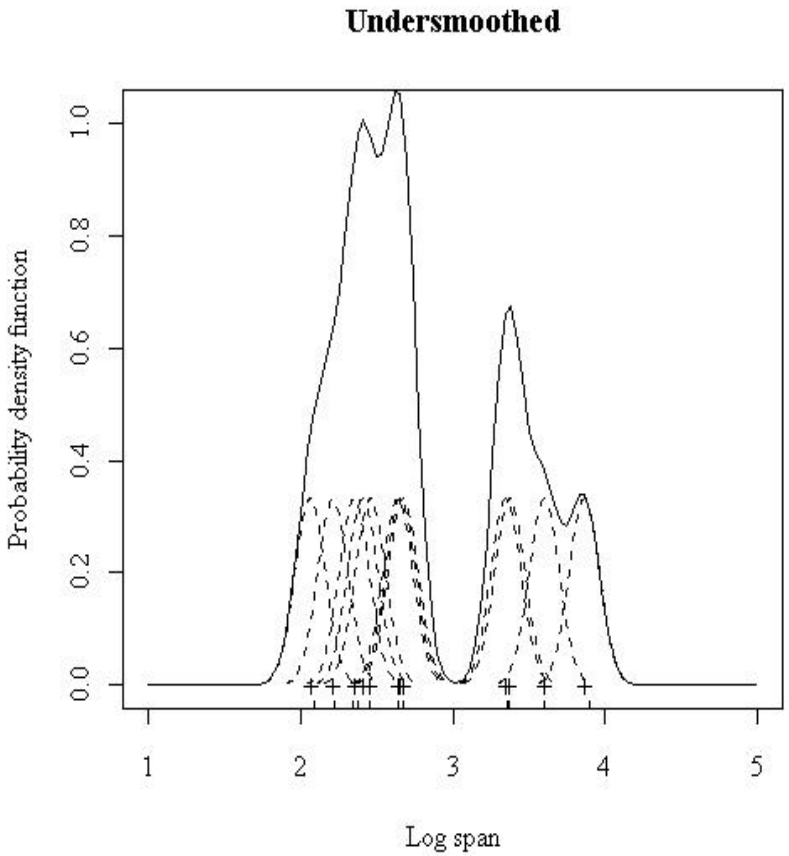
where $\int K(t)dt = 1$ to ensure that the estimates f(x) integrates to 1 and where the kernel function *K* is usually chosen to be a smooth unimodal function with a peak at 0. Even though Gaussian kernels are the most often used, there are various choices among kernels as shown in the table below.

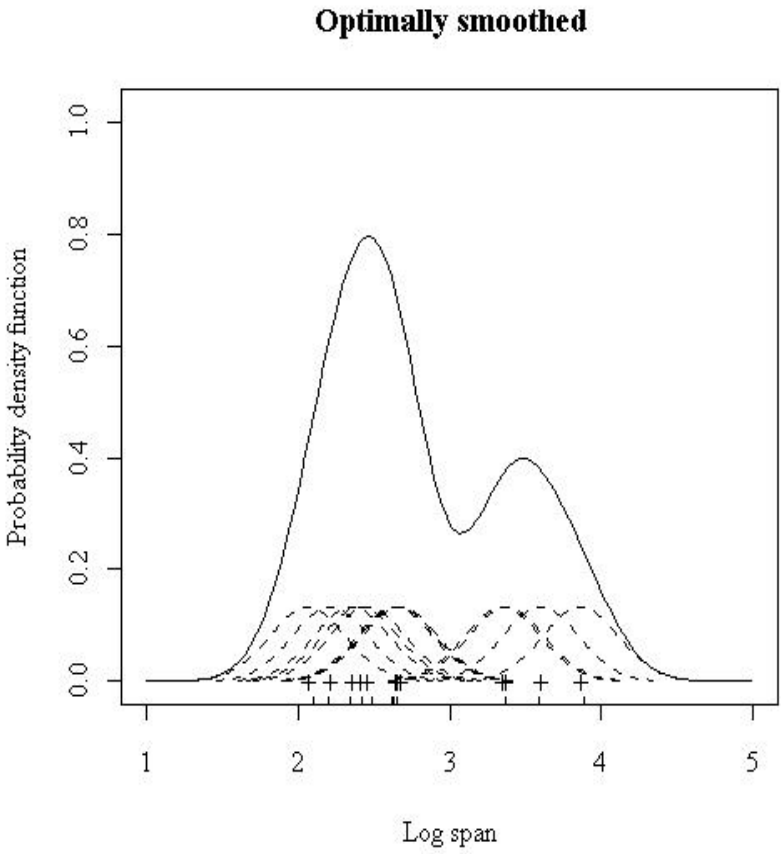
Kernel	K(u)
Uniform	$\frac{1}{2}I(\mid u \mid \leq 1)$

Triangle	$(1-\mid u \mid)\ I(\mid u \mid \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\mid u \mid \leq 1)$
Quartic	$\frac{15}{16}(1-u^2)^2 I(\mid u \mid \leq 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\mid u \mid \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}u^2)$
Cosinus	$\frac{\pi}{4} cos(\frac{\pi}{2}u) I(\mid u \mid \leq 1)$

The quality of a kernel estimate depends less on the shape of the *K* than on the value of *its bandwidth h*. It's important to choose the most appropriate bandwidth as a value that is too small or too large is not useful. Small values of h lead to very spiky estimates (not much smoothing) while larger h values lead to “oversmoothing”. The following three figures show the effect of 3 different bandwidths. When the bandwidth is 0.1 (very narrow) then the kernel density estimate is said to “*undersmoothed”* as the bandwidth is too small.


Oversmoothed
Probability density function
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Log span


Undersmoothed
Probability density function
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Log span



We can try to alleviate it by increasing the bandwidth of the kernel to a larger value (0.5). But, now we obtain a much flatter estimate with only one mode in place of the earlier four. This situation is said to be *oversmoothed*as we have chosen a bandwidth that is too large and have obscured most of the structure of the data.

A common method to choose the optimal bandwidth is to use the bandwidth that minimizes the AMISE (Asymptotic Mean Integrated Squared Error).

so, $h_{opt} = arg_{min}$ AMISE

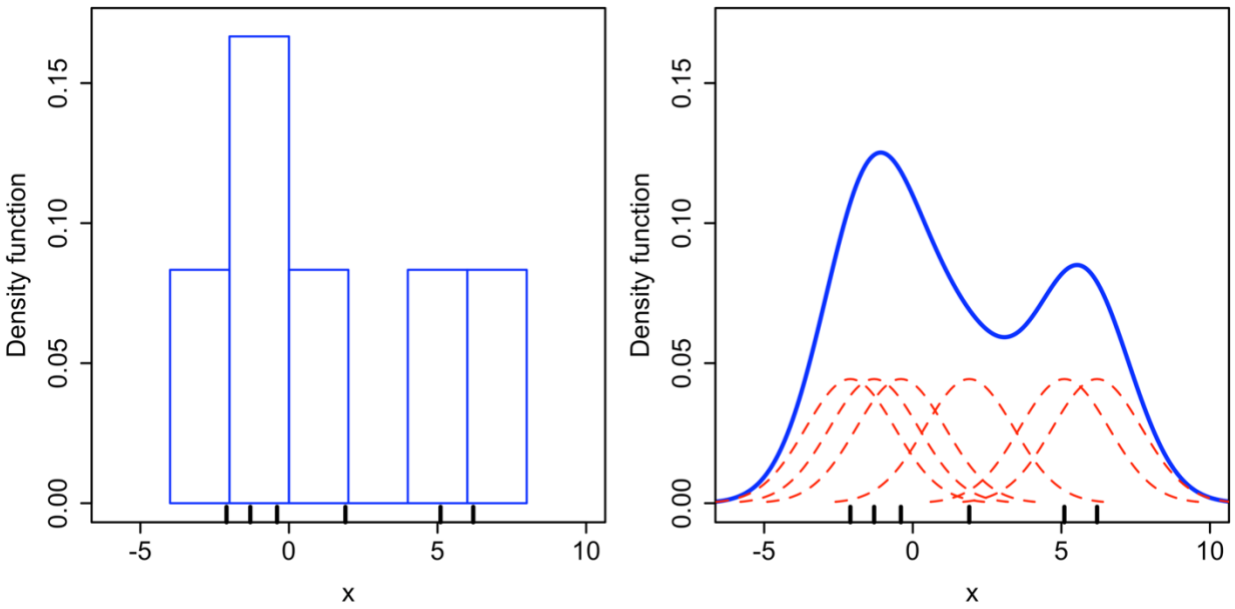
AMISE still depends on the true underlying density and so we need to estimate the AMISE from our data. This means that the chosen bandwidth is an estimate of an asymptotic approximation. It sounds that it's too far away from the true optimal value, but it turns out that this particular choice of bandwidth recovers all the important features whilst maintaining smoothness.

What do the results tell you?

Kernel Density estimators are closely related to histograms but, can be endowed with properties such as smoothness or continuity by using a suitable kernel. To interpret results of the Kernel density estimator, let’s look at an example given a 6 points data set:

Sample	1	2	3	4	5	6
Value	-2.1	-1.3	-0.4	1.9	5.1	6.2

For the histogram, first the horizontal axis is divided into sub-intervals or bins which cover the range of the data: In this case, six bins each of width 2. Whenever a data point falls inside this interval, a box of height 1/12 is placed there. If more than one data point falls inside the same bin, the boxes are stacked on top of each other.



For the kernel density estimate, a normal kernel with standard deviation 2.25 (indicated by the red dashed lines) is placed on each of the data points x_i. The kernels are summed to make the kernel density estimate (solid blue curve). The smoothness of the kernel density estimate (compared to the discreteness of the histogram) illustrates how kernel density estimates converge faster to the true underlying density for continuous random variables.

References

School of Informatics. The University of Edinburgh. *Kernel Density Estimators.* http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0405/MISHRA/kde.html

Wikipedia, 2021. *Kernel density estimation.* https://en.wikipedia.org/wiki/Kernel_density_estimation