

# First Draft KDE

Owain Morgan

29/11/2021

## Kernel Density Estimation

Kernel Density Estimation seeks to ask an important question, given

Suppose we have  $n$  independent, identically distributed random variables  $X_1, X_2, \dots, X_n$  (known as samples) from a distribution whose density function,  $f(x)$  is not known but is assumed to exist, can we infer the nature of the probability density

Then Kernel Density Estimation (also known as Parzen Windows or the empirical probability density function) provides an easy, if computationally intensive [Dekking et al., 2005], way of estimating the true density function and as shown by Parzen in 1962, will converge to the true distribution as the number of samples  $n \rightarrow \infty$ .

### Historical context

Kernel Density Estimation was only seriously considered in research beginning in the late 1950's. At its core, KDE seeks to infer the nature of the probability density function (pdf) of a random variable given a sample from that random variable. Prior to the development of KDE by Rosenblatt [Rosenblatt, 1956] & Parzen [Parzen, 1962] in the late 1950's and early 60's, inference of pdfs from data was limited to histograms which by their very nature are unsmooth and as the bin width narrows, become less reliable. The lack of research was in spite of the admission by Parzen [Parzen, 1962] of "the obvious importance of these [estimation] problems."

### The Method

Intuitively, one expects areas with a high density of samples to have a higher density function, and those areas which are more sparse to have a lesser density function.

To that end, a kernel function  $K(y)$  is introduced which in general is radially symmetric and unimodal.  $\langle ! \rightarrow$  The kernel functions are evaluated at each data point and then summed together, with an appropriate normalisation constant to give the empirical probability density function. That is to say, given each sample  $(X_i)_{i=1}^n$  and assuming a bandwidth  $h$ , the empirical probability density function,  $\hat{f}(x)$  is [Sheather and Jones, 1991]

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

### Choices and parameters

Fundamentally, there are two key choices to be made when implementing KDE, viz. the choice of kernel function and the choice of bandwidth. We consider both in turn.

#### Choice of Kernel Function

The two most common choices of kernel function are the rectangular and the Gaussian.

The simplest choice of kernel function is the rectangular (or window) function which has the formula:

$$K_r(x) = \begin{cases} 1 & \text{if } |x - x_i| \leq \frac{h}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

and can be seen in Figure 1 on the left. This kernel function has the disadvantage of weighing all points within a certain distance (the bandwidth) of the data point equally. It would be natural to assume that we should weigh points closest to the data point the heaviest.

The Gaussian kernel function is precisely that, it fits a Gaussian (or normal) function about each data point and then sums these together to give the empirical density function. The formula for this type of density function is given (in 1-dimension) by

$$K_G(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{1}{2h}(x - x_i)^2\right)$$

The shape of this kernel function is the familiar bell-curve as can be seen on the right hand side of Figure 1

```
plot(density.default(10, bw=2, kernel = "rectangular"))
plot(density.default(10, bw=2, kernel = "gaussian"))
```

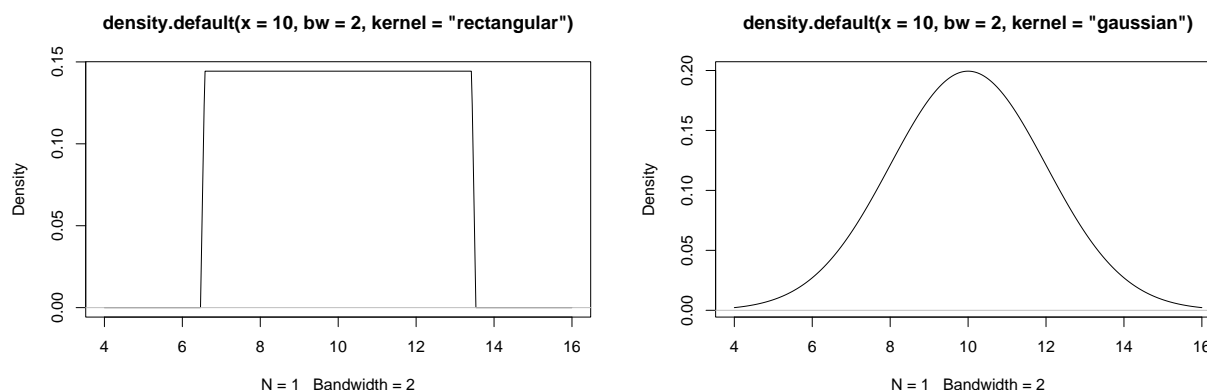


Figure 1: The window and Gaussian Kernel Functions, both centred at 10 with bandwidth  $h = 2$ .

## Examples

In this section of the report, we will consider some results about KDE. [Parzen, 1962]

For reproducibility, we set the seed and include some R libraries as follows.

```
set.seed(184)
library(tidyverse)
library(bbmle)
```

### A Limited Number of Samples from a (Unimodal) Normal Distribution

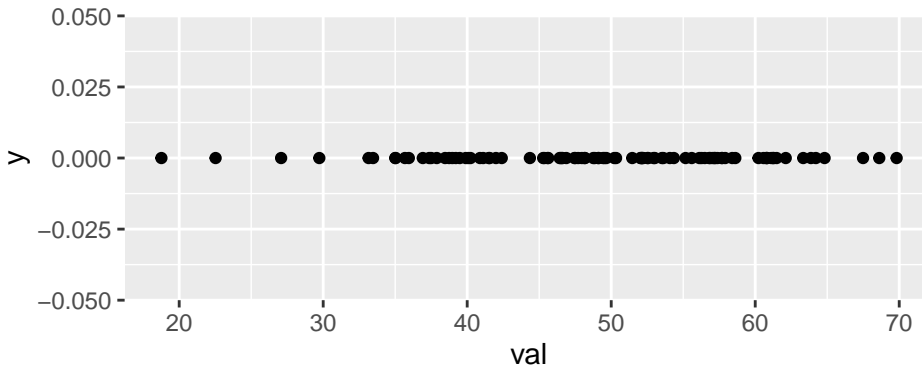
We first demonstrate how KDE looks in practice and the importance of the choice of kernel function.

We consider a random variable,  $X \sim N(50, 10)$  of which we take 100 i.i.d. samples, namely  $X_1, X_2, \dots, X_{100}$ .

```
normSample <- rnorm(100, 50, 10)
```

We plot these values by creating a tibble.

```
normSampDF <- tibble(val = normSample)
ggplot(data = normSampDF) +
  geom_point(aes(val, 0))
```



As can be seen, the data points are more concentrated towards the mean as expected. We now plot the kernel density estimate using both the window and Gaussian kernel functions, and include the true density function since we know we took a sample from a  $N(50, 10)$  random variable.

```
ggplot(data = normSampDF) +
  xlim(10,90) + geom_point(aes(val, 0)) +
  geom_density(aes(val), kernel = "rectangular") +
  geom_function(fun = dnorm, args = list(mean = 50, sd = 10), linetype = "dashed")
ggplot(data = normSampDF) +
  xlim(10,90) + geom_point(aes(val, 0)) + geom_density(aes(val)) +
  geom_function(fun = dnorm, args = list(mean = 50, sd = 10), linetype = "dashed")
```

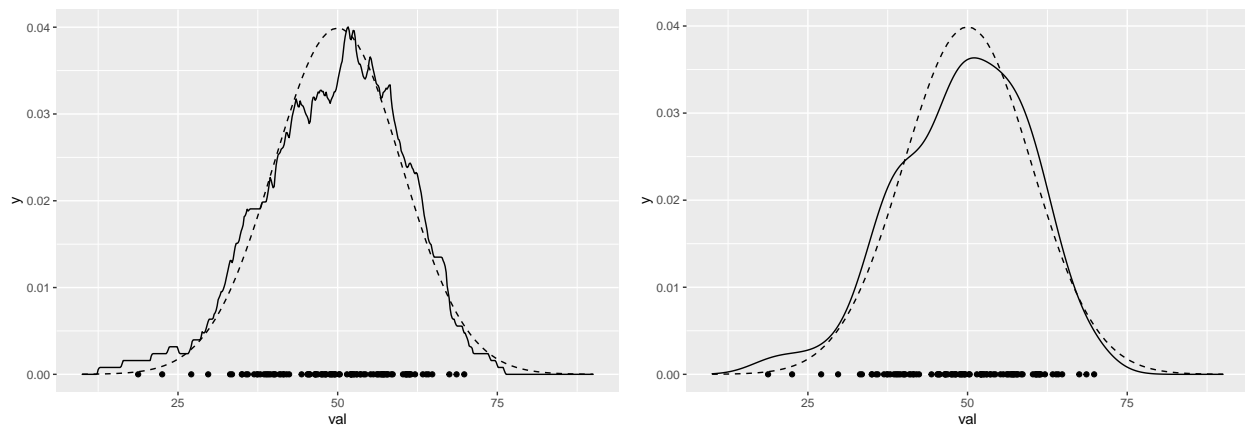


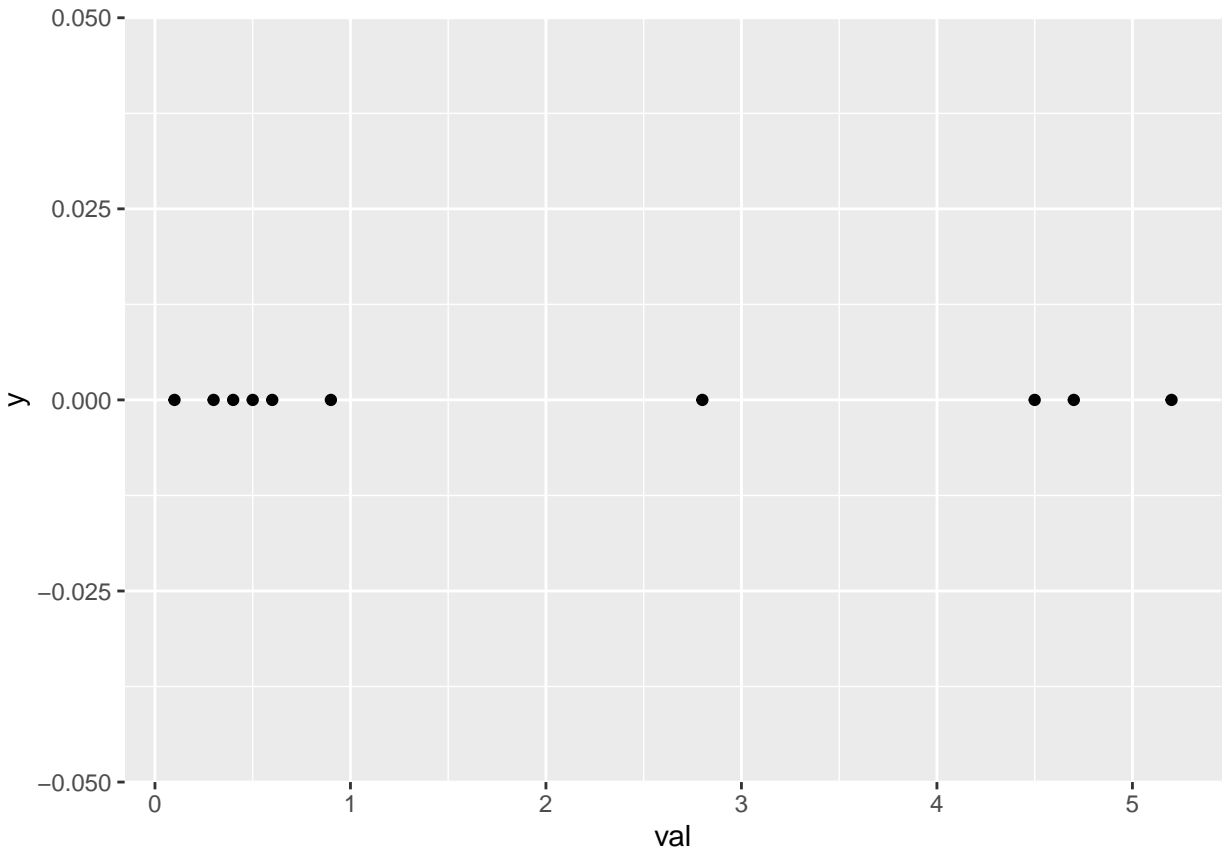
Figure 2: Evaluating the empirical probability density using the window (left) and Gaussian (right) functions.

Clearly, the Gaussian kernel gives a much smoother empirical pdf and in this case is more appropriate, however it is more computationally expensive and has an unlimited support.

## Bandwidth Selection

Now suppose we have the data set  $(0.1, 0.3, 0.4, 0.5, 0.6, 0.9, 2.8, 4.5, 4.7, 5.2)$  which we can graph as follows:

```
interest<- c(0.1,0.3,0.4,0.5,0.6,0.9,2.8,4.5,4.7,5.2)
interestingDF <- tibble(val = interest)
ggplot(data = interestingDF) +
  geom_point(aes(val, 0))
```



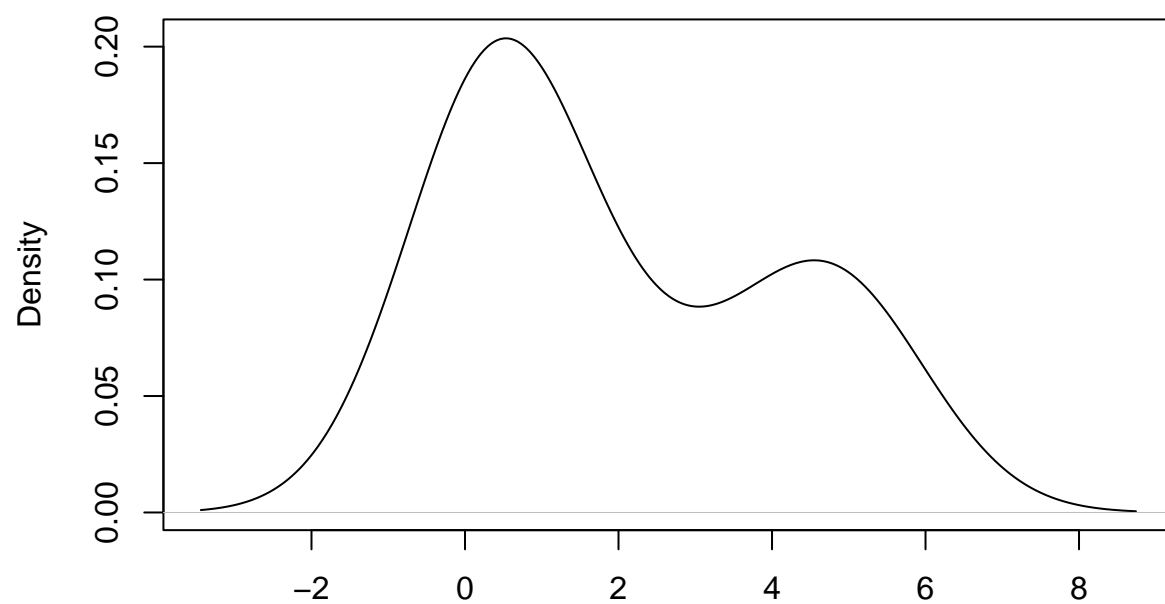
Clearly, this data set is bimodal. Let us perform kernel density estimation to illustrate the importance of bandwidth selection. We will use the Gaussian kernel for this section.

R, by default will use the method described in [Sheather and Jones, 1991] that seeks to “[choose] the bandwidth to (approximately) minimize good quality estimates of the mean integrated squared error”.

Using this bandwidth, the kernel density estimate is as follows.

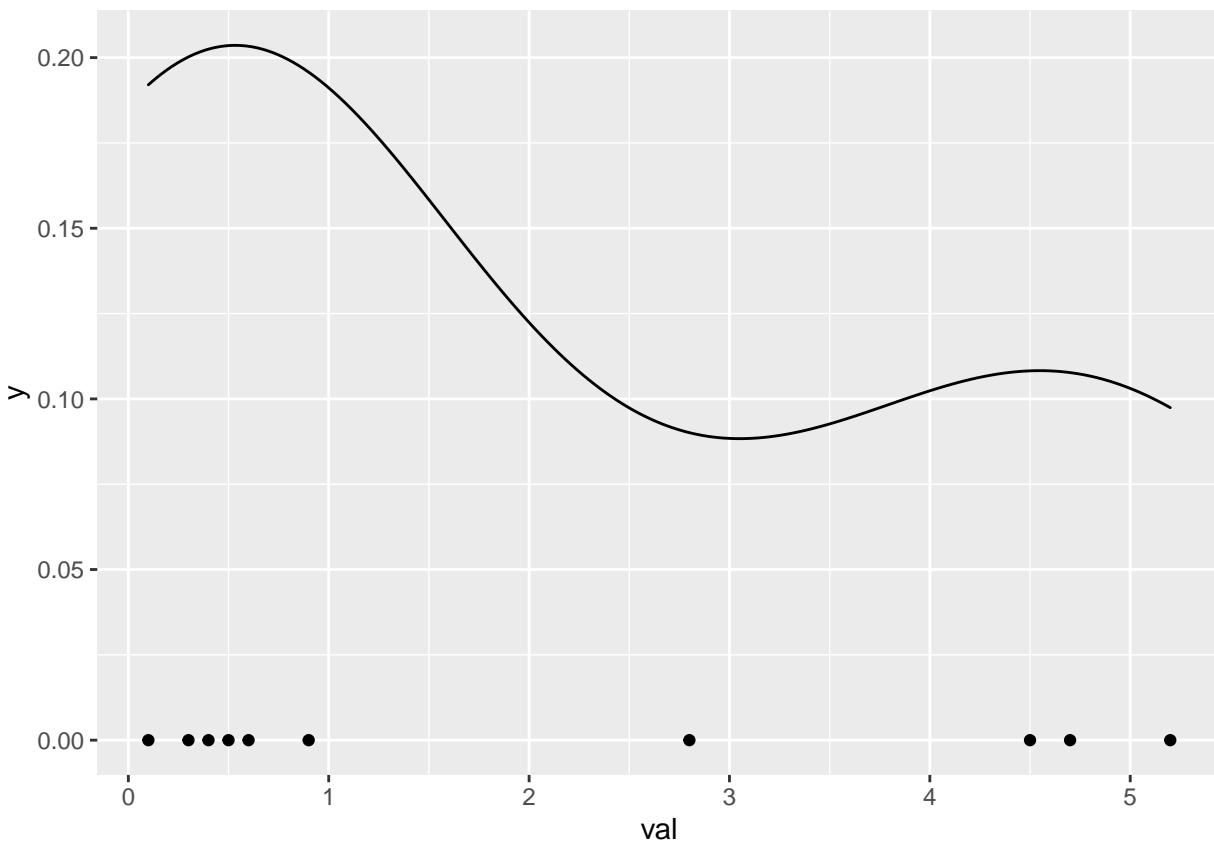
```
intGauss <- density.default(interest)
plot(intGauss)
```

**density.default(x = interest)**



N = 10 Bandwidth = 1.181

```
ggplot(data = interestingDF) +  
  geom_point(aes(val, 0))+  
  geom_density(aes(val))
```



In this case, R estimates the best bandwidth to be 1.1805817.

Let us now consider two extremes, a very narrow bandwidth and a very wide bandwidth.

```
ggplot(data = interestingDF) +
  xlim(-1,7) + geom_point(aes(val,0)) +
  geom_density(aes(val), bw = 0.05)
ggplot(data = interestingDF) +
  xlim(-5,15) + geom_point(aes(val,0)) +
  geom_density(aes(val), bw = 3)
```

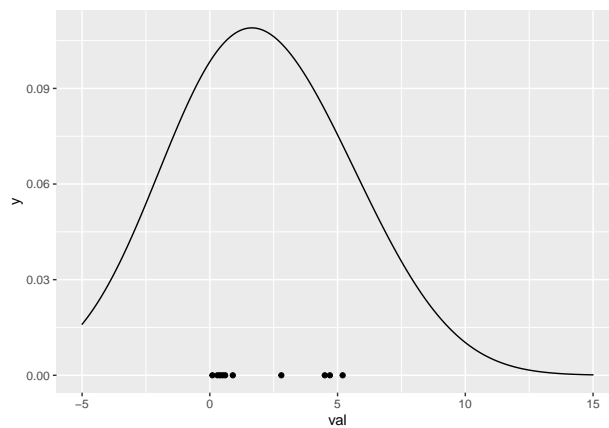
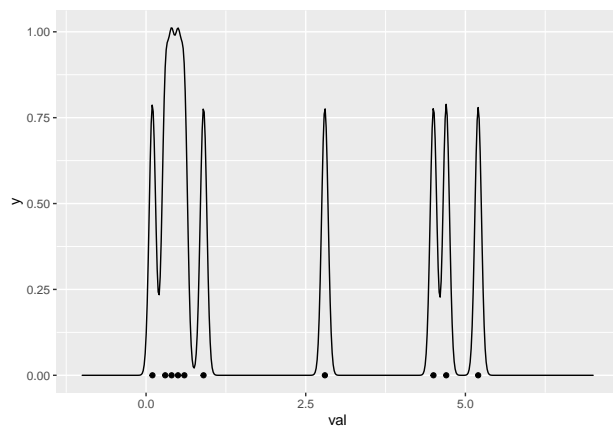


Figure 3: Evaluating the empirical probability density using a bandwidth of 0.05 (left) and 3 (right).

Both bandwidths in Figure 3 are totally inappropriate. In the first case with a narrow bandwidth, the density is totally concentrated where the data points are and so is overfitted and thus is unlikely to generalise to other examples of the same distribution. On the other hand, using a very wide bandwidth of 3 totally masks the underlying nature of the data.

## References

- [Dekking et al., 2005] Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., and Meester, L. E. (2005). *A Modern Introduction to Probability and Statistics : Understanding Why and How*. Springer Texts in Statistics. Springer-Verlag London Limited, London.
- [Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.
- [Rosenblatt, 1956] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.
- [Sheather and Jones, 1991] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B, Methodological*, 53(3):683–690.