# Chapter 5: Density Estimation and Smoothing

## 1 Density Estimation

**Density Estimation** is the problem of reconstructing the probability density function using a set of given data points. Suppose we observe $X_1, \cdots, X_n$ and we want to recover the underlying probability density function generating our dataset.

Here, we assume that the data comes from a continuous distribution. In a parametric setting, the density function relies on its parameters. Hence to estimate the density, it is enough to estimate the parameters involved in the density function. But these estimates of density will be valid only when data actually follows the assumed distribution.

It is very likely that someone who wants to know the distribution does not know ahead of time that the density function belongs to a certain class of parametric distribution functions. In such cases, we need non-parametric estimation. For a detailed reading, see [1].

### 1.1 Estimating CDF

Assume that $X$ follows a continuous distribution $f$ with CDF $F$, then for $h > 0$

$$P\left(x - \frac{h}{2} < X < x + \frac{h}{2}\right) = \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y)\mathrm{d}y \tag{1}$$

Now if $f$ is smooth and $h$ is small, then:

$$\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y)\mathrm{d}y \approx h \cdot f(x) \Rightarrow \hat{f}(x) = \frac{F\left(x+\frac{h}{2}\right) - F\left(x-\frac{h}{2}\right)}{h}$$

Hence we have now moved to the problem of estimating CDF of an unknown distribution. The following are some of the ways of estimating the cumulative density of an unknown continuous distribution.

[1] Textbook: Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

### Empirical CDF

If $X_1 \cdots X_n$ are IID random variables with distribution function $F$, then the empirical CDF is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right), \tag{2}$$

where $I$ is the Indicator function. Note that here we approximate the unknown distribution with a discrete distribution function. It is a step function estimate for $F$. From Eqs. (1) and (2), we can show that

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} I\left(x - \frac{h}{2} \leq X_i \leq x + \frac{h}{2}\right)}{nh} \tag{3}$$

Some properties of empirical distribution function (EDF) are as follows.

1. The empirical distribution function $(\hat{F}_n)$ is an unbiased estimator of the true underlying distribution $(F)$.

   *Proof.*

   $$\begin{aligned} E(\hat{F}(x)) = E\left(\frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right)\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(I\left(X_i \leq x\right)\right) &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} I(y \leq x) f(y) dy \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{x} f(y) dy \\ &= \frac{1}{n}\sum_{i=1}^{n} F(x) = F(x). \end{aligned}$$

   □

2. $MSE = \mathrm{Var}\left(\hat{F}_n\right) = \frac{1}{n}F(x)(1 - F(x))$

   *Proof.*

   $$\mathrm{Var}\left(\hat{F}_n(x)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left[E\left(\left\{I\left(X_i \leq x\right)\right\}^2\right) - \left\{E\left(X_i \leq x\right)\right\}^2\right] \tag{4}$$

   where $E\left(X_i \leq x\right) = F(x)$ and $E\left(\left\{I\left(X_i \leq x\right)\right\}^2\right) = F(x)$ and hence this proves the result. □

3. $\hat{F}_n(x) \longrightarrow F(x)$ in probability.

*Proof.* This result can be proved using Chebyshev's inequality: $P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$ □

In fact, it can be shown that $\hat{F}_n(x) \longrightarrow F(x)$ almost surely and uniformly.

Drawbacks: There are two main drawbacks for this estimate:

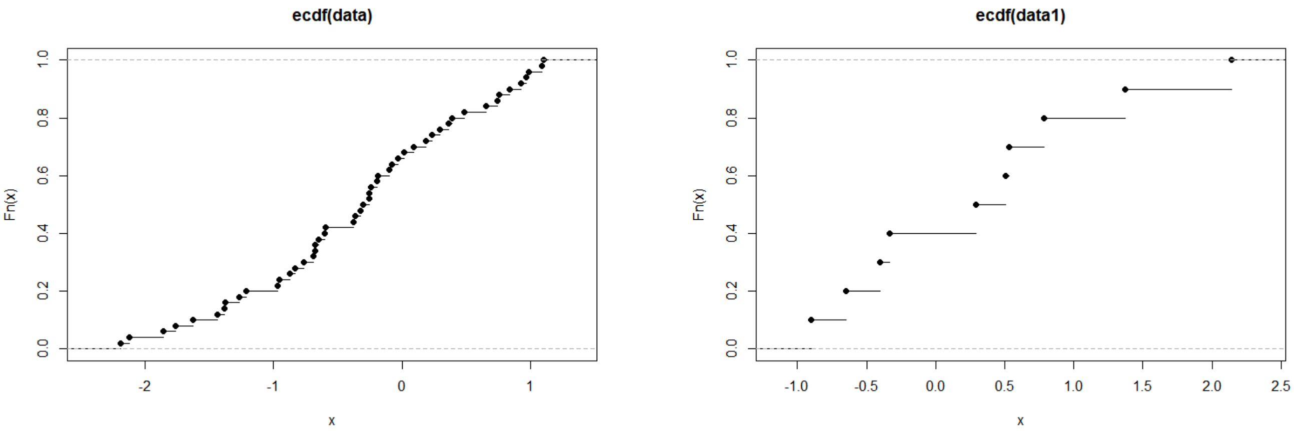
1. A continuous distribution is being defined by a discrete non-continuous function.
2. If our data set is small, then we cannot expect a good estimate. In order to get a good estimate in this case, we need a large number of data points.

---

The following R-code shows how to plot an ecdf in R.

```
data <- rnorm(50 , 0 , 1) ## store 50 normal random variables in variable "d
F50 <- ecdf(data) # f stores the emperical distribution function
plot.ecdf(F50) # See Figure 1.(a)
data1 <- rnorm(10)
F10 <- ecdf(data1)
plot.ecdf(F10) # See Figure 1.(b)
```

---



(a) ***ECDF for 50 points drawn from normal distribution***

(b) ***ECDF for 10 points drawn from normal distribution***

*Figure 1 :* ***The distribution is approximated better with a greater number of points.***

### 1.2 Histogram and Centred histogram

A histogram is a well-known and popular density estimate for a continuous distribution. Formally, it can be defined in the following mathematical manner:

Let X be the random sample of size $n$ from a continuous population then the histogram density estimate is given by

$$\hat{h}_n(x) = \frac{\text{no. of } X_i \in B_j}{nh} \text{ when } x \in B_j \tag{5}$$

where $B_j$ are different partitions of $[0,1]$. Without loss of generality, we can assume that each $X_i$ is in $[0,1]$, then

$$B_1 = \left[0, \frac{1}{M}\right); B_2 = \left[\frac{1}{M}, \frac{2}{M}\right) \cdots B_M = \left[\frac{M-1}{M}, 1\right) \text{ where } h = \frac{1}{M} \text{ is called bin width.}$$

It can be shown that the expectation of this estimate is:

$$E\left(\hat{h}_n(x)\right) \frac{M}{n} \sum_{i=1}^{n} P\left(X_i \in B_j\right) = MP\left(X_i \in B_j\right) = M\left[F\left(\frac{j}{M}\right) - F\left(\frac{j-1}{M}\right)\right] \tag{6}$$

Using the fact that $\frac{1}{M} = \frac{j}{M} - \frac{j-1}{M}$ and then applying mean value theorem to $F$, we get that there exists $x^*$ such that

$$E\left(\hat{h}_n(x)\right) = \frac{\left[F\left(\frac{j}{M}\right) - F\left(\frac{j-1}{M}\right)\right]}{\frac{j}{M} - \frac{j-1}{M}} = h\left(x^*\right); \text{ for } x^* \in B_j \tag{7}$$

Applying mean value theorem to $h$, we get the existance of $x^{**}$ such that

$$\frac{h\left(x^*\right) - h(x)}{x^* - x} = h'\left(x^{**}\right) \tag{8}$$

From Eqs. (7) and (8), we can show that

$$\text{Bias } = E\left(\hat{h}_n(x) - h(x)\right) \leq \frac{\left|h'\left(x^{**}\right)\right|}{M} \tag{9}$$

$$\text{Var}\left(\hat{h}_n(x)\right) = MSE \leq \frac{\left(h'\left(x^{**}\right)\right)^2}{M^2} + \frac{Mh\left(x^*\right)}{n} + \frac{\left(h\left(x^*\right)\right)^2}{n}. \tag{10}$$

Observations: Using similar calculations as shown above, we can show that from the inequalities (9) and (10), we observe that as $M$ increases (number of bins increases), the chance of over-estimating or under-estimating the data reduces. Hence increasing the bin width increases the bias but reduces the variability. This is referred to as **over-smoothing**. Conversely, if the bin width reduces, our bias reduces but variability increases. This is referred to as **under-smoothing**.

Drawbacks: The drawbacks of using histogram as density estimate are:

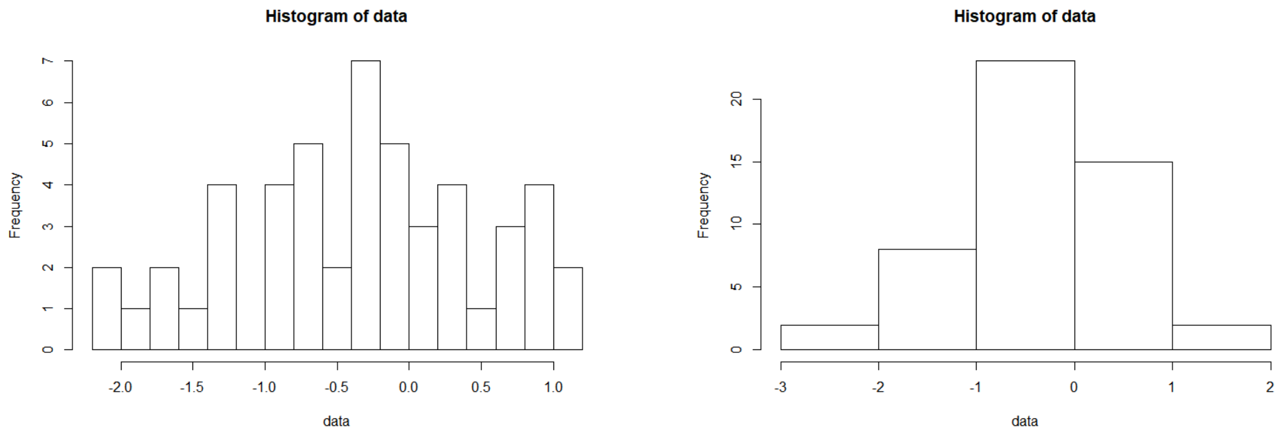
1. The estimate of a continuous distribution is not continuous
2. The density estimator depends on the width and end points of certain fixed intervals which are chosen beforehand.
3. Increasing the number of bins might reduce the bias but it will also increase the chances of getting regions of 0 probability in the histogram.

The estimate's dependency on the bin width and its endpoints can be removed by defining a **centered histogram**. In a normal histogram, all $B_j$ are fixed and predetermined. But in the case of the centered histogram, the bandwidth is kept fixed at $h$ but the intervals are not. The pdf estimate at $x$ in this case is

$$\hat{h}_c(x) = \frac{\text{no. of } X_i \in \left(x - \frac{h}{2}, x + \frac{h}{2}\right)}{nh}.$$

The R-codes below show how to plot a histogram and a centered histogram in R.

```
hist(data, breaks = 20, col = NULL) ## to get a histogram with 20 bins
hist(data, breaks = 3, col = NULL) #over-smoothed, see Figure 2.
```

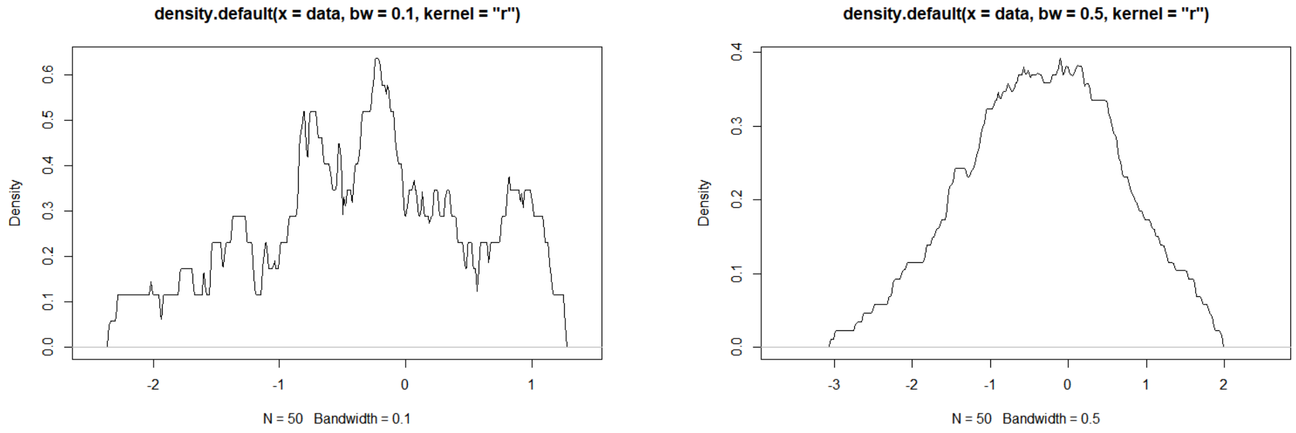


*(a)* ***High variance due to many bins*** *(b)* ***Oversmoothed estimate***

*Figure 2 :* ***The distribution is not approximated well with too many or too few points.***

Now, using the R-code below, we plot centered histograms and observe that an increase in bandwidth increases the smoothness of the estimate.

```
x <- density(data, kernel = "r", bw = 0.1)
z <- density(data, kernel = "r", bw = 0.5)
plot(x) # see Figure 3.(a)
plot(z) # see Figure 3.(b)
```



*(a)* $\text{bw} = 0.1$ ***(high variance)*** *(b)* ***bw*** $= 0.5$ ***(over-smoothed)***

*Figure 3 :* ***Plotting a centered histograms with*** $\text{bw} = 0.1$ ***and*** $0.5$**.**

## 1.3 Kernel Density Estimates

A function $K$ is called a kernel function if

- $K(x) \geq 0$
- $K(-x) = K(x)$
- $\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$

**Example 1.3.1.** *Examples of some of the commonly used kernels are:*

1. $K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ *is called the Gaussian kernel*
2. $K(x) = \frac{3}{4} \max\left\{1 - x^2, 0\right\}$
3. $K(x) = \begin{cases} 1 & \textit{if } -0.5 \leq x \leq 0.5 \\ 0 & \textit{otherwise} \end{cases}$ *is called the box-kernel.*

Disadvantages of histogram have motivated defining the **kernel density estimates** (KDE). Not only does kernel density estimate solve the problem of dependency on the choice of bins but also solves the continuity problem i.e., KDE of a continuous function is continuous provided we choose a smooth kernel.

**Remark 1.1.** *The order in which the estimates have developed is as given below:*
*Histogram* $\longrightarrow$ *Centred Histogram* $\longrightarrow$ *Kernel Density Estimate*
*Moving from histogram to KDE, we solved the problem of dependencies on the bin and then we generalized the centered histogram to obtain KDE which solved the continuity problem as well.*

Given $X_1, \cdots X_n$ and a kernel $K$, the KDE at point $x$ is given by:

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right) \text{ where } h= \text{ bandwidth} \tag{11}$$

**Theorem 1.1.** *The definitions histogram, centered histogram and box kernel are all equivalent.*

*Proof.* Simple manipulation of the definitions can prove the above statement. □

**Remark 1.2.** *From Remark 1.1 above, we can view KDE as a generalization of a centered histogram. Changing the kernels gives different estimates for the function.*

### Properties of KDE

Let $X_1 \cdots X_n$ be an IID sample from an unknown population following a density function "$p(x)$". Initially, let us just consider a single point $x_0$. We analyze the quality of the estimate:

$$\hat{p}_n\left(x_0\right)=\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)$$

### Bias of the KDE

Using the basic definition of expectation (using integral) of the function of random variables, we can show that

$$E\left(\hat{p}_n\left(x_0\right)-p\left(x_0\right)\right)=\left\{\frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-x_0}{h}\right) p(x)\mathrm{d}x\right\}-p\left(x_0\right) \tag{12}$$

Applying the change of variable $y=\frac{x-x_0}{h}$ in the Eq. (12) and using the following Taylor expansion for p will help us simplify Eq. (12).

$$p\left(x_0+hy\right)=p\left(x_0\right)-(hy)p'\left(x_0\right)+\frac{1}{2}(hy)^2 p''\left(x_0\right)+\mathcal{O}\left(h^2\right) \tag{13}$$

We use kernel properties, change of variable, and Eq. (13) to simplify Eq. (12), to the following equation:

$$\text{Bias }=\frac{h^2}{2}p''\left(x_0\right)\int_{-\infty}^{\infty} K(y)y^2\,\mathrm{d}y+\mathcal{O}\left(h^2\right)=\frac{1}{2}h^2 p''\left(x_0\right)\mu_k+\mathcal{O}\left(h^2\right) \tag{14}$$

$\mu_k$ in the above line is a $\int_{-\infty}^{\infty} y^2 K(y)\mathrm{d}y$ from Eq. (14), we can say that if the bias is large then $p''\left(x_0\right)$ value is large $\Longrightarrow$ there is more rate of change of slope $\Longrightarrow$ the function p(x) will be more curved at $x_0$ (like peaks). Hence KDE tries to smoothen the peaks of a distribution function.

**Variance of KDE**

$$\operatorname{Var}\left(\hat{p}_n\left(x_0\right)\right)=\frac{1}{n^2h^2}\sum_{i=1}^{n}\operatorname{Var}\left[K\left(\frac{X-X_i}{h}\right)\right]\leq\frac{1}{nh^2}\int_{-\infty}^{\infty}K^2\left(\frac{X-X_i}{h}\right)p(x)\mathrm{d}x \quad (15)$$

Again using the change of variable $y=\frac{x-x_0}{h}$ and applying Eq. (13) as before to Eq. (15) along with some kernel properties, we get can simplify Eq. (15) as follows:

$$\operatorname{Var}\left(\hat{p}_n\left(x_0\right)\right)\leq\frac{1}{nh}p\left(x_0\right)\sigma_k^2+\mathcal{O}\left(\frac{1}{nh}\right)\text{ where }\sigma_k^2=\int_{-\infty}^{\infty}K^2(y)\mathrm{d}y \quad (16)$$

What we can tell from the inequality in Eq. (16) is that at a given point $x_0$ where density value $p\left(x_0\right)$ is large, the variance is also large.

$$\text{Now, } MSE=\left(\frac{h^2}{2}p''\left(x_0\right)\mu_k\right)^2+\frac{1}{nh}p\left(x_0\right)\sigma_k^2+\mathcal{O}\left(h^4\right)+\mathcal{O}\left(\frac{1}{nh}\right) \quad (17)$$

Then the term on the right-hand side in Eq. (17) is called **Asymptotic Mean Square Error** (AMSE). Minimizing AMSE with respect to the bandwidth $h$ gives us the following optimal value of $h_{\text{opt}}\left(x_0\right)$ for a given $x_0$.

$$h_{opt}\left(x_0\right)=\left[\frac{p\left(x_0\sigma_k^2\right)}{n\left|p''\left(x_0\right)\right|^2\mu_k^2}\right]^{\frac{1}{5}} \quad (18)$$

**Remark 1.3.** *However, the major problem is that the above optimal h is just a theoretical minimum. We cannot use this in real life because we do not know the distribution $p(x)$.*

In all the above analysis, we considered only one point $x_0$, but in general, we want to control the overall $MSE$ of the entire function. For one point $x_0$,

$$\mathrm{MSE}=\mathrm{E}\left[\left(\hat{p}_n\left(x_0\right)-p\left(x_0\right)\right)^2\right].$$

So for all points $x$, we have:

$$\mathrm{MISE}=\mathrm{E}\left[\int_{-\infty}^{\text{infty}}\left(\hat{p}_n\left(x_0\right)-p\left(x_0\right)\right)^2\ \mathrm{d}x\right] \quad (19)$$

MISE is called the mean integrated square error. Now some calculations will let us show that

$$\mathrm{MISE}\left(\hat{p_n}\right)=\left\{\frac{1}{4}h^4\mu_k^2\int_{-\infty}^{\infty}\left|p''(x)\right|\mathrm{d}x\right\}+\frac{\sigma_k^2}{nh}+\mathcal{O}\left(h^4\right)+\mathcal{O}\left(\frac{1}{nh}\right) \quad (20)$$

The dominating term of Eq. (20) is called as the **asymptomatic mean integrated square error**. Then the optimal smoothing bandwidth can be obtained by minimizing AMISE w.r.t $h$ is given by:

$$h_{opt} = \left[ \frac{\sigma_k^2}{\mu_k^2 \left( \int_{-\infty}^{\infty} |p''(x)| \, \mathrm{d}x \right) n} \right]^{\frac{1}{5}} \tag{21}$$

**Remark 1.4.** *Again Eq. (21) can't be used for practical purposes, because p is unknown.*

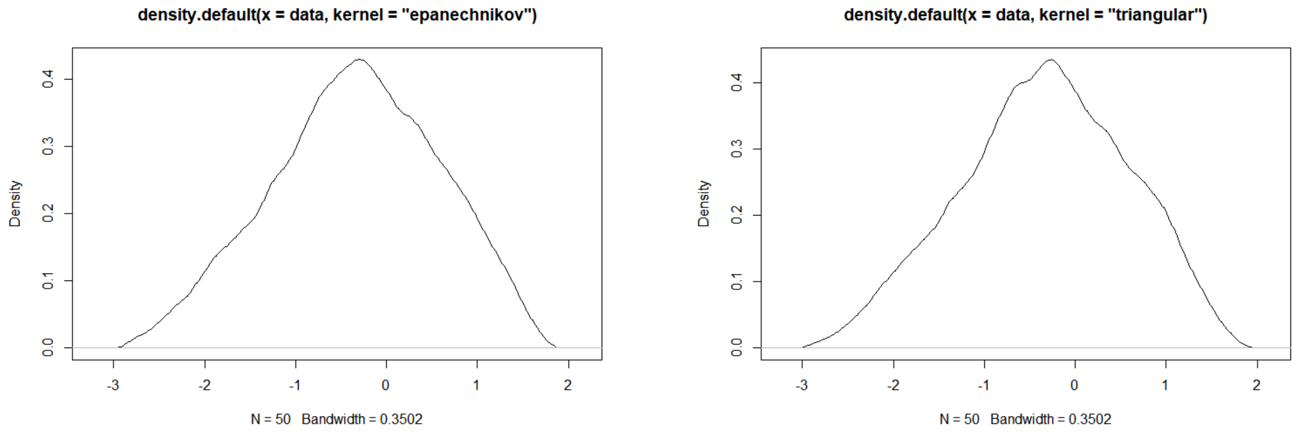
- *In-fact the problem about how to choose "h " is still unsolved and is known as* ***bandwidth selection problem***
- *There are modifications to this procedure where we can use a variable bandwidth. This has not been discussed here.*

---

The following R-codes plot KDEs. We observe two things from these plots. Fig. 4 shows that the choice of the kernel slightly affects the shape of the density estimate. Fig. 5 demonstrates that increasing the bandwidth increases the smoothness of the estimate.

```
plot(density(data, kernel = "gaussian"))
plot(density(data, kernel = "epanechnikov")) #smoother the kernel, smoother t
plot(density(data, kernel = "triangular"))
plot(density(data, kernel = "gaussian", bw= 0.3 ))
plot(density(data, kernel = "gaussian", bw= 0.6 ))
```

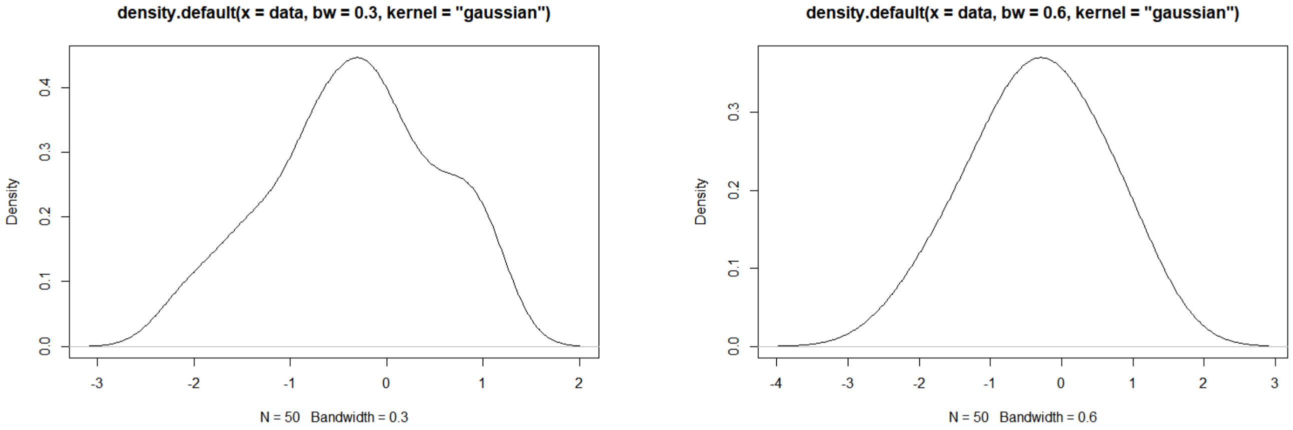
---



(a) ***KDE with kernel = epanechnikov kernel***

(b) ***KDE with kernel = triangular kernel***

*Figure 4 :* ***Change in kernel does not effect the estimate so much***



*(a)* ***Kernel*** $=$ ***gaussian , bandwidth*** $= 0.3$ *(b)* ***Kernel*** $=$ ***gaussian, bandwidth*** $= 0.5$

*Figure 5 :* ***Increase in bandwidth smoothens the estimate***

## 2 Smoothing

A smoothing algorithm is a summary of trend in $Y$ as a function of $X_1 \cdots X_n$. Smoother takes the data and returns a function called **smooth**. In the bivariate case, a smoother is a procedure that is applied to the bivariate data $(x_1, y_1) \cdots (x_n, y_n)$ that produces a decomposition $y_i = s(x_i) + \epsilon_i$ where $s$ is called the smooth function, also called as smooth.

In this chapter, we discuss a couple of smoothing methods. But all smoothers considered will be linear. The following are the assumptions for this chapter:

1. There are $n$ pairs of observations $(x_1, y_1) \cdots (x_n, y_n)$ and without loss of generality, assume that $x_1 \leq x_2 \leq \cdots \leq x_n$.
2. All observations are related through the expression $y_i = f(x_i) + \epsilon_i$ for $i = 1, 2, \cdots n$
3. $\epsilon_i$'s are IID from a continuous distribution centered at 0.

### 2.1 Local Averaging (Friedman)

This linear smoother is given by the below expression where $x_j$ are such that $f(x_j) = y_j$ and $x_j$ for $j = 1, 2 \cdots s$ are "$s$" points in the "neighbourhood" of $x_i$.

$$\hat{f}(x_i) = \frac{\sum_{i=1}^{s} y_j}{s}.$$

Now, the neighbourhood of $x_i$ is the smallest symmetric window about $x_i$ containing $s$ observations. The number of points in the window is called the span. Note that the window size changes for different values of $x_i$ but always includes $s$ data points.

A larger span over-smooths the data and smaller spans provide an under-smoothed estimate. Friedman proposed using a cross-validation method to choose the span.

### Cross-Validation

Let $\hat{g}_\lambda(x)$ be an estimate for $g(x)$. Then the **predictive squared error** (pse) of $\hat{g}_\lambda(x)$ is given by

$$\text{pse}(\lambda) = E\left[\left(y^* - \hat{g}_\lambda(x)\right)^2\right]$$

where $y^*$ is the **new** response value associated with the predictor $x$ i.e. $y^*$ and $y_i$ are independent of each other for all $i$.

Cross-validation is an idea in regression where we estimate pse($\lambda$) by CV($\lambda$) where

$$\text{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{g}_{\lambda,-i}\left(x_i\right)\right)^2 \qquad (22)$$

where $\hat{g}_{\lambda,-i}$ is the estimate for the $g$ using the data points $x_1 \cdots x_{i-1}, x_{i+1}, \cdots x_n$. Eliminating the $i^{\text{th}}$ point while estimating $g$ makes sure that $y_i$ is independent of all the other response values.

### Choosing a span

Let $\hat{y}_{(i)}$ be the estimate of $y_i$ determined by using all observed data points except $(x_i, y_i)$. If $s$ is the span, $e_i(s)$ is defined as $y_i - \hat{y}_{(i)}$, then $s$ is chosen such that $\frac{\sum_{i=1}^{n} e_i(s)^2}{n}$ is minimized over a set $S$ of possible span values. The span selected this way is called as the *global span* because this span is used for every point $x_i$. There are procedures for obtaining variable spans but that hasn't been discussed here.

---

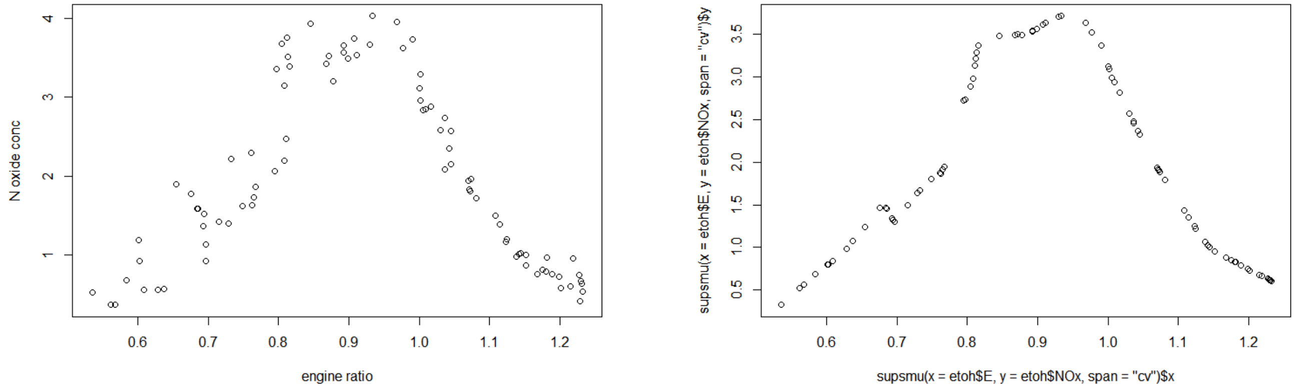
Now, we see an R-data example for the above theory. This data is about nitrogen oxide concentrations found in engine exhaust for ethanol engines. There are 88 pairs of data in this data set. We wish to smooth the data using "Friedman's local averaging".

```
##########  Friedman Local Averaging ############
library(lattice)
etoh <- lattice::ethanol #ethanol data
head(etoh)
plot(x = etoh$E , y = etoh$NOx , xlab = "engine ratio", ylab = "N oxide conc"

## data we want to smooth
# use "supsmu" to smooth the data using local averaging method
# use span = "cv" to use the cross validated variable span

plot(supsmu(x = etoh$E, y = etoh$NOx , span = "cv"))

# cv span related smoothing might have a better balance of variance and bias
```



(a) ***The ethanol data we wish to smooth***

(b) ***Ethanol data smoothed using friedman's averaging and*** cv ***span***

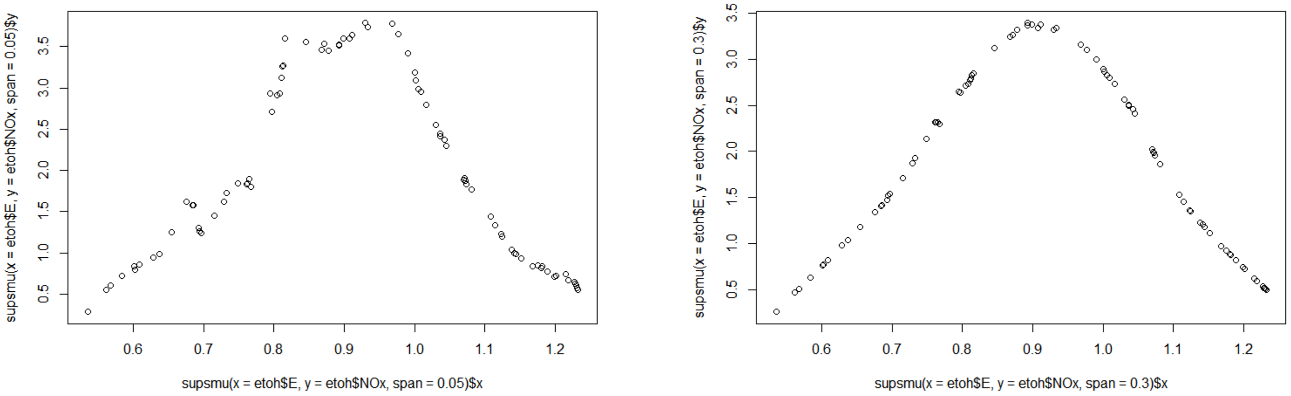
*Figure 6*

The data we wanted to smooth is given in Fig. 6 a while Fig. 6 b has been smoothed using cross-validated span.

```
# if span = p then it smooths the data using a constant span of size pn

plot(supsmu(x = etoh$E, y = etoh$NOx , span = 0.05))
plot(supsmu(x = etoh$E, y = etoh$NOx , span = 0.30))

# appears to be too smoothed. So biased.
```



*(a)* ***Ethanol data smoothed using friedman's averaging and span*** $= 0.05$

*(b)* ***Ethanol data smoothed using friedman's averaging and*** $\text{span} = 0.3$

*Figure 7*

We observe how a change in bandwidth changes how smoothed the data gets from Fig. 7 . Clearly, increasing the span increases the smoothness, and hence bias increases.

## 3 Kernel smoothing (Nadaraya and Watson)

Here we want to estimate a general function "f" rather than a density function. Nadaraya and Watson independently introduced the following kernel regression estimate:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} = \sum_{i=1}^{n} y_i w_i \text{ where } w_i = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} \tag{23}$$

Clearly, this estimate is linear in the observed data $y_i$ and the weights $w_i$ depend on kernel $(K)$, bandwidth $(h)$, and distance between $x$ and $x_i$. Note that this is not the nearest neighbor method, unlike the previous two.

### 3.1 Deriving Nadaya-Watson estimator

Let $(X_i, Y_i)$ be independent pairs of random variables such that we have $Y_i = m(X_i) + \epsilon_i$ given that $E(\epsilon_i \mid X_i = x) = 0$

$$\hat{m}(x) = E(Y \mid X = x) = \int y f_{Y|X}(y \mid x) \mathrm{d}y = \frac{\int y f_{X,Y}(x,y) \mathrm{d}y}{f_X(x)} \tag{24}$$

The marginal $(f_X)$ and joint $(f_{X,Y})$ densities are estimated using the following kernel density estimates.

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \text{ and } \hat{f}_{X,Y}(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right) \tag{25}$$

Now use Eq. (25) in Eq. (24) and simplify in order to get the following

$$\hat{m}(x) = \frac{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\int_{\mathbb{R}} yK\left(\frac{y-y_i}{h}\right)\mathrm{d}y}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} \tag{26}$$

Now to simplify the integral, we use a change of variable $\phi = \frac{y-Y_i}{h}$ to get

$$\int_{\mathbb{R}} yK\left(\frac{y - y_i}{h}\right)\mathrm{d}y = \int_{\mathbb{R}} h\left(h\phi + Y_i\right) K(\phi)\mathrm{d}\phi = \int_{\mathbb{R}} \left(h^2 + hY_i\right) K(\phi)\mathrm{d}\phi = 0 + hY_i \tag{27}$$

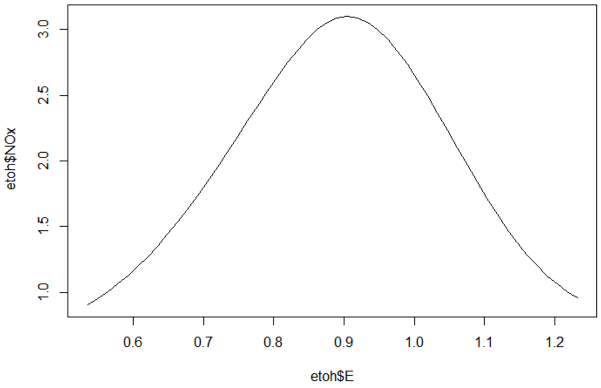
The step in Eq. (27) follows from kernel properties. Substituting Eq. (27) in Eq. (26) gives the desired result.

---

Now, we have R-codes using which we will see how to smooth the ethanol data using N-W estimator and a Gaussian kernel. The smoothed data is in Fig. 8 a.

```r
# "npreg" command implements the N-W kernel regression estimator
etoh$NOx <- etoh$NOx[order(etoh$E)]
etoh$E <- sort(etoh$E) # "npreg" requires x variable data to be sorted
library(np)
etoh.npreg <- npreg(bws = 0.09 , txdat = etoh$E , tydat = etoh$NOx)
plot(etoh.npreg)
```

---



(a) ***Ethanol data smoothed using N-W estimator and a Gaussian kernel***

*Figure 8*