Kernal density estimation :- (KDE)

* KDE is a non parametric way to estimate the probability density function of a random Variable.

* Non parametric way means the distribution will not follow any specified parameter such as mean, median etc..., simply defined as distribution - free way.

* KDE is a fundamental data smoothing problem where inferences abt the population were made based on a finite data sample.

what is Kernal ?

* Kernal of PDFs or PMFs is the form of PDFs ~~as~~ and PMFs in which any factors that are not functions of any of the Variables in the domain are omitted.

* Simply it means ~~as~~ in PDFs the factors which doesn't involved the domain that is the Variable will removed because it will create unnecassary PDFs.

* Kernal is a weighing function

PDFs of normal distribution: $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Its associated Kernal is

$$p(x \mid \mu; \sigma^2) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

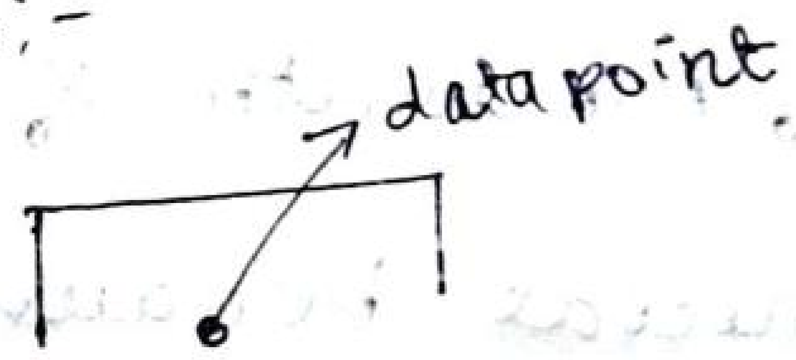
Here $1/\sqrt{2\pi\sigma^2}$ is removed ~~bee~~ even though it has σ^2, ~~it~~ because it is not the function of x.

* A Kernal is a non-negative real valued integerable function K.

If ~~It~~ K is symetric $K(x) = K(-x)$

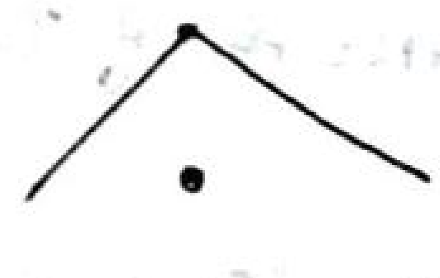
Some Kernal functions:-

$K(u) = 1/2$



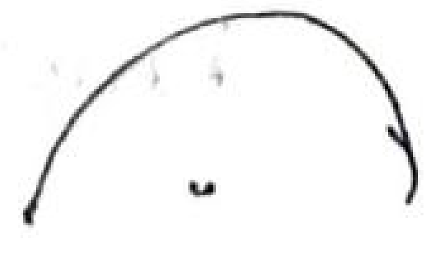
Box car

$K(u) = (1 - |u|)$



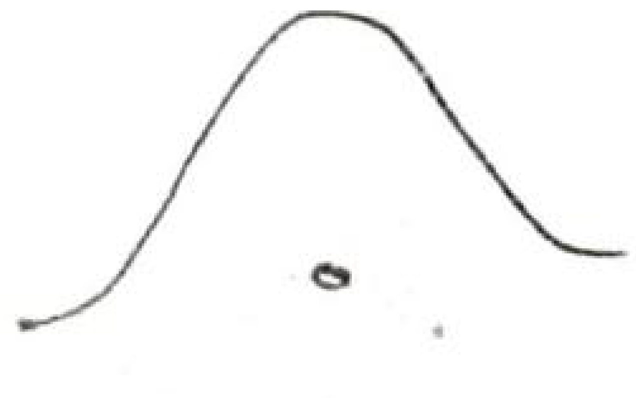
Triangular

$K(u) = \frac{3}{4}(1 - u^2)$



Epanechnikov

$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$



Gaussian

and many more

Graphical Explanation with respect to data

* Suppose there are 5 data points in the dataset (Just for example)

* on every data point x_i, we place a Kernal function K

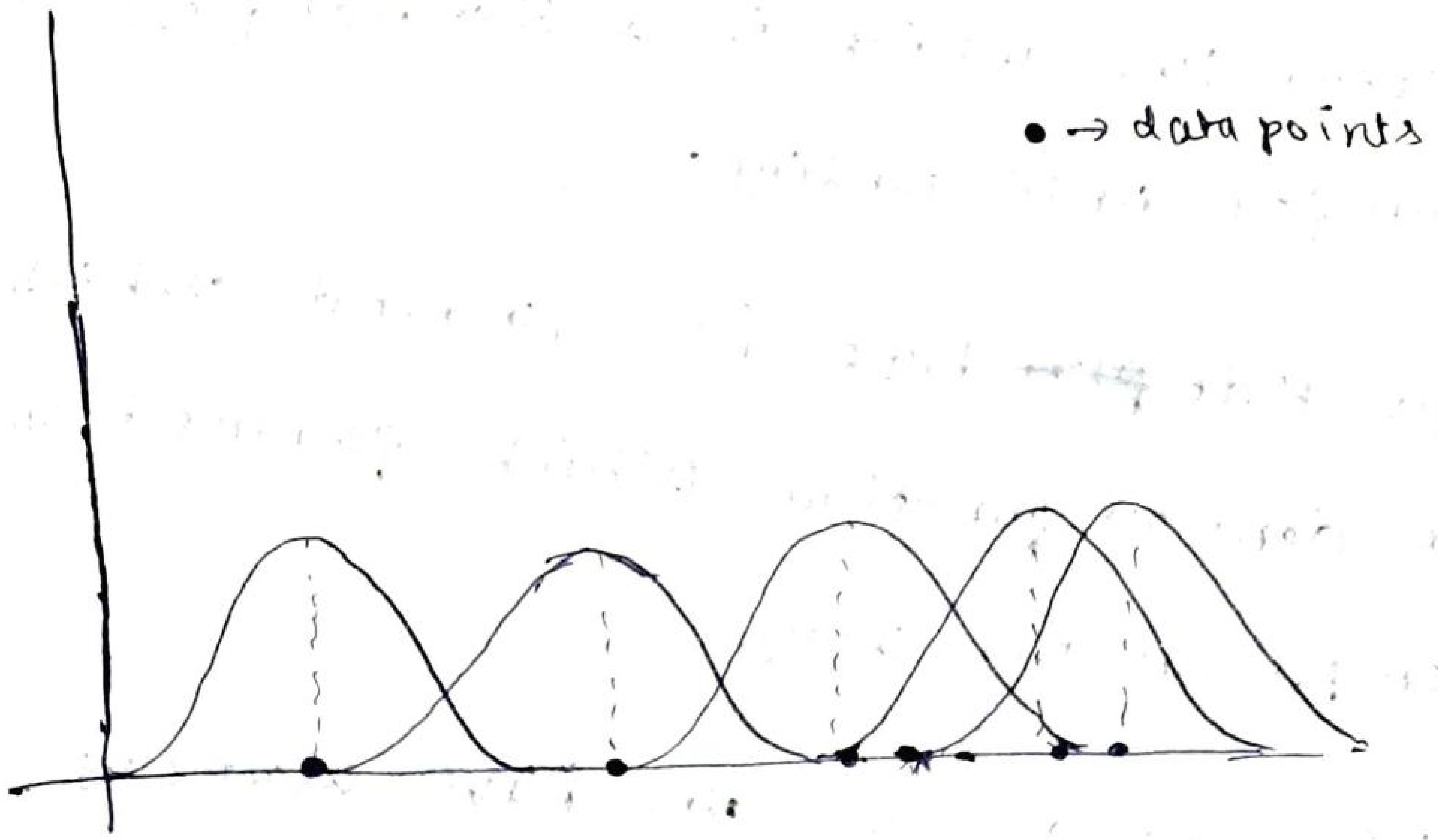
* The Kernal density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i)$$

* Since there are more Kernal functions like Box car, gaussian, triangular etc..., the choosing of the Kernal is not crucial.

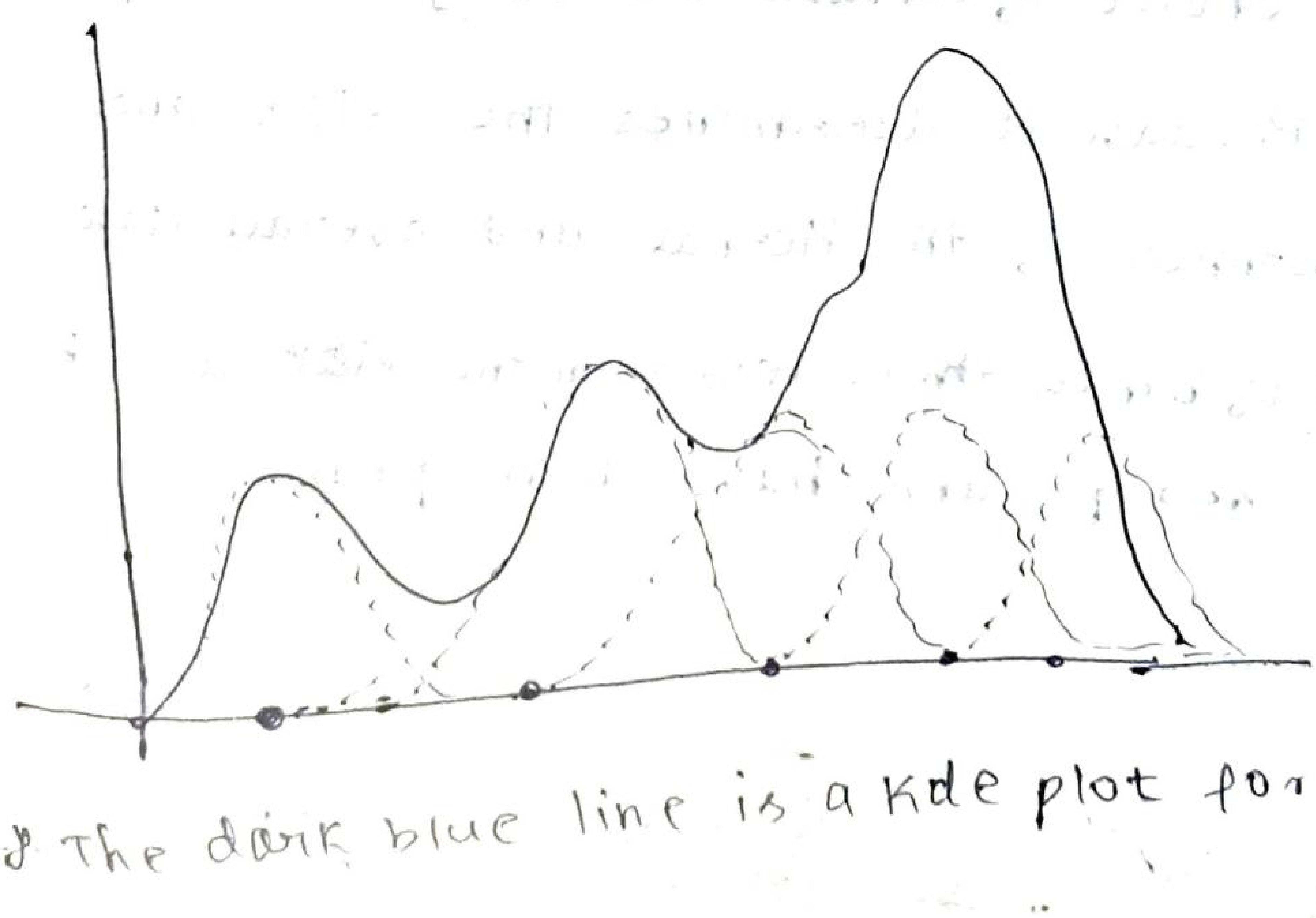
* But choosing of bandwidth of Kernal is more important and crucial because it determines the shape and smoothness of the Kde.

* ~~there~~ But we don't know exactly for ~~reason~~ what reason choosing of Kernal is not crucial.



→ Since i have said for example 5 data points I have centered a Gaussian Kernal function for each datapoint. The Kernal have some bandwidth h.

→ Now ~~it~~ all the Kernals or Kernal functions should be joined are added up to form a Kernal density estimate



→ The dark blue line is a kde plot for a feature.

* dotted line which is a Kernal, just shown for understanding.

* The kde ~~plot~~ plot is formed which is a density function which formed in a non parametric way.

* This kind of plot we have already seen in distplot (Histogram + Kdeplot).

* Simple if we join the edges or centres of Histogram we will get kde-plot.

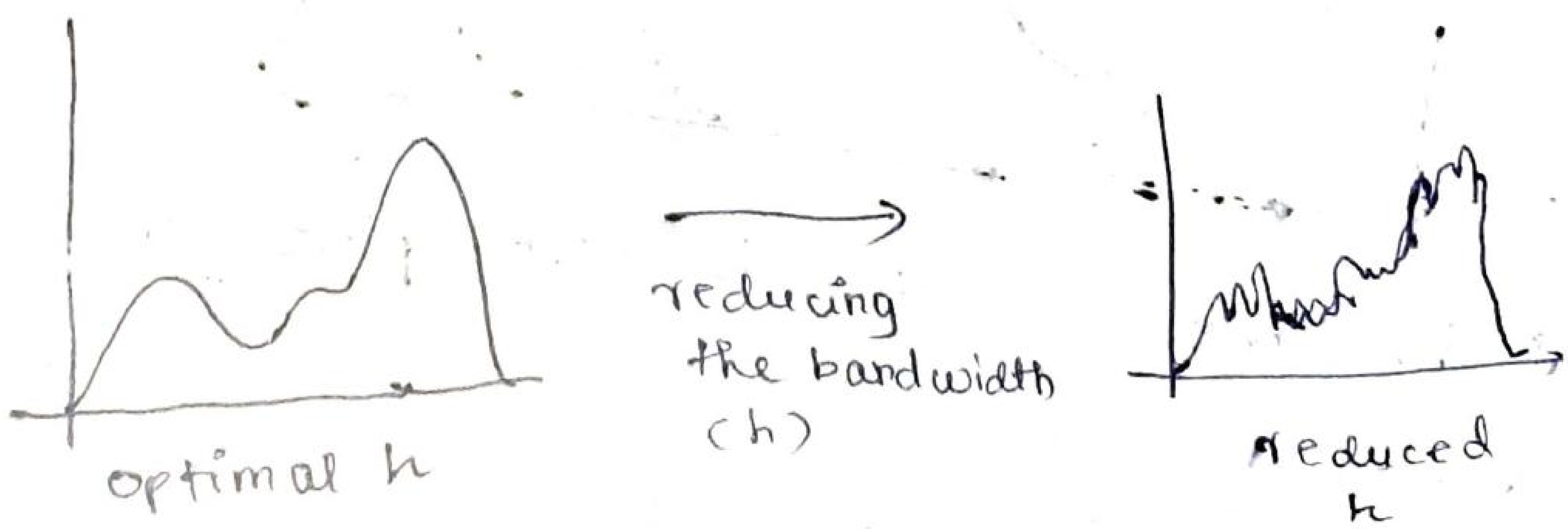
* It is mainly known for its smoothness.

Choice of bandwidth of Kernals :-

* Choice of bandwidth is very crucial point.

* Because it determines the size and smoothness of the Kernal and overall Kde.

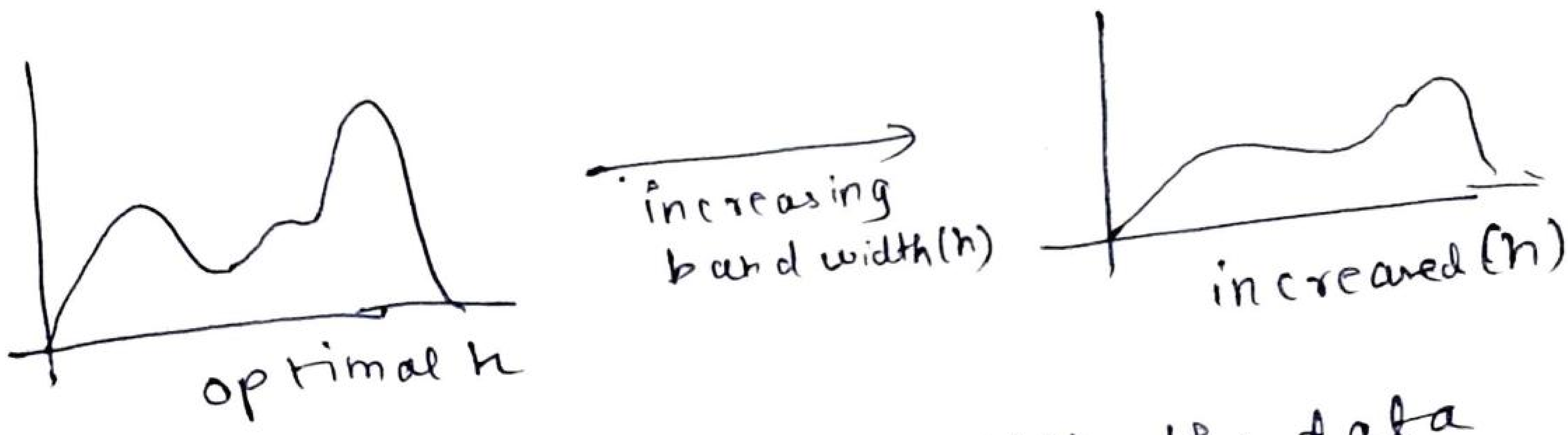
* If bandwidth is less then the Kde will be more sharps and have more peaks



Scanned by TapScanner

* If I reduce the bandwidth it ~~may~~ will increase variance of Kde plot.

* If I increase the bandwith it will smooth the plot



* If I increase the bandwidth the data may loose its significance because due ~~the~~ to the ~~high bar~~ increased bandwidth the data~~set~~ may loose its modality since its peaks are smoothening.

* The optimal bandwidth h can be computed using Silverman's thumb rule. ~~It~~

* It assumes the data is normally distributed.

Scanned by TapScanner