A Simulation Study on Kernel Density Estimation

Arijit Naskar
Debarshi Chakraborty
Spandan Ghoshal

Stat Math Unit (SMU),Indian Statistical Institute ,Delhi

May 03,2022

Contents

What is Density Estimation?

- One of the most fundamental and important concept in statistics is the *probability density function* of a random variable X.

What is Density Estimation?

- One of the most fundamental and important concept in statistics is the *probability density function* of a random variable X.
- In practice,when we deal with data in real life, it is too much unrealistic to think that the density f will be known.

What is Density Estimation?

- One of the most fundamental and important concept in statistics is the *probability density function* of a random variable X.
- In practice,when we deal with data in real life, it is too much unrealistic to think that the density f will be known.
- Then a natural question arises that in this kind of scenario , what to do?

What is Density Estimation?

- One of the most fundamental and important concept in statistics is the *probability density function* of a random variable X.
- In practice,when we deal with data in real life, it is too much unrealistic to think that the density f will be known.
- Then a natural question arises that in this kind of scenario , what to do?
- The main motive of this presentation is an attempt to answer this question.

What is Density Estimation?

- One of the most fundamental and important concept in statistics is the *probability density function* of a random variable X.
- In practice,when we deal with data in real life, it is too much unrealistic to think that the density f will be known.
- Then a natural question arises that in this kind of scenario , what to do?
- The main motive of this presentation is an attempt to answer this question.
- Let's see ☺.

What is Density Estimation?

- Suppose,we have a set of observed data points assumed to be a sample from a distribution with unknown pdf f .

What is Density Estimation?

- Suppose,we have a set of observed data points assumed to be a sample from a distribution with unknown pdf f .
- Density Estimation is a method to construct an estimate of this unknown density f .

What is Density Estimation?

- Suppose,we have a set of observed data points assumed to be a sample from a distribution with unknown pdf f .
- Density Estimation is a method to construct an estimate of this unknown density f .
- One approach can be **parametric density estimation** i.e. we assume that the distribution from where the data is drawn,is a known parametric family and we just have to find the estimates of the parameters which characterize the distribution.

What is Density Estimation?

- Suppose,we have a set of observed data points assumed to be a sample from a distribution with unknown pdf f .
- Density Estimation is a method to construct an estimate of this unknown density f .
- One approach can be **parametric density estimation** i.e. we assume that the distribution from where the data is drawn,is a known parametric family and we just have to find the estimates of the parameters which characterize the distribution.
- But, the form of the density is also seldom known to us.

What is Density Estimation?

- Suppose,we have a set of observed data points assumed to be a sample from a distribution with unknown pdf f .
- Density Estimation is a method to construct an estimate of this unknown density f .
- One approach can be **parametric density estimation** i.e. we assume that the distribution from where the data is drawn,is a known parametric family and we just have to find the estimates of the parameters which characterize the distribution.
- But, the form of the density is also seldom known to us.
- Hence,we make our assumptions less rigid about the distribution of the observed data.

What is Density Estimation?

- Suppose,we have a set of observed data points assumed to be a sample from a distribution with unknown pdf f .
- Density Estimation is a method to construct an estimate of this unknown density f .
- One approach can be **parametric density estimation** i.e. we assume that the distribution from where the data is drawn,is a known parametric family and we just have to find the estimates of the parameters which characterize the distribution.
- But, the form of the density is also seldom known to us.
- Hence,we make our assumptions less rigid about the distribution of the observed data.
- This approach is **nonparametric density estimation.**

What is Density Estimation?

- Suppose,we have a set of observed data points assumed to be a sample from a distribution with unknown pdf f .
- Density Estimation is a method to construct an estimate of this unknown density f .
- One approach can be **parametric density estimation** i.e. we assume that the distribution from where the data is drawn,is a known parametric family and we just have to find the estimates of the parameters which characterize the distribution.
- But, the form of the density is also seldom known to us.
- Hence,we make our assumptions less rigid about the distribution of the observed data.
- This approach is **nonparametric density estimation.**
- The focus of discussion will be on this nonparametric approach.

An Approach to Visualize - Histograms

- To get an idea about any kind of data,it is very helpful if we can visualize it properly.

An Approach to Visualize - Histograms

- To get an idea about any kind of data,it is very helpful if we can visualize it properly.
- In our problem,we can serve that purpose with *Histograms.*

An Approach to Visualize - Histograms

- To get an idea about any kind of data,it is very helpful if we can visualize it properly.
- In our problem,we can serve that purpose with *Histograms.*
- Infact,histogram is the oldest and most widely used density estimator.

An Approach to Visualize - Histograms

- To get an idea about any kind of data,it is very helpful if we can visualize it properly.
- In our problem,we can serve that purpose with *Histograms.*
- Infact,histogram is the oldest and most widely used density estimator.
- Let , our origin $= x_0$ and bin width $= h$,define the bins of our histogram to be the intervals $[x_0 + mh, x_0 + (m + 1)h)$,for $+$ve and -ve integers m.

An Approach to Visualize - Histograms

- To get an idea about any kind of data,it is very helpful if we can visualize it properly.
- In our problem,we can serve that purpose with *Histograms.*
- Infact,histogram is the oldest and most widely used density estimator.
- Let , our origin $= x_0$ and bin width $= h$,define the bins of our histogram to be the intervals $[x_0 + mh, x_0 + (m+1)h)$,for $+$ve and -ve integers m.
- The histogram is hence defined as :

$$\hat{f}(x) = \frac{1}{nh}(\text{no of } X_i \text{ in the same bin as x})$$

An Approach to Visualize - Histograms

- To get an idea about any kind of data,it is very helpful if we can visualize it properly.
- In our problem,we can serve that purpose with *Histograms*.
- Infact,histogram is the oldest and most widely used density estimator.
- Let , our origin $= x_0$ and bin width $= h$,define the bins of our histogram to be the intervals $[x_0 + mh, x_0 + (m+1)h)$,for +ve and -ve integers m.
- The histogram is hence defined as :

$$\hat{f}(x) = \frac{1}{nh}(\text{no of } X_i \text{ in the same bin as x})$$

- To construct the histogram,we have to choose both origin x_0 and binwidth h , but it is the binwidth h that plays the key role in the amount of smoothing.

An Approach to Visualize - Histograms

- To get an idea about any kind of data,it is very helpful if we can visualize it properly.
- In our problem,we can serve that purpose with *Histograms.*
- Infact,histogram is the oldest and most widely used density estimator.
- Let , our origin $= x_0$ and bin width $= h$,define the bins of our histogram to be the intervals $[x_0 + mh, x_0 + (m+1)h)$,for $+$ve and -ve integers m.
- The histogram is hence defined as :

$$\hat{f}(x) = \frac{1}{nh}(\text{no of } X_i \text{ in the same bin as x})$$

- To construct the histogram,we have to choose both origin x_0 and binwidth h , but it is the binwidth h that plays the key role in the amount of smoothing.
- We will see this fact verified shortly through our simulation.

Histograms by varying number of bins

- A sample $X_1, X_2, \ldots, X_n$ of size $n = 200$ from the density

$$f(x) = \phi(2(x-1)) + \frac{1}{2}\phi(x-5)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is drawn.

Histograms by varying number of bins

- A sample $X_1, X_2, \ldots, X_n$ of size $n = 200$ from the density
$$f(x) = \phi(2(x-1)) + \frac{1}{2}\phi(x-5)$$
where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is drawn.
- Then we plot the histograms of this sampled data for different number of bins.

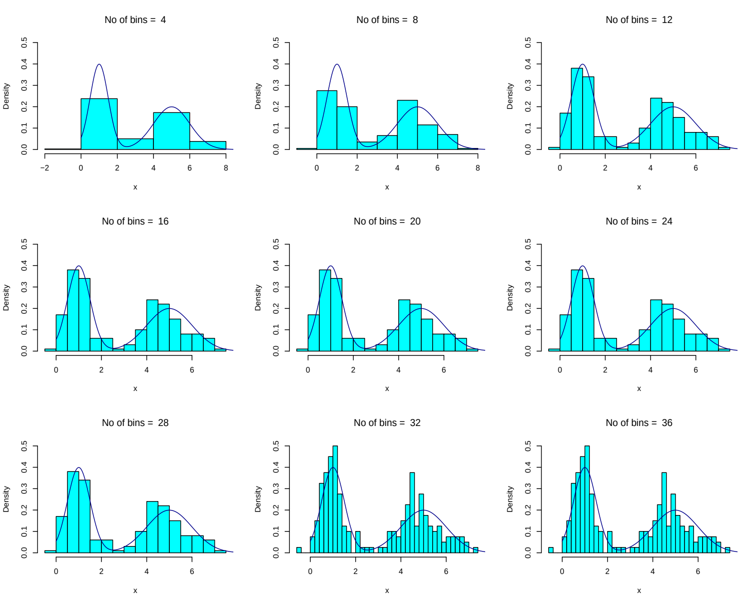
Histograms by varying number of bins

- A sample $X_1, X_2, ..., X_n$ of size $n = 200$ from the density

$$f(x) = \phi(2(x-1)) + \frac{1}{2}\phi(x-5)$$

  where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is drawn.
- Then we plot the histograms of this sampled data for different number of bins.
- See all the plots for varying *number of bins* or equivalently *varying binwidth* on the same panel to catch the difference:-

Histograms for varying number of bins



Observations

- If number of bins is too small (equivalently binwidth is large),the histogram is not able to estimate or *mimic* the true distribution very well , in our case , it underestimates the values of the true density at some points.

Observations

- If number of bins is too small (equivalently binwidth is large),the histogram is not able to estimate or *mimic* the true distribution very well , in our case , it underestimates the values of the true density at some points.
- Similarly, if we take the number of bins to be too large , then also the histogram does not perform good enough.In contrary to the previous situation,it overestimates the values of the true density at some points.

Observations

- If number of bins is too small (equivalently binwidth is large),the histogram is not able to estimate or *mimic* the true distribution very well , in our case , it underestimates the values of the true density at some points.
- Similarly, if we take the number of bins to be too large , then also the histogram does not perform good enough.In contrary to the previous situation,it overestimates the values of the true density at some points.
- If we take a look at the plots in the middle,they give the best estimate among all the plots available.

Observations

- If number of bins is too small (equivalently binwidth is large),the histogram is not able to estimate or *mimic* the true distribution very well , in our case , it underestimates the values of the true density at some points.
- Similarly, if we take the number of bins to be too large , then also the histogram does not perform good enough.In contrary to the previous situation,it overestimates the values of the true density at some points.
- If we take a look at the plots in the middle,they give the best estimate among all the plots available.
- Hence,our takeaway from this visualization is that neither we should decrease the binwidth arbitrarily nor should we make the bins too thick.

Observations

- If number of bins is too small (equivalently binwidth is large),the histogram is not able to estimate or *mimic* the true distribution very well , in our case , it underestimates the values of the true density at some points.
- Similarly, if we take the number of bins to be too large , then also the histogram does not perform good enough.In contrary to the previous situation,it overestimates the values of the true density at some points.
- If we take a look at the plots in the middle,they give the best estimate among all the plots available.
- Hence,our takeaway from this visualization is that neither we should decrease the binwidth arbitrarily nor should we make the bins too thick.
- Thus,one of our main challenges is to find the **best choice of** h or **optimum binwidth** h_{opt} for our estimator.

Observations

- If number of bins is too small (equivalently binwidth is large),the histogram is not able to estimate or *mimic* the true distribution very well , in our case , it underestimates the values of the true density at some points.
- Similarly, if we take the number of bins to be too large , then also the histogram does not perform good enough.In contrary to the previous situation,it overestimates the values of the true density at some points.
- If we take a look at the plots in the middle,they give the best estimate among all the plots available.
- Hence,our takeaway from this visualization is that neither we should decrease the binwidth arbitrarily nor should we make the bins too thick.
- Thus,one of our main challenges is to find the **best choice of** h or **optimum binwidth** h_{opt} for our estimator.
- We will somehow try to achieve this goal through different approaches.

Drawbacks of Histogram

- Subjective choice of origin.

Drawbacks of Histogram

- Subjective choice of origin.
- Bin width can be subjective.

Drawbacks of Histogram

- Subjective choice of origin.
- Bin width can be subjective.
- Discontinuity.

Drawbacks of Histogram

- Subjective choice of origin.
- Bin width can be subjective.
- Discontinuity.
- Derivatives of density function cannot be estimated.

Drawbacks of Histogram

- Subjective choice of origin.
- Bin width can be subjective.
- Discontinuity.
- Derivatives of density function cannot be estimated.
- Difficult to extend this idea to high dimensions.

The Naive Estimator

- By definition,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

The Naive Estimator

- By definition,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- For any given h,it is quite intuitive to estimate $P(x - h < X < x + h)$ by the proportion of sample points falling in the interval $(x - h, x + h)$.

The Naive Estimator

- By definition,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- For any given h,it is quite intuitive to estimate $P(x - h < X < x + h)$ by the proportion of sample points falling in the interval $(x - h, x + h)$.
- Hence, a natural estimator is

$$\hat{f}(x) = \frac{1}{2hn}[\text{no. of } X_i \text{ falling in } (x - h, x + h)]$$

The Naive Estimator

- By definition,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- For any given h,it is quite intuitive to estimate $P(x - h < X < x + h)$ by the proportion of sample points falling in the interval $(x - h, x + h)$.
- Hence, a natural estimator is

$$\hat{f}(x) = \frac{1}{2hn}[\text{no. of } X_i \text{ falling in } (x - h, x + h)]$$

- We shall call this estimator the *Naive Estimator.*

The Naive Estimator

- By definition,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- For any given h,it is quite intuitive to estimate $P(x - h < X < x + h)$ by the proportion of sample points falling in the interval $(x - h, x + h)$.
- Hence, a natural estimator is

$$\hat{f}(x) = \frac{1}{2hn}[\text{no. of } X_i \text{ falling in } (x - h, x + h)]$$

- We shall call this estimator the *Naive Estimator.*
- $x - h < X_i < x + h \Rightarrow -1 < \frac{x - X_i}{h} < 1$

The Naive Estimator

- By definition,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- For any given h,it is quite intuitive to estimate $P(x - h < X < x + h)$ by the proportion of sample points falling in the interval $(x - h, x + h)$.
- Hence, a natural estimator is

$$\hat{f}(x) = \frac{1}{2hn}[\text{no. of } X_i \text{ falling in } (x - h, x + h)]$$

- We shall call this estimator the *Naive Estimator.*
- $x - h < X_i < x + h \Rightarrow -1 < \frac{x - X_i}{h} < 1$
- So,to define our naive estimator more mathematically,define

$$w(x) = \frac{1}{2}\boldsymbol{I}_{\{|x|<1\}}$$

The Naive Estimator

- By definition,

$$f(x) = \lim_{h\to 0} \frac{1}{2h} P(x-h < X < x+h)$$

- For any given h,it is quite intuitive to estimate $P(x-h < X < x+h)$ by the proportion of sample points falling in the interval $(x-h, x+h)$.
- Hence, a natural estimator is

$$\hat{f}(x) = \frac{1}{2hn}[\text{no. of } X_i \text{ falling in } (x-h, x+h)]$$

- We shall call this estimator the *Naive Estimator.*
- $x-h < X_i < x+h \Rightarrow -1 < \frac{x-X_i}{h} < 1$
- So,to define our naive estimator more mathematically,define

$$w(x) = \frac{1}{2}\boldsymbol{I}_{\{|x|<1\}}$$

- Then,the naive estimator is defined as:-

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

The Naive Estimator

- By definition,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- For any given h,it is quite intuitive to estimate $P(x - h < X < x + h)$ by the proportion of sample points falling in the interval $(x - h, x + h)$.
- Hence, a natural estimator is

$$\hat{f}(x) = \frac{1}{2hn}[\text{no. of } X_i \text{ falling in } (x - h, x + h)]$$

- We shall call this estimator the *Naive Estimator.*
- $x - h < X_i < x + h \Rightarrow -1 < \frac{x - X_i}{h} < 1$
- So,to define our naive estimator more mathematically,define

$$w(x) = \frac{1}{2} \boldsymbol{I}_{\{|x| < 1\}}$$

- Then,the naive estimator is defined as:-

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

- Due to some drawbacks hence we try to generalize this concept.

Any Improvement??

- Takes care of the origin issue.

Any Improvement??

- Takes care of the origin issue.
- All other drawbacks remain as they are.

Kernel Density Estimator

- We just replace the function w by a **kernel function** K **satisfying** $\int_{-\infty}^{\infty} K(x)dx = 1$.

Kernel Density Estimator

- We just replace the function w by a **kernel function** K **satisfying** $\int_{-\infty}^{\infty} K(x)dx = 1$**.**
- Usually (not always) , K will be a symmetric pdf.

Kernel Density Estimator

- We just replace the function w by a **kernel function** K **satisfying** $\int_{-\infty}^{\infty} K(x)dx = 1$.
- Usually (not always) , K will be a symmetric pdf.
- Assume $\sup_x K(x) \leq M, |x|K(x) \to 0$ as $|x| \to \infty$

Kernel Density Estimator

- We just replace the function w by a **kernel function** K **satisfying** $\int_{-\infty}^{\infty} K(x)dx = 1$**.**
- Usually (not always) , K will be a symmetric pdf.
- Assume $\sup_x K(x) \le M, |x|K(x) \to 0$ as $|x| \to \infty$
- Assume $K(x) = K(-x); x \in \mathbb{R}, \int_{-\infty}^{\infty} x^2 K(x)dx < \infty$

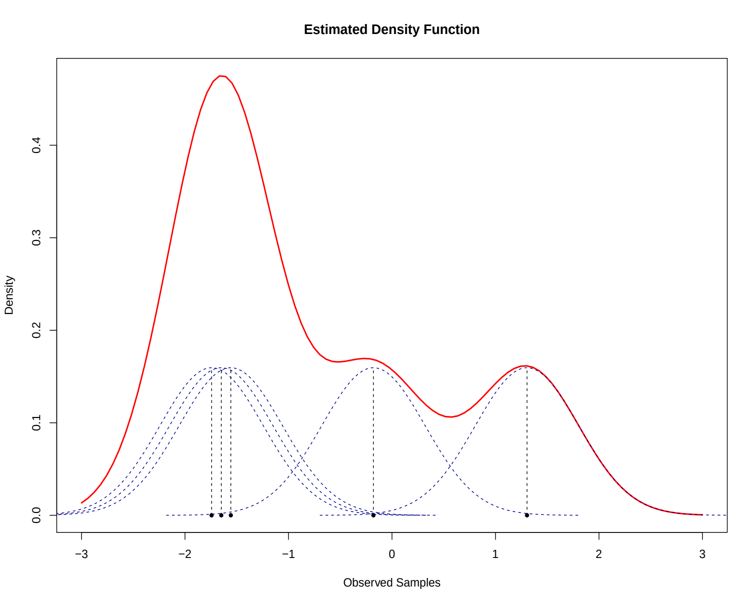
Kernel Density Estimator

- We just replace the function w by a **kernel function** K **satisfying** $\int\limits_{-\infty}^{\infty} K(x)dx = 1$**.**
- Usually (not always) , K will be a symmetric pdf.
- Assume $\sup_x K(x) \leq M, |x|K(x) \to 0$ as $|x| \to \infty$
- Assume $K(x) = K(-x); x \in \mathbb{R}, \int\limits_{-\infty}^{\infty} x^2 K(x)dx < \infty$
- The *Kernel Type Estimator* is thus given by :

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x - X_i}{h_n})$$

$h_n \to 0$ as $n \to \infty$

Demonstrating how KDE Works

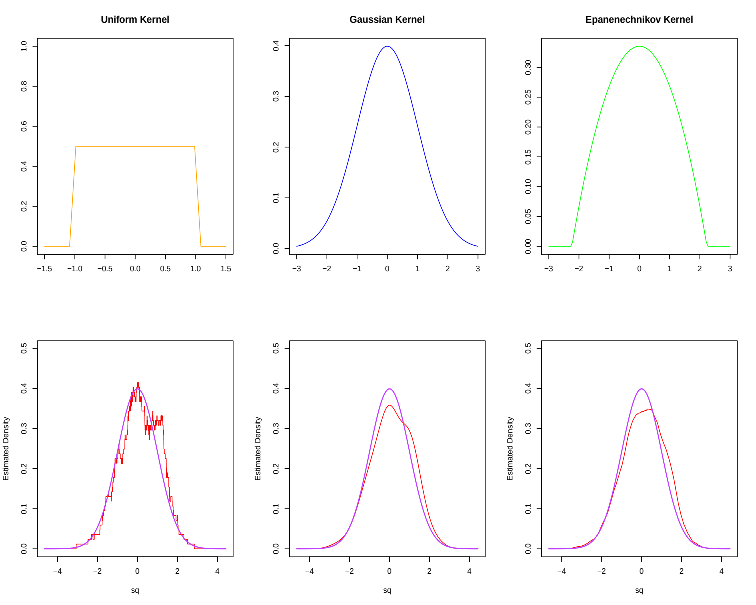


Some examples of Kernel Functions

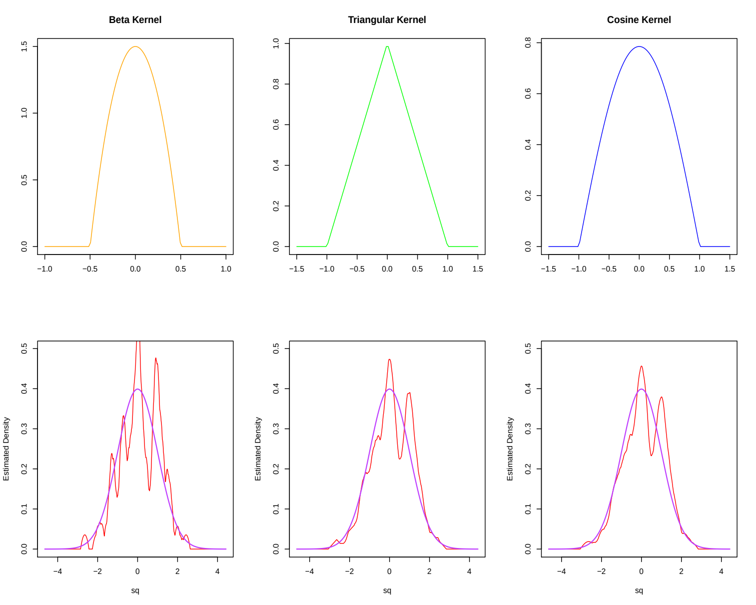
Kernel Names	Functional Form $(K(x))$	Efficiency relative to Epanechnikov Kernel
Uniform	$\frac{1}{2}\boldsymbol{I}_{\lvert x\rvert\leq 1}$	92.9%
Triangular	$(1-\lvert x\rvert)\,\boldsymbol{I}_{\lvert x\rvert\leq 1}$	98.5%
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1-\frac{t^2}{5}\right)\boldsymbol{I}_{\lvert x\rvert\leq\sqrt{5}}$	1
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\boldsymbol{I}_{x\in\mathbb{R}}$	95.1%
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)\boldsymbol{I}_{\lvert x\rvert\leq 1}$	99.9%

Here,we have done a comparative study between these kernels to show our results.To compare how different choices of kernels influence the estimated density, we make the KDE plots for the same observed data and different kernels.

Comparison between different kernels



Comparison between different kernels



Asymmetric kernels-how they behave?

- Though it's usually assumed in KDE literature that the kernel function $K(x)$ is symmetric, it's of no harm to observe what happens if we take $K(x)$ to be asymmetric also.

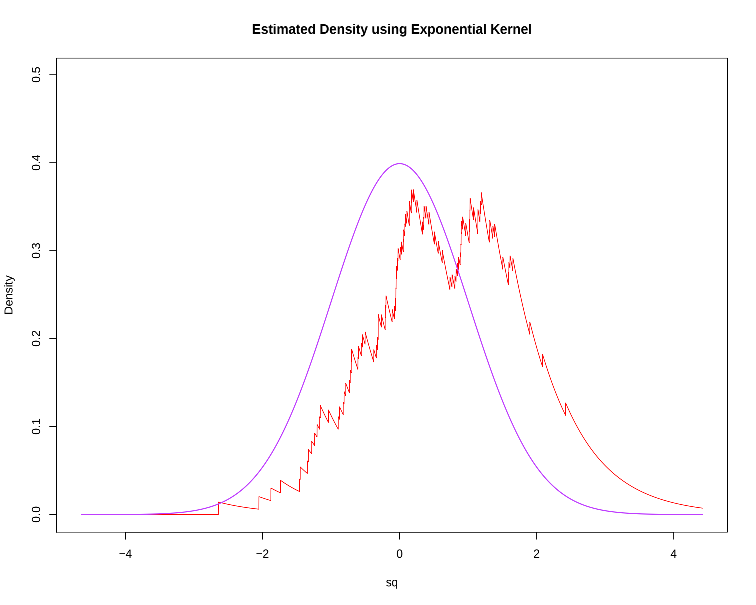
Asymmetric kernels-how they behave?

- Though it's usually assumed in KDE literature that the kernel function $K(x)$ is symmetric, it's of no harm to observe what happens if we take $K(x)$ to be asymmetric also.
- For e.g we can choose our kernel as $K(x) = e^{-x}\boldsymbol{I}_{x\geq 0}$

Asymmetric kernels-how they behave?

- Though it's usually assumed in KDE literature that the kernel function $K(x)$ is symmetric, it's of no harm to observe what happens if we take $K(x)$ to be asymmetric also.
- For e.g we can choose our kernel as $K(x) = e^{-x}\boldsymbol{I}_{x\geq 0}$
- Let's take a look at the plot of the estimated density.

Asymmetric kernels-how they behave?



Observation

- In the previous diagram,we have compared our estimated density to the true density.

Observation

- In the previous diagram,we have compared our estimated density to the true density.
- We have a very very interesting observation to make here.

Observation

- In the previous diagram,we have compared our estimated density to the true density.
- We have a very very interesting observation to make here.
- The estimated kernel underestimates the true density for smaller values of x and for larger values of x, it overestimates the density.

Observation

- In the previous diagram,we have compared our estimated density to the true density.
- We have a very very interesting observation to make here.
- The estimated kernel underestimates the true density for smaller values of x and for larger values of x, it overestimates the density.
- Why does this happen?

Observation

- The reason behind this is the form of the kernel density estimate at any point x is defined as

$$\begin{aligned}
\widehat{f}_n(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \\
&= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\left\{\frac{x-X_i}{h}\geq 0\right\}} \\
&= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\{x\geq X_i\}}
\end{aligned}$$

Observation

- The reason behind this is the form of the kernel density estimate at any point x is defined as

$$
\begin{aligned}
\widehat{f}_n(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \\
&= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\left\{\frac{x-X_i}{h}\geq 0\right\}} \\
&= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\{x\geq X_i\}}
\end{aligned}
$$

- Note that, the estimated value at x depends on only the observations that lie on the left of x.

Observation

- The reason behind this is the form of the kernel density estimate at any point x is defined as

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$
$$= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\left\{\frac{x-X_i}{h}\geq 0\right\}}$$
$$= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\{x\geq X_i\}}$$

- Note that, the estimated value at x depends on only the observations that lie on the left of x.
- If we move from left to right in the plot, number of $X_i's$ increase.

Observation

- The reason behind this is the form of the kernel density estimate at any point x is defined as

$$\begin{aligned}
\widehat{f}_n(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \\
&= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\left\{\frac{x-X_i}{h}\geq 0\right\}} \\
&= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\{x\geq X_i\}}
\end{aligned}$$

- Note that, the estimated value at x depends on only the observations that lie on the left of x.
- If we move from left to right in the plot, number of $X_i's$ increase.
- As a consequence they contribute more and more to the estimated value.

Observation

- The reason behind this is the form of the kernel density estimate at any point x is defined as

$$\begin{aligned}\widehat{f}_n(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\\ &= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\left\{\frac{x-X_i}{h}\geq 0\right\}}\\ &= \frac{1}{nh}\sum_{i=1}^{n} e^{-\frac{x-X_i}{h}} \boldsymbol{I}_{\{x\geq X_i\}}\end{aligned}$$

- Note that, the estimated value at x depends on only the observations that lie on the left of x.
- If we move from left to right in the plot, number of $X_i's$ increase.
- As a consequence they contribute more and more to the estimated value.
- On the left portion, lesser number of points contribute which results in such a type of estimated density.

Kernel-Smoothed Cumulative Distribution Function

Since, we define the kernel density estimator as :-

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

where $K(.)$ is the kernel function, an obvious extension of this idea is to make smoothed estimators of CDF as :-

$$\begin{aligned}
\widehat{F}_n(x) &= \int_{-\infty}^{x} \widehat{f}_n(t)\, dt \\
&= \int_{-\infty}^{x} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) dt \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{x} \frac{1}{h} K\left(\frac{t - X_i}{h}\right) dt \\
&= \frac{1}{n}\sum_{i=1}^{n} \widetilde{K}\left(\frac{x - X_i}{h}\right) dt
\end{aligned}$$

Kernel Smoothed CDF Estimators

- where $\widetilde{K}(x) = \int\limits_{-\infty}^{x} K(t)\,dt$ is the cumulative function formed using the kernel function.

Kernel Smoothed CDF Estimators

- where $\widetilde{K}(x) = \int\limits_{-\infty}^{x} K(t)\,dt$ is the cumulative function formed using the kernel function.
- We plot the kernel smoothed estimate CDF for a sample from standard normal distribution for different choices of smoothing parameter h.

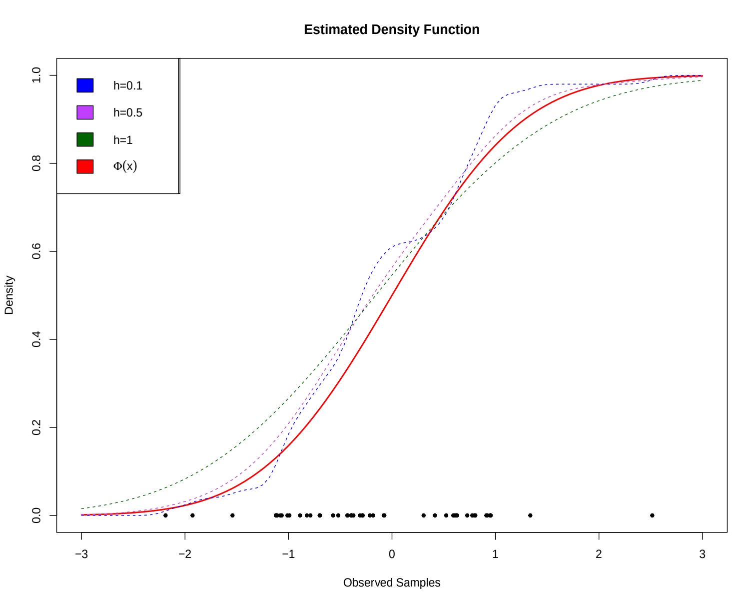
Kernel Smoothed CDF Estimators

- where $\widetilde{K}(x) = \int\limits_{-\infty}^{x} K(t)\,dt$ is the cumulative function formed using the kernel function.
- We plot the kernel smoothed estimate CDF for a sample from standard normal distribution for different choices of smoothing parameter h.
- We also plot the population CDF for standard normal distribution with this.

Kernel Smoothed CDF Estimators

- where $\widetilde{K}(x) = \int\limits_{-\infty}^{x} K(t)\,dt$ is the cumulative function formed using the kernel function.
- We plot the kernel smoothed estimate CDF for a sample from standard normal distribution for different choices of smoothing parameter h.
- We also plot the population CDF for standard normal distribution with this.
- Let's take a look at the plot for different choices of bandwidths.

Kernel Smoothed CDF Estimators



Subjective Choice

- A natural method for choosing the smoothing parameter is to plot out several curves and choose the estimate that is most in accordance with one's prior ideas about the density.

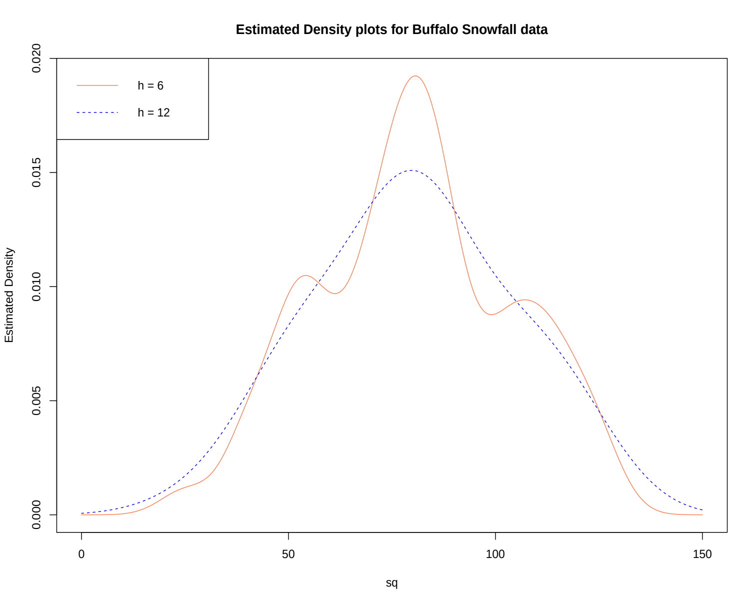
Subjective Choice

- A natural method for choosing the smoothing parameter is to plot out several curves and choose the estimate that is most in accordance with one's prior ideas about the density.
- For many applications this approach will be perfectly satisfactory. Indeed, the process of examining several plots of the data, all smoothed by different amounts, may well give more insight into the data than merely considering a single automatically produced curve.

Subjective Choice

- A natural method for choosing the smoothing parameter is to plot out several curves and choose the estimate that is most in accordance with one's prior ideas about the density.
- For many applications this approach will be perfectly satisfactory. Indeed, the process of examining several plots of the data, all smoothed by different amounts, may well give more insight into the data than merely considering a single automatically produced curve.
- Consider, as an example, the estimate given in following plot, the data underlying these estimates are the amounts of winter snowfall (in inches) at Buffalo, New York, for each of the 63 winters from 1910/11 to 1972/73.

Kernel estimates for annual snowfall data



Observation

- It can be seen from the plot that varying the smoothing parameter yields essentially two possible explanations of the data, either a roughly normal distribution or a trimodal curve suggesting a mixture of three populations approximately in the ratio 1:3:1.

Observation

- It can be seen from the plot that varying the smoothing parameter yields essentially two possible explanations of the data, either a roughly normal distribution or a trimodal curve suggesting a mixture of three populations approximately in the ratio 1:3:1.
- For many purposes, particularly for model and hypothesis generation, it is by no means unhelpful for the statistician to supply the scientist with a range of possible presentations of the data. A choice between the two alternative models suggested by our figures is a very useful step forward from the enormous number of possible explanations that could conceivably be considered.

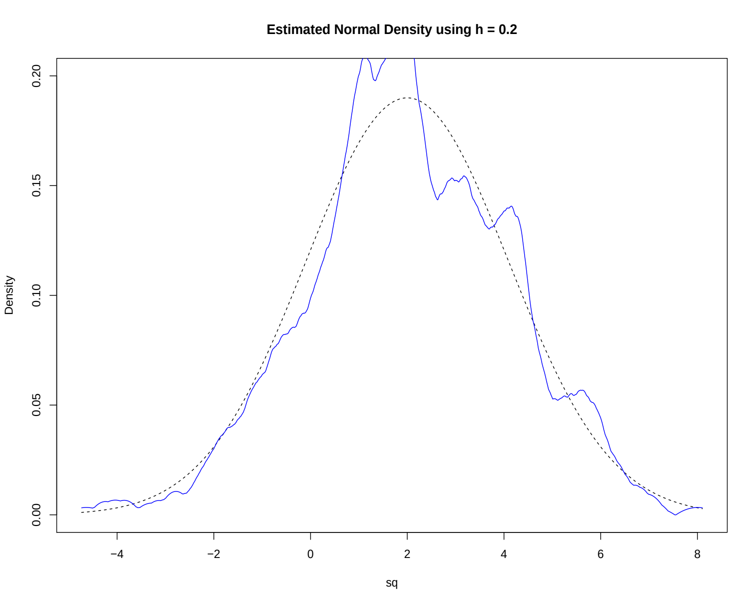
Reference to a standard distribution

If we assume a standard family of distributions while estimating the density, then we can obtain an exact expression of the optimal bin width h_{opt}. For example if we assume the density to be normally distributed with variance σ^2, and if we use gaussian kernel for estimation (i.e. $K(x) = \phi(x)$), then we get :-

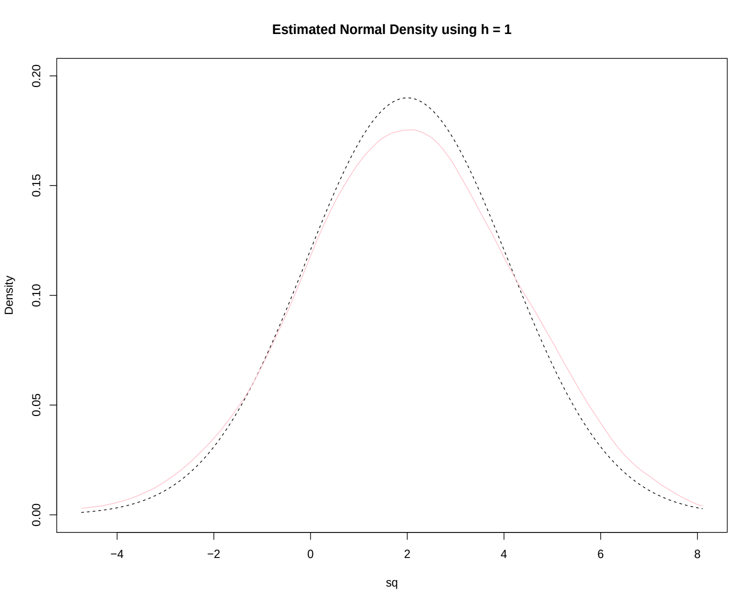
$$
\begin{aligned}
h_{\text{opt}} &= (4\pi)^{-1/10} \left(\frac{3}{8} \pi^{-1/2} \right)^{-1/5} \sigma n^{-1/5} \\
&= \left(\frac{4}{3} \right)^{1/5} \sigma n^{-1/5} = 1.06 \sigma n^{-1/5}
\end{aligned}
$$

We draw a sample of size $n = 500$ from a $N\left(0, \sigma^2 = 4.41\right)$ distribution which gives $h_{\text{opt}} \approx 0.628$.

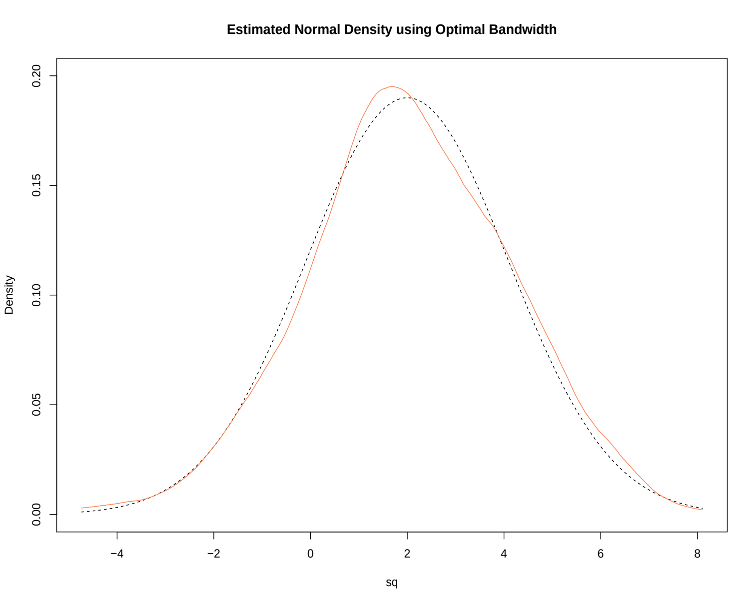
Effect of Undersmoothing



Effect of Oversmoothing



Using optimal bandwidth



Least Squares cross validation

- Assume $X_1, X_2, ..., X_n$ are iid random variable having unknown density f.

Least Squares cross validation

- Assume $X_1, X_2, ..., X_n$ are iid random variable having unknown density f.
- Let $\widehat{f}(x)$ is any estimator of the density $f(x)$.

Least Squares cross validation

- Assume $X_1, X_2, ..., X_n$ are iid random variable having unknown density f.
- Let $\widehat{f}(x)$ is any estimator of the density $f(x)$.
- The integrated squared error of $\widehat{f}(x)$ is

$$\int (\hat{f}(x) - f(x))^2 dx = \int (f\hat{(}x))^2 dx - 2 \int \hat{f}(x) f(x) dx + \int (f(x))^2 dx$$

Least Squares cross validation

- Assume $X_1, X_2, ..., X_n$ are iid random variable having unknown density f.
- Let $\widehat{f}(x)$ is any estimator of the density $f(x)$.
- The integrated squared error of $\widehat{f}(x)$ is

$$\int (\hat{f}(x) - f(x))^2 dx = \int (\hat{f(x)})^2 dx - 2 \int \hat{f}(x) f(x) dx + \int (f(x))^2 dx$$

- The last term doesn't depend on $\hat{f}(x)$ so we want to minimize,

$$R(\hat{f}) = \int (\hat{f(x)})^2 dx - 2 \int \hat{f}(x) f(x) dx$$

Least Squares cross validation

- Assume $X_1, X_2, ..., X_n$ are iid random variable having unknown density f.
- Let $\widehat{f}(x)$ is any estimator of the density $f(x)$.
- The integrated squared error of $\widehat{f}(x)$ is

$$\int (\hat{f}(x) - f(x))^2 dx = \int (f\hat{(x)})^2 dx - 2 \int \hat{f}(x) f(x) dx + \int (f(x))^2 dx$$

- The last term doesn't depend on $\hat{f}(x)$ so we want to minimize,

$$R(\hat{f}) = \int (f\hat{(x)})^2 dx - 2 \int \hat{f}(x) f(x) dx$$

- The basic principle of least square cross validation is to construct a estimator of $R(\hat{f})$ and minimise the estimator with respect to h.

Least Squares cross validation

- $M_0(h)$ is an estimator of $R(\hat{f})$, Which is defined as

$$M_o(h) = \int f(\hat{x})^2 dx - \frac{2}{n} \Sigma \hat{f}_{-i}(X_i)$$

Least Squares cross validation

- $M_0(h)$ is an estimator of $R(\hat{f})$, Which is defined as

$$M_o(h) = \int f(\hat{x})^2 dx - \frac{2}{n} \Sigma \hat{f}_{-i}(X_i)$$

- Since we can think of the histogram as an estimator of the density function as :-

$$\widehat{f_n}(x) = \frac{1}{nh} (\text{no of } X_i \text{ in the same bin as } x)$$

Least Squares cross validation

- $M_0(h)$ is an estimator of $R(\hat{f})$, Which is defined as

$$M_o(h) = \int f(\hat{x})^2 dx - \frac{2}{n} \Sigma \hat{f}_{-i}(X_i)$$

- Since we can think of the histogram as an estimator of the density function as :-

$$\widehat{f_n}(x) = \frac{1}{nh} (\text{no of } X_i \text{ in the same bin as } x)$$

- we can also find an approximate value of the optimal binwidth for histogram using cross validation method.

Least Squares cross validation

- $M_0(h)$ is an estimator of $R(\hat{f})$, Which is defined as

$$M_o(h) = \int f(\hat{x})^2 dx - \frac{2}{n}\Sigma \hat{f}_{-i}(X_i)$$

- Since we can think of the histogram as an estimator of the density function as :-

$$\widehat{f_n}(x) = \frac{1}{nh}(\text{no of } X_i \text{ in the same bin as } x)$$

- we can also find an approximate value of the optimal binwidth for histogram using cross validation method.
- Since, the cross-validation estimator of risk is :-

$$M_0(h) = \int \widehat{f_n^2}(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f_{-i}}(X_i)$$

Least Squares cross validation

- $M_0(h)$ is an estimator of $R(\hat{f})$, Which is defined as

$$M_o(h) = \int f(\hat{x})^2 dx - \frac{2}{n}\Sigma\hat{f}_{-i}(X_i)$$

- Since we can think of the histogram as an estimator of the density function as :-

$$\widehat{f_n}(x) = \frac{1}{nh}(\text{no of } X_i \text{ in the same bin as } x)$$

- we can also find an approximate value of the optimal binwidth for histogram using cross validation method.
- Since, the cross-validation estimator of risk is :-

$$M_0(h) = \int \widehat{f_n^2}(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f_{-i}}(X_i)$$

- We will find our optimum value of h by minimising this quantity.

Least Squares cross validation

- We demonstrate the fact using a simulation study where we use our old sample of size 200 from the density

$$f(x) = \phi(2(x-1)) + \frac{1}{2}\phi(x-5)$$

Least Squares cross validation

- We demonstrate the fact using a simulation study where we use our old sample of size 200 from the density

$$f(x) = \phi(2(x-1)) + \frac{1}{2}\phi(x-5)$$

- where $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ is the standard normal density.

Least Squares cross validation

- We demonstrate the fact using a simulation study where we use our old sample of size 200 from the density

$$f(x) = \phi(2(x-1)) + \frac{1}{2}\phi(x-5)$$

- where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the standard normal density.
- Without loss of generality, we transform this observations linearly so that they lie in the range $[0, 1]$ by doing the transformation $y_i = \frac{x_i - \min\limits_{1 \le j \le n}(x_j)}{\max\limits_{1 \le j \le n}(x_j) - \min\limits_{1 \le j \le n}(x_j)}$.

Least Squares cross validation

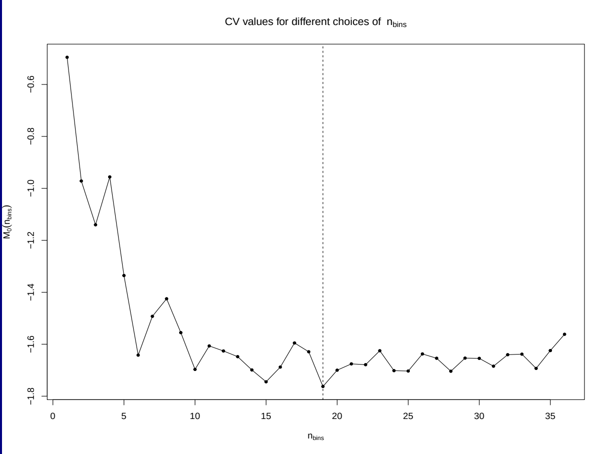
- We demonstrate the fact using a simulation study where we use our old sample of size 200 from the density

$$f(x) = \phi(2(x-1)) + \frac{1}{2}\phi(x-5)$$

- where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the standard normal density.
- Without loss of generality, we transform this observations linearly so that they lie in the range $[0,1]$ by doing the transformation $y_i = \frac{x_i - \min\limits_{1\leq j\leq n}(x_j)}{\max\limits_{1\leq j\leq n}(x_j) - \min\limits_{1\leq j\leq n}(x_j)}$.
- We take a range of values of n_{bins} as $\{1,\ldots,40\}$ and then, for each of the choices we calculate $M_0(h)$ where $h = \frac{1}{n_{bins}}$ is the binwidth since we have normalized the range to $[0,1]$. (Hence, $n_{bins}h = 1$.)

Least Squares cross validation

- After calculating the risk estimates, we plot them and the plot looks like this :-



Least Squares cross validation

- So from the plot we can see that the estimated risk is minimum when $n_{bins} = 19 \implies h \approx 0.0526$ and indeed it is quite similar to the heuristic observations, we made earlier.

Likelihood Cross-validation

- $X_1, X_2, .., X_n$are iid observations with density f.$\hat{f}_n$ is a kernel density estimator of f with using kernel K.

Likelihood Cross-validation

- $X_1, X_2, .., X_n$are iid observations with density f.$\hat{f}_n$ is a kernel density estimator of f with using kernel K.
- Suppose we have an observation Y from density f.Then the log likelihood of f is $log f(Y)$.

Likelihood Cross-validation

- $X_1, X_2, .., X_n$are iid observations with density f.$\hat{f}_n$ is a kernel density estimator of f with using kernel K.
- Suppose we have an observation Y from density f.Then the log likelihood of f is $log f(Y)$.
- $\hat{f}_n$ is a parametric family of densities depending on h.

Likelihood Cross-validation

- $X_1, X_2, .., X_n$ are iid observations with density f.$\hat{f}_n$ is a kernel density estimator of f with using kernel K.
- Suppose we have an observation Y from density f.Then the log likelihood of f is $log f(Y)$.
- $\hat{f}_n$ is a parametric family of densities depending on h.
- The log likelihood of h is $log \hat{f}_n(Y)$.

Likelihood Cross-validation

- $X_1, X_2, .., X_n$are iid observations with density f.$\hat{f}_n$ is a kernel density estimator of f with using kernel K.
- Suppose we have an observation Y from density f.Then the log likelihood of f is $log f(Y)$.
- $\hat{f}_n$ is a parametric family of densities depending on h.
- The log likelihood of h is $log \hat{f}_n(Y)$.
- Since Y is not available, we can omit i^{th} of the observation (X_i) from the sample and reconstruct the density estimate $\hat{f}_{-i}$ and the use X_i as Y. This would give the log likelihood as $log \hat{f}_{-i}(X_i)$. We can evaluate $log \hat{f}_{-i}(X_i)$ by omitting $X_1, X_2, ..., X_n$ and average out.

Likelihood Cross-validation

- $X_1, X_2, .., X_n$ are iid observations with density f. $\hat{f}_n$ is a kernel density estimator of f with using kernel K.
- Suppose we have an observation Y from density f. Then the log likelihood of f is $log f(Y)$.
- $\hat{f}_n$ is a parametric family of densities depending on h.
- The log likelihood of h is $log \hat{f}_n(Y)$.
- Since Y is not available, we can omit i^{th} of the observation (X_i) from the sample and reconstruct the density estimate $\hat{f}_{-i}$ and the use X_i as Y. This would give the log likelihood as $log \hat{f}_{-i}(X_i)$. We can evaluate $log \hat{f}_{-i}(X_i)$ by omitting $X_1, X_2, ..., X_n$ and average out.
- The score function is

$$CV(h) = n^{-1} \sum_i log \hat{f}_{-i}(X_i)$$

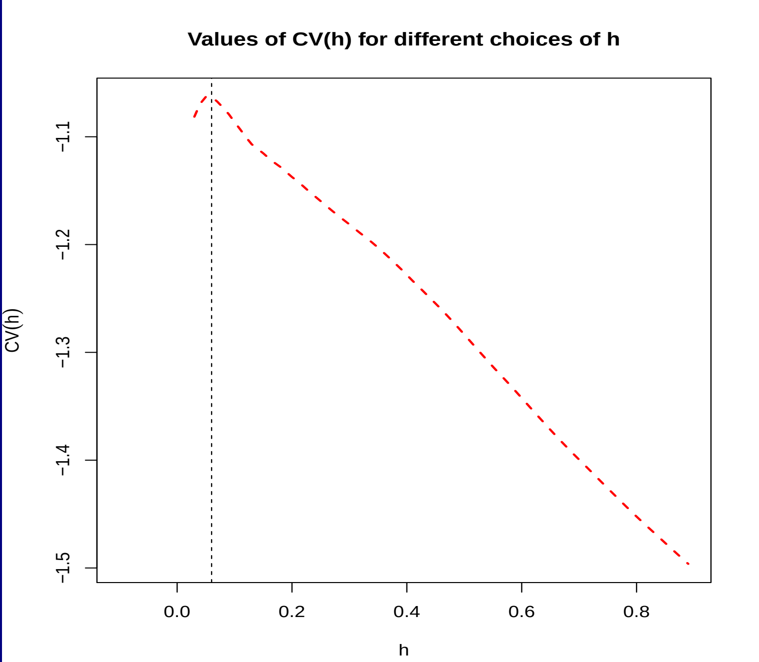
Likelihood Cross-validation

- $X_1, X_2, .., X_n$are iid observations with density f.$\hat{f}_n$ is a kernel density estimator of f with using kernel K.
- Suppose we have an observation Y from density f.Then the log likelihood of f is $log f(Y)$.
- $\hat{f}_n$ is a parametric family of densities depending on h.
- The log likelihood of h is $log \hat{f}_n(Y)$.
- Since Y is not available, we can omit i^{th} of the observation (X_i) from the sample and reconstruct the density estimate $\hat{f_{-i}}$ and the use X_i as Y. This would give the log likelihood as $log \hat{f}_{-i}(X_i)$. We can evaluate $log \hat{f}_{-i}(X_i)$ by omitting $X_1, X_2, ..., X_n$ and average out.
- The score function is

$$CV(h) = n^{-1} \sum_i log \hat{f}_{-i}(X_i)$$

- We can find the optimal value of h by maximising $CV(h)$.

Likelihood Cross-validation



Optimal Bandwidth in case of an exponential data

- In this simulation study, we first draw a random sample of size 100 from a exponential population with mean 1.

Optimal Bandwidth in case of an exponential data

- In this simulation study, we first draw a random sample of size 100 from a exponential population with mean 1.
- Then we use the likelihood cross validation technique to find out the optimal value of h.

Optimal Bandwidth in case of an exponential data

- In this simulation study, we first draw a random sample of size 100 from a exponential population with mean 1.
- Then we use the likelihood cross validation technique to find out the optimal value of h.
- We obtain $h_{opt} \approx 0.06$ and then using this optimal bandwidth we plot the estimated kernel density.

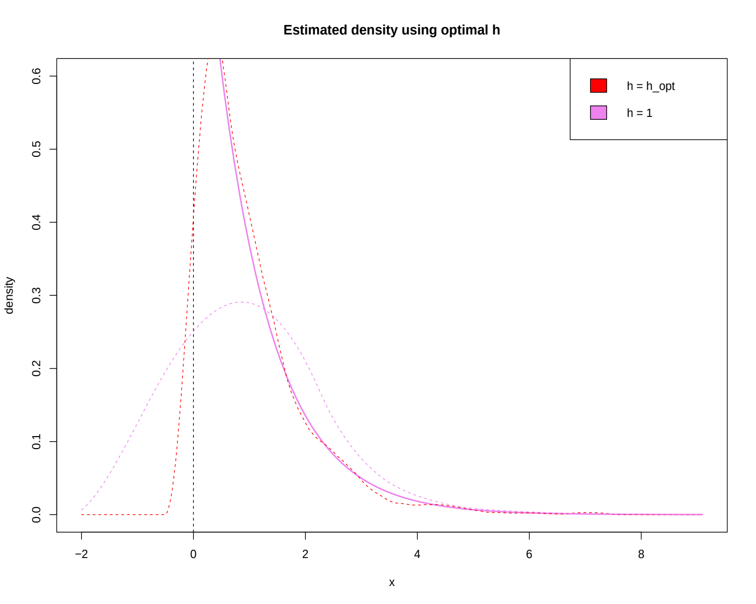
Optimal Bandwidth in case of an exponential data

- In this simulation study, we first draw a random sample of size 100 from a exponential population with mean 1.
- Then we use the likelihood cross validation technique to find out the optimal value of h.
- We obtain $h_{opt} \approx 0.06$ and then using this optimal bandwidth we plot the estimated kernel density.
- In order to understand the importance of choosing optimal bandwidth, we make the plots for some other choice of h like $h = 1$.

Optimal Bandwidth in case of an exponential data

- In this simulation study, we first draw a random sample of size 100 from a exponential population with mean 1.
- Then we use the likelihood cross validation technique to find out the optimal value of h.
- We obtain $h_{opt} \approx 0.06$ and then using this optimal bandwidth we plot the estimated kernel density.
- In order to understand the importance of choosing optimal bandwidth, we make the plots for some other choice of h like $h = 1$.
- Let us have a look at them.

Optimal Bandwidth in case of an exponential data



Observation

- We can clearly see how drastically the estimated density changes if we take choices of h other than the optimal one as it fails to capture the asymptotic nature of the density near 0.

Observation

- We can clearly see how drastically the estimated density changes if we take choices of h other than the optimal one as it fails to capture the asymptotic nature of the density near 0.
- Another important drawback to mention here is that the estimated density is non-zero for negative values of x as well which is occuring due to positive weight that is being assigned due to observations whose values are near zero.

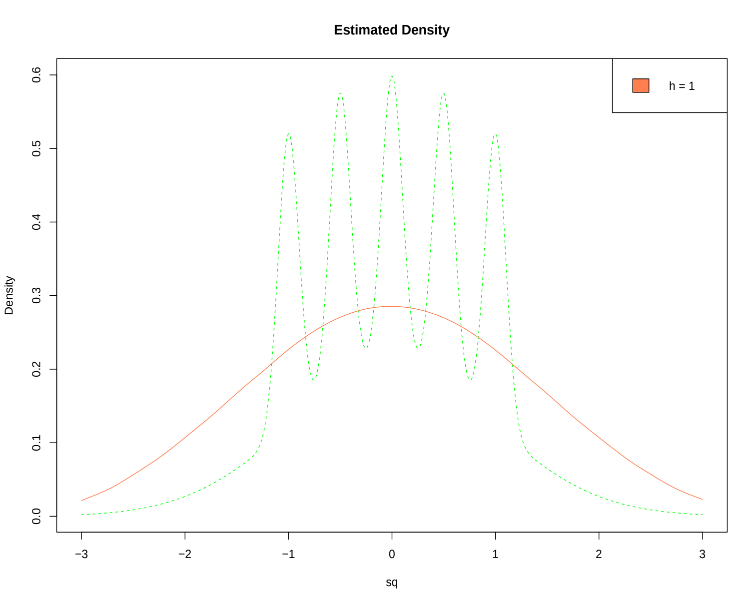
Demonstrating importance of choosing h optimally

Suppose we generate a random sample of size 1000 from a density $f(x)$ of the form :-

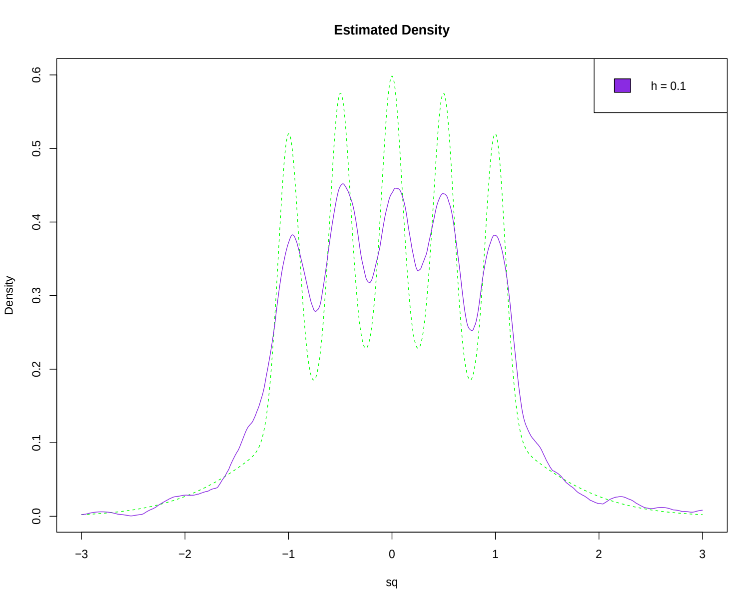
$$f(x) = \frac{1}{2}\phi(x;0,1) + \sum_{i=1}^{4} \phi\left(x;\left(\frac{i}{2}-1\right),\frac{1}{10}\right)$$

We have estimated this "claw shaped" density known as "Bart Simpson" density using **epanechnikov** kernel based on the sample observations.

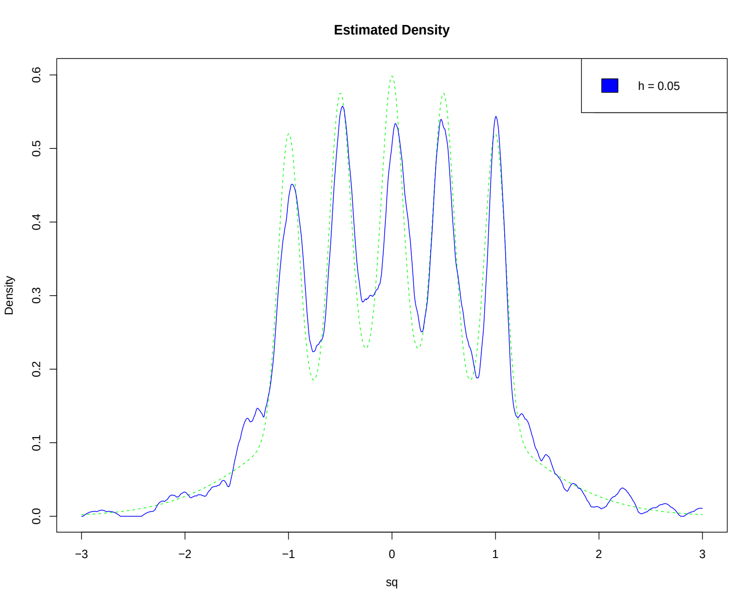
Demonstrating importance of choosing h optimally



Demonstrating importance of choosing h optimally



Demonstrating importance of choosing h optimally



Observation

- For large bin width the kernel density is undersmoothed. The estimated density doesn't resemble the true density.

Observation

- For large bin width the kernel density is undersmoothed. The estimated density doesn't resemble the true density.
- For a particular value of bin width the kernel density estimates the true density precisely.

Observation

- For large bin width the kernel density is undersmoothed. The estimated density doesn't resemble the true density.
- For a particular value of bin width the kernel density estimates the true density precisely.
- For small bin width, the estimated kernel density is over smoothed and it doesn't capture the true nature of the density.

Observation

- For large bin width the kernel density is undersmoothed. The estimated density doesn't resemble the true density.
- For a particular value of bin width the kernel density estimates the true density precisely.
- For small bin width, the estimated kernel density is over smoothed and it doesn't capture the true nature of the density.
- So, it is very crucial to find the optimal bin width for estimating the density using kernel estimation.

Limiting Distribution of V_n

- If $K(.)$ is a kernel function of bounded variation and the series $\sum_{n=1}^{\infty} e^{-\gamma n h_n^2}$ converges $\forall\ \gamma > 0$ where h_n denotes the bandwidth. Then

$$V_n = \sup_x \left| \widehat{f}_n(x) - f(x) \right| \to 0 \text{ w.p. } 1$$

Limiting Distribution of V_n

- If $K(.)$ is a kernel function of bounded variation and the series $\sum_{n=1}^{\infty} e^{-\gamma n h_n^2}$ converges $\forall\ \gamma > 0$ where h_n denotes the bandwidth. Then
$$V_n = \sup_x \left|\widehat{f}_n(x) - f(x)\right| \to 0 \text{ w.p. } 1$$
- **Assumption 1** For $h_n > 0$ and $h_n \to 0$ as $n \to \infty$.

Limiting Distribution of V_n

- If $K(.)$ is a kernel function of bounded variation and the series $\sum_{n=1}^{\infty} e^{-\gamma n h_n^2}$ converges $\forall\ \gamma > 0$ where h_n denotes the bandwidth. Then
$$V_n = \sup_x \left| \widehat{f_n}(x) - f(x) \right| \to 0 \text{ w.p. } 1$$
- **Assumption 1** For $h_n > 0$ and $h_n \to 0$ as $n \to \infty$.
- **Assumption 2** $K(.)$ is a probability probability density on R such that $K(.)$ is right continuous, of bounded variation on R,

Limiting Distribution of V_n

- If $K(.)$ is a kernel function of bounded variation and the series $\sum_{n=1}^{\infty} e^{-\gamma n h_n^2}$ converges $\forall\ \gamma > 0$ where h_n denotes the bandwidth. Then

$$V_n = \sup_x \left| \widehat{f_n}(x) - f(x) \right| \to 0 \text{ w.p. } 1$$

- **Assumption 1** For $h_n > 0$ and $h_n \to 0$ as $n \to \infty$.
- **Assumption 2** $K(.)$ is a probability probability density on R such that $K(.)$ is right continuous, of bounded variation on R,
- and

$$\int_{-\infty}^{\infty} |u| K(u)\, du < \infty$$

Limiting Distribution of V_n

- If $K(.)$ is a kernel function of bounded variation and the series $\sum_{n=1}^{\infty} e^{-\gamma n h_n^2}$ converges $\forall\ \gamma > 0$ where h_n denotes the bandwidth. Then
$$V_n = \sup_x \left| \widehat{f}_n(x) - f(x) \right| \to 0 \text{ w.p. } 1$$
- **Assumption 1** For $h_n > 0$ and $h_n \to 0$ as $n \to \infty$.
- **Assumption 2** $K(.)$ is a probability probability density on R such that $K(.)$ is right continuous, of bounded variation on R,
- and
$$\int_{-\infty}^{\infty} |u| K(u)\, du < \infty$$
- **Assumption 3** $K(.)$ satisfies the following two conditions
$$\int_{-\infty}^{\infty} uK(u)\, du = 0, \int_{-\infty}^{\infty} u^2 K(u)\, du < \infty$$

Convergence Rate of V_n

- Under these conditions it can be shown that[58]:-

Convergence Rate of V_n

- Under these conditions it can be shown that[58]:-
- $$\sup_x \left| \widehat{f}_n(x) - f(x) \right| = O\left[n^{-1/2} h_n^{-1} (\log\log n)^{1/2} \right]$$

Convergence Rate of V_n

- Under these conditions it can be shown that[58]:-
- $$\sup_x \left|\widehat{f_n}(x) - f(x)\right| = O\left[n^{-1/2} h_n^{-1} (\log\log n)^{1/2}\right]$$
- and if we choose $h_n = n^{-1/4}$then we have :-

$$\sup_x \left|\widehat{f_n}(x) - f(x)\right| = O\left[n^{-1/4} (\log\log n)^{1/2}\right]$$

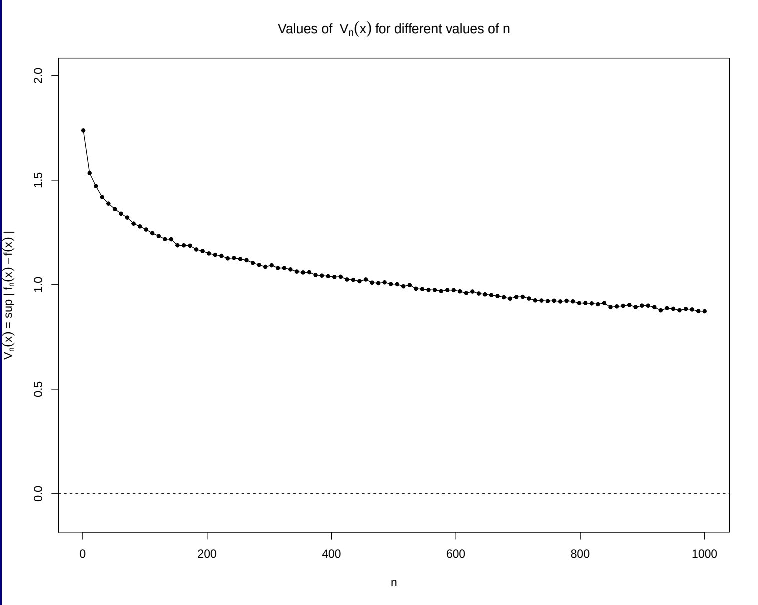
Convergence Rate of V_n

- Under these conditions it can be shown that[58]:-
- $$\sup_x \left|\widehat{f}_n(x) - f(x)\right| = O\left[n^{-1/2} h_n^{-1} (\log\log n)^{1/2}\right]$$
- and if we choose $h_n = n^{-1/4}$then we have :-

$$\sup_x \left|\widehat{f}_n(x) - f(x)\right| = O\left[n^{-1/4} (\log\log n)^{1/2}\right]$$

- which indicates that the rate at which $\sup_x \left|\widehat{f}_n(x) - f(x)\right|$ goes to 0 is very slow which can also be verified from the plots we made using simulation

Asymptotic distribution of V_n



Asymptotic normality of kernel density estimator

- We assume that f is continuous at x.

Asymptotic normality of kernel density estimator

- We assume that f is continuous at x.
- $\hat{f}_n(x)$ is the estimated density of f.

$$\hat{f}_n(x) = n^{-1} \sum_{k=1}^{n} Z_{nk}$$

Asymptotic normality of kernel density estimator

- We assume that f is continuous at x.
- $\hat{f}_n(x)$ is the estimated density of f.

$$\hat{f}_n(x) = n^{-1} \sum_{k=1}^{n} Z_{nk}$$

- Where,

$$Z_{nk} = h_n^{-1} K[(x - X_k)/h_n]$$

Asymptotic normality of kernel density estimator

- We assume that f is continuous at x.
- $\hat{f}_n(x)$ is the estimated density of f.

$$\hat{f}_n(x) = n^{-1} \sum_{k=1}^{n} Z_{nk}$$

- Where,

$$Z_{nk} = h_n^{-1} K[(x - X_k)/h_n]$$

- are i.i.d random variables. A necessary and sufficient condition for

$$\{\hat{f}_n(x) - E(\hat{f}_n(x))\}\{var(\hat{f}_n(x)\}^{-\frac{1}{2}} \xrightarrow{\mathcal{L}} N(0,1)$$

Asymptotic normality of kernel density estimator

- We assume that f is continuous at x.
- $\hat{f}_n(x)$ is the estimated density of f.

$$\hat{f}_n(x) = n^{-1} \sum_{k=1}^{n} Z_{nk}$$

- Where,

$$Z_{nk} = h_n^{-1} K[(x - X_k)/h_n]$$

- are i.i.d random variables. A necessary and sufficient condition for

$$\{\hat{f}_n(x) - E(\hat{f}_n(x))\}\{var(\hat{f}_n(x)\}^{-\frac{1}{2}} \xrightarrow{\mathcal{L}} N(0,1)$$

- is that, for every $\varepsilon > 0$.

$$nP[|Z_{n1} - E(Z_{n1})|\{var(Z_{n1})\}^{-\frac{1}{2}} \geq \varepsilon n^{1/2}] \longrightarrow 0 \quad as \text{ n} \longrightarrow \infty$$

Asymptotic normality of kernel density estimator

- A sufficient condition for

$$\frac{\hat{f}_n(x) - E(\hat{f}_n(x))}{\sqrt{var(\hat{f}_n(x))}} \xrightarrow{\mathcal{L}} N(0,1)$$

is that, for some $\delta > 0$

$$\frac{E|Z_{n1} - E(Z_{n1})|^{2+\delta}}{n^{\delta/2}[var(Z_{n1})]^{1+\delta/2}} \longrightarrow 0 \;\; as \text{ n} \to \infty$$

which is satisfied by the Epanechnikov kernel.

Asymptotic normality of kernel density estimator

- A sufficient condition for

$$\frac{\hat{f}_n(x) - E(\hat{f}_n(x))}{\sqrt{var(\hat{f}_n(x))}} \xrightarrow{\mathcal{L}} N(0,1)$$

is that, for some $\delta > 0$

$$\frac{E|Z_{n1} - E(Z_{n1})|^{2+\delta}}{n^{\delta/2}[var(Z_{n1})]^{1+\delta/2}} \longrightarrow 0 \;\; as \text{ n} \to \infty$$

which is satisfied by the Epanechnikov kernel.
- Now, we use simulation to verify the result stated above.

Asymptotic normality of kernel density estimator

- A sufficient condition for

$$\frac{\hat{f}_n(x) - E(\hat{f}_n(x))}{\sqrt{var(\hat{f}_n(x))}} \xrightarrow{\mathcal{L}} N(0,1)$$

is that, for some $\delta > 0$

$$\frac{E|Z_{n1} - E(Z_{n1})|^{2+\delta}}{n^{\delta/2}[var(Z_{n1})]^{1+\delta/2}} \longrightarrow 0 \;\; as \text{ n} \to \infty$$

which is satisfied by the Epanechnikov kernel.

- Now, we use simulation to verify the result stated above.
- Suppose, we have drawn a samples from standard normal distribution.Then $f(x) = \phi(x)$ which is a continuous function.

Asymptotic normality of kernel density estimator

- A sufficient condition for

$$\frac{\hat{f}_n(x) - E(\hat{f}_n(x))}{\sqrt{var(\hat{f}_n(x))}} \xrightarrow{\mathcal{L}} N(0,1)$$

  is that, for some $\delta > 0$

$$\frac{E|Z_{n1} - E(Z_{n1})|^{2+\delta}}{n^{\delta/2}[var(Z_{n1})]^{1+\delta/2}} \longrightarrow 0 \;\; as \text{ n} \to \infty$$

  which is satisfied by the Epanechnikov kernel.
- Now, we use simulation to verify the result stated above.
- Suppose, we have drawn a samples from standard normal distribution.Then $f(x) = \phi(x)$ which is a continuous function.
- We have estimated the true density using the **epanechnikov kernel** with a fixed bandwidth.

Asymptotic normality of kernel density estimator

- We plot the asymptotic distribution of

$$\frac{\hat{f}_n(x) - E(\hat{f}_n(x))}{\sqrt{var(\hat{f}_n(x))}}$$

  for $x = 0$ where we estimate $E(\hat{f}_n(0))$ and $var(\hat{f}_n(0))$ and put them and plot the histogram from the simulated data.

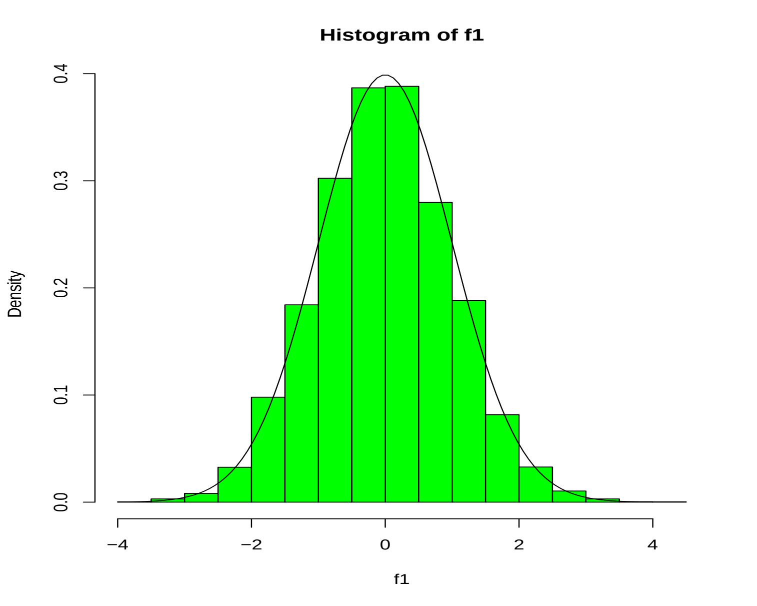
Asymptotic normality of kernel density estimator

- We plot the asymptotic distribution of

$$\frac{\hat{f}_n(x) - E(\hat{f}_n(x))}{\sqrt{var(\hat{f}_n(x))}}$$

  for $x = 0$ where we estimate $E(\hat{f}_n(0))$ and $var(\hat{f}_n(0))$ and put them and plot the histogram from the simulated data.
- We can clearly see that the histogram closely resembles that of a standard normal density as expected.

Asymptotic normality of kernel density estimator



Asymptotic Distribution

- Since, for large n ,

$$Z_n(x) = \frac{\widehat{f}_n(x) - E\widehat{f}_n(x)}{\sqrt{V\widehat{f}_n(x)}} \xrightarrow{\mathcal{L}} N(0,1)$$

where if we choose $K(u) = \phi(u)$, we get a closed form expression for the expectation term as,

$$E\widehat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h_n} K_n\left(\frac{x-u}{h_n}\right) f(u)\,du$$

$$= \frac{1}{2\pi h_n} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-u}{h_n}\right)^2} e^{-\frac{1}{2}u^2} du = (K_n * f)(x)$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{h_n^2+1}} e^{-\frac{1}{2}\frac{x^2}{h_n^2+1}} = \frac{1}{\sqrt{h_n^2+1}} \phi\left(\frac{x}{\sqrt{h_n^2+1}}\right)$$

Asymptotic Distribution

- and,

$$
\begin{aligned}
\sigma_n^2(x) &= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(\frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)\right) \\
&\approx \frac{f(x)}{nh_n} \int_{-\infty}^{\infty} K^2(u)\,du \\
&= \frac{f(x)}{nh_n} \frac{1}{2\sqrt{\pi}}
\end{aligned}
$$

Asymptotic Distribution

- and,

$$\sigma_n^2(x) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right)\right)$$
$$\approx \frac{f(x)}{nh_n}\int_{-\infty}^{\infty} K^2(u)\,du$$
$$= \frac{f(x)}{nh_n}\frac{1}{2\sqrt{\pi}}$$

- While doing simulation, here we know the actual pdf $f(x)$, hence we can use this property to visualize the CLT property of kernel density estimates by drawing upper and lower α points for each of the values of x that gives us two nice curves like :-

$$P\left(L_\alpha(x) \le \widehat{f}_n(x) \le U_\alpha(x)\right) \approx 1-\alpha$$

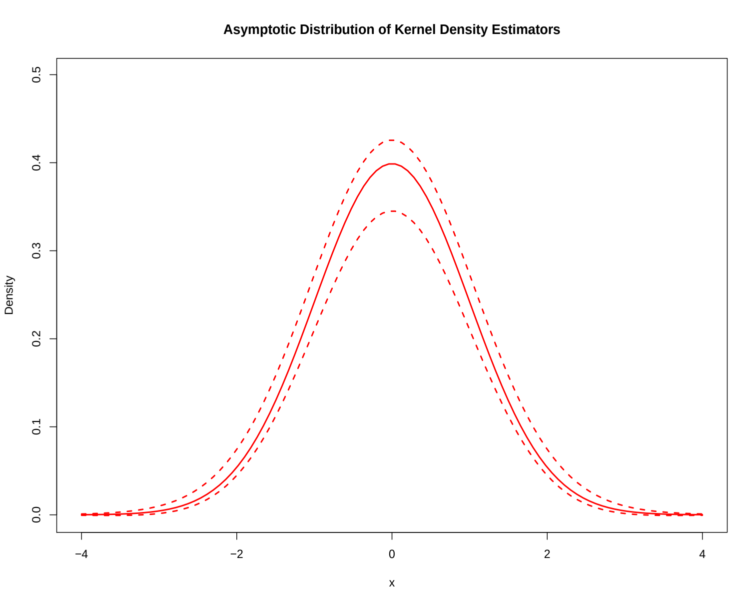
Asymptotic Distribution

- where ,

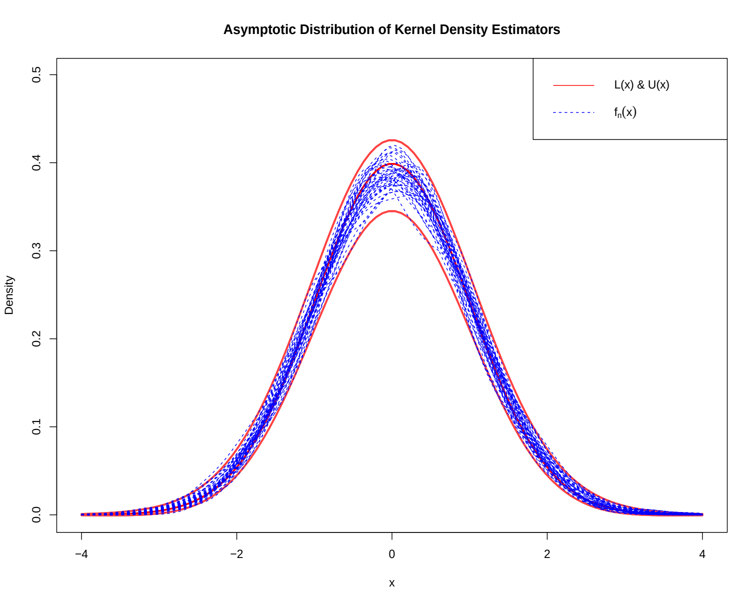
$$L_\alpha(x) = E\widehat{f_n}(x) - \tau_{\alpha/2}\sigma_n(x)$$
$$U_\alpha(x) = E\widehat{f_n}(x) + \tau_{\alpha/2}\sigma_n(x)$$

- if we draw these lines around the true density of the sample $(\phi(x))$, we get something like this :-

Asymptotic Distribution



Asymptotic Distribution



References

- Density Estimation for Statistics and Data Analysis by BW Silverman

References

- Density Estimation for Statistics and Data Analysis by BW Silverman
- Nonparametric Functional Estimationby B.L.S Prakasa Rao

References

- Density Estimation for Statistics and Data Analysis by BW Silverman
- Nonparametric Functional Estimationby B.L.S Prakasa Rao
- All of Nonparametric Statistics by Larry Wasserman

References

- Density Estimation for Statistics and Data Analysis by BW Silverman
- Nonparametric Functional Estimationby B.L.S Prakasa Rao
- All of Nonparametric Statistics by Larry Wasserman
- Elements of Statistical Learning by Trevor Hastie,Robert Tibshirani,Jerome H Friedman

References

- Density Estimation for Statistics and Data Analysis by BW Silverman
- Nonparametric Functional Estimationby B.L.S Prakasa Rao
- All of Nonparametric Statistics by Larry Wasserman
- Elements of Statistical Learning by Trevor Hastie,Robert Tibshirani,Jerome H Friedman
- Another helpful read in this genre

References

- Density Estimation for Statistics and Data Analysis by BW Silverman
- Nonparametric Functional Estimationby B.L.S Prakasa Rao
- All of Nonparametric Statistics by Larry Wasserman
- Elements of Statistical Learning by Trevor Hastie,Robert Tibshirani,Jerome H Friedman
- Another helpful read in this genre
- The Wikipedia page for KDE

Softwares and Tools used

- Google Colab - R version - for coding

Softwares and Tools used

- Google Colab - R version - for coding
- Overleaf - online LaTeX editor

Softwares and Tools used

- Google Colab - R version - for coding
- Overleaf - online LaTeX editor
- Both are great for collaborative works/group projects.

Acknowledgements

We express our gratitude to our professor Dr.Isha Dewan ma'am for giving us the opportunity to do such a beautiful project which not only made us verify and feel the essence of our theoritical knowledge/results practically but also exposed us to a lot of new perspectives of statistics as a subject in modern research and we got to explore some things beyond our coursework too.

Acknowledgements

THANK YOU

☺