A Brief Overview of Density Estimation

(Project of MTH516A)

Submitted by:

Sagnik Dey (201397)
Saumyadip Bhowmick (201408)
Soumyadip Sarkara (201431)

Supervised by,

Dr. Dootika Vats

Contents

1 Abstract

Kernel density estimation is a technique for estimating probability density function that is a must-have enabling a user to better analyse the studied probability distribution than when using traditional histogram.Unlike the traditional histogram kernel technique produces smooth estimates of the pdf, uses all sample points' location and more convincingly suggest multi-modality.In Kernel estimation 2 factors play important roles viz kernel function shape and the smoothing parameter.Although in Kernel density estimation is nowadays standard technique to explore density function there is a big dispute to assess the quality of the estimate and which choice of bandwidth is optimal. This project summarises one of the most important criteria in this regard which is integrated squared error (ISE) and gives an overview over the existing bandwidth selection method based on this i.e. Least square Cross Validation method (LSCV).

2 Introduction

Out of all probabilty distribution fuctions, probability density function best shows how the whole 100% probability mass is distributed over x-axis. However the oldest pdf empirical representation - a histogram- is a highly subjective structure as its shape depends on the subjective choice of number of class intervals to which the range of a sample is divided, and on the choice of initial point. The histogram suffers from an obvious problem: Data binning, which deprives the data of their individual location replacing their locations with a bin(interval) location.This causes the histogram shape to become discontinuous, and flat in each bin.
Kernel estimate of probability density function does not have these drawbacks. It produces a smooth empirical pdf based on individual locations of all sample data. Such pdf estimate seems to better represent the true pdf of a continuous variable.
Kernel estimation is not a quite new technique; it was originated more than a half century ago by Rosenblatt and Parzen. With the development in computer technology this method is developing rapidly and vastly.

2.1 Histogram

The simplest method to estimate a density f from an iid sample $X_1, X_2, ..., X_n$ is the histogram. From an analytical point of view the idea is to aggregate the data in intervals of the form $[x_0, x_0 + h)$and then use their relative frequency to approximate the density at $x \in [x_0, x_0 + h), f(x)$, by the estimate of

$$f(x_0) = F'(x_0) = \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0)}{h} = \lim_{h \to 0} \frac{\mathbf{P}(x_0 < X < x_0 + h)}{h}$$

More precisely given an origin t_0 and a bandwidth $h > 0$ the histogram builds a piecewise constant function in the intervals $\{B_k := [t_k, t_{k+1}) : t_k = t_0 + hk, k \in \mathbb{Z}\}$ by counting the number of sample points inside each of them. These constant length intervals are also called bins. The fact that they are of constant length h is important: we can easily standardise the counts on any bin by h in order to have relative frequencies per length in the bins. The histogram at a point x is defined as

$$\hat{f}_H(x; t_0, h) := \frac{1}{nh} \sum_{i=1}^{n} 1_{\{X_i \in B_k : x \in B_k\}}$$

The analysis of $\hat{f}_H(x; t_0, h)$ as a random variable is simple once one recognises that the bin counts are distributed as $B(n, p_k)$ with $p_k = \mathbb{P}[X \in B_k] = \int_{B_k} f(t)dt$. If f is continuous then by mean value theorem, $p_k = hf(\epsilon_{k,h})$ for an $\epsilon_{k,h} \in (t_k, t_{k+h})$. Assume that $k \in \mathbb{Z}$ is such that $x \in B_k = [t_k, t_{k+h})$. Therefore,

$$E[\hat{f}_H(x; t_0, h)] = \frac{np_k}{nh} = f(\epsilon_{k,h})$$

$$Var[\hat{f}_H(x; t_0, h)] = \frac{np_k(1 - p_k)}{n^2h^2} = \frac{f(\epsilon_{k,h})(1 - hf(\epsilon_{k,h}))}{nh}$$

The results above yield some interesting results:

1. if $h \to 0$, then $\epsilon_{k,h} \to x$,resulting in $f(\epsilon_{k,h}) \to f(x)$, and thus becomes asymptotically unbiased estimator of f(x).

2. But if $h \to 0$, variance increases. For decreasing the variance $nh \to \infty$ is required.

3. The variance is directly dependent on $f(\epsilon_{k,h})(1 - hf(\epsilon_{k,h})$(as $h \to 0$),Hence there is more variability at regions with higher density.

Clearly the shape of histogram depends on:

- t_0, since the separation between bins happens at $t_0 + kh, k \in \mathbb{Z}$

- h, which controls the bin size and the effective number of bins for aggregating the sample.

Therefore the subjectivity introduced by the dependence of t_0 is something we would like to get rid of. We can do so by allowing the bins to be dependent on x(the point at which we want to estimate f(x)), rather than fixing them beforehand.

2.2 Moving Histogram

An alternative to avoid the dependence on t_0 is the moving histogram, also known as naive density estimator. The idea is to aggregate the sample $X_1, X_2, ..., X_n$ in the intervals of the form $(x-h, x+h)$ and then use its relative frequency in $(x-h, x+h)$ to approximate the density at x, which can be expressed as,

$$\begin{aligned} f(x) &= F'(x) \\ &= \lim_{h \to 0^+} \frac{F(x+h) - F(x-h)}{2h} \\ &= \lim_{h \to 0^+} \frac{P(x-h < X < x+h)}{2h} \end{aligned}$$

The basic difference with the histogram is that the interval depends on the evaluation point x and are centered about it. This allows us to directly estimate f(x) using the symmetric derivative of F, instead of employing an estimate based on the forward derivative of F at x_0. More precisely given a bandwidth h 0, the naive density estimator builds a piecewise constant function by considering the relative frequency of $X_1, X_2, ..., X_n$ inside $(x-h, x+h)$:

$$\hat{f}_N(x;h) := \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{x-h<X<x+h\}}$$

From the figures also it is clearly revealed that there is a remarkable improvement in case of moving histogram with respect to histograms shown when estimating the underlying density. Analogous to histogram,the analysis of $\hat{f}_N(x;h)$ as a random variable follows from realising that,

$$\sum_{i=1}^{n} 1_{\{x-h<X<x+h\}} \sim B(n, p_{x,h})$$

$$p_{x,h} := P[x-h < X < x+h] = F(x+h) - F(x-h)$$

Then:

$$E[\hat{f}_N(x;h)] = \frac{F(x+h) - F(x-h)}{2h}$$

$$Var[\hat{f}_N(x;h)] = \frac{F(x+h) - F(x-h)}{4nh^2} - \frac{(F(x+h) - F(x-h))^2}{4nh^2} \tag{1}$$

Results provide interesting insights into the effect of h:

1. if $h \to 0$, then $E[\hat{f}_N(x;h)] \to f(x)$ and thus an asymptotically unbiased estimator of f(x)
2. if $h \to 0$, $\text{Var}[\hat{f}_N(x;h)] \approx \frac{f(x)}{2nh} - \frac{f(x)^2}{n} \to \infty$
3. if $h \to \infty,$ then $E[\hat{f}_N(x;h)] \to 0$
4. if $h \to \infty,$ then $Var[\hat{f}_N(x;h)] \to 0$
5. The variance shrinks to 0 if $nh \to \infty$. So both the bias and the variance can be reduced if $n \to \infty, h \to 0,$ and $nh \to \infty$, simultaneously.
6. The variance is almost proportional to f(x).

We are estimating $f(x) = F'(x)$ by estimating $\frac{F(x+h) - F(x-h)}{2h}$ thorough the relative frequency of $X_1, X_2, ..., X_n$ in the interval (x-h,x+h). It therefore seems reasonable that the data points closer to x are more important to assess the infinitesimal probability of x than ones further away.

2.3 Kernel Density Estimation

The moving histogram can be equivalently written as,

$$\begin{aligned}
\hat{f}_N(x;h) &= \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{x-h<X<x+h\}} \\
&= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} 1_{\{-1<\frac{x-X_i}{h}<1\}} \\
&= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad (*)
\end{aligned}$$

With $K(z) = \frac{1}{2} 1_{\{-1<z<1\}}$. Interestingly, K is a uniform density in (-1,1). This means when approximating $P(x-h < X < x+h) = P[-1 < \frac{x-X}{h} < 1]$ by (*) we are giving equal weight to all the points $X_1, X_2, ..., X_n$. So from generalisation of (*) to non-uniform weighting , we can replace K by any arbitrary density. Then K is known as Kernel. This generalisation provides the definition of the Kernel Density Estimator(kde):

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

A common notation is $K_h(z) = \frac{1}{h}K(\frac{z}{h})$ the so called scaled kernel, so that the kde can be written as

$$\hat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

2.3.1 Properties of Kernel

1. Kernel functions are symmetric
2. Since these are densities $\int_{-\infty}^{\infty} K(z)dz = 1$

2.3.2 Examples of Kernels

The following tables shows some popularly used kernels to estimate density.

Names	Density
Rectangular/Uniform	$\frac{1}{2}1_{\{\lvert z\rvert<1\}}$
Triangular	$(1-\lvert z\rvert)1_{\{\lvert z\rvert<1\}}$
Epanechnikov	$\frac{3}{4}(1-z^2)1_{\{\lvert z\rvert<1\}}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$

2.4 Bias and variance of Kernel Density Estimate

In order to compute expectation and variance of Kernel Density Estimate we use linearity of KDE and definition of Convolution. The convolution of 2 functions f and g is defined

$$(f * g)(x) = \int f(x-y)g(y)dy$$

So now we want to find the expression for $E(\hat{f}(x;h))$

$$\begin{aligned} E(\hat{f}(x;h)) &= \frac{1}{n}\sum_{i=1}^{n} E[K_h(x - X_i)] \\ &= \int K(x-y)f(y)dy \\ &= (k * f)(x) \end{aligned}$$

For the bias we consider the change of variables $z = \frac{x-y}{h} \implies y = x - hz,\ dy = -hdz$. The integral limits flip and we have,

$$E(\hat{f}(x;h)) = \int K(z)f(x-hz)dz$$

Since $h \to 0$ we apply second order Taylor series expansion that gives,

$$f(x - hz) = f(x) - hzf'(x) + \frac{h^2z^2}{2}f''(x) + o(h^2z^2)$$

Substituting it in the expression of $E(\hat{f}(x;h))$ and bearing in mind that K is symmetric about 0, we have

$$\begin{aligned}\int K(x-y)f(y)dy &= \int K(z)\{f(x) - hzf'(x) + \frac{h^2z^2}{2}f''(x) + o(h^2z^2)\}dz \\ &= f(x) + \frac{1}{2}\mu_2(K)f''(x)h^2 + o(h^2)\end{aligned}$$

So,

$$bias[\hat{f}(x;h)] = E(\hat{f}(x;h)) - f(x) = \frac{1}{2}\mu_2(K)f''(x)h^2 + o(h^2)$$

where $\mu_j(L) = \int p^j L(p)dp$ For variance of kernel Density Estimate,

$$\begin{aligned}Var[\hat{f}(x;h)] &= \frac{1}{n^2}\sum_{i=1}^{n} Var[K_h(x - X_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(E[K_h^2(x-X)] - E^2[K_h(x-X_i)]\right)\end{aligned}$$

The second term is already computed, so we focus on the first. Using the previous change of variables and a first-order Taylor expansion, we have

$$\begin{aligned}E[K_h^2(x-X)] &= \frac{1}{h}\int K^2(z)f(x-hz)dz \\ &= \frac{1}{h}\int K^2(z)\{f(x) + O(hz)\}dz \\ &= \frac{R(K)}{h}f(x) + O(1)\end{aligned}$$

Where $R(L) = \int L^2(p)dp$
Which gives us the expression of variance,

$$\begin{aligned}Var[\hat{f}(x;h)] &= \frac{1}{n}\left(\frac{R(K)}{h}f(x) + O(1) - O(1)\right) \\ &= \frac{R(K)}{nh}f(x) + O(n^{-1}) \\ &= \frac{R(K)}{nh}f(x) + o((nh)^{-1})\end{aligned}$$

since $n^{-1} = o((nh)^{-1})$

2.4.1 Interpretations of bias and variance of KDE

- The bias at x is directly proportional to $f''(x)$. This has some interesting interpretations:
  1. The bias is negative where f is concave, $f" < 0$. These regions correspond to the peaks and modes of f, where the kde underestimates f.
  2. Conversely the bias is positive where f is convex, $f" > 0$.These regions correspond to valleys or tails of f, where kde overestimates f.
  3. The wilder is the curvature f", the harder is to estimate f. Flat density regions are easier to estimate than wiggling regions with high curvatures.
- The variance depends directly on f(x). The higher the density the more variable is the kde. Interestingly, the variance decreases as a factor of $(nh)^{-1}$.

2.5 Bandwidth

The bandwidth of the kernel(h) is a parameter which exhibits a strong influence on the resulting estimate. Basically the bandwidth h is the standard deviation of the kernel function. This is because the implementation contains many kernels some with finite support and some without,and using standard deviation to quantify the bandwidth allows easy comparison. Choice of bandwidth in KDE plays more important role than the choice of kernel. Small values of h makes the estimated density curve wiggly whereas large values of h smooth out the estimated density. We can illustrate its effect by the following figure.

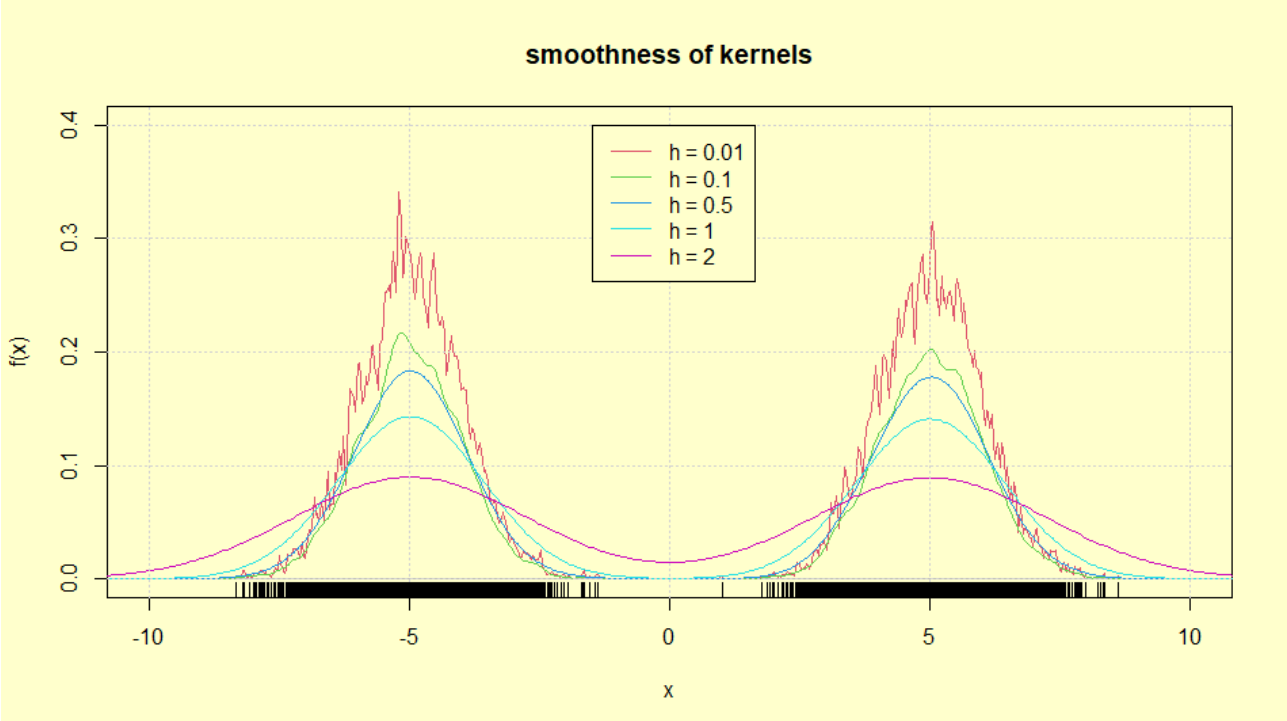

Figure 1: Estimated Density using Different Kernels

3 How "good" is the estimator

Since Kernel density estimation critically depends on bandwidth employed, the purpose of this section is to introduce objective and automatic bandwidth selectors that attempt to minimise the estimation error of the target density f.
The first step is to define a global rather than local error criterion. The Integrated Squared Error (ISE) is defined as,

$$ISE(\hat{f}(x;h) := \int (\hat{f}(x;h) - f(x))^2 dx$$

This is the squared distance between the kde and target density.The ISE is a random quantity since it depends directly on the sample $X_1, X_2, ..., X_n$. As a consequence looking for an optimal-ISE bandwidth is a hard task, since the optimality is dependent on sample itself and not on population and n. To avoid this problem it is usual to compute Mean Integrated Squared Error (MISE).

$$\begin{aligned} MISE[\hat{f}(.;h)] &= E[ISE[\hat{f}(.;h)]] \\ &= E[\int (\hat{f}(x;h) - f(x))^2 dx] \\ &= \int E[(\hat{f}(x;h) - f(x))^2] dx \\ &= \int MSE[\hat{f}(.;h)] dx \end{aligned}$$

A further well known criterion is derive from the asymptotic analysis of MISE(h). Note that MISE(h) has the presentation,

$$\begin{aligned} MISE(h) &= \int E[\hat{f}(x;h) - f(x)]^2 dx \\ &= \int E[\hat{f}(x;h) - E[\hat{f}(x;h)] + E[\hat{f}(x;h)] - f(x)]^2 dx \\ &= \int E[\hat{f}(x;h) - \hat{f}(x;h)] + [E[\hat{f}(x;h)] - f(x)]^2 dx \\ &= \int Var\hat{f}(x;h) dx + \int bias^2 \hat{f}(x,h) dx \end{aligned}$$

Some straight forward analysis and by change of variables and Taylor expansion of f shows that MISE(h) turns out to be,

$$MISE(\hat{f}(.;h)) = \frac{1}{4}\mu_2^2(K) R(f'') h^4 + \frac{R(k)}{nh} + o(h^4 + (nh)^{-1})$$

The dominating part of MISE is denoted by AMISE which stands for Asymptotic MISE:

$$AMISE(\hat{f}(.,h)) = \frac{1}{4}\mu_2^2(K) R(f'') h^4 + \frac{R(k)}{nh}$$

Where $\mu_j(L) = \int x^j L(x)dx$ and $R(L) = \int L^2(x)dx$

So the bandwidth that minimises the AMISE is:

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2^2(K)R(f'')n}\right]^{\frac{1}{5}}$$

The objective should be to find h that minimises these expressions of ISE,MISE or AMISE.

4 Bandwidth selection method

There are different bandwidth selection methods depending on which criterion viz ISE,MISE or AMISE is being minimised. In this project we restrict our study upto minimising ISE and look into corresponding optimal bandwidth selection method viz Cross-Validation method in detail.

4.1 Cross-validation method

In this method of bandwidth selection we directly attempt to minimise the ISE.The idea is to use the sample twice: one for computing the kde and the other for evaluating its performance on estimating f. To avoid the clear dependence on the sample,we do this evaluation in a cross-validatory way: the data used for computing the kde is not used for its evaluation.
Firstly expansion of ISE expression looks like,

$$\begin{aligned} ISE[\hat{f}(x;h)] &= \int [\hat{f}(x;h) - f(x)]^2 dx \\ &= \int [\hat{f}^2(x;h) - 2\hat{f}(x;h)f(x) + f(x)^2]dx \\ &= \int \hat{f}^2(x;h)dx - 2\int \hat{f}(x;h)f(x)dx + \int f(x)^2 dx \\ &= R(\hat{f}(x;h)) - 2E_{f(x)}[\hat{f}(x;h)] + R(f(x)) \end{aligned}$$

Here the last term is independent of h and hence minimising ISE becomes equivalent to minimisation of first 2 terms.
This quantity is unknown but can be estimated unbiasedly as,

$$LSCV(h); = \int \hat{f}(x;h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(Xi;h)$$

And here $\hat{f}_{-i}(.,h)$ is the leave-one-out kernel density estimate and is based on the sample with the X_i removed

$$\hat{f}_{-i}(x;h) = \frac{1}{n-1}\sum_{j=1;j\neq i}^{n} K_h(x - X_j)$$

Now LSCV(h) is an estimator for ISE(h)-R(f). And it is easy to check that:

$$\begin{aligned} E(LSCV(h)) &= E(ISE(h)) - R(f) \\ &= MISE(h) - R(f) \end{aligned}$$

Thus the least square cross validation function is an unbiased estimator for ISE(h)-R(f).At the end h_{LSCV} is chosen such that

$$h_{LSCV} = \arg\min_{h>0} LSCV(h)$$

5 Practical issue

In case of any density with a support with boundary, for example a LN(0,1) in $(0, \infty)$, kde runs into trouble. Here what happens is that since kde is defined over entire real line, it is spreading probability mass outside the support of the distribution, resulting in a severe negative bias about 0. As a result the kde does not integrate to 1 in the support of the data. This is known as Boundary issue in the context of KDE.

5.1 Remedies

5.1.1 Reflection Method

A simple but ingenious idea is to reáect the data points $X_1, ..., X_n$ at origin and then to work with the rv's,

$$Y_i = \begin{cases} -X_j, & j = 1, ..., n \\ X_{2n-j} & j = n+1, .., 2n \end{cases}$$

This not only yields a twice as large sample size but most importantly yields a sample from a density with unbounded support.
Therefore, a standard kernel estimator can be applied to the data which is now of sample size 2n, i.e.

$$f^*_{refl}(x) = \frac{1}{2nh} \sum_{j=1}^{2n} K\left(\frac{x - X_j}{h}\right), x \in \mathbb{R}$$

This is the standard kernel density estimator. Moreover it is also easy to see that this estimate is symmetric around the origin. Thus, the natural way to get an estimate with support $[0, \infty)$,

i.e. we take reflection again and the final estimate becomes,

$$\hat{f}^*_{refl}(x) = \begin{cases} 2f^*_{refl}(x), & x \geq 0 \\ 0, & x \leq 0 \end{cases}$$

5.1.2 Transformation Method

A simple approach to deal with the boundary bias is to map a non real supported density f into a real supported density g which is simpler to estimate by means of a transformation t:

$$f(x) = g(t(x))t'(x)$$

The transformation kde is obtained by replacing g with usual kde, but acting on transformed data $t(X_1), t(X_2), ..., t(X_n)$:

$$\hat{f}_T(x; h, t) := \frac{1}{n} \sum_{i=1}^{n} K_h(t(x) - t(X_i))t'(x)$$

A table with some common transformations is the following:

Data in	t(x)	t'(x)
(a,∞)	log(x-a)	$\frac{1}{x-a}$
(a,b)	$\Phi^{-1}(\frac{x-a}{b-a})$	(b-a)$\phi(\frac{\phi^{-1}(x-a)}{b-a})^{-1}$
$(-\lambda_1, \infty)$	$(x+\lambda_1)^{\lambda_2} sign(\lambda_2)$ if $\lambda_2 \neq 0$	$\lambda_2(x+\lambda_1)^{\lambda_2 - 1} sign(\lambda_2)$ if $\lambda_2 \neq 0$

6 Simulation studies

6.1 Histogram

A random sample is taken from $N(0, 10^2)$. We then plotted histograms on this data varying t_0 and h.

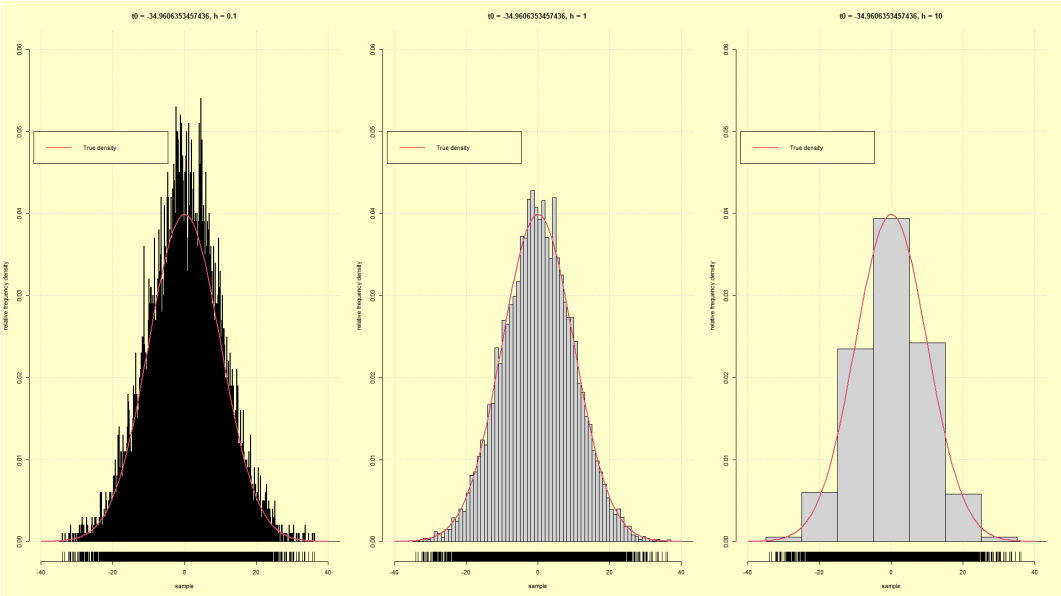

Figure 2: Histograms using fixed t_0 = min(sample) - 1 and h = 0.1, 1, 10

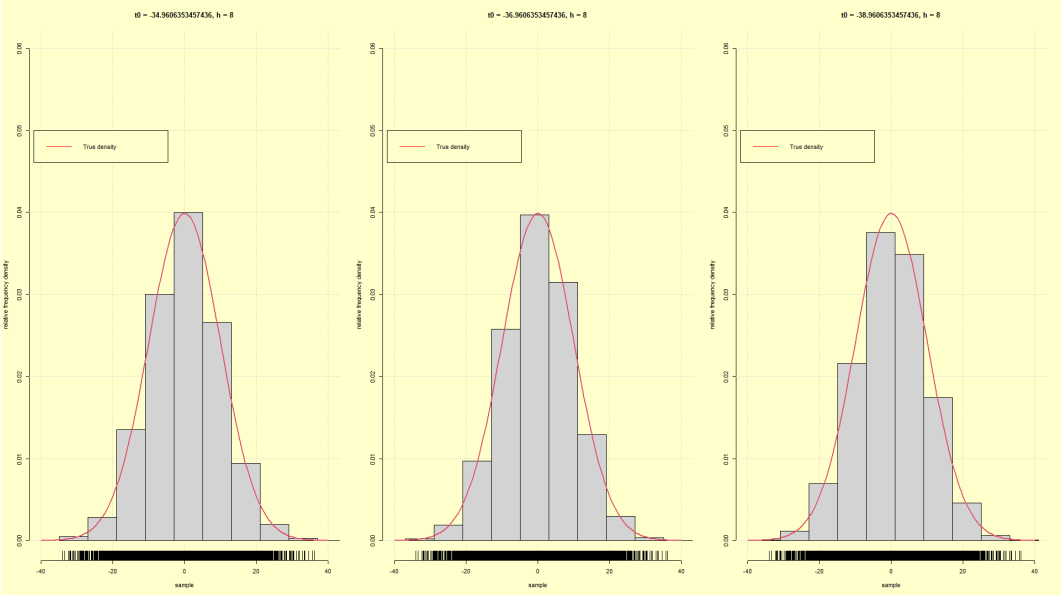

Figure 3: Histograms using varying t_0 = min(sample) - 1 , min(sample) - 3, min(sample) - 5 and fixed h = 8

A random sample is taken from $exp(20)$. We then plotted histograms on this data varying t_0 and h.

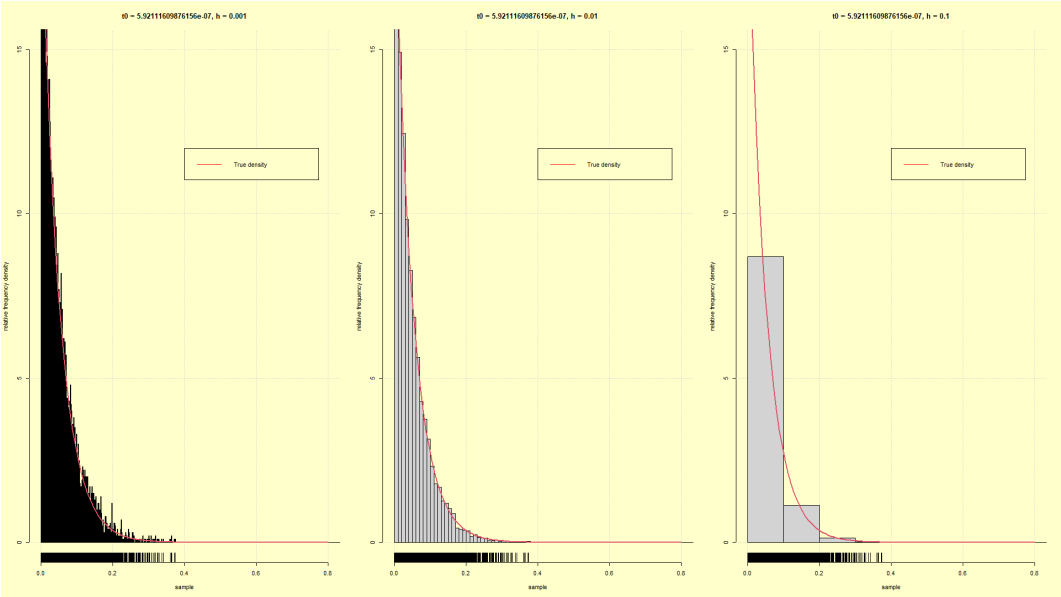

Figure 4: Histograms using fixed t_0 = min(sample) - 1 and h = 0.001, 0.01, 0.1

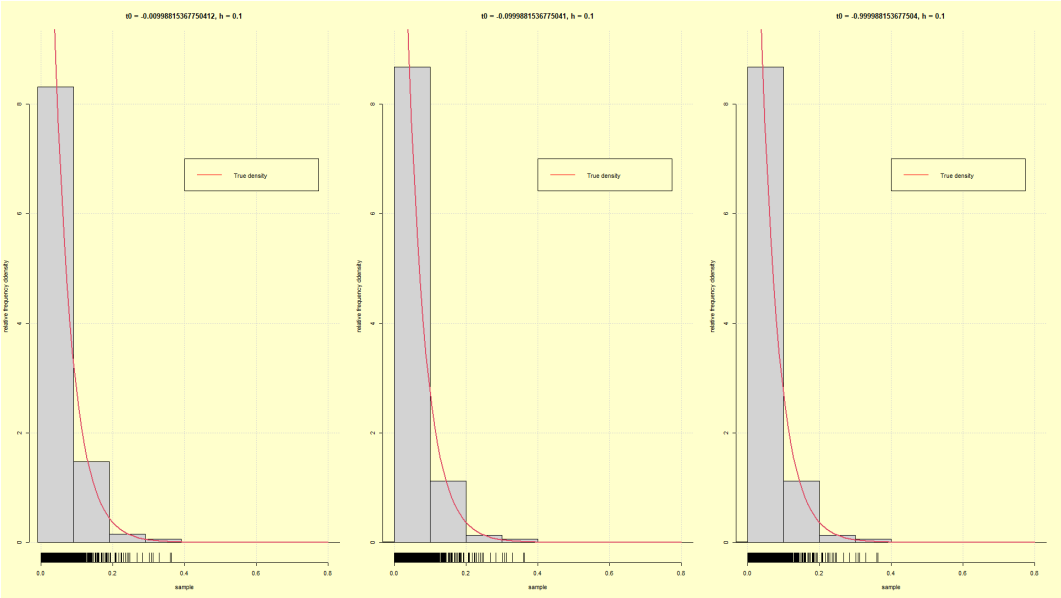

Figure 5: Histograms using varying t_0 = min(sample) - 0.01 , min(sample) - 0.1, min(sample) - 1 and fixed h = 0.1

A random sample is taken from a Gaussian Mixture distribution i.e. $0.5N(-5,1)+0.5N(5,1)$. We then plotted histograms on this data varying t_0 and h.

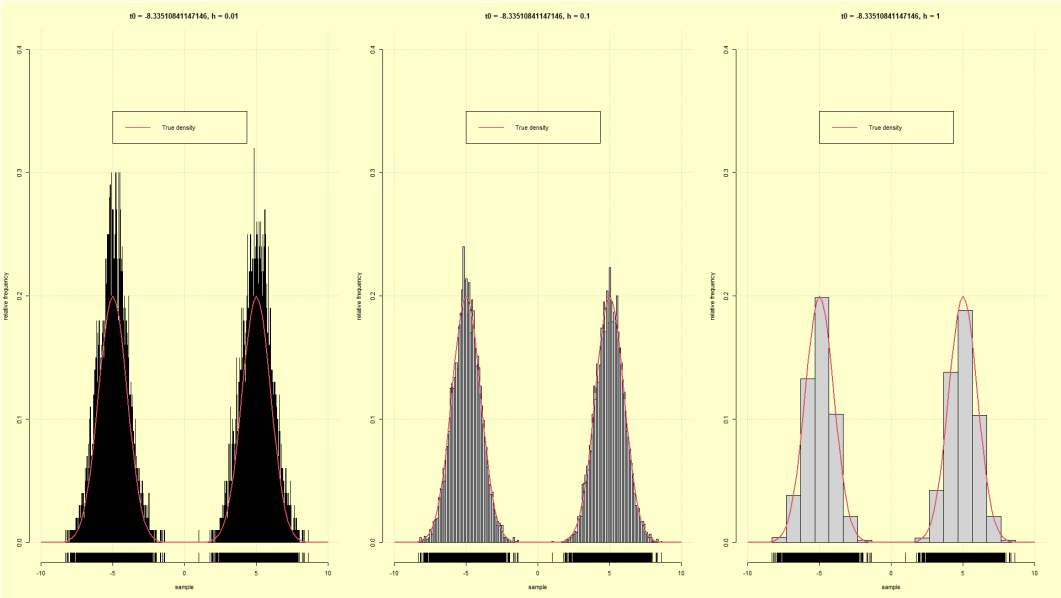

Figure 6: Histograms using fixed t_0 = min(sample) - 1 and h = 0.01, 0.1, 1

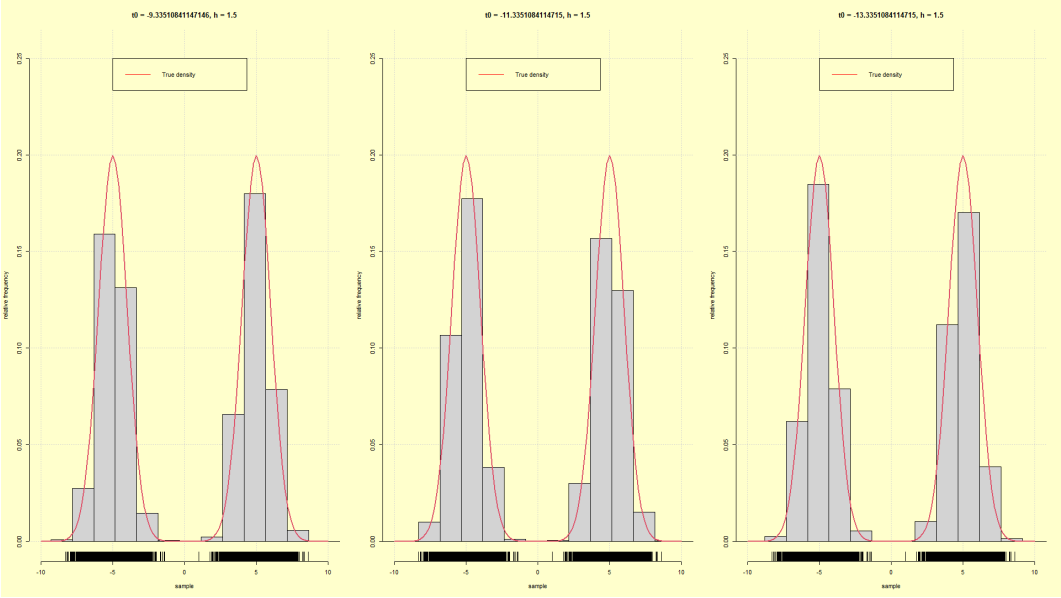

Figure 7: Histograms using varying t_0 = min(sample) - 1 , min(sample) - 3, min(sample) - 5 and fixed h = 1.5

6.1.1 Insights

- Clearly the shape of histogram depends on t_0 and h, irrespective of which distribution is chosen.

6.2 Moving Histogram

A random sample is taken from $N(0, 10^2)$. We then plotted Moving Histograms on this data varying h.

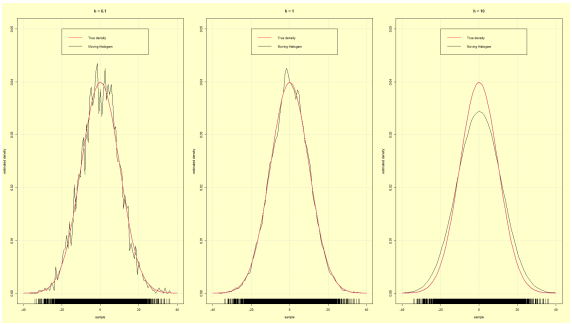

Figure 8: Moving Histograms using h = 0.1, 1, 10

A random sample is taken from $exp(20)$.We then plotted Moving Histograms on this data varying h.

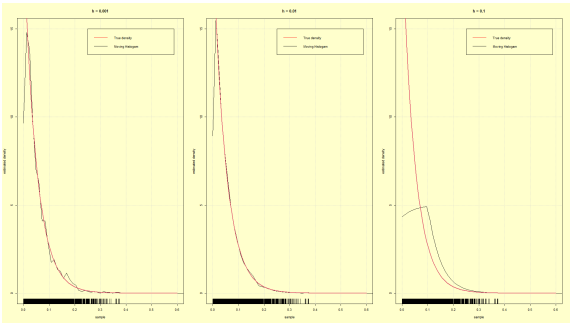

Figure 9: Moving Histograms using h = 0.001, 0.01, 0.1

A random sample is taken from $0.5N(-5, 1) + 0.5N(5, 1)$. We then plotted Moving Histograms on this data varying h.

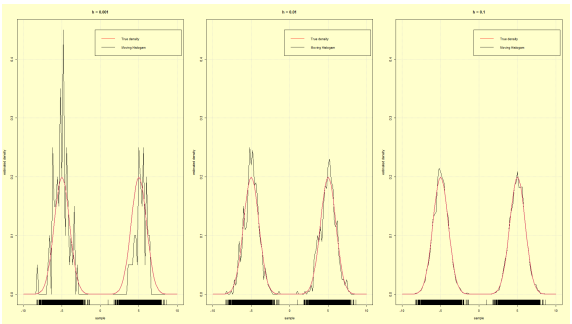

Figure 10: Moving Histograms using h = 0.001, 0.01, 0.1

6.2.1 Insights

- Clearly the moving histogram is estimating density better than histogram, still it depends on h.

6.3 Kernel Density Estimation

A random sample is taken from $N(0, 10^2)$. We then plotted Moving Histograms on this data varying h with Rectangular, Gaussian, Epanechnikov kernels.

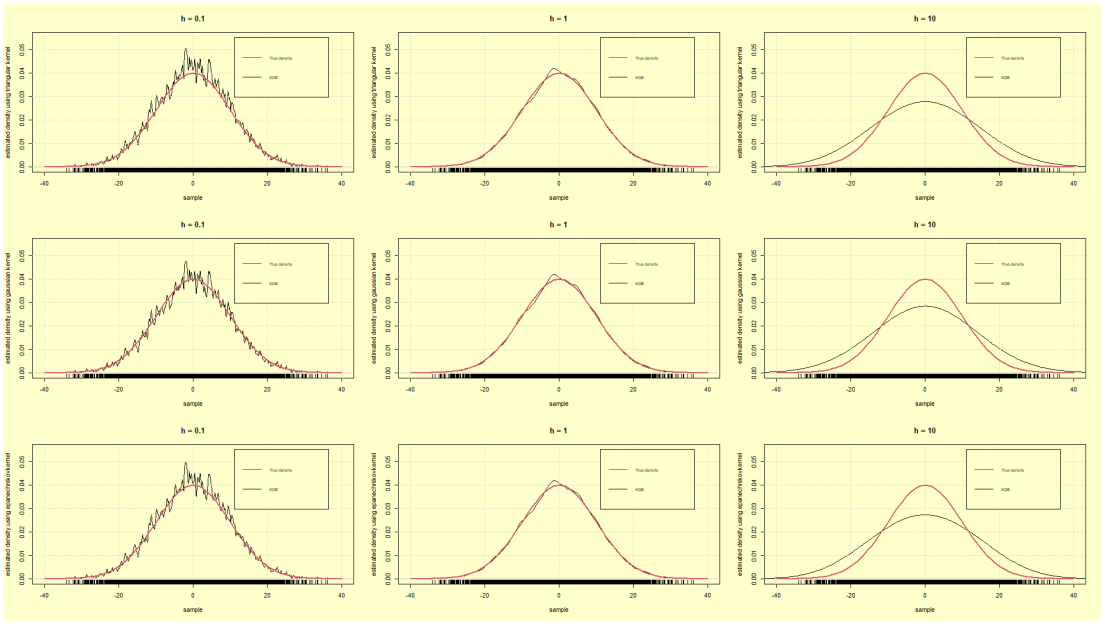

Figure 11: Kernel Density Estimates using h = 0.1, 1, 10, and 3 different kernels

A random sample is taken from $exp(20)$. We then plotted Moving Histograms on this data varying h with Rectangular, Gaussian, Epanechnikov kernels.

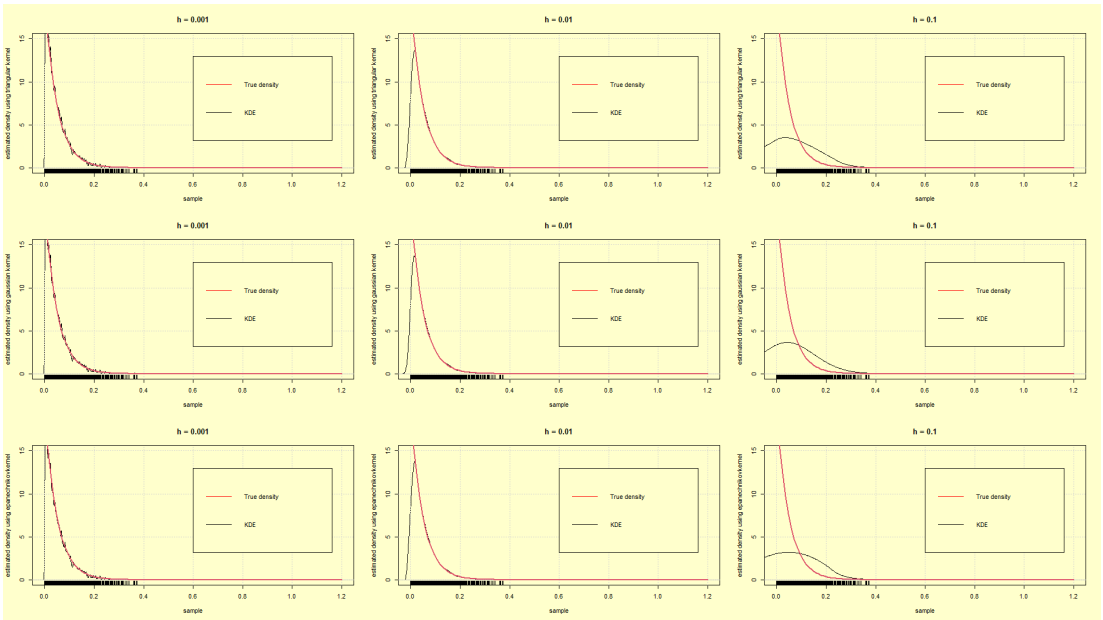

Figure 12: Kernel Density Estimates using h = 0.001, 0.01, 0.1 and 3 different kernels

A random sample is taken from Gaussian Mixture Distribution i.e. $0.5N(-5,1)+0.5N(5,1)$. We then plotted Moving Histograms on this data varying h with Rectangular, Gaussian, Epanechnikov kernels.

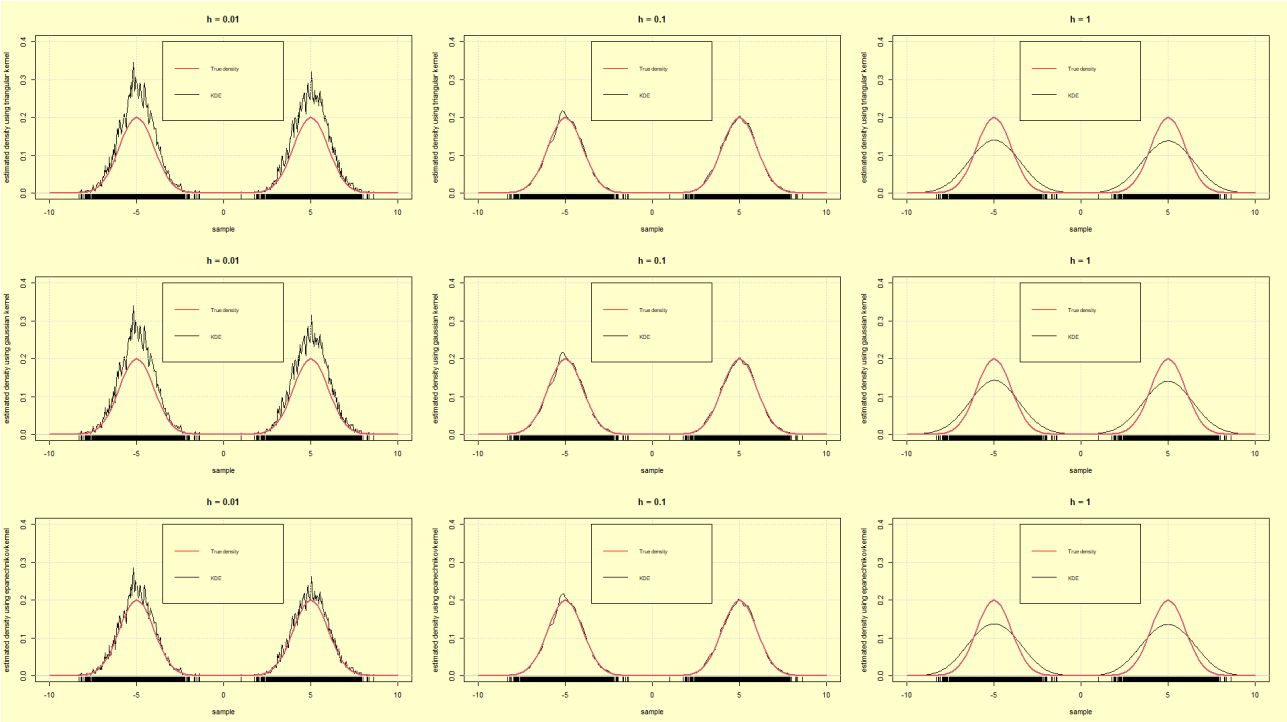

Figure 13: Kernel Density Estimates using h = 0.01, 0.1, 1 and 3 different kernels

6.3.1 Insights

- Clearly the KDE is estimating density better than histogram and moving histograms, still it depends on h.
- Once a certain smoothness is guaranteed, the choice of Kernels is not that important, i.e. at h = 0.01, the estimates are different from each other, but for higher values of h, i.e., for h = 0.1, 1, the estimates are more or less are similar for different kernels.

6.4 Bias and Variance of KDE

Data is simulated 1000 times from $0.7N(-3, 0.3) + 0.3N(3, 0.6)$ distribution. For different nh values, the 1000 Kernel Density Estimates are plotted.

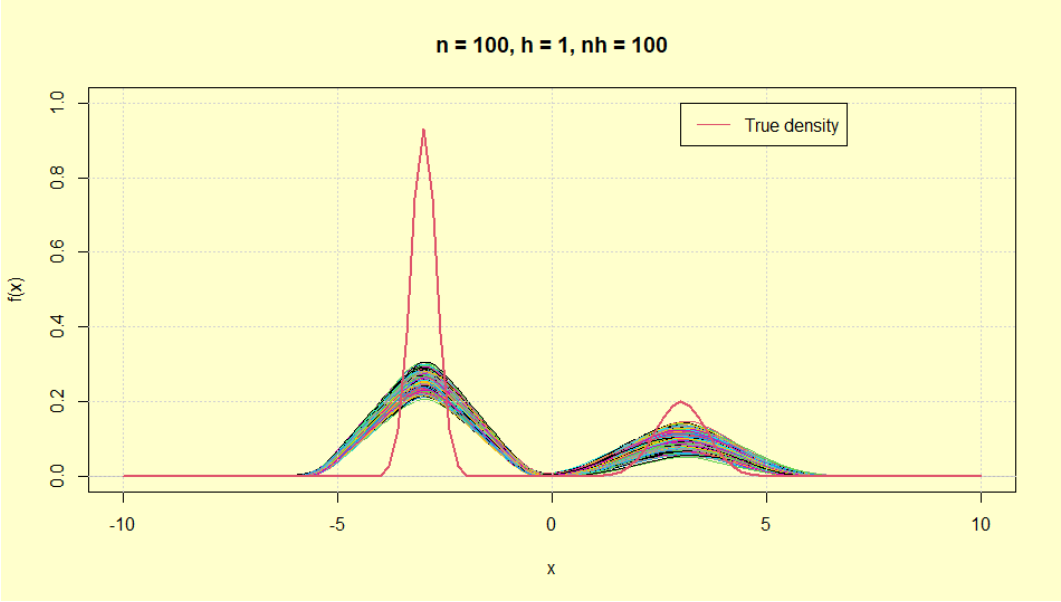

Figure 14: 1000 KDEs for n = 100, h = 1, nh = 100

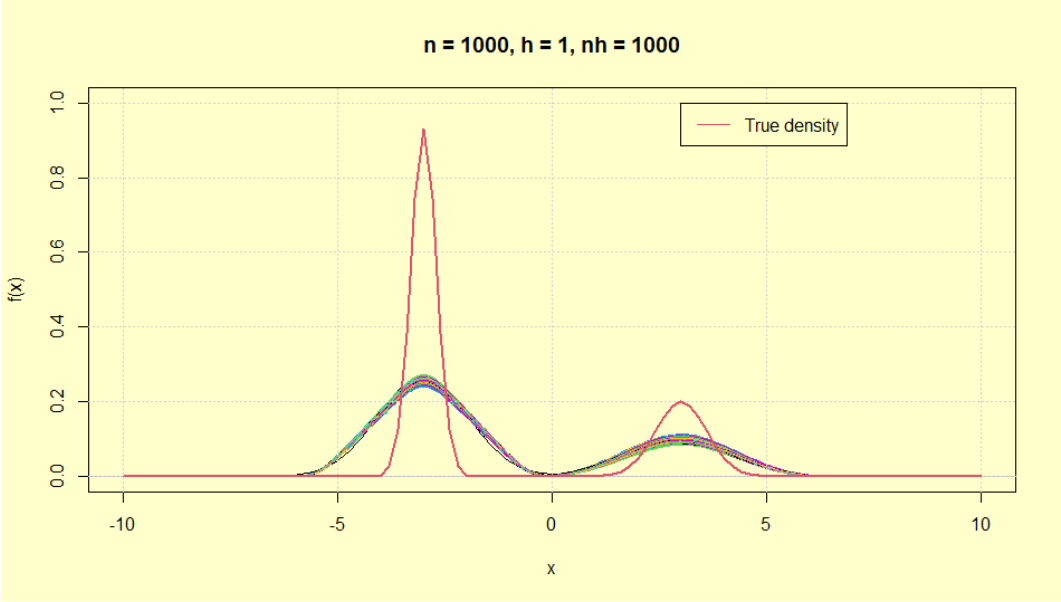

Figure 15: 1000 KDEs for n = 1000, h = 1, nh = 1000

6.4.1 Insights

- at x = 3,-3, f" < 0, f is concave, so bias is negative and near x = -4,4, f" > 0, f is convex, so bias is positive.
- x = 3, -3, f(x) is high so the variance is also high, x = 4,-4 f(x) is low, so the variance is lower.
- As nh inrceases, variance of the density estimates decreases and bias remains similar, as though nh increases, h remains same, and bias only depends on h, not n.

6.5 Bandwidth Selection for $N(0, 10^2)$

A random sample is taken from $N(0, 10^2)$ the LSCV(h) values are plotted for different values of h, and $\hat{h}_{LSCV}$ is taken as the value of h for which LSCV(h) is minimum.

6.5.1 LSCV plot

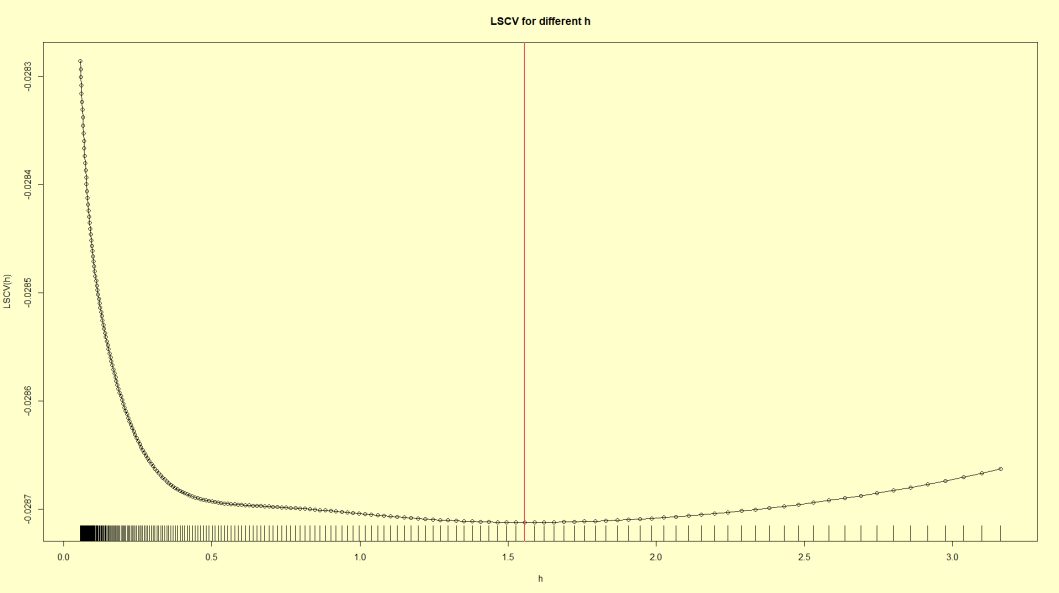

Figure 16:

We see that a clear minima is occuring here for a particular h, using that $\hat{h}_{LSCV}$ value and using Rectangular, Gaussian, Epanechnikov kernels, density is estimated.
It is evident from the plot that for each of the kernels, the KDE here more or less coincides with the actual density, as h is optimally chosen.

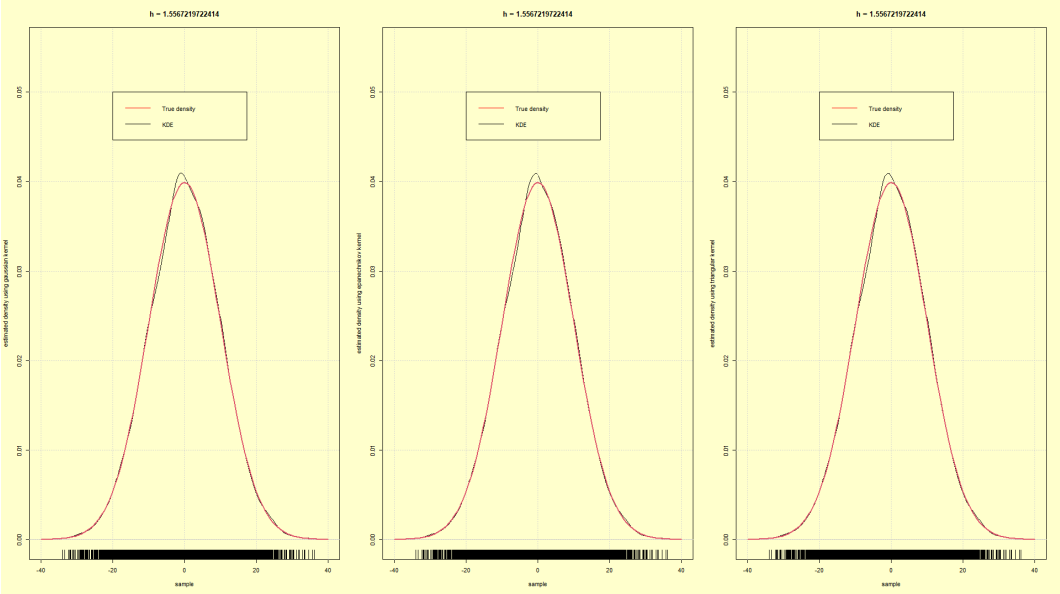

Figure 17:

6.6 Bandwidth Selection for $exp(1)$

A random sample is taken from $exp(1)$ the LSCV(h) values are plotted for different values of h, and h_{LSCV} is taken as the value of h for which LSCV(h) is minimum.

6.6.1 LSCV plot

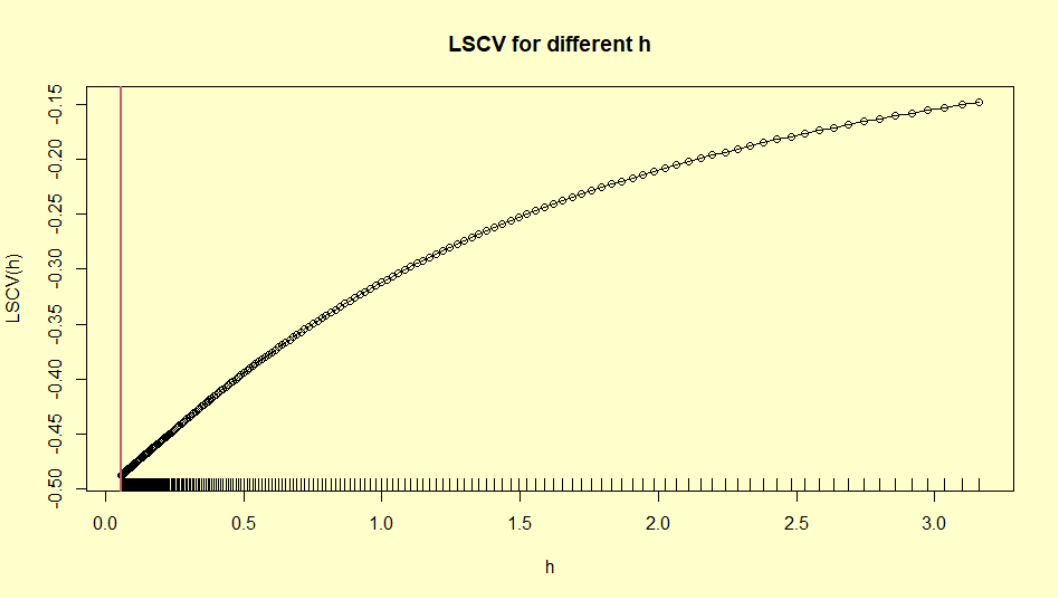

Figure 18:

Here we see that the graph is stritctly increasing, hence there is no global minima, so $\min_h\{LSCV(h)\}$ does not exist, still we a value of h_{LSCV} near to 0, and Using that h_{LSCV} value and using Rectangular, Gaussian, Epanechnikov kernels, density is estimated.

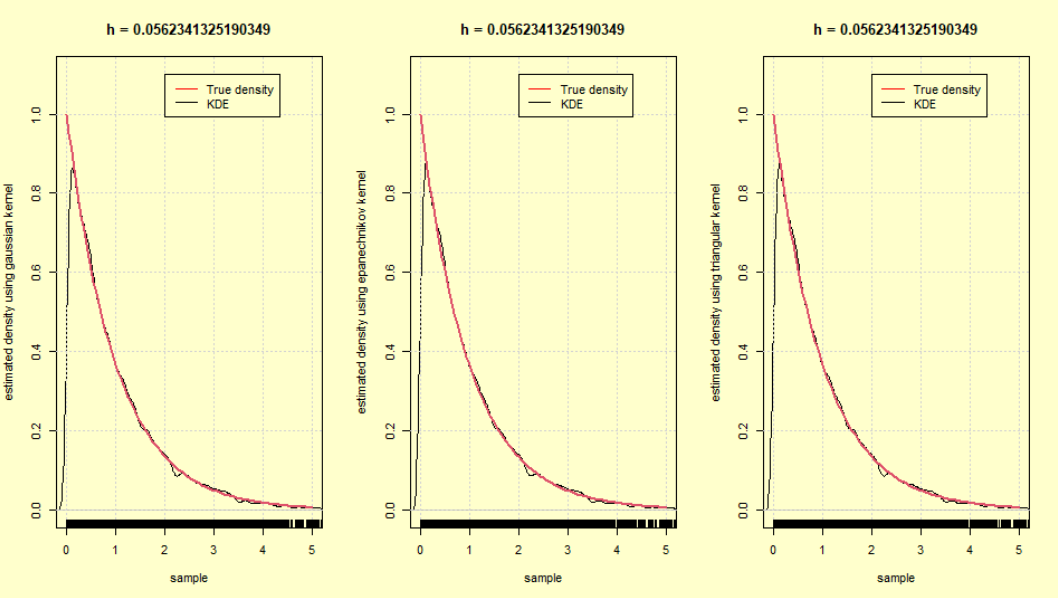

Figure 19:

It is evident from the plot that for each of the kernels, the KDE here more or less coincides with the actual density in the interval $(0.5, \infty)$, but in the neighborhood of 0, the KDE fails to estimate the density well.

6.7 Bandwidth Selection for $0.5N(-5,1) + 0.5N(5,1)$

A random sample is taken from $0.5N(-5,1) + 0.5N(5,1)$ the LSCV(h) values are plotted for different values of h, and h_{LSCV} is taken as the value of h for which LSCV(h) is minimum.

6.7.1 LSCV plot

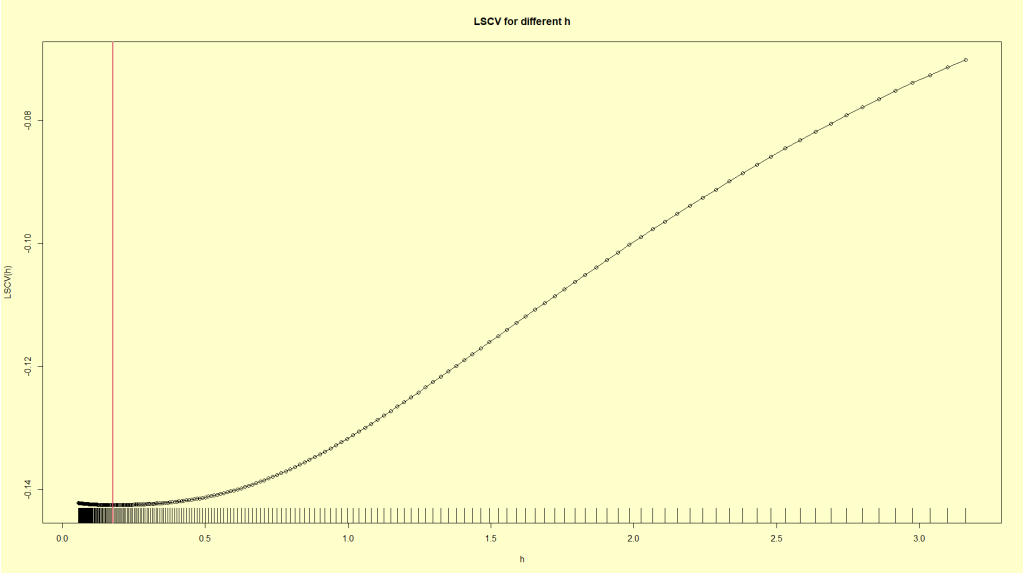

Figure 20:

We see that a clear minima is occuring here for a particular h, using that h_{LSCV} value and using Rectangular, Gaussian, Epanechnikov kernels, density is estimated.

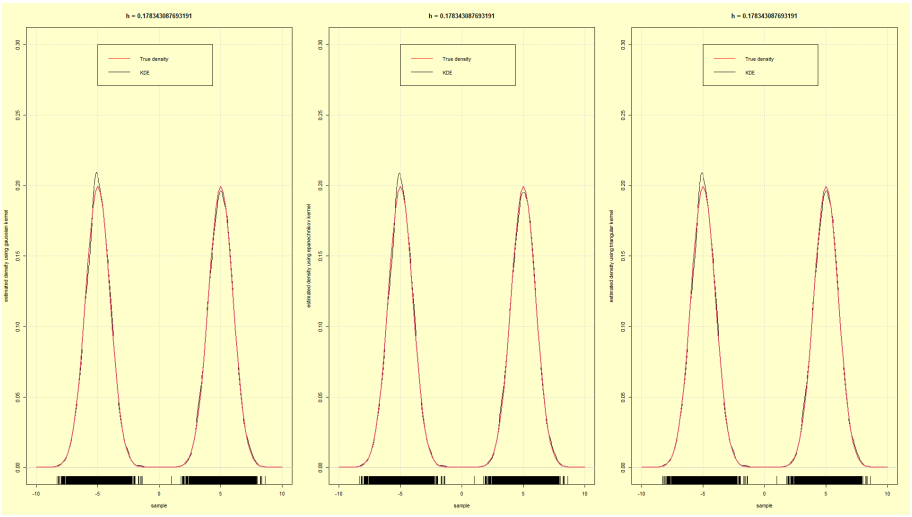

Figure 21:

6.7.2 Insights

- It is evident from the plot that for each of the kernels, the KDE here more or less coincides with the actual density, as h is optimally chosen.

6.8 Reflection Method

A random sample is taken from $exp(1)$, and Boundary correction is applied using reflection method.

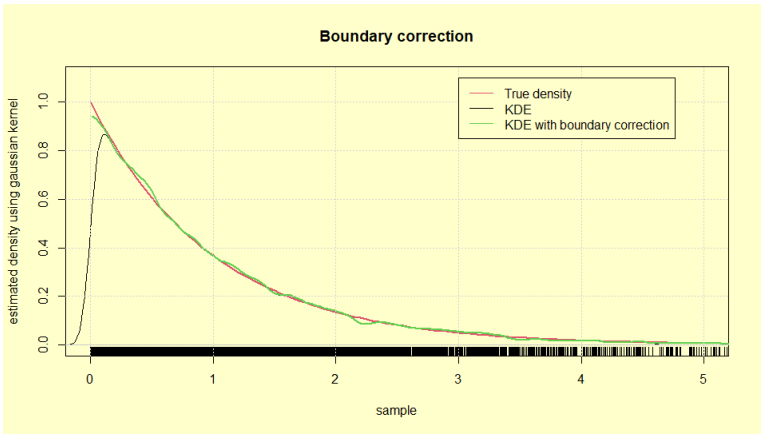

Figure 22:

6.8.1 Insights

- It is evident from the plot that the KDE with boundary correction is estimating the density in the neighborhood also.

7 Supplementary

Interested readers may visit this link for the R codes used in this project :
https://drive.google.com/drive/folders/1BCtkHVwcjrJeFsGP7m5FgyuTIfyygHOy?usp=sharing

8 References

[1] Turlach, Berwin. (1999). Bandwidth Selection in Kernel Density Estimation: A Review. Technical Report.
[2] Kernel density estimation and its application. Stanisław Weglarczyk. ITM Web Conf. (2018).
[3] https://bookdown.org/egarpor/NP-UC3M/kde-i.html
[4] On Boundary Correction in Kernel Estimation. NECIR Abdelhakim. YAHIA Djabrane. YOUSFATE Abderrahmane. SAYAH Abdallah. BRAHIMI Brahim.2016