Kernel Density Estimation

Kernels, Bandwidth, MSE, MISE

Jit Mondal & Sayan Dewanjee

Indian Statistical Institute

April 6, 2025

Introduction to Kernel Density Estimation (KDE)

Motivation

- **Problem:** How can we estimate an unknown probability density function (PDF) from observed data?
- **Histograms:** Simple but discontinuous and sensitive to bin choice.
- **Goal:** A smooth, non-parametric estimator that adapts to the data.

What is KDE?

- Places a *kernel function* (e.g., Gaussian, Epanechnikov) at each data point.
- Sums these kernels, scaled by a *bandwidth* h (controls smoothness).

Key Questions

- How does kernel choice affect the estimate?
- How do we select the optimal bandwidth h?
- How do we measure performance (MSE, MISE)?

Mathematical Formulation

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

Common Kernel Functions in KDE

Kernel Requirements:

- Symmetric: $K(-u) = K(u)$
- Non-negative: $K(u) \geq 0$
- Integrates to 1: $\int_{-\infty}^{\infty} K(u)\,du = 1$

Kernel	Formula $K(u)$	Support	Efficiency
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$	$(-\infty, \infty)$	95.1%
Epanechnikov	$\frac{3}{4\sqrt{5}}(1 - \frac{u^2}{5})$	$\|u\| \leq \sqrt{5}$	100%
Rectangular	$\frac{1}{2}$	$\|u\| \leq 1$	93.0%
Biweight	$\frac{15}{16}(1 - u^2)^2$	$\|u\| \leq 1$	99.4%
Triangular	$1 - \|u\|$	$\|u\| \leq 1$	98.6%
Cauchy	$\frac{1}{\pi(1+u^2)}$	$(-\infty, \infty)$	83.7%

Table 1: Kernel functions and their properties. Efficiency is relative to the Epanechnikov kernel.

Key Insight

The Epanechnikov kernel is *theoretically optimal* (minimizes MISE), but Gaussian is often preferred for smoothness and differentiability.

Error Metrics in KDE: MSE and MISE

Mean Squared Error (MSE) at a Point

The MSE measures the expected squared deviation of the estimator $f_n(x)$ from the true density $f(x)$ at a point x:

$$\mathrm{MSE}(x) = \mathbb{E}\left[\left(f_n(x) - f(x)\right)^2\right]$$

Decomposition:

$$\mathrm{MSE}(x) = \underbrace{\left(\mathbb{E}[f_n(x)] - f(x)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(f_n(x) - \mathbb{E}[f_n(x)]\right)^2\right]}_{\text{Variance}}$$

Key Insights

MSE captures the trade-off between:

- Bias: Systematic error from smoothing
- Variance: Sensitivity to random fluctuations

Mean Integrated Squared Error

Integrates MSE over all x:

$$\text{MISE} = \mathbb{E}\left[\int \left(f_n(x) - f(x)\right)^2 dx\right]$$

Decomposed form:

$$\text{MISE} = \int \text{Bias}^2(f_n(x))\,dx + \int \text{Var}(f_n(x))\,dx$$

Bandwidth Impact

- **Small** h:
  - Low bias
  - High variance
  - Noisy estimate

- **Large** h:
  - High bias
  - Low variance
  - Oversmoothed

Kernel Efficiencies in KDE

Efficiency of a Kernel

The efficiency of a kernel K is:

$$\text{Eff}(K) = \left(\frac{C(K_\epsilon)}{C(K)}\right)^{\frac{5}{4}} \times 100\%$$

where K_ϵ is the optimal Kernel &

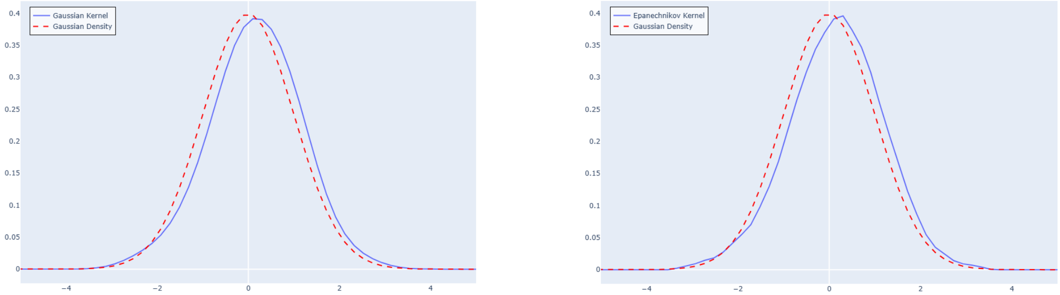
$$C(K) = \left\{\int_{-\infty}^{\infty} K^2(u)\, du\right\}^{\frac{4}{5}}$$

Kernel	**Efficiency**
Epanechnikov	100%
Biweight	99.4%
Triangular	98.6%
Gaussian	95.1%
Rectangular	93.0%
Cauchy	83.7%

Table 2: Efficiencies and key constants for common kernels

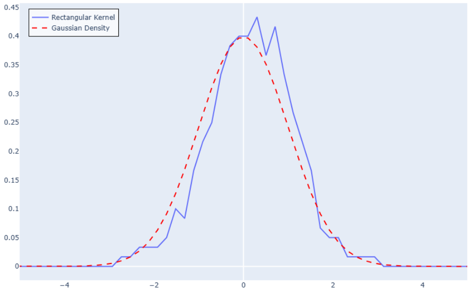
Kernel Density Estimation with 4 different kernels

Random sample(sample size = 60) from standard normal distribution is considered with smoothing parameter h=0.5

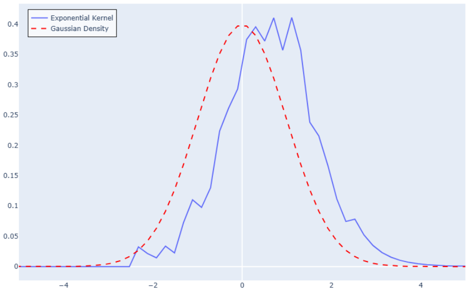


(a) Gaussian Kernel (b) Epanechnikov Kernel

Figure 1: Kernel Estimates using Different Kernels



(c) Rectangular Kernel



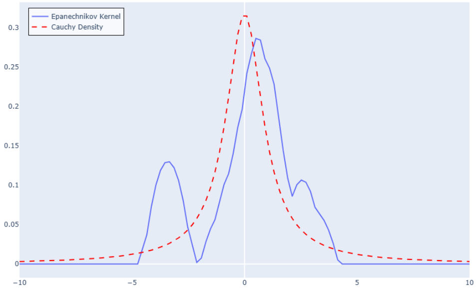
(d) Exponential Kernel

Figure 1: Kernel Estimates using Different Kernels

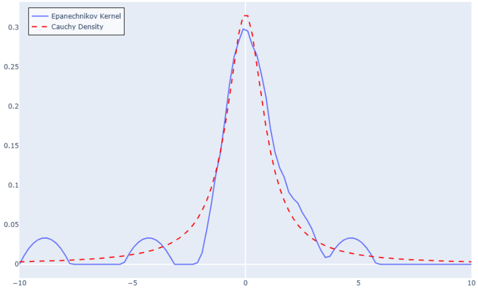
For first 3 kernels, the estimation is almost same for any kernels. In the last case , we used exponential kernel and it resulted almost same with a shift to the right. Possible reason is exponential is not symmetric kernel and takes value only in non-negative real line.

Kernel Density Estimation for Varying sample size

Random sample from standard cauchy distribution is considered with Epanechnikov kernel ($h_n = 0.5$)

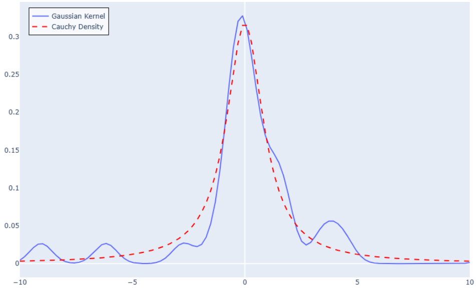


(a) $n = 10$

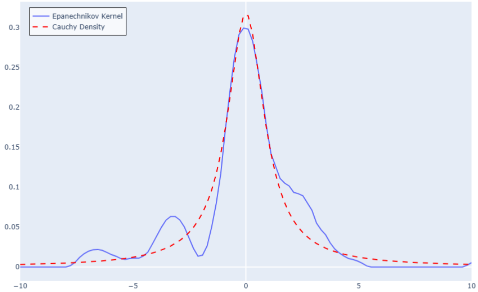


(b) $n = 20$

Figure 2: Kernel Density Estimation using Different Sample sizes



(c) $n = 30$



(d) $n = 60$

Figure 2: Kernel Density Estimation using Different sample sizes

Here we can see that with increase in sample size the estimation is getting better.

Bimodal Density estimation : Old Faithful Geyser Data

We used Old faithful Geyser data to show bimodal density estimation.

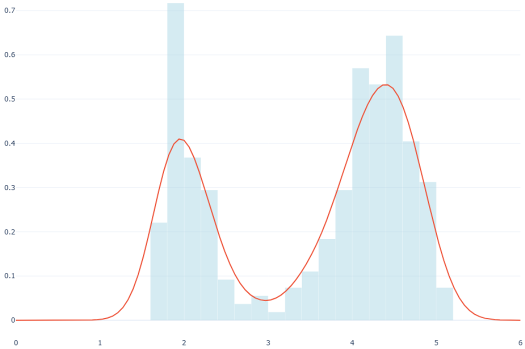
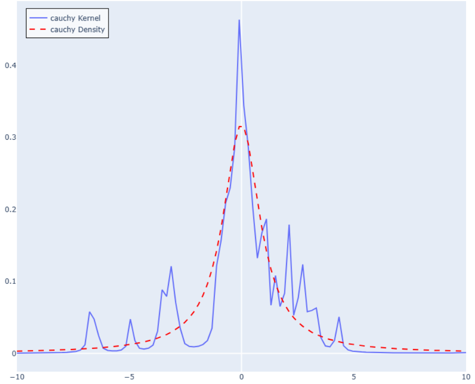


Figure 3: Kernel Density Estimation with Gaussian Kernel, $h = 0.25$ for old Faithful dataset

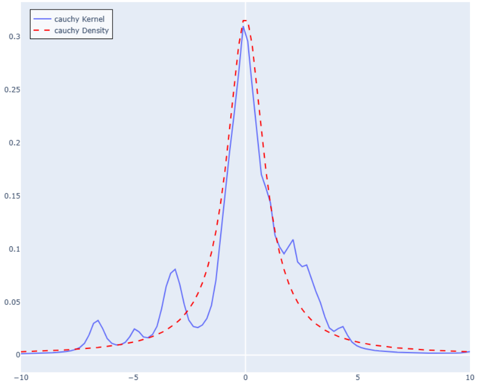
Here by using Gaussian kernel and suitable smoothing parameter we are able to fit a density with keeping its bimodal nature unharmed. Hence, with suitable kernel we can estimate bimodal and trimodal data too.

Different smoothing parameter

Random sample(sample size = 60) from standard cauchy distribution is considered with standard cauchy kernel

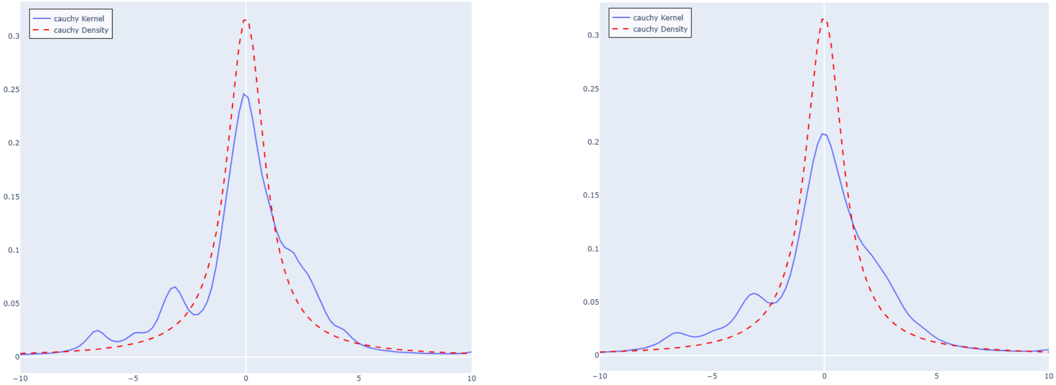


(a) $h = 0.1$



(b) $h = 0.3$

Figure 4: Kernel Density Estimation using Different choices of h

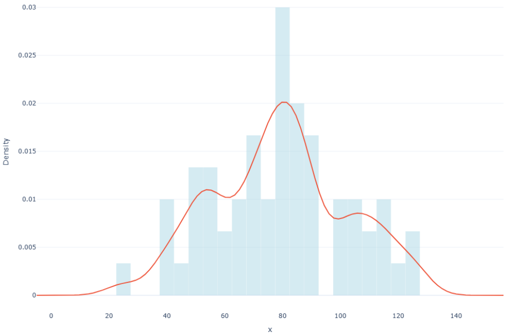


(c) $h = 0.5$ (d) $h = 0.7$

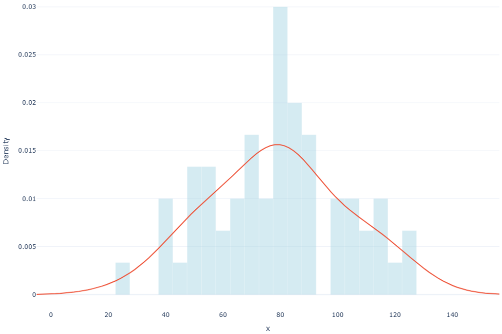
Figure 4: Kernel Density Estimation using Different choices of h

When h is small (near 0.1) there are several bumps indicating undersmoothing. Again when h is large(near 0.7) then real nature of data is obscured. From these 4 plots, we can say that optimal smoothing parameter is between 0.3 and 0.5

Annual Snowfall Data, Buffalo



(a) $h = 6$



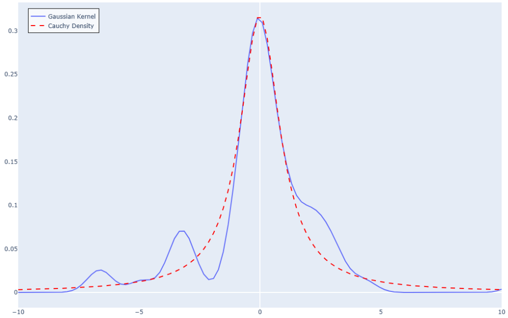
(b) $h = 12$

Figure 5: Kernel Desity Estimation of Snowfall Data Using Gaussian Kernel

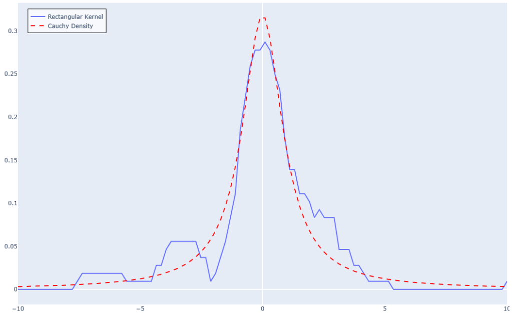
These two plots showed that choosing smoothing parameter is very crucial. Varying smoothing parameters results into two possible explainations, either a normal curve or a trimodal curve. In most of the cases, several smoothed plot is considered as sample serves several explanation for data.

Good Smoothing Parameter Depends on Kernel

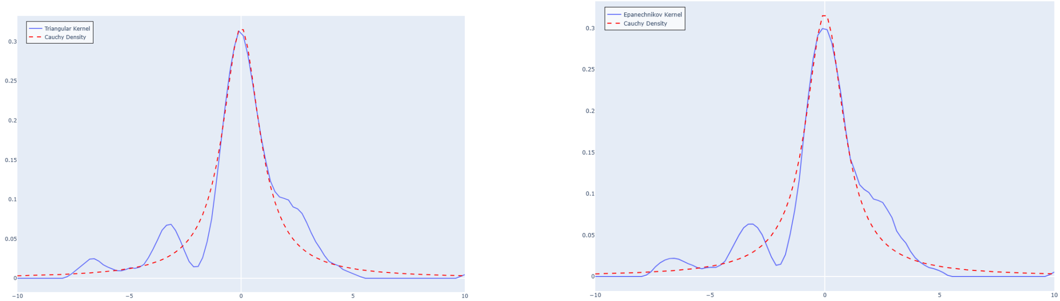
Random sample(sample size = 60) from Cauchy(0,1) Distribution is considered with Gaussian, Epanechnikov, Rectangular and Triangular kernel.



(a) Gaussian ($h = 0.5$)



(b) Rectangular ($h = 0.9$)



(c) Triangular ($h = 1.2$)

(d) Epanechnikov ($h = 0.5$)

Figure 6: Dependence of h_{opt} on Kernel

Optimal h_n with Reference to a Standard Distribution

The ideal value of h_n, from the point of view of minimizing the approximate MISE

$$= \frac{1}{4} h_n^4 \int_{-\infty}^{\infty} \left(f''(x)\right)^2 dx + \frac{1}{nh_n} \int_{-\infty}^{\infty} (K(t))^2 \, dt$$

can be shown by simple calculus to the equal to h_{opt}, where

$$h_{opt} = \left[\int_{-\infty}^{\infty} (K(t))^2 \, dt\right]^{\frac{1}{5}} \left[\int_{-\infty}^{\infty} \left(f''(x)\right)^2 dx\right]^{-\frac{1}{5}} n^{-\frac{1}{5}}$$

The above is a function of unknown $f''(x)$. Now a very easy and natural approach is to use a standard family of distributions to assign a value to this term in the above expression for the ideal bandwidth.

For example, if we take the standard normal distribution, then setting ϕ to be the standard normal density, we get

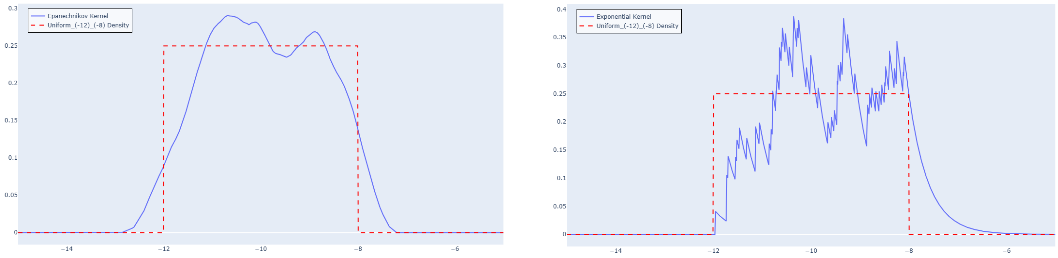
$$\int_{-\infty}^{\infty}\left(f^{''}(x)\right)^2 dx = \int_{-\infty}^{\infty}\left(\phi^{''}(x)\right)^2 dx = \frac{3}{8\sqrt{\pi}}$$

If a Gaussian Kernel is used, then the bandwidth obtained would be -

$$h_{opt} = \left(\frac{4}{3n}\right)^{\frac{1}{5}}$$

Using Assymetric Kernel

Random sample(sample size = 60) from Uniform(-12,-8) Distribution is considered with Epanechnikov kernel and exponential kernel.



(a) Epanechnikov Kernel

(b) Exponential Kernel

Figure 7: Kernel Estimates when $h = 0.4$

For exponential kernel, notice i's left side has no boundary bias but right hand side has boundary bias which is changing exponentially. Using non-negative kernel has the tendency for such right side boundary bias.

Using Negative Kernel

Consider now we are dropping the non-negativity condition from our kernel and let's assume $K(x)$ satisfies :

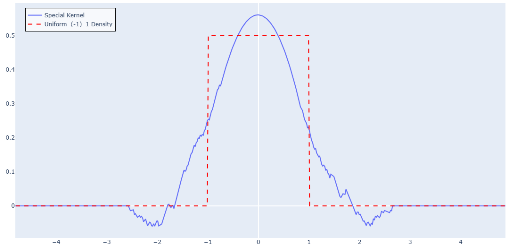
- Symmetric: $K(-t) = K(t)$
- Integrates to 1: $\int_{-\infty}^{\infty} K(t)\, dt = 1$
- $\int_{-\infty}^{\infty} t^2 k(t)\, du = 0$
- $\int_{-\infty}^{\infty} t^4 k(t)\, du = k_4 \neq 0$

Now minimizing asymptotic variance subject to these constraints lead us to a negative and discontinuous kernel:

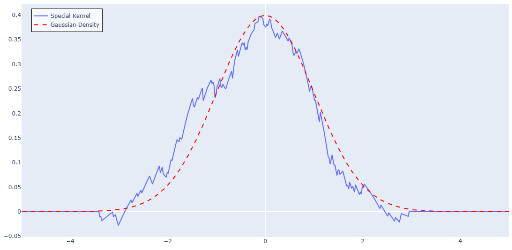
$$k(x) = \frac{3}{8}(3 - 5x^2) I_{|x| \leq 1}$$

This kernel density does not satisfy non-negative criterion and discontinuous at 1 and 1.

Here we have used this negative kernel for density estimation.



(a) Sample from Uniform (-1,1)



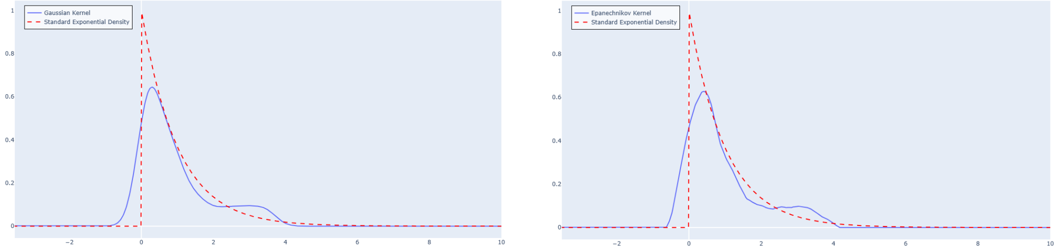
(b) Sample from Normal(0,1)

Figure 8: Kernel Estimates with negative kernel

Both the plots indicate our estimated function takes negative value which is impractical. It happens as we used a negative kernel. Again for bounded density on plot a) probability leakage happens.

Probability Leakage

Random sample(sample size = 60) from standard Exponential Distribution is considered with Gaussian kernel and Epanechnikov kernel.



(a) Gaussian Kernel (b) Epanechnikov Kernel

Figure 9: Kernel Estimates for $h = 0.3$

As here data is bounded, KDE can struggle near boundaries. The kernel functions placed near the boundaries extend beyond the data domain, causing probability mass to "leak" out. On both these plots, there are positive mass on negative side of real line causing "Probability Leakage".

Boundary Effects

Random sample(sample size = 60) from Uniform(-1,1) Distribution is considered.

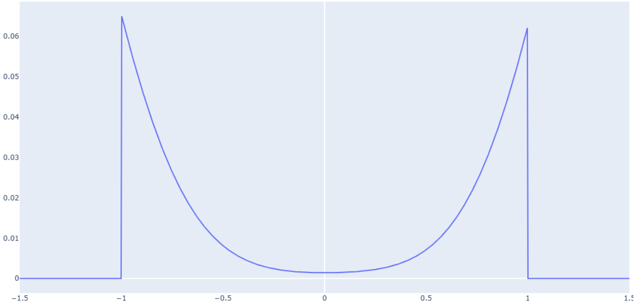
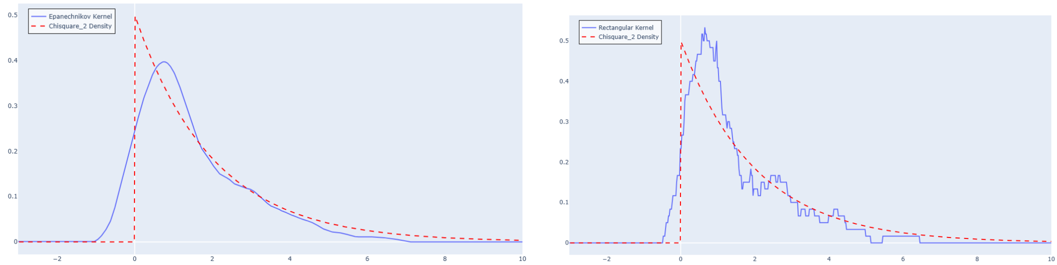


Figure 10: Plot of MSE with gaussian kernel ($h = 0.5$)

The standard KDE estimator assumes that the data extends infinitely beyond the observed boundaries, which is not true for bounded data. This assumption leads to a bias near the boundaries, as the estimator "tries" to spread the density beyond the actual boundaries. Note MSE increases as we further move to boundaries.

Random sample(sample size = 60) from Chi square Distribution(d.f. 2) is considered with Rectangular kernel and Epanechnikov kernel.



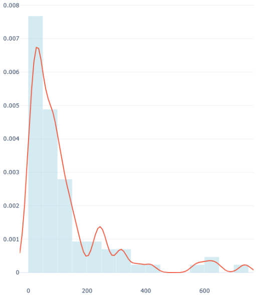
(a) Epanechnikov Kernel

(b) Rectangular Kernel

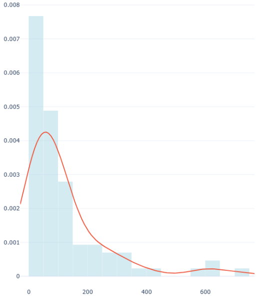
Figure 11: Kernel Density Estimates for $h = 0.5$

As we try to estimate a distribution with non-negative support, here again we can notice Boundary Bias. So a major drawback of KDE is that it fails to perform on boundary region.

Long Tailed Distribution : Suicide Data



(a) $h = 20$



(b) $h = 60$

Figure 12: Kernel Density Estimates for Suicide Study Data

We estimated density from Suicide Data. The estimate shown in Fig 12a is noisy in right hand tail.In Fig 12b we can notice still a slight bump in tail.

As smoothing parameter is constant across the entire sample,there is tendency for spuriou noise to appear in the tail for sample from long tail distribution. Again if estimates are smoothed sufficiently, important information will be masked (Fig. 6.b).

Though suicide data is non-negative, both plots have mass on negative part of real line causing "Probability Leakage".

Boundary Bias : A problem to KDE

Limitation of KDE

- KDE approximates the data distribution but extends support beyond actual data range.
- Non-zero density assigned to regions with no observations.
- Leads to inaccurate estimation when true support is known.

Remedy

In such cases we can use reflection boundary correction strategy, proposed by Jones 1993.

- In step 1, we will reflect the KDE tails against the boundaries.
- In step 2, we will sum the reflected tails with the rest of the KDE so that the total density area is 1.

Illustration: Step 1

Random sample (size = 100) from Uniform(0,1) is considered.

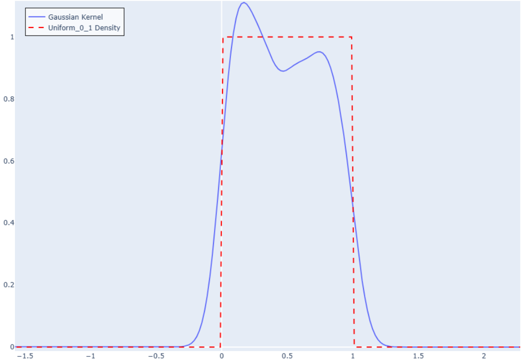


Figure 13: Kernel density estimation with gaussian kernel, $h = 0.1$

Figure shows that estimated density has mass outside $(0, 1)$, causing Boundary Bias.

Illustration: Step 2

In step 2, we reflected KDE against the line x = 0 and x = 1.

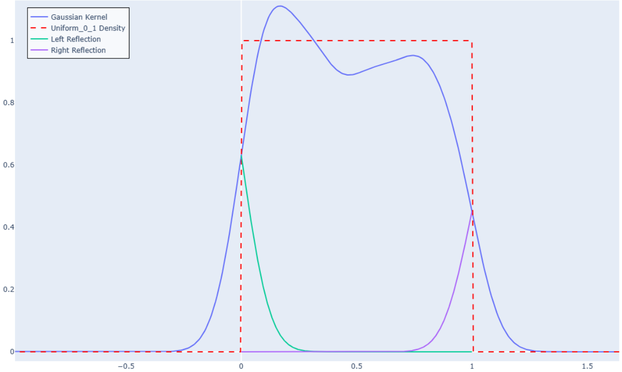


Figure 14: Reflecting kernel with boundary lines

Illustration: Step 3

In step 3, we sum the reflected tails with the rest of the KDE .

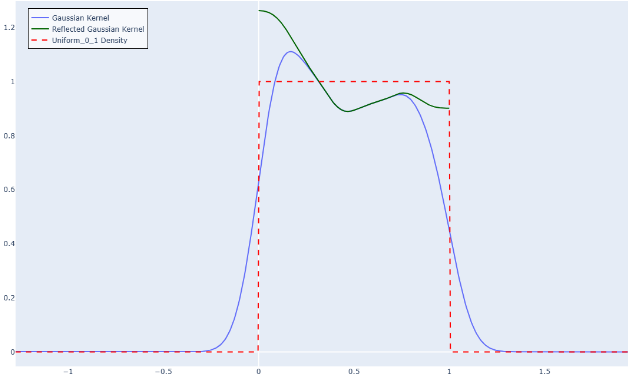


Figure 15: Final estimate after using Reflection Strategy

Now we end up with a valid density estimation.

Illustration

Random sample (sample size=100) from standard exponential distribution is considered.

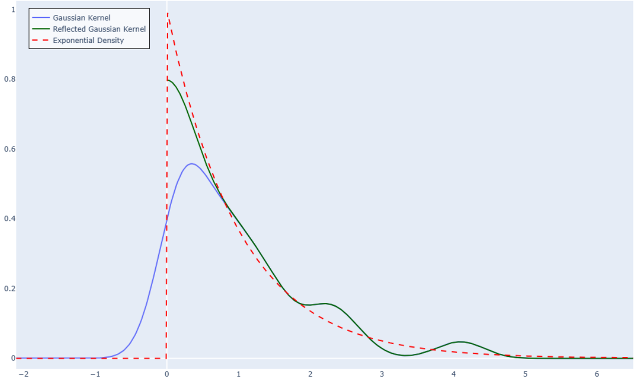


Figure 16: Final estimate after using Reflection Strategy

As we can see, a significant part of the KDE is placed in the negative region, which does not make much sense since the exponential distribution supports only non-negative numbers. If we specify bounds $[0, \infty]$, we can get the bounded KDE which covers only positive numbers.

Reference:

- Density Estimation for Statistics and Data Analysis by B. W. Silverman
- Jones, M. C. (1993). Simple boundary correction for kernel density estimation. Statistics and Computing, 3(3), 135-146.
- Kernel estimation of cumulative distribution function of a Random Variable with bounded support,Aleksandra Baszczyska1
- Kernel density estimation boundary correction: reflection - website by Andrey Akinshin

Thank you for listening !

Why be chunky when you can be classy?