2024

DENSITY ESTIMATION

Ganeshi Umayanagana
s15669
2020s17913

REPORT

TABLE OF CONTENT

LIST OF FIGURES

INTRODUCTION

Density estimation is a statistically important technique. In simple terms, density estimation involves estimating how likely different values are to occur within a dataset. Density estimation is calculating the likelihood that various values will occur inside a dataset. It offers a method for simulating the data point distribution and gaining an understanding of the underlying structures and patterns. Tasks like anomaly detection, classification, and the creation of new data points can benefit from this information.

Density estimation has various approaches, such as parametric and nonparametric approaches. Here only discuss the nonparametric approach with R. Because R offers a wide range of nonparametric density estimation tools that make it easier to explore data point distributions without making strong assumptions. Practitioners can simulate the underlying probability distribution and obtain a visual representation of the likelihood landscape of the data by using techniques like R's kernel density estimation (KDE). This helps consumers understand the inherent unpredictability and gives them the capacity to make judgments based on a thorough understanding of the dataset.

TYPES OF THE DENSITY ESTIMATIONS

Density estimations can be broadly classified into two groups.

1. **Parametric density estimation**

   data is from a known family it means parametric methods assume specific mathematical form for given distribution.

   - Normal density estimation
   - Weibull distribution estimation

2. **Nonparametric density estimation**

   which attempts to flexibly estimate unknown distribution. Its mean nonparametric is a more flexible method and making fewer assumption about data structure.

   - Histogram
   - Kernel density estimation
   - Frequency polygon density estimate

UNIVARIATE DENSITY ESTIMATION

1. Histograms

The classical nonparametric estimator of a density is the histogram. It offers estimates that are piecewise constant and discontinuous. The R function density () implements a significant class of smooth density estimators that include kernel approaches. In essence, these estimators are locally weighted averages.

Histogram density estimate,

$$\hat{f} = \frac{V_k}{nh} \quad t_k \leq x \leq t_{k+1}$$

h -number of points in the interval V_k

n – sample size

In histogram basically, we face two types of problems,

1. How to decide the number of bins

- Square Root Rule

  Calculate the square root of the total number of data points and use that as the number of bins.

$$Number\ of\ bins\ =\ sqrt(n)$$

- Sturges' Rule

  The optimal width of class intervals, $\frac{R}{1+ log_2 n}$

R is a sample range, based on normal assumptions, it is not good for skewed distribution or multiple distribution. Tends to over smooth.

- Scott's Rule

  Squared error (ISE ($\hat{f}$)) = $\int \left(\hat{f}(x) - f(x)\right)^2 dx$

  MISE ($\hat{f}$) = E [ISE ($\hat{f}$)]

  Scott show,

  MISE = $\frac{1}{nh} + \frac{h^2}{12}\int f'(x)^2 dx + O(\frac{1}{n} + h^3)$

  Optimal choice of bin width is, $h_n^* = \left(\frac{6n}{\int f'(x)^2 dx}\right)^{\frac{1}{3}}$

  The Scott's normal reference rule in normal distribution $\hat{h} = 3.49\ \hat{\sigma} n^{\frac{-1}{3}}$

- Freedom- Diaconis Rule

  Considers the interquartile range (IQR). It is less sensitive to outliers. Trends to under smooth.

  Optimal bin width, $\hat{h} = 2(IQR)n^{\frac{-1}{3}}$

- Custom choice

  Depending on the properties of your data and the insights you hope to obtain, choose the number of bins.

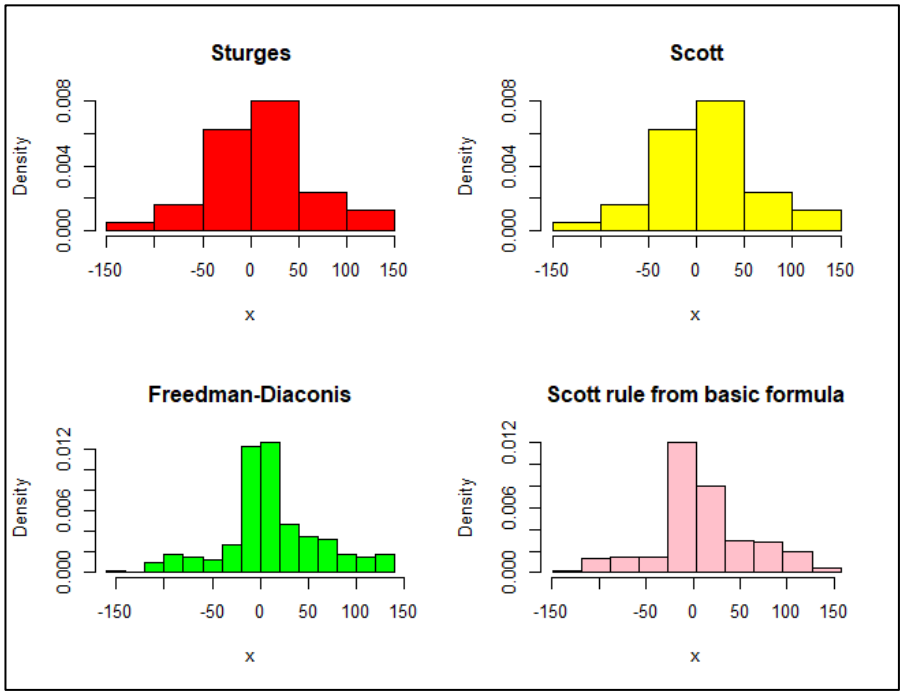

Figure 1- Density of varies number of bits

2. ***Boundaries and width of class intervals.***

Limitations To establish the boundaries of your histogram, find the lowest and highest values in your dataset. Ensure that all data points fall within the specified range.

Your histogram's interpretation depends heavily on the width of each bin, or class interval. While narrow bins can highlight specifics, they can also cause noise or overfitting. Extensive bins could over smooth the data and obscure significant trends.

- Over smoothing (Wide Bin Width) – reduced sensitivity to noise and outliers, Highlights broader trends and general patterns.
- Under smoothing (Narrow Bin Width) – capture fine details and variations in the data. Suitable for datasets with intricate patterns.

2. Frequency polygon density estimate.

The frequency polygon continues using linear interpolation. It doesn't need to make any conclusions about the data's underlying probability distribution. Rather, it depends on presenting the observed frequency of intervals or values.

Optimal frequency polygon bin width, $h_n^{fp} = 2[\frac{49}{50} \int f"(x)^2 dx]^{-\frac{1}{5}} n^{-\frac{1}{5}}$

For normal density, $h_n^{fp} = 2.15\ \sigma\ n^{-\frac{1}{5}}$

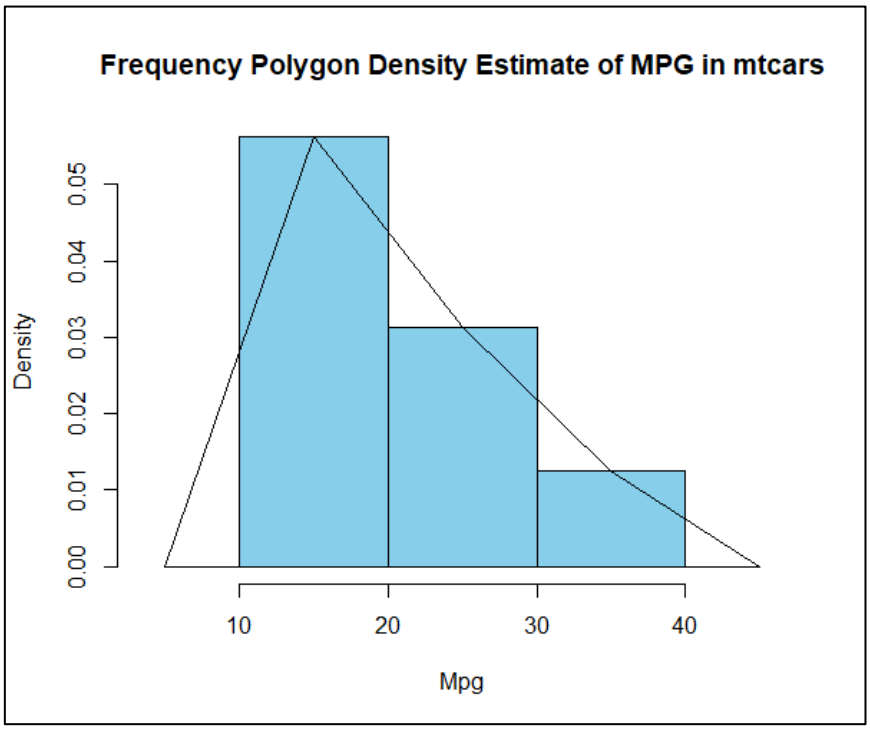

Figure 2- Polygon Density Estimate

3. The average shifted histogram.

This method is data-driven and flexible, and it works especially well with multimodal or complicated distributions. Histograms are shifted and averaged using the ASH method to get the underlying density.

$$\widehat{f_{ASH}}(x) = \frac{1}{m}\sum_{j=1}^{m} \hat{f}_j(x)$$

In here, class boundaries for estimate $\hat{f}_j(x)$ are shifted by h/m from the boundaries for $\hat{f}_j(x)$

The optimal bin width for the naïve ASH estimate of a normal density,

$$h^* = 2.576\,\sigma\, n^{-\frac{1}{5}}$$

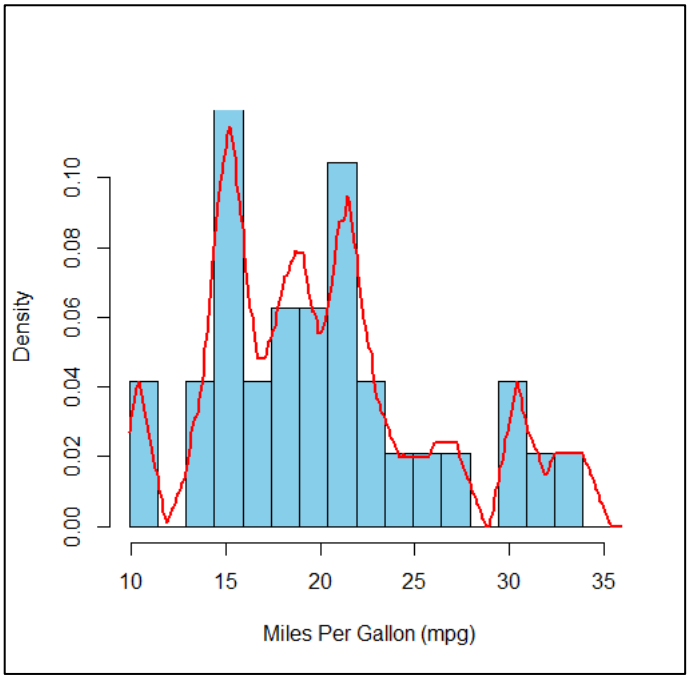

Figure 3- Average Shifted Histogram

KERNEL DENSITY ESTIMATION

A non-parametric statistical technique called kernel density estimation (KDE) is used to calculate the probability density function (PDF) of a continuous random variable. KDE is adaptable and versatile, making it appropriate for a broad range of data types and distributions, in contrast to parametric approaches that require a particular functional form for the underlying distribution.

The basic idea behind KDE is to use a kernel function on each data point and add them together to depict the data as a smooth continuous curve. Each data point is given a weight by the kernel, which is usually a smooth, symmetric function that adds to the total estimate of the probability density. The Gaussian (normal) kernel, the Epanechnikov kernel, and other popular kernels are examples; each has unique properties that affect how the predicted density is shaped.

A crucial parameter in KDE is the bandwidth, denoted as h. The bandwidth controls the kernel's width and, in turn, the degree of data smoothing that is done. While a greater bandwidth produces a smoother estimate that might miss fine-scale features, a lower bandwidth produces an estimate that is more variable and detailed, catching local changes. Selecting the appropriate bandwidth is essential and frequently requires making a trade-off between preserving the estimate's noise level and capturing intricate features.

Suppose a sample $x_1, x_2, \ldots, x_n$ creates a histogram with bin width 2h.

If x is a center of a region, the density estimator is, $\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right)$

Where $w(t) = \frac{1}{2} I(|t| < 1)$ is a weight function and also this estimation called **naïve density estimator**.

In kernel density estimation, a kernel function $K(x)$ is substituted for the weight function w(t) in the naïve estimator so that $\int_{-\infty}^{\infty} K(t)dt = 1$. In here weight function $w(t) = \frac{1}{2} I(|t| < 1)$ is a kernel function and it called as a **rectangular kernel**.

A univariate kernel density function with a kernel function $K(x)$ is,

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

KDE is widely used in exploratory data visualization and data analysis. It is useful for finding underlying patterns, seeing various modes, and visually examining the shape of the data since it gives a smooth and continuous picture of the distribution of the data. Numerous statistical software programs, such as R, incorporate KDE. The density() function is frequently used for KDE, including options for bandwidth selection techniques, visualization, and kernel selection.

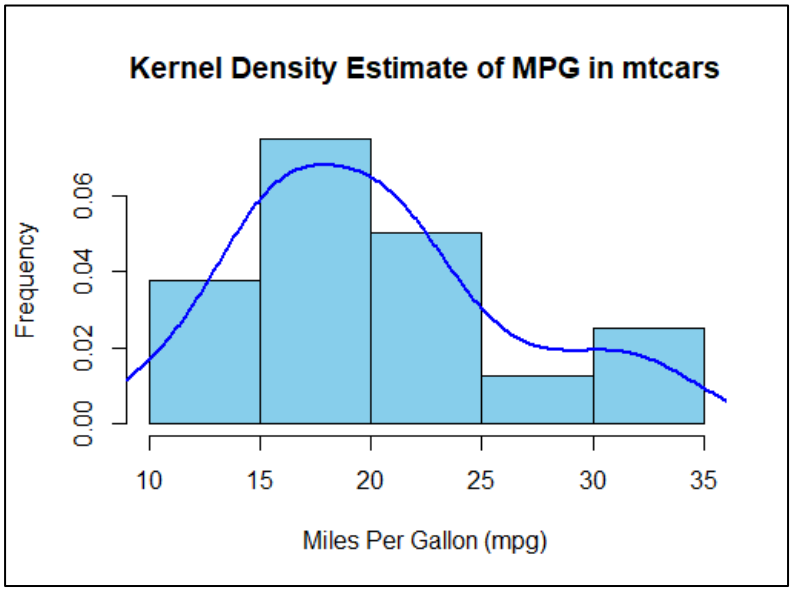

Figure 4- kernel Density Estimation

Even though KDE is a strong tool, problems at data borders must sometimes be addressed by carefully choosing the kernel and bandwidth parameters. In other situations, boundary correction techniques should also be taken into account. To discover the ideal bandwidth values, cross-validation techniques can be used to balance the estimate process's variance and bias.

Using univariate kernel density estimator $\hat{f}_K(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$,

Where h > 0 is the smoothing bandwidth, which regulates the degree of smoothing, and K(x) is the kernel function, which is often a symmetric, smooth function, such as a Gaussian. The KDE basically smoothest each data point Xi into a small density bump and adds all of these bumps to reach the final density estimate. The sample below shows the KDE and each tiny bump it produces.

As previously mentioned above,To ensure that the KDE $\hat{f}(x)$ is a probability density function, in particular, the second criterion is required. In actual use, the density estimator does alter slightly when the kernel function is used. Following, we examine the most popular kernel functions.

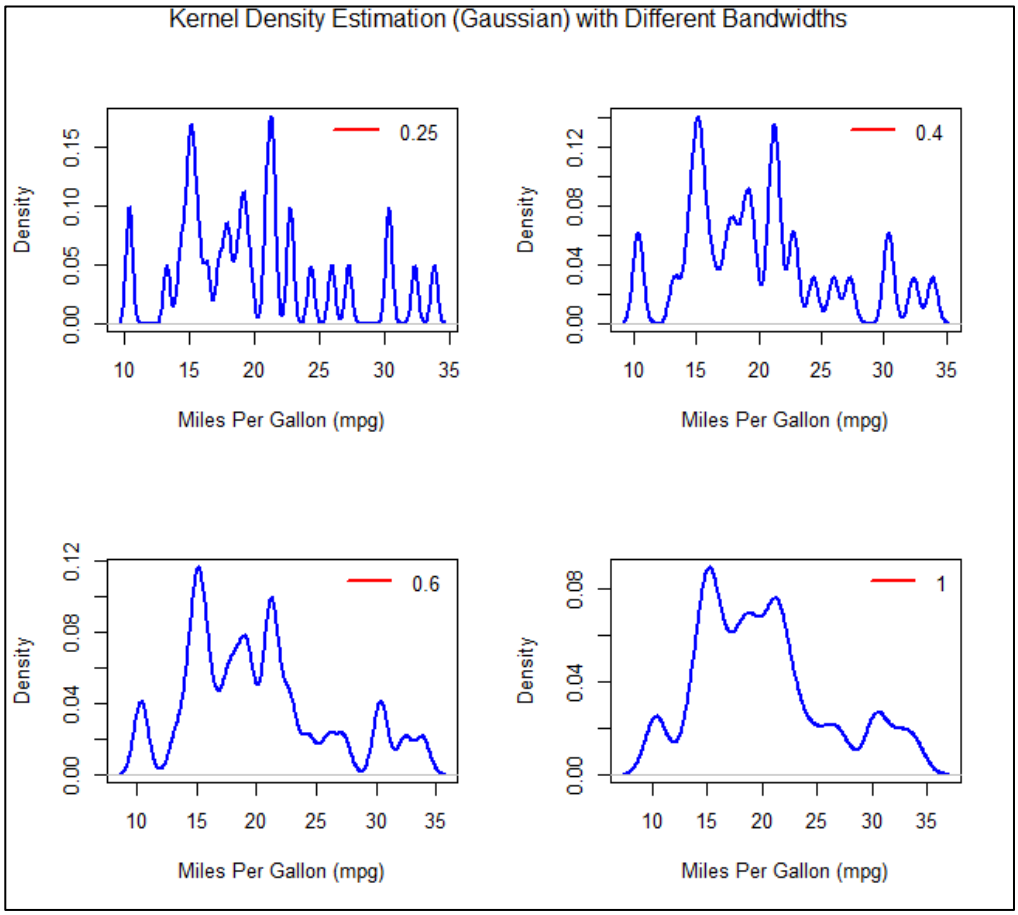

Figure 5-Kernel Density Estimation With Different Bandwidth

By considering these separate plots easily see when the bandwidth increases then the number of waves in given range decreases.

So, we can see the different of the bandwidths using the all waves in one plot, so that bellow plot shows different bandwidth and the waves of each bandwidths.

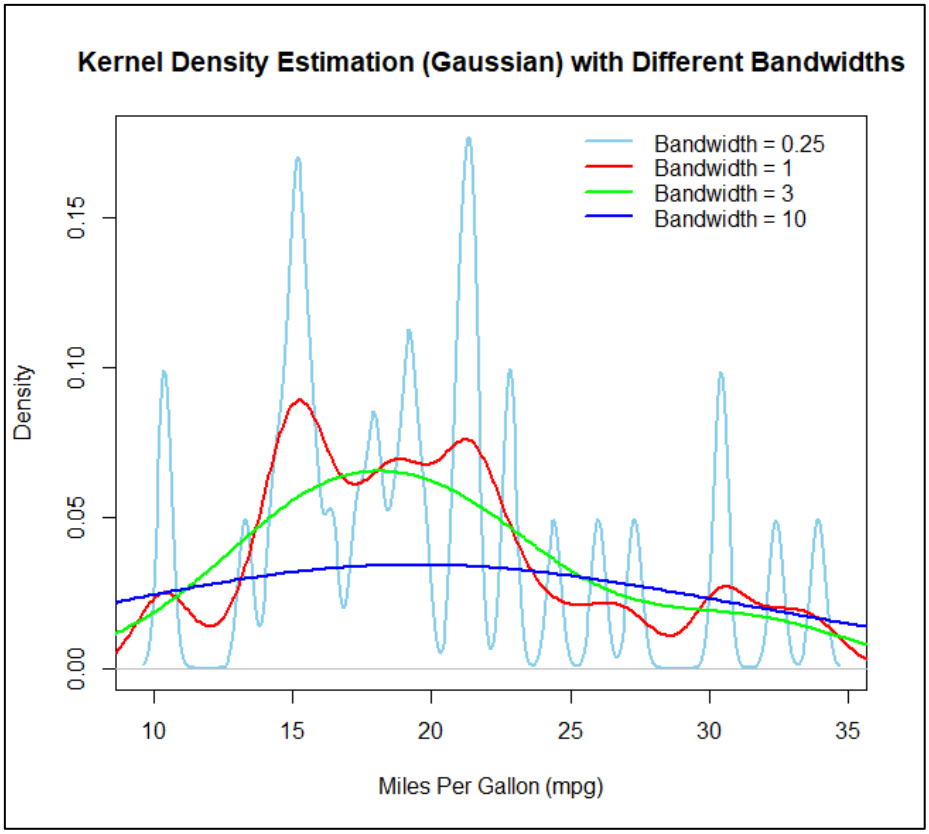

Figure 6-Comparing Different Bandwidth

We can plainly see that our density curve has a lot of wavy features when h is too little (the sky blue curve). This is an indication of undersmoothing, where there is not enough smoothing and some of the structures found by our approach may just be the result of randomness. On the other hand, the two bumps are smoothed out when h is too large (the light blue curve). Over smoothing is the term used to describe this circumstance, in which a significant number of smoothing hides some crucial structures.

Popular kernel functions

Name	$K(t)$	support	σ_K^2	Efficiency
Gaussian	$(1/\sqrt{2\pi})\exp(-\frac{1}{2}t^2)$	$\mathbb{R}$	1	1.051
Epanechnikov	$\frac{3}{4}(1-t^2)$	$\lvert t\rvert<1$	1/5	1
Rectangular	$1/2$	$\lvert t\rvert<1$	1/3	1.076
Triangular	$1-\lvert t\rvert$	$\lvert t\rvert<1$	1/6	1.014
Biweight	$\frac{15}{16}(1-t^2)^2$	$\lvert t\rvert<1$	1/7	1.006
Cosine	$\frac{\pi}{4}\cos\frac{\pi}{2}t$	$\mathbb{R}$	$1-8/\pi^2$	1.001

Figure 7- Popular Kernel Functions

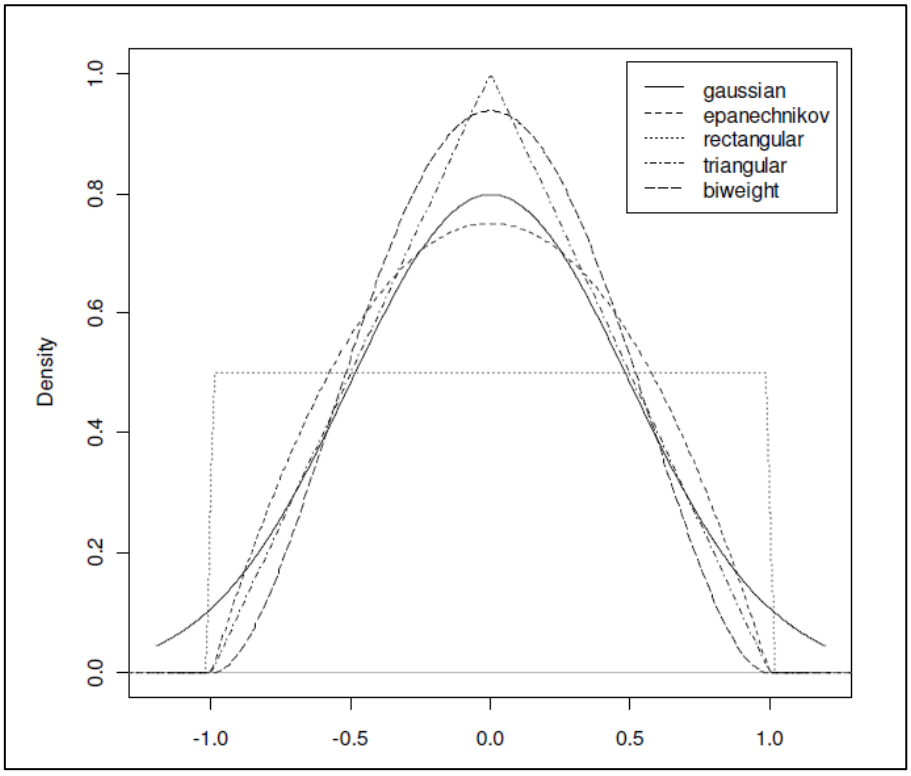

Figure 8-Comparison of Popular kernel Functions

- Gaussian kernel and a normal distribution , Optimized IMSE is $h = 1.06\sigma n^{-\frac{1}{5}}$
- Density not unimodal, bandwidth tends to over smooth,
- The Silverman's rule of thumb : $h = 0.9 \min(s, \frac{IQR}{1.34}) n^{-\frac{1}{5}}$

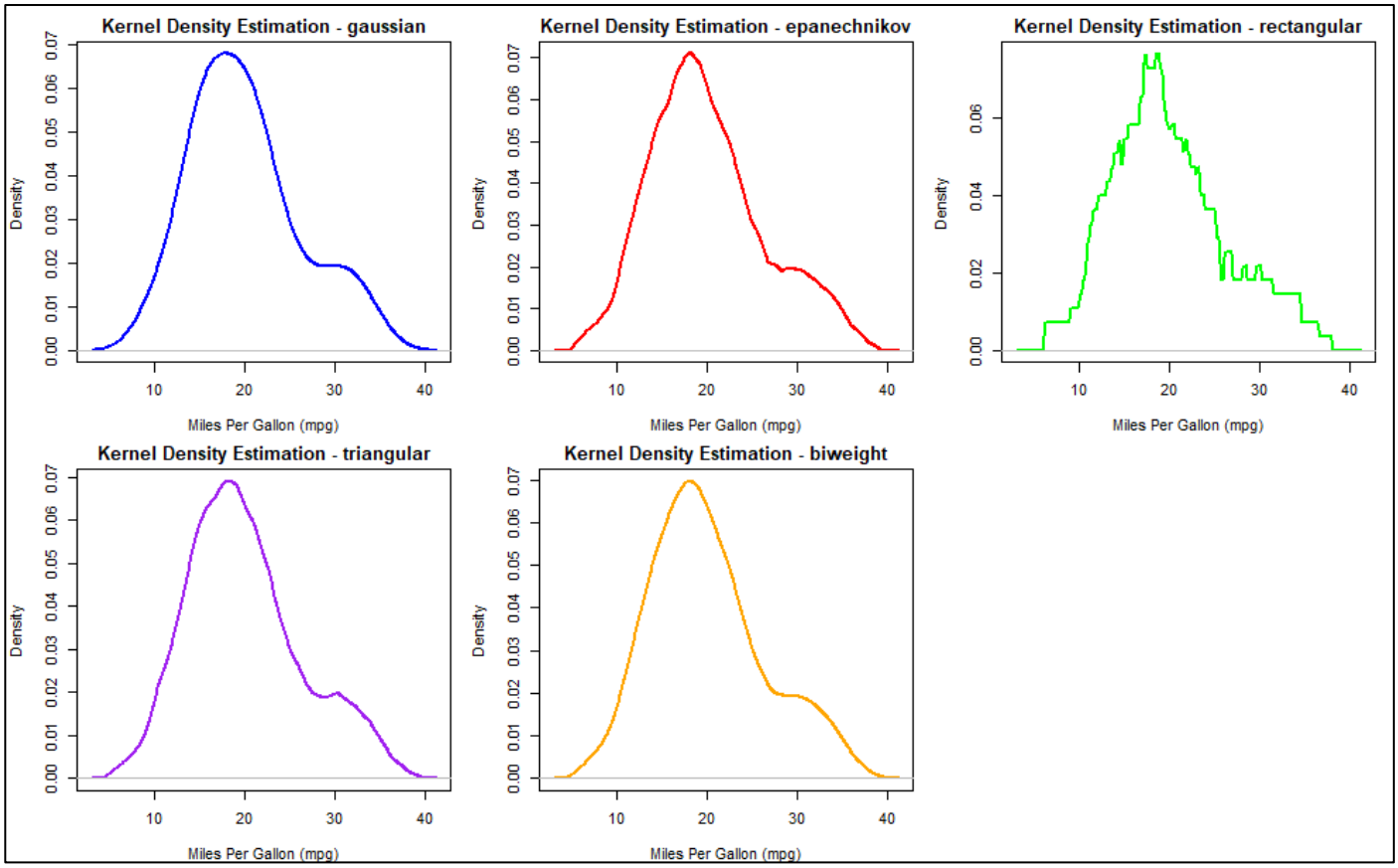

Figure 9- Density of Kernel Functions

Boundary kernels

Boundary kernels When determining the limits of a density's support, it is less accurate.

Any practically important kernel function can naturally continue onto the boundary due to these border kernels, which are solutions of the same variational issue as the kernels in the interior.

In the kernel estimate doesn't fit well at the discontinuity points.

If the discontinuity occurs at the origin, we use reflection boundary technique.

Then we add reflection of the sample $-x_1, -x_2, \ldots, -x_n$ to the data. So here we use 2n points for estimate the density but used only n to determine the bandwidth.

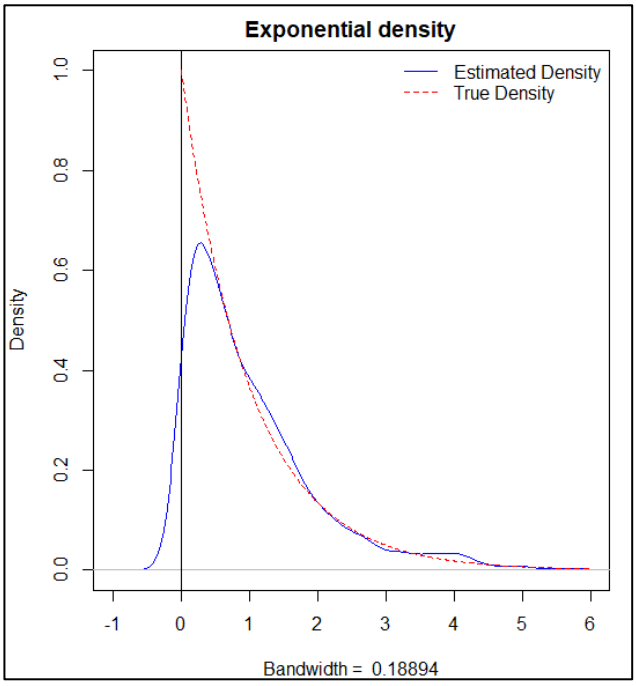

Figure 10- Exponential Density

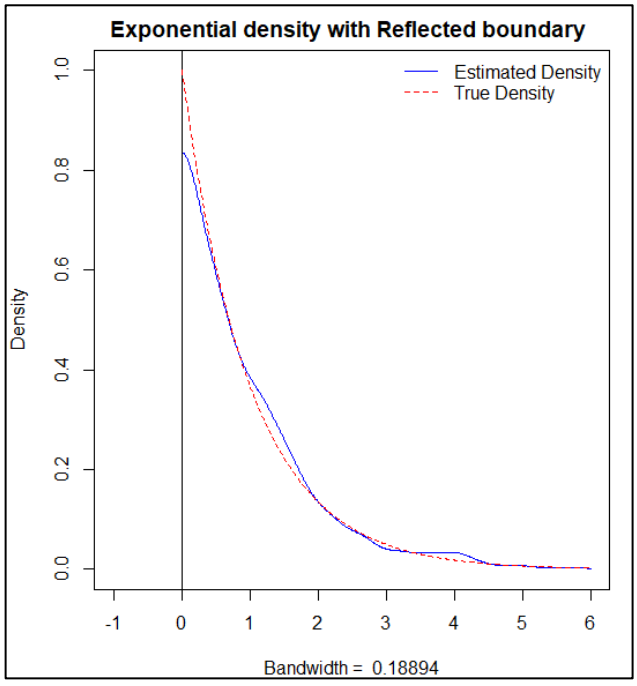

Figure 11 - Exponential density With Reflected Boundary

BIVARIATE AND MULTIVARIATE DENSITY ESTIMATION

Bivariate frequency polygon

Additionally, the bivariate frequency polygon is investigated. The frequency polygon and alternative density estimators are compared. In here we need to define two-dimensional bins and count the number of observations in each bin before constructing a bivariate frequency polygon.

In bivariate frequency polygon gives graphical representation, it's shows the visually encapsulates the relationships and trends between two variables.

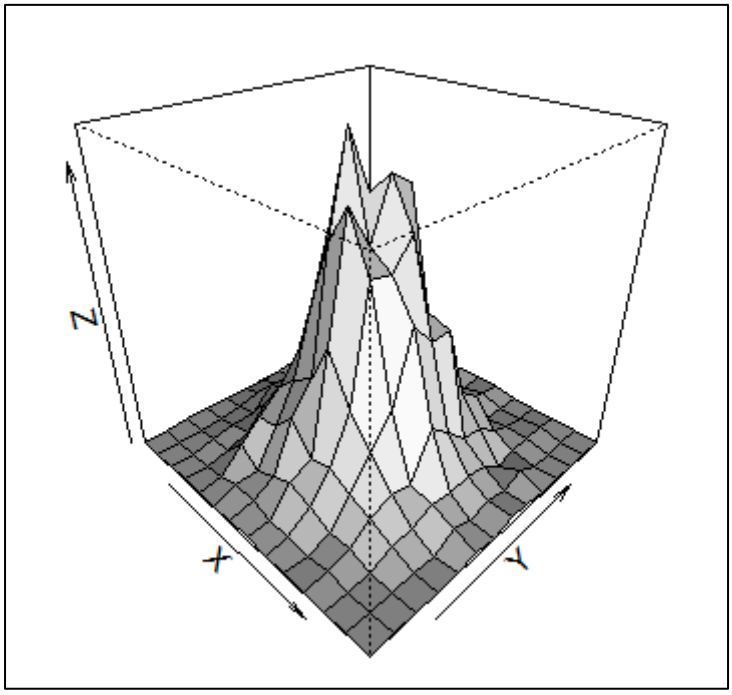

Figure 12- Bivariate Frequency Polygon

Bivariate ASH

A nonparametric technique for determining the joint probability density function of a bivariate dataset is the bivariate average shifted histogram, or ASH. This is a two-dimensional adaptation of the univariate Average Shifted Histogram. When working with data points that don't fit into a particular parametric distribution, the approach is especially helpful.

Suppose {x,y} are the bivariate data have been sorted into an array bins. The parameter $m = (m_1, m_2)$ is the number of shifted histograms of each axis used in the estimate.

The bivariate ASH estimate is,

$$\hat{f}_{ASH}(x,y) = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \hat{f}_{ij}(x,y)$$

Multidimensional kernel methods

The concept of "multidimensional kernel methods" describes statistical approaches that apply the idea of kernel density estimation (KDE) to multidimensional datasets. Multidimensional kernel methods enable the estimation of probability density functions in higher-dimensional spaces, whereas classic kernel density estimation is often used to univariate or bivariate data.

And also this method allow statisticians and data scientists to extend the principles of kernel density estimation to datasets with multitude with variables.

General Kernel Function

Suppose $x = (x_1, x_2, \dots, x_d)$- random vector of d dimension, $K(x)$- d dimensional kernel function.

The bandwidth is equal in all dimensions,

$\hat{f}_K(x) = \frac{1}{nh_1^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}\right)$ here $h1$- smoothing parameter.

Product Kernel Function

In this product kernel function estimate with smoothing parameter $h = (h_1, h_2, \dots, h_d)$ is,

$$\hat{f}(x) = \frac{1}{n(h_1 \dots h_d)} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{x_i - X_{ij}}{h_j}\right)$$

For uncorrelated multivariate normal data, the optimal bandwidths are

$$h_j^* = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \sigma_i n^{-\frac{1}{d+4}}$$

By minimizing the MISE.

APENDIXS

- https://drive.google.com/file/d/1D5l4rVz9LoSfYJXcCSvBp9tsk72xO7ds/view?usp=drive_link

REFERENCES

- https://machinelearningmastery.com/probability-density-estimation/
- https://cswr.nrhstat.org/densit
- https://vita.had.co.nz/papers/density-estimation.pdf
- https://en.wikipedia.org/wiki/Multivariate_kernel_density_estimation
- https://bookdown.org/epeterson_2010/docs/introduction-to-kernel-density-estimation.html
- https://www.geeksforgeeks.org/non-parametric-density-estimation-methods-in-machine-learning/