Application of Nonparametric Kernel Density Estimation in Colombo stock market

(S&P SL 20)

for the Bachelor of Science Honours Degree

in Financial Mathematics and Industrial Statistics

Supervisor: Dr. M.K. Abeyratne

GROUP NO: 02

By

G.H.V.C.Darshani – SC/2020/11781

Department of Mathematics

University of Ruhuna Matara

January 26, 2024

Acknowledgement

In performing our case study, we had to get the help and guidelines from some respected persons, whom we should be most grateful. We are quite happy to have finished successfully this task.

Firstly, we would want to express our appreciation to, Supervisor of MFM3151 case study 1, providing us with an excellent task guideline after multiple chats. As well as we would like to thank for all of team members of the supervising team Senior Professor Leslie Jayasekara, Dr. W. A. R. De Mel, Dr. K.D.Prasangika, Dr. L. W. Somathilake, Dr. D.M. Samarathunga, Dr. N. Yapage, Ms. K. C. N. Shanthidevi ,Ms. S.D.L.Geeganage, Dr. H L Jayetileke, Ms. K.G.P. Hansani and Ms. S.D.M. Dilshani for helping us throughout and motivating us as a team.

In addition, we would like to express our sincere appreciation to everyone who helped us to write this case study, both directly and indirectly. We were inspired to improve our case study by the many insightful comments and ideas that were provided on this case study, particularly from our classmates. We thank all the people for their help directly and indirectly to complete our case study.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background of the Study

Financial markets, being dynamic and influenced by multifaceted factors, demand sophisticated analytical approaches. KDE, a nonparametric method, holds promise in revealing hidden patterns in stock prices. As the Colombo Stock Market experiences substantial price fluctuations, the application of nonparametric kernel methods becomes particularly relevant for accurate modeling and prediction.

When considering the Sri Lanka's financial environment, Colombo Stock Market (CSM) is doing a major role as responsible institution of the country, which investors can trade shares of listed companies. The number of variables, including exchange rate fluctuations, economic conditions and political developments affect the CSM's behavior and performance. Investors, regulators, and policymakers can evaluate the stability and efficiency of the market as well as the risk and return of their investments by having a thorough understanding of the distribution as well as the fluctuations of the CSM returns. Using advanced statistical methods becomes necessary considering this in order to improve the reliability of predictions and risk assessments.

This research proposal evaluates the use of nonparametric kernel density estimation in an effort to fill in the gaps in the techniques currently used in the study of the CSM. Conventional approaches frequently include assumptions about certain distributions and parametric structures, which may cause them to ignore the variety of non-linear features that are present in stock movements. A more detailed comprehension of the basic trends and distributions is made possible by nonparametric techniques, especially KDE, which offer a flexible and data-driven substitute.

Traditional parametric approaches are still used extensively in financial market analysis, although it is still unclear how well they capture the real characteristics of the Colombo Stock Market. Assuming distributional forms and parametric structures could have drawbacks that make it difficult to accurately describe the dynamics of the underlying data. As such, it is imperative to investigate other approaches that are more suited to the special features of the Colombo Stock Market.

1.2 Research approach

In here the study we use Kernal density estimation for the given Colombo stock market data set applying nonparametric kernel methods in the context of the Colombo Stock Market involves leveraging these statistical techniques to analyze and model financial data.

1.3 Research design

In here the data set (Colombo stock market) that has more changing price. To predict a valuable Kernal density estimator we want get their return so respect to the last one year to make a best model we found the return of the stock price.

$$R_{(t+1)} = \log(\, p_{(t+1)}\,) + \log\,(\, P_{(t)}\,)$$

$R_{(t+1)}$ = Rate of return

$P_{(t+1)}$ = Current trading price (Closing price)

$p_{(t)}$ = Previous trading price (Closed price)

1.4 Research Question

The Colombo Stock Market presents a unique challenge in terms of price dynamics, necessitating an exploration of advanced statistical techniques for meaningful analysis. The central problem addressed in this study is how nonparametric kernel density estimation can enhance the analysis of stock prices within the Colombo Stock Market.

This study proposal asks the following questions to address the previously described problems.

- In what ways might nonparametric kernel density estimation improve the Colombo Stock Market stock price analysis?
- What special patterns of the data from the Colombo Stock Market make nonparametric methods potentially useful?
- What is the difference between the outcomes of classic parametric approaches and nonparametric kernel density estimation?

1.5 Research Aim

This research aims to investigate the use of non-parametric KDE as a new analytical method in reference to the S&P SL20. By means of a comprehensive investigation of non-parametric techniques, the research aims to augment our comprehension of the fundamental patterns in stock prices, investigate the degree to which KDE can be tailored to the particularities of the S&P SL20 index prices, and combine the outcomes with non-parametric approach.

1.6 Significance of the Study

The significance of this study lies in its contribution to advancing analytical methodologies within financial markets. By exploring the application of nonparametric KDE in the Colombo Stock Market, it seeks to enhance our understanding of stock price patterns, providing insights that can inform investment decisions and risk management.

Chapter 2

Literature Review

Classic parametric approaches provide solid basis for financial market analysis, But the problem of using parametric approaches is the underlying assumptions have sparked interest in nonparametric alternatives. When considering Kernel Density Estimation, the importance of this work, which attempts to provide fresh perspectives into the examination of stock prices in this financial context, is highlighted by the lack of research using these methodologies to the S&P SL20.

The work by (Wang, Y., & Wang, J. (2011).) reflects the application of nonparametric KDE in the Hong Kong stock market, laying a foundation for similar applications in other financial environments. The study emphasizes the flexibility and adaptability of KDE, enabling a detailed comprehension of stock price distributions without assuming rigid parametric structures.

Moreover, (Wijayasiri, M.P.A., & Abeyratne, M.K. (2015).) Trading in financial services are at the core of Sri Lankan market in Colombo Stock Exchange. This is measured using two indicator, the ASPI and the S&PSL-20. In as much as it is a serious concern, the availability of data about the distribution density of stock index fluctuations are nowadays pivotal aspect in financial markets. Distribution of Returns in a relation to Aspi Index and S & P SL 20 as indexes in a Colombo Stocks Markets.

This research endeavors to extend the principles applied in the Hong Kong stock market to the unique landscape of the CSM, aiming to enhance decision-making processes and refine financial evaluations.

Chapter 3

Methodology

3.1 Nonparametric kernel density estimation

Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. The basic idea behind KDE is to place a kernel (a smooth, symmetric function) at each data point and then sum up these kernels to obtain a smooth estimate of the underlying distribution.

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

where;

$\hat{f}(x)$ = Estimated density at point 'x'

n = number of data points

h = the bandwidth (a smoothing parameter that determines the width of the kernel),

x_i =the data points,

K(u) = the kernel function, which is a symmetric, non-negative function that integrates.

3.7. Kernel Density Estimation Process

Kernel Density Estimation (KDE) process is a non-parametric way to estimate the Probability Density Function (PDF) of a random variable. Let's move to the process of the KDE.

3.7.1 Selecting the Kernel function

According to the smoothness of estimated density function, we can choose appropriate Kernel function among Gaussian kernel, Epanechnikov kernel, Biweight kernel, rectangular kernel, Triangular kernel etc. Epanechnikov kernel is chosen according to the density function we obtained. Here is the Epanechnikov kernel function $K(u)$,

$$K(u) = \begin{cases} \frac{4}{3}(1 - u^2) & \text{for} - 1 \leq \text{u} \leq 1 \\ 0 & otherwise \end{cases}$$

The Epanechnikov kernel is favored for its optimal properties in terms of mean integrated squared error (MISE) when compared to other kernels, such as the Gaussian kernel. The kernel is symmetric, non-negative, and has a unique shape with a flat top. Its compact support makes it particularly suitable for applications where the data is limited to a specific range.

3.7.2 Optimal Bandwidth Selection

The bandwidth (h) determines the width of the kernel and, consequently, the smoothness of the estimated density function. The optimal bandwidth is crucial for obtaining a reliable and accurate density estimate. One common approach to selecting the optimal bandwidth is through cross-validation.

Cross-validation involves dividing the dataset into training and testing sets iteratively. The idea is to find the bandwidth that minimizes the mean integrated squared error (MISE) or a similar criterion across different folds.

By choosing the bandwidth that minimizes the LSCV score, you aim to find the optimal balance between oversmoothing and undersmoothing, leading to a more accurate estimate of the underlying probability density function.

3.7.3 Choosing the optimal bandwidth

The smoothness of the density and the width of the kernel are controlled by bandwidth parameter. So it's important to choose most appropriate bandwidth. We can use "Cross Validation method" to find the optimal bandwidth. For the Epanechnikov kernel, the cross-validation process involves finding the bandwidth that minimizes a criterion like the Least Squares Cross Validation (LSCV) score.

In this study, we adopt classical approaches for bandwidth selection, with a particular focus on Cross Validation (CV), a widely utilized method known for its simplicity and applicability in both univariate and multivariate density estimation. CV is chosen over plug-in methods, such as those presented by Sheather and Jones (1991), due to its effectiveness without requiring intricate derivations or preliminary estimates. CV has been demonstrated to provide advantages for multivariate densities as well (Sain et al., 1994), as highlighted by Loader (1999).

Two notable CV methods, Likelihood Cross Validation (LCV) and Least Squares Cross Validation (LSCV), are considered. Habbema et al. (1974) introduced LCV, defined as the maximization of the log-likelihood function.

$$LCV(h) = \frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_{-i}(h)$$

Here, $\hat{f}_{-i}(h)$ denotes the leave-one-out density estimate, given by $\hat{f}_{-i}(h) = \frac{1}{n-1}\sum_{j\neq i} K_h(X_i - X_j)$. An alternative method, LSCV, was proposed by Rudemo (1982) and Bowman (1984). It seeks to minimize the integrated squared density.

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

Despite its popularity, LSCV is known for its high variability and a tendency to undersmooth data. Additionally, it involves computational expenses, especially in the case of multivariate densities, due to the computation of the integrated squared density.

It is essential to note that LCV, while avoiding some computational challenges, may exhibit sensitivity to extreme observations and the tail heaviness of the underlying distribution (Schuster and Gregory, 1981; Hall, 1987).

$$MISE(\widehat{f}) = \int \left(\hat{f}(x) - f(x)\right)^2 dx$$

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

$$\widehat{MCV}\left(\hat{f}\right) = \frac{1}{n}\sum\nolimits_{i=1}^{n} \int \hat{f}_i(x)^2\, dx - \frac{2}{n}\sum\nolimits_{i=1}^{n} \hat{f}_{-i}(x_i)$$

Optimal bandwidth can be chosen which minimizes the value of $\widehat{MCV}\left(\hat{f}\right)$. Here we can see how smoothness changes with the bandwidth in Epanechnikov Kernel.

Chapter 4

Data Analysis and Interpretation

4.1 Identification of key features

- **Closing Prices**

Considered a fundamental indicator, closing prices reflect the final valuation of a stock for a specific trading day. Changes in closing prices can signal trends and market sentiment.

- **Trading Volumes**

Volume measures the number of shares traded during a given period. High trading volumes often accompany significant price movements and can indicate the strength of a trend.

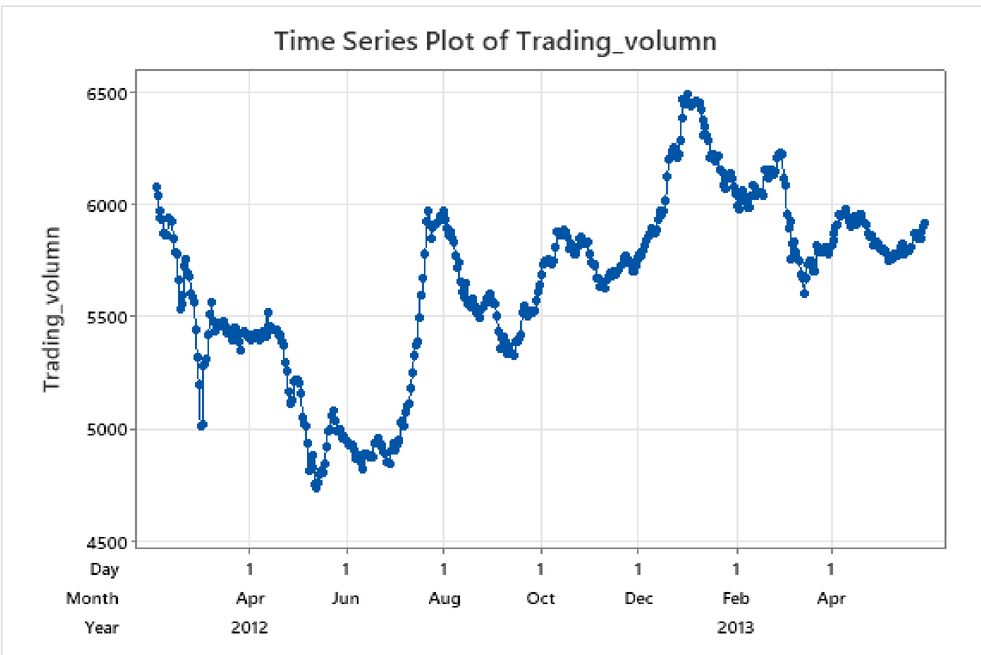

Figure 1: Time series plot for variation of trading volume

- **Moving Averages**

Utilize moving averages to smooth out price fluctuations and identify trends over time. Short-term and long-term moving averages may provide insights into short-term volatility and long-term market trends.

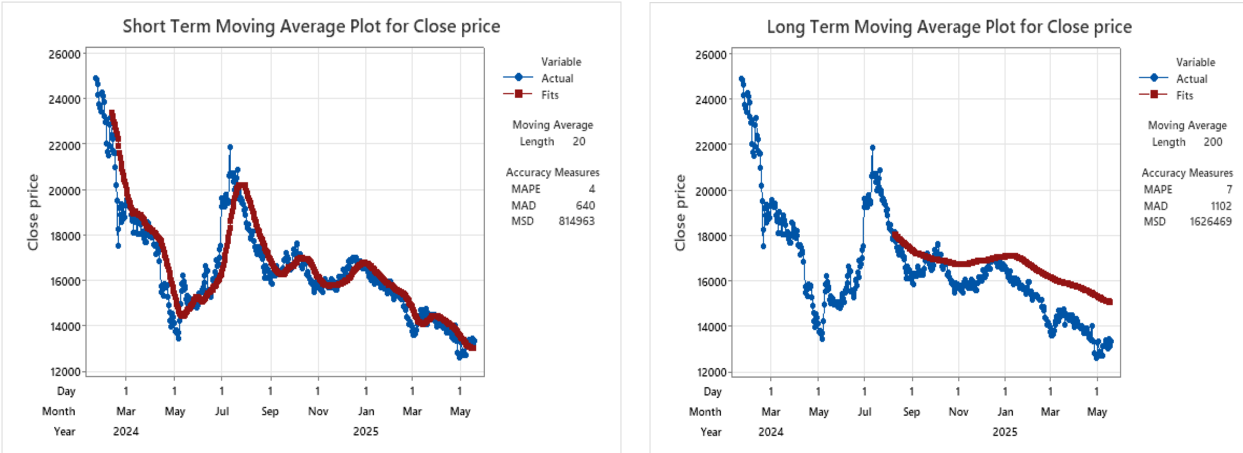

Figure 2: Short term MA and long-term MA for closing prices

- **Volatility Measures**

Incorporate volatility indicators (e.g., standard deviation of returns) to assess the degree of variation in stock prices. Volatility can impact trading strategies and risk management.

4.2 Data Preprocessing

The data analysis began with data preprocessing to ensure the reliability of the Colombo Stock Market dataset. The steps involved loading the raw data from the Excel file, which included critical information such as date, closing price, log price, and returns. To handle missing values, any rows with either missing dates or closing prices were removed. Additionally, linear interpolation was applied to fill in missing values, and returns were computed for subsequent analysis.

4.2.1 Dataset

Gather historical data for the S&P SL 20 index, including daily stock prices, trading volumes, and other relevant financial indicators. The dataset spans a sufficiently long period to capture diverse market conditions.

4.2.2 Data dictionary

Table 1: Data dictionary of each variable

Variables	Variable type	Measurement Scale
Date	Temporal	Ordinal
Close price	Numerical	Ratio
Rate of Return	Numerical	Ratio

4.3. Exploratory data analysis

4.3.1 Numerical Summary

Exploratory Data Analysis (EDA) is a critical step in kernel density estimation to identifying patterns, clusters, or anomalies within the stock market data, understanding the distribution of stock returns or prices is crucial for risk assessment, discerning the differences or similarities in the behavior of various stocks or market segments.

Table 2: Summary Statistics

Mean	Median	Standard Deviation	Minimum value	Maximum Value	JB value
-0.0005595	-0.00075525	0.007839	-0.047331	0.041608	0.0010

The mean (-0.0005595) and median (-0.00075525) values are close, suggesting a relatively symmetrical distribution without significant skewness. The standard deviation (0.007839) measures the dispersion of data points around the mean. A higher value indicates more variability in the data. The data range from a minimum value of -0.047331 to a maximum value of 0.041608, showcasing the spread of the dataset. The Jarque-Bera (JB) value of 0.0010 is a test for normality. Lower JB values indicate a closer fit to a normal distribution.

4.3.2 Graphical Summary

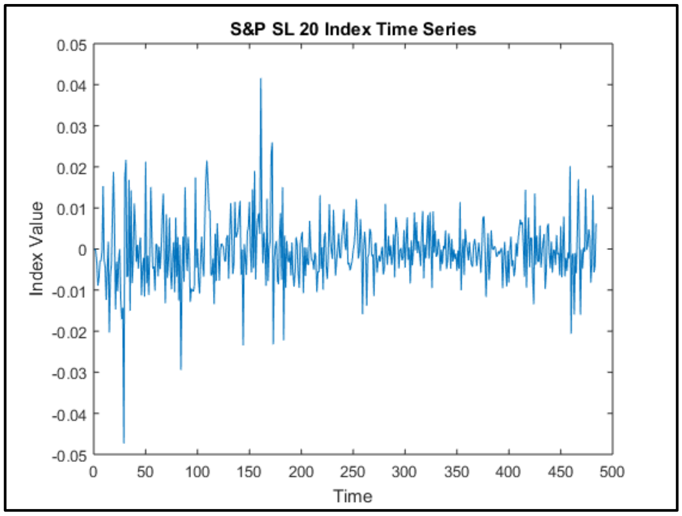

Figure 3: Time series plot of returns

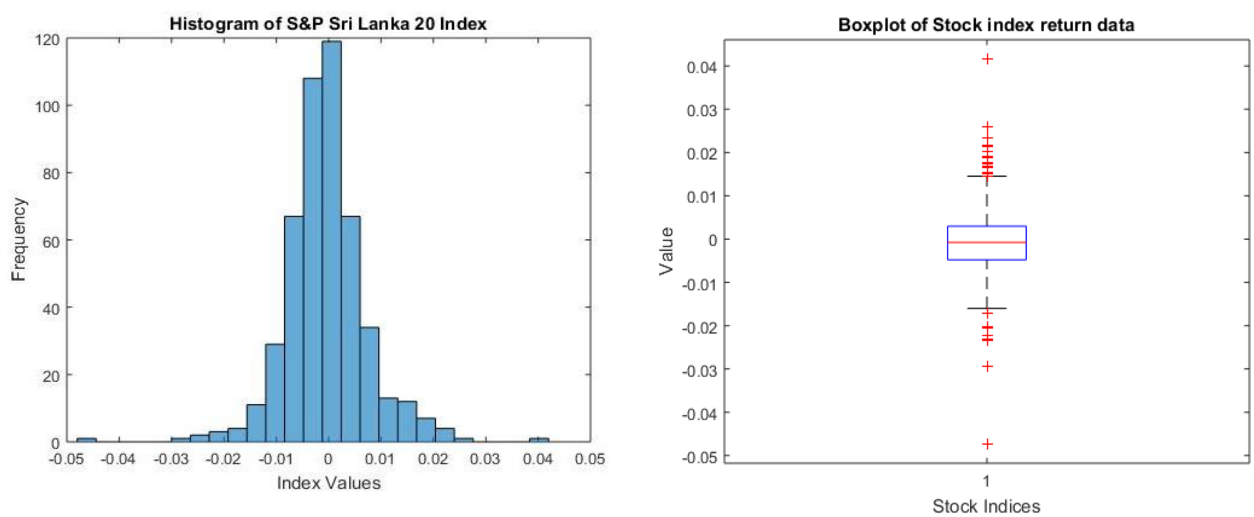

Figure 4:Histogram and boxplot of S&P SL 20 index

For this research model we expect to use the Jarque-Bera Test to check the normality.

H_0: The data is normally distributed.

H_1 : The data does not come from a normal distribution.

According to this we will check the normality in future research. After below we will be going to apply Kernel density estimator for Colombo stock market data.

The data appears to be relatively symmetrically distributed around a mean close to zero, showing moderate variability. The JB value suggests that the data might conform reasonably well to a normal distribution.

4.4 Kernel Density Estimation (KDE)

The focus of the analysis was on understanding the distribution of returns in the Colombo Stock Market. The nonparametric kernel density estimation (KDE) method, utilizing the Epanechnikov kernel, was employed. This method involves placing a kernel at each data point and summing up these kernels to create a smooth estimate of the underlying distribution. The KDE was visualized through a plot, showcasing the distribution of returns.

The selection of the Epanechnikov kernel over the Gaussian kernel in kernel density estimation is driven by several advantageous properties.

- ✓ The Epanechnikov kernel exhibits computational efficiency due to its compact support, being nonzero only within a specific range, which is particularly beneficial for large datasets.
- ✓ Known for its optimality in mean squared error among certain kernels, the Epanechnikov kernel strikes a balance between bias and variance, providing a robust compromise between oversmoothing and undersmoothing.
- ✓ Its symmetric, non-negative, and straightforward mathematical form contributes to simplicity in both computation and interpretation.

Here is the Epanechnikov kernel density function we have chosen,

$$K(u) = \begin{cases} \frac{4}{3}(1 - u^2) & \text{for} -1 \le u \le 1 \\ 0 & otherwise \end{cases}$$

The following figures indicate the kernel density plot at different bandwidth values to obtain the understanding of distribution of returns.

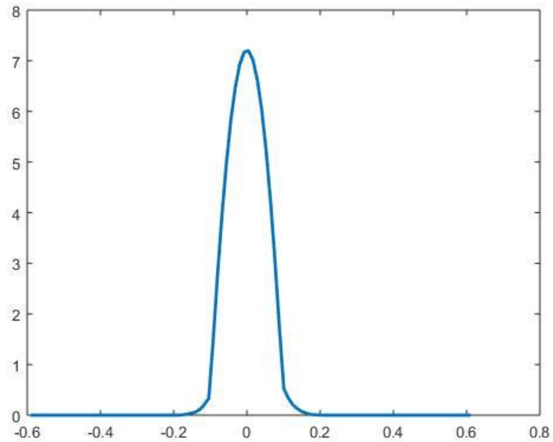

Figure 5: Bandwidth = 0.1

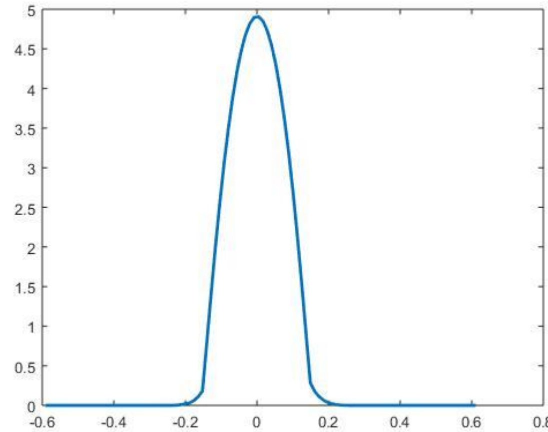

Figure 6: Bandwidth = 0.15

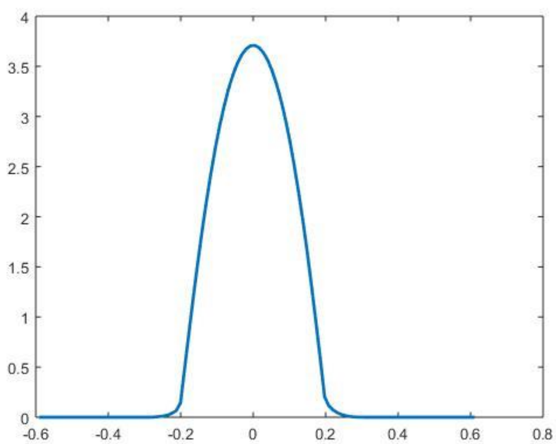

Figure 7: Bandwidth = 0.2

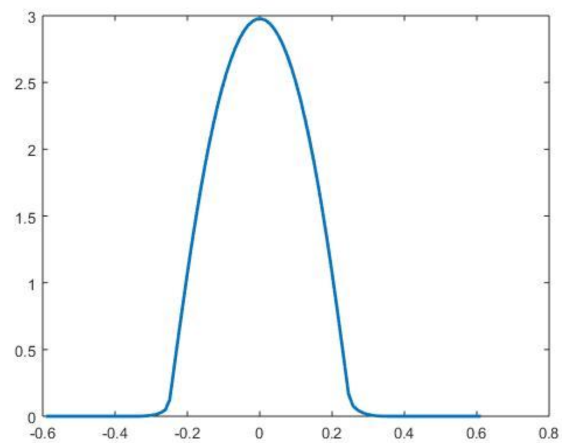

Figure 8: Bandwidth = 0.25

These graphs represent kernel density plots of the Colombo Stock Market returns, but they differ in the bandwidth used. The bandwidth determines the smoothness of the curve and how much detail it captures.

Figure 4 (Bandwidth = 0.15 has a narrow peak and steeper sides, indicating a tight distribution of returns around the central value. It suggests that most returns fall close to the mean, with fewer outliers on either side. This could indicate relatively low volatility in the market.

Figure 5 (Bandwidth = 0.2) compared to Figure 4, this plot has a wider peak and gentler slope, leading to a smoother curve. This suggests a broader distribution of returns, with more data point spread out across a wider range. This could indicate slightly higher volatility compared to Figure 4.

Figure 6 (Bandwidth = 0.1) plot has a narrower peak and sharper slopes than Figure 4, potentially overfitting the data. This could be overly sensitive to small fluctuations and might not accurately represent the overall distribution.

Figure 7 (Bandwidth = 0.25) plot has a wider peak and flatter slope than Figure 5, potentially undersmoothing the data. This could miss out on subtle details in the distribution and create a less accurate representation.

Observing the changes in the curves with different bandwidths helps us understand the distribution of returns.

4.5 Bandwidth Selection and Cross Validation

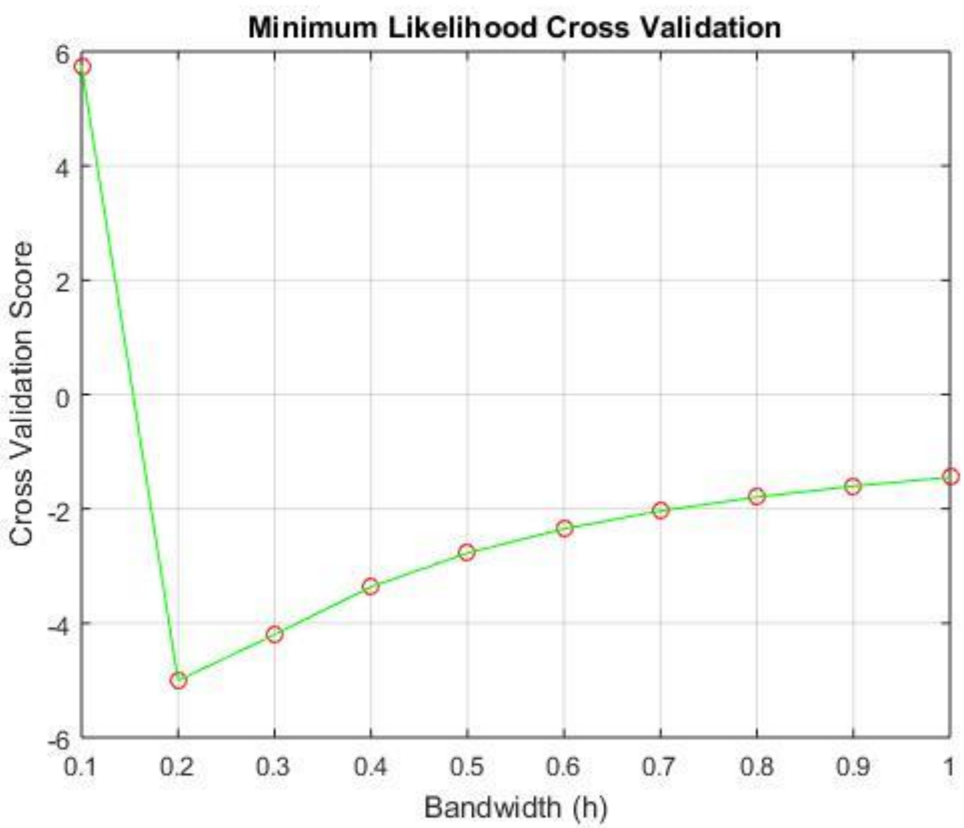

Figure 9: CV(h) vs h for optimal bandwidth

According to the figure 9 the results were visualized as the optimal bandwidth value is 0.2.

The optimal bandwidth for the KDE was crucial for obtaining an accurate and reliable density estimate. The range of bandwidth values was explored, and cross-validation was performed to assess the effectiveness of each bandwidth. The Least Squares Cross Validation (LSCV) score was used as a criterion for evaluating the performance of different bandwidths.

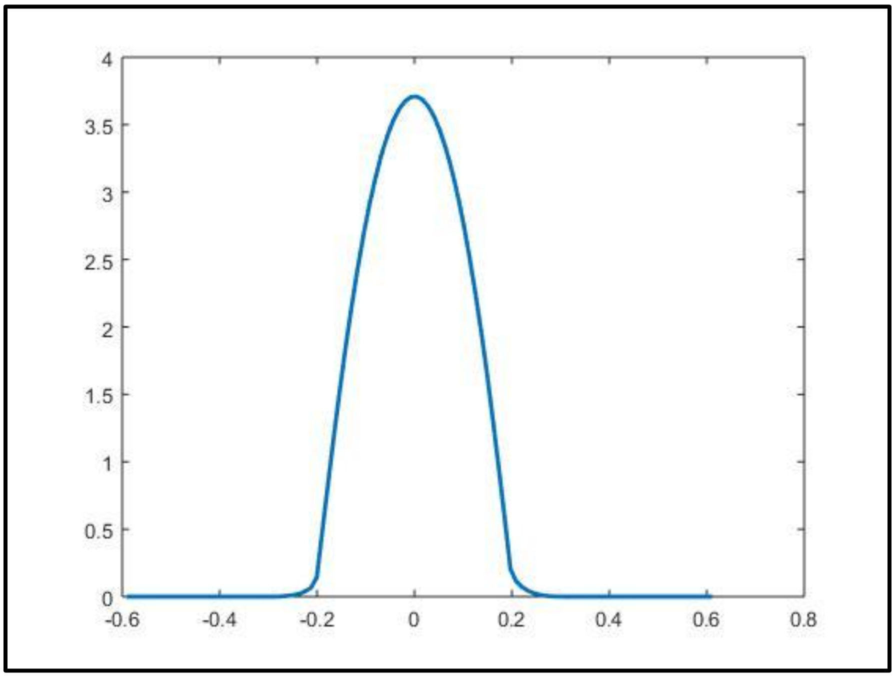

Figure 10: Kernel density estimation for optimal bandwidth

Above figure 9 represents the kernel density plot of the S&P SL 20 index returns with the optimal bandwidth at h=0.2, which is the bandwidth that minimizes the LSCV score.

In summary, this chapter provides a comprehensive analysis of the Colombo Stock Market data. The KDE method allowed us to uncover the underlying distribution of returns, and the process of bandwidth selection through cross-validation ensured the robustness of the analysis.

Chapter 5

Discussion & Conclusion

The application of nonparametric kernel density estimation (KDE) in the context of the Colombo Stock Market (S&P SL 20) reveals valuable insights into the distribution of stock returns. The choice of the Epanechnikov kernel, characterized by its symmetric, non-negative, and optimal properties, ensures a robust estimation process. The research aims to address specific questions regarding the improvement of stock price analysis, the utility of nonparametric methods in capturing unique data patterns, and the differences between traditional parametric approaches and KDE.

In the methodology section, the research design involves the use of KDE to analyze the S&P SL 20 data, considering closing prices, trading volumes, moving averages, and volatility measures. The data preprocessing steps ensure the reliability of the dataset, and exploratory data analysis provides a numerical and graphical summary of key features. The application of KDE is presented with varying bandwidths, demonstrating the impact on the smoothness and detail capture of the density curve.

The optimal bandwidth selection is a critical aspect of KDE, and cross-validation methods, such as Least Squares Cross Validation (LSCV), are employed to determine the bandwidth that minimizes errors in density estimation. The discussion of different bandwidth scenarios and their implications on the kernel density plots contributes to a comprehensive understanding of the Colombo Stock Market returns.

Conclusion

A new method for comprehending the distribution of stock returns has been provided by the research effort on the use of nonparametric kernel density estimation in the Colombo Stock Market (S&P SL 20). The reliability and usefulness of the results are enhanced by the selection of the Epanechnikov kernel, careful data analysis, and an optimal bandwidth selection made via cross-validation.

The main conclusions show that the Colombo Stock Market displays specific patterns and behaviors that nonparametric techniques such KDE may effectively capture. The methodology provides a data-driven and adaptable substitute for classical parametric methodologies, especially in financial market analysis where distributional assumptions might not hold true.

This study is important because it advances the use of analytical techniques in the financial markets. By leveraging nonparametric KDE, the research enhances our understanding of stock price patterns, providing insights that can inform investment decisions and risk management in the dynamic Colombo Stock Market environment.

Overall, the project demonstrates the applicability and benefits of nonparametric kernel density estimation in the financial domain, opening avenues for further research and exploration of advanced statistical techniques in similar contexts.

Appendix

%kernel.m

```
function k=kernel(u)
if ((u<1)&(u>-1))
    k = (3.0/4)*(1-u*u);
else
    k=0.0;
end
```

%pc_density.m

```
function [p] = pc_density(x, xe, h)

p = zeros(size(x));

for i = 1:length(x)
    temp = 0;

    for j = 1:length(xe)
        if (i ~= j)  % Only calculate temp if i is not equal to j
            temp = temp + kernel((x(i) - xe(j)) / h);
        end
    end

    p(i) = temp / (length(xe) * h);
end

end
```

%main.m

```
clc
data = xlsread('data_SP.xlsx');
currentDirectory = pwd;
disp(currentDirectory);

data = readtable('data_SP.xlsx', 'Sheet', 'Sheet2');
data.date = datetime(data.date, 'InputFormat', 'dd/MM/yyyy');
dataclosing_price = str2double(data.closing_price);

missingRows = isnat(data.date) | isnan(data.closing_price);
data(missingRows, :) = [];
fullDates = min(data.date):max(data.date);

% Linear interpolation to fill missing values
interpolatedPrices = interp1(datenum(data.date), data.closing_price,
datenum(fullDates), 'linear');
length(interpolatedPrices)
n=length(interpolatedPrices);
xe=zeros(length(n));
```

```
for k=1:n-1
     %xe(k)=log10(data.closing_price(k+1))-log10(data.closing_price(k));
     xe(k)=(interpolatedPrices(k+1)-
interpolatedPrices(k))./interpolatedPrices(k);
end
n=n-1;

% Density Estimation in a region of [min(xe)-0.3 max(xe)+0.3]
% at 100 points with given bandwidth h=0.02
xmin=min(xe);
xmax=max(xe);
m=100;
x=linspace(xmin-0.5,xmax+0.5,m);
h=0.1;
p_hat = pc_density(x,xe,h);

figure(1)
plot(x,p_hat,'LineWidth',2);
%%
r=10;
CV=zeros(r,1);
h=zeros(r,1);
for q = 1:r
 h(q)=(1/r)*q;
 p = pc_density(xe,xe,h(q));
 T = trapz(xe,p.^2);
 CV(q)=min(T-(2.0/n)*sum(p));
end
CV
%%
figure(2)
plot(h,CV,'ro');
hold on;
plot(h,CV,'g-');
grid on;
xlabel('Bandwidth (h)');
ylabel('Cross Validation Score');
title('Minimum Likelihood Cross Validation');
hold off
```

References

- Wang, Y., & Wang, J. (2011). Application of Nonparametric Kernel Density Estimation in Hong Kong Stock Market. *Applied Mechanics and Materials*, Vols 55-57, pp. 209-214. Trans Tech Publications, Switzerland. DOI: 10.4028/www.scientific.net/AMM.55-57.209

- Wijayasiri, M.P.A., & Abeyratne, M.K. (2015). An Application of Nonparametric Kernel Density Estimators in Colombo Stock Market Indices. *Proceedings of 2nd Ruhuna International Science & Technology Conference*, University of Ruhuna, Matara, Sri Lanka, January 22-23, 2015.

- Wu, X. (2018). Robust Likelihood Cross-Validation for Kernel Density Estimation. *Journal of Business*. DOI: 10.1080/07350015.2018.1424633

- B. W. Silverman. Financial News and Analysis. Financial Times. 2013. URL https://www.ft.com.

- Colombo Stock Market, Anual Report. CSM, 2012,2013. URL https://www.cse.lk.