COMPUTATIONAL METHODS FOR A COPULA-BASED MARKOV CHAIN MODEL

LEE CHIN YEE

UNIVERSITI TUNKU ABDUL RAHMAN

COMPUTATIONAL METHODS FOR A COPULA-BASED MARKOV CHAIN MODEL

LEE CHIN YEE

A project report submitted in partial fulfilment of the requirements for the award of Bachelor of Science (Honours) in Applied Mathematics with Computing.

Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

September 2024

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature :

Name : LEE CHIN YEE

ID No. : 2200494

Date : 01/08/2024

APPROVAL FOR SUBMISSION

I certify that this project report entitled **"COMPUTATIONAL METHODS FOR A COPULA-BASED MARKOV CHAIN MODEL"** prepared by **LEE CHIN YEE** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Science (Honours) Applied Mathematics with Computing at Universiti Tunku Abdul Rahman.

Approved by,

Signature : ____________________

Supervisor : Dr Tan Wei Lun

Date : 05/ 09/ 2024

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Dr. Tan Wei Lun for his invaluable advice, guidance and his enormous patience throughout the development of the research.

In addition, I would also like to express my gratitude to my loving parents and friends who had helped and given me encouragement.

ABSTRACT

Copula-based Markov models have gained recognition as powerful tools for capturing intricate dependence structures in time series datasets. This study focuses on estimating parameters and assessing the performance of Clayton and Gaussian copulas in modelling Laplace distributed time series data. The Clayton and Gaussian copulas were chosen due to the Clayton copula's capability to model tail dependencies and the Gaussian copula's alignment with the data's pseudo-observations. The ten-year daily log return of the SPX500 index is used in this study as the preliminary analysis revealed that it follows a Laplace distribution rather than the traditionally used *t*-distribution for modelling tail behaviour. Parameters were estimated using Maximum Likelihood Estimation (MLE) and the inversion of Kendall's Tau, yielding feasible results for both copulas. The model's performance was evaluated using the Root Mean Square Error (RMSE), with the Clayton copula achieving a lower RMSE of 0.01332 compared to 0.01541 for the Gaussian copula, indicating a better fit to the data. This study underscores the importance of selecting appropriate copulas, marginal distributions and estimation methods, demonstrating that the Clayton copula, combined with MLE, offers superior performance for modelling the Laplace distributed SPX500's daily log returns.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ ABBREVIATIONS

$c(u,v)$	probrability distribution function of bivariate Copula
$C(u,v)$	cumulative distribution function of bivariate Copula
d	dimension
H	dependency structure
θ	parameter of Copula
Φ	standard normal cumulative distribution function
M	correlation matrix of Gaussian copula
I	identity matrix
w	bivariate data in the 1x2 matrix form
D_t	test statistic of Komolgorov-Smirnov test at time t
μ	location of distribution
α	scale of distribution
L	likelihood function
ℓ	log-likelihood function
$\hat{P}_t$	estimated parameters at time t
τ	Kendall's Tau
V	inverse of Copulas
s_t	simulated data at time t
a_t	actual data at time t

CDF	cumulative distribution function
PDF	probrability distribution function
SPC	statistical process control
MLE	maximum likelihood estimation
AR	autoregressive model
MA	moving average
ARMA	autoregressive moving average
ARIMA	autoregressive integrated moving average
MSE	mean square error
RMSE	root mean square error
GARCH	generalized autoregressive conditional heteroskedasticity

CHAPTER 1

INTRODUCTION

1.1 General Introduction

"Copula" is derived from Latin meaning "join, bond, tie, or link," and was initially introduced by Sklar in a mathematical context to describe the relationships between variables. It serves as a multivariate cumulative distribution function (CDF) that connects multiple distribution functions to their respective one-dimensional marginal distribution functions. (Sun *et al*., 2020) This characteristic enables copulas to effectively capture and simulate complex dependencies between data points without imposing restrictions on marginal distributions. Thus, it is widely applied in the real world to predict the multivariate dependencies in diverse fields such as finance, risk management, healthcare, and statistical process control (SPC).

Through decades of theoretical exploration and practical application, various types of copulas have emerged, each adept at capturing distinct structures of dependence. Notably, the Archimedean and Elliptical copula families have played significant roles in real-world scenarios. Additionally, Darsow *et al*. (1992) introduced copula-based Markov chain models tailored to fit serially dependent time series data. This methodology has since been widely embraced by researchers for analyzing both continuous and discrete datasets.

1.2 Importance of the Study

In this digital age, data and big data have become the driving forces behind technological advancements. The ability to analyze data, make inferences, and draw conclusions is a vital skill set for the current and future generations. By developing statistical models that fit the available data, researchers can make predictions about future trends and behaviours, providing invaluable insights for various industries.

In the realm of statistical modelling, copula-based Markov models have emerged as a particularly potential tool for capturing the intricate dependence structures present within complex data sets. These models combine the principles of copula models, which describe the dependence structure, with stochastic Markov models, which capture the transition probabilistic behaviour of systems evolving. The integration of these two concepts has led to a robust framework for modelling intricate, time-dependent data with complex dependence patterns.

The significance of studying computational methods for copula-based Markov models lies in their broad applicability across diverse domains. For instance, in finance and economics, these models are instrumental in analyzing and forecasting financial data, offering insights into market trends and risk assessment. (Chen *et al*., 2022; Dewick and Liu, 2022; D'Amico *et al*., 2019) Similarly, in environmental sciences, it enables the joint modelling of multiple environmental variables, such as temperature, sandstorm count, and air quality, enabling more accurate predictions of climate patterns and natural disasters (Pepi *et al*., 2024; Alqawba and Diawara, 2020).

Moreover, these models have applications in areas like healthcare, where they can model the survival data and dependent terminal events like death (Huang *et al*., 2020). In manufacturing and quality control, copula-based Markov models aid in monitoring and optimizing production processes by capturing the intricate relationships between process variables and product quality (Kim *et al*., 2019; Huang and Emura, 2022).

In short, the significance of studying computational methods for copula-based Markov models lies in their potential to unlock the power for a wide range of real-world applications, spanning diverse sectors and offering valuable insights for decision-making and problem-solving.

1.3 Problem Statement

The Gaussian copula is blamed by some experts for the financial crash of 2007 and 2008. (Dewick and Liu, 2022) However, it's important to recognize that different copulas are designed to capture distinct dependency structures suitable for various types of data. The issue lies not with the use of the Gaussian copula itself, but rather with modelers who may have failed to thoroughly understand its properties and applicability. Hence, the crucial aspect is the selection of an appropriate model tailored to fit the specific characteristics of the data.

Furthermore, the critical importance of accurately specifying time-series parameters to enable the estimation of a correctly-specified copula Markov model. Variations in marginal distributions, including symmetries, asymmetries, fat tails, and structural breaks, significantly influence the estimation outcomes when modelling dependency structures. In cases where the marginal distributions are unknown, semi-parametric methods can be employed. Otherwise, the parametric methods can be used after the model diagnostic is carried out.

1.4 Aim and Objectives

1. To determine the marginal distribution of the Log return of SPX500.
2. To estimate the parameters of copula-based Markov models using the Maximum Likelihood Method and inversion of Kendall's Tau method.
3. To evaluate the performance of the Clayton copula and Gaussian copula.

1.5 Scope and Limitation of the Study

This study focuses on the computational method of the copula-based Markov model under two widely used copulas: the Clayton copula and the Gaussian copula. The choice of these two copulas is driven by their prevalence in academic research. Numerous studies have studied the performance of the Clayton and Gaussian copulas in various applications, providing a solid foundation for further exploration.

Additionally, the data used for analysis is the SPX500 index. The stock return is heavy-tailed distributed due to large deviation from the average value. Thus, the ability of Clayton to capture tail dependence might show a better fit to financial data, which often have heavy-tailed distributions and skewness (Sun *et al*., 2018).

For parameter estimation, this study employs the Maximum Likelihood Estimation (MLE) and inversion of Kendall's Tau method. Research by Long and Emura (2014) indicates that parametric methods, such as MLE, tend to outperform semi-parametric approaches. This preference arises from the higher likelihood of achieving a better model-data fit with well-estimated parameters.

While the computational methods developed in this study aim to address the challenges associated with parameter estimation and model fitting for copula-based Markov models, it is essential to recognize their limitations. One significant limitation is the assumption of stationarity in the underlying Markov process, which may not hold for certain real-world scenarios where the financial time series data exhibits non-stationarity or structural breaks. Moreover, the choice of parameter estimation can also impose constraints on the range of dependence structures that can be modelled accurately.

Though the MLE are versatile and widely applicable, they may not capture certain intricate dependence patterns present like structural breaks and change points. In such cases, a novel change point estimator may be employed, which involves the three-state copula model proposed by Emura *et al*. (2023). However, the computational tools for this change point method have not been developed yet.

1.6 Contribution of the Study

This study might have notable contribution by combining the Copula-based Markov model with a Laplace marginal distribution, a relatively underexplored approach recently in this field. The selection of copula and marginal distribution is crucial, as any misalignment between the two can lead to poor performance. Thus, the existing R package *Copula.Markov* is not performing well for dataset following distribution other than Normal and Binomial. While there has been one prior financial paper utilizing this combination, it lacks detailed computational methodology, making this study a valuable addition by filling this gap and providing insights for future research. (Nadaf *et al*., 2022)

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This section primarily delved into a review of the literature concerning copulas, copula-based Markov models, and the latest computational methods. The focus was examining the fundamental concepts, definitions, and properties of these models. Such an exploration serves as a foundational step towards identifying viable solutions for subsequent sections of the research.

2.2 Review of Copula Models

As aforementioned, the idea of a copula was initiated by Abe Sklar in 1969. All the copula proposed afterwards were modified based on his concepts. There are many families of copula, such as the Elliptical Copulas, Archimedean Copulas, Marshall-Olkin Copulas and Extreme-Value Copulas. Their corresponding C functions, parameters, generators, and inverse generators, are introduced in the following table.

Table 2.1: Most important copulas from the Archimedean family

Name	**Copula Function, $C_\theta(u,v)$**	**Parameter, θ**	**Generator, $\phi_\theta(t)$**	**Inverse Generator, $\phi_\theta^{-1(t)}$**
Clayton	$max(u^{-\alpha} - v^{-\alpha} - 1,0)^{-1/\alpha}$	$\alpha \in [-1,+\infty)\backslash\{0\}$	$(t^{-\alpha} - 1)/\alpha$	$(1+\alpha t)^{-1/\alpha}$
Joe	$1 - [(1-u)^\alpha + (1-u)^\alpha - (1-u)^\alpha(1-v)^\alpha]^{1/\alpha}$	$\alpha \in [1,\infty)$	$-log(1-(1-t)^\alpha)$	$1-(1-e^{-t})^{1/\alpha}$
Frank	$-\frac{1}{\theta} log\left[1 + \frac{(exp(-\theta u)-1)(exp(-\theta v)-1)}{exp(-\theta)-1}\right]$	$\theta \in \mathbb{R}\{0\}$	$-log\left(\frac{exp(-\theta t)-1}{exp(-\theta)-1}\right)$	$-\frac{1}{\theta}(1+exp(-t)(exp(-\theta)-1))$
Gumbel	$exp\left[-\left((-log(u))^\theta + (-log(v))^\theta\right)^{1/\theta}\right]$	$\theta \in [1,\infty)$	$(-log(t))^\theta$	$exp(-t^{1/\theta})$
Independence	uv		$-log(t)$	$exp(-t)$

The pursuit of a "perfect model" for time-series data remains a focal point for analysts across various sectors, particularly in finance and economics. While existing models like AR, MA, ARMA, and ARIMA have been extensively studied and utilized, recent advancements in the application of copulas have garnered attention as a promising avenue in financial modelling. (Huang *et al*., 2020)

These models primarily address linear dependencies within financial time series data. This assumption can be overly restrictive when dealing with real-life data, where dependency structures are often complex and challenging to capture. The copula Markov model emerges as a viable alternative, offering the capability to capture nonlinear and tail dependencies inherent in financial time series data (Chen *et al*., 2022).

Moreover, copula-based Markov models provide flexibility in exploring various types of dependencies within time-series data. These may include serial dependencies within univariate data, dependencies between variables, or a conditional combination of both. (Kim *et al*., 2019) Despite the limitations of these models and the flexibility offered by copula-based Markov models, it may not always represent the optimal choice. The process of utilizing copula-based models can be relatively cumbersome, and uncertainties surrounding parameter estimation may pose challenges.

2.3 Marginal Distribution

Though not usually mentioned this various studies, determining the marginal distribution of each variable is indeed a crucial step in constructing a copula model (Nadaf *et al*., 2022) This process informs the types of parameters to be estimated later on which aids in ensuring the accuracy of the proposed model.

Huang and Emura (2019) introduced a novel package designed to assist users in conducting model diagnostic tests, particularly for fitting a copula Markov model later on. The stringent normality assumption avoids poor performance of the authors' proposed model, which following a Normal marginal distribution, underscores the importance of thorough diagnostic testing.

However, a notable issue observed in various studies is the practice of determining marginal distributions without conducting diagnostic tests. For instance, Sun *et al*. (2018) analyzing the log return of SPX500 data concluded that it follows a *t*-distribution solely based on the violation of normality and the presence of fat-tail characteristics. Since the paper focuses on comparing the performance of the method of parameter estimation, the performance of the copula is not mentioned.

In contrast, Nadaf *et al*. (2022) suggest that the Laplace distribution provides a better fit for financial data. Nevertheless, it acknowledges that copulas can perform adequately as long as a fat-tailed marginal distribution is utilized for fitting the data. Hence, the distribution test of these two will be performed in this study.

2.4 Parameter Estimation Method

Researchers have proposed various parameter estimation methods, which are typically categorized into parametric and semi-parametric estimations. Methods include but not limited to rank regression, Bayesian inference methods, inversion of Kendall’s tau, Spearman’s rho, and Optimization-based method.

In a study by Sun *et al*. (2018), it was found that parametric estimators, particularly MLE, outperform semi-parametric estimators and provide asymptotically efficient estimates under certain regularity conditions. However, it's crucial to acknowledge the limitations of MLE. This includes challenges with change points, complex models, and small sample sizes, leading to issues like bias, convergence problems, and computational challenges. Thus, the Bayesian inferences method is proposed as marginal *t*-distribution is used, and a maximization problem will emerge if MLE is used.

Conversely, Chen *et al*. (2020) claimed that semi-parametric estimators perform equally well for both stationary and non-stationary data, suggesting their overall robustness compared to parametric methods, which are constrained by the filtration and estimation of marginal distributions.

In short, while copula-based Markov models offer a promising approach to capturing complex dependencies in financial time series data, analysts must carefully consider their suitability and weigh the trade-offs against the potential benefits, ensuring that the chosen modelling approach aligns with the specific characteristics and objectives of the analysis.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This section briefly introduced the research flow and the model formulations for specific methods. The flow chart included in the end of this section shows the research methodology of this study. There are four main phases including data preprocessing, data modelling, simulation and result evaluation. The detailed steps in each phase were illustrated and explained in each subsection below.

3.2 Copula Formulation

Copula is a multivariate distribution function of the unit hypercube $[0,1]^d$ in a d-dimensional case ($d \in \mathbb{N}$) where the marginal distributions follow uniform distributions in the range of [0,1] (Durante and Sempi, 2015). Thus, the original copula defined by Sklar is as follows:

$$\mathrm{H}(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{d}) = \mathrm{C}\big(\mathrm{F}_1(\mathrm{x}_1), \mathrm{F}_2(\mathrm{x}_2), \dots, \mathrm{F}_\mathrm{d}(\mathrm{x}_\mathrm{d})\big),$$

$$\forall (x_1, x_2, \dots, x_d) \in \mathbb{R}^d \tag{3.1}$$

$$C(u_1, u_2, \dots, u_d) = H\left(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_d^{-1}(u_d)\right)$$

Some properties of C need to be satisfied. To minimize the complexity, the bivariate copula is introduced here as an example to illustrate its properties.

$$H(\mathrm{x}, y) = C\big(\mathrm{F}(\mathrm{x}), \mathrm{G}(\mathrm{y})\big), \qquad \forall (x, y) \in \mathbb{R}^2$$

$$x = F^{-1}(u), y = G^{-1}(v), \tag{3.2}$$

$$C(u, v) = H\big(F^{-1}(u), G^{-1}(v)\big)$$

Properties for a bivariate copula function, $C: [0,1]^2 \rightarrow [0,1]$:

(P1) $C(u, 0) = C(0, v) = 0, \ \ for\ 0 \le u, v \le 1.$

(P2) $C(u, 1) = u \ \ and \ \ C(1, v) = v, \ \ for\ 0 \le u, v \le 1.$

(P3) $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \ge 0, \ \ for\ \ 0 \le u_1 \le u_2 \le 1 \ \ and\ \ 0 \le v_1 \le v_2 \le 1.$

The properties (P1) and (P2) proof that the marginal distributions of u and v follow uniform distribution in the range of [0,1]. The property (P3) shows that the probability mass on $(u_1, u_2] \times (v_1, v_2]$ is always positive, which satisfied the conditions of a cumulative distribution function. (Sun *et al*., 2020)

3.2.1 Clayton and Gaussian Copula

Clayton copula and Gaussian copula belong to two different family, which is Archemedian and Elliptical respectively. Thus, both of them have a totally different charactheristic in modelling the dependency structures.

Clayton is widely used to model lower tailed and asymmetric dependence. Hence, it is ideal for data showing extreme negative events, like a sudden temperature drop. While Gaussian assumes the dependency structure is normally distributed, regardless of how the marginal are distributed. Hence, it is used to model symmetric dependence without any tail dependence, where it treat both extreme high and extreme low values equally. So, the purpose of choosing these two models, other than what had mentioned in the other section, is to see whether the SPX500 is more to a normal condition or have extreme lower dependency.

In order to derive the log-likelihood function for the parameter estimation afterwards, the PDF of both bivariate Copulas is derived as follow:

CDF of Clayton Copula:

$$C(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}} \tag{3.3}$$

where

$$\theta \in [-1,0) \cup (0, \infty).$$

PDF of Clayton Copula:

$$c(u, v) = (\theta + 1)(uv)^{-(\theta+1)}\left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}-2} \tag{3.4}$$

CDF of Gaussian Copula:

$$C(u,v) = \Phi_2\big(\Phi^{-1}(u), \Phi^{-1}(v)\big) \tag{3.5}$$

where

$\theta \in [-1,1]$,

Φ is the standard normal CDF,

Φ_2 is the normal CDF with covariance matrix 2x2.

PDF of Gaussian Copula:

$$c(u,v) = -\frac{1}{\sqrt{\det(\mathrm{M})}} exp\left(-\frac{1}{2}(\boldsymbol{w}^T(M^{-1} - I)\boldsymbol{w})\right) \tag{3.6}$$

where

I is the identity matrix,

M is the correlation matrix,

$w = \big(\Phi^{-1}(u), \Phi^{-1}(v)\big)$.

3.2.2 Copula-based Markov Model

A Markov process is a stochastic process that satisfies the Markov property. This property implies that the transition to the next state is determined only by the present state, without any influence from previous states, given the current state. Formally, a discrete stochastic process $X(t), t \in T$,where $t_0 < t_1 < \cdots < t_{n+1}$, is said to be a Markov process the following condition holds:

$$\begin{aligned} &P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &\qquad = P(X_{n+1} = x_{n+1} | X_n = x_n) \end{aligned} \tag{3.7}$$

The copula-based Markov Chain model is the combination of Copula function and the Markov process. The equations below show the general equations for a bivariate copula-based Markov Chain model,

$$P(\mathrm{X_n} \leq \mathrm{x_n},\ X_{\mathrm{n}-1} \leq \mathrm{x_{n-1}}) = C_{n,n-1}\big(\mathrm{F(x_n)}, \mathrm{F(x_{n-1})}\big) \tag{3.8}$$

This model maps the copula function to the univariate time series model to describe the serial dependencies of a given Markov process.

3.3 Data

For this study, the widely recognized SPX500 index, which is the Standard and Poor's 500, was used as the data source. It is the stock performance of 500 largest companies in the United States including Apple, Microsoft, Amazon, NVIDIA, etc. Approximately 10 years of daily closing data, spanning from January 1, 2010, to December 31, 2019, were considered.

3.3.1 Data Preprocessing

Due to the extensive time frame required, data is sourced from multiple online platforms like Kaggle and FRED. Thus, this process of combining various files, ensuring data format alignment, and checking for null or duplicate entries. To ensure the reliability of the data, cross-checking with yahoo finance data is carried out before proceeding to the next step. Subsequently, around 10 years span of data is extracted for further modelling purposes and performance evaluation. Following this, the logarithmic return of the closing price is computed using the training data. The equation for logarithmic return is outlined below.

$$R = ln\left(\frac{Close\ Price_t}{Close\ Price_{t-1}}\right) \tag{3.9}$$

3.3.2 Diagnostic Test

When considering parameter estimators over semi-parametric estimators, this step becomes particularly crucial. The diagnostic tests such as normality, stationarity, and distribution tests are performed using visualization tools and statistical analysis as follows,

1. **Quantile-Quantile plot (QQ-plot)**

   QQ plot is normally used to checked the normality of a distribution. It plot the quantiles of the normal distribution against the one of data.If the graph showed a diagonal straight line, means that the data is normally distributed, otherwise, skewness or heavy tails. It can be used for distribution other than normal as well.

2. **Density plot**

   The density plot plots the PDF of a continuous distribution. It is a smoothed and perfect version of the distribution thus widely use to compare with the real data. Though sometimes it is similar at the first glance, it might not necessary following the distribution, further investigation like KS test needs to be carry out.

3. **Augmented Dickey-Fuller (ADF) test**

   ADF is a unit root test widely used to determined whether a time series is non-stationary or not. The hypothesis and rejection criteria are as follows:

   H_0: The series is non-stationary

   H_1: The series is stationary

   a. If the test statistic is smaller than the critical value.
   b. If the p-value is smaller than 0.05.

4. **Kwiatkowski-Philips-Schmidt-Shin (KPSS) test**

   While KPSS is another test that has the same purpose as ADF, the hypothesis is contradict with ADF.

   H_0: The series is trend stationary

   H_1: The series is non-stationary

   Both the ADF and KPSS tests are carried out to with the purpose of investigate more than just the stationarity of data. This is because the combination of rejecting or not rejecting the null hypothesis will hold different meanings:

   a. Both concluded stationary.
   b. Both concluded non-stationary.
   c. Only the ADF test indicated stationarity, suggesting difference stationarity and warranting further investigation.
   d. Only the KPSS test indicated stationarity, implying trend stationarity.

5. **Kolmogorov-Smirnov (KS) test**

   KS test is a nonparametric method used to investigate the distribution of given data. This method utilise the charactheristic of CDF which is always monotonic by comparing it to the computed CDF. The test statistic is computed as follow:

$$D_t = \max_{1<=t<=N}\left(\left|F(x_t) - \frac{i-1}{N}\right|, \left|\frac{i}{N} - F(x_t)\right|\right) \quad (3.10)$$

If the maximum of the test stastistic is larger than the critical value, means that there is a significant difference between the data and the distribution. In this case, the null hypothesis of following specific distribution will be rejected.

This process is essential for capturing unique characteristics such as skewness or heavy tails and for validating the stationarity, normality, and distribution of the data. It aids in selecting appropriate marginal distributions for the time series if ideal results are obtained. Additionally, it ensures the reliability and robustness of the parameter estimation process.

3.3.3 Copulas Selection

Before training and modelling the data, the pseudo-observation of the data is visualized which allowed us to further confirm the decision of copula used. The pair plot and pseudo-observation do not show any obvious relationship between the 't' data and 't-1' data compared to the example of various copulas. Thus, the Gaussian and Clayton copula are chosen to model the SPX500 data.

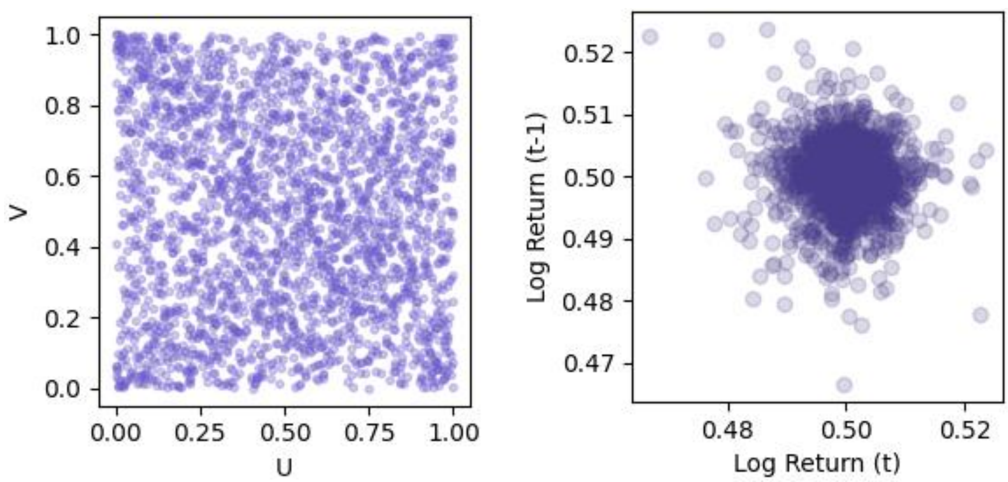

Figure 3.1 Pair plot and Pseudo-observation

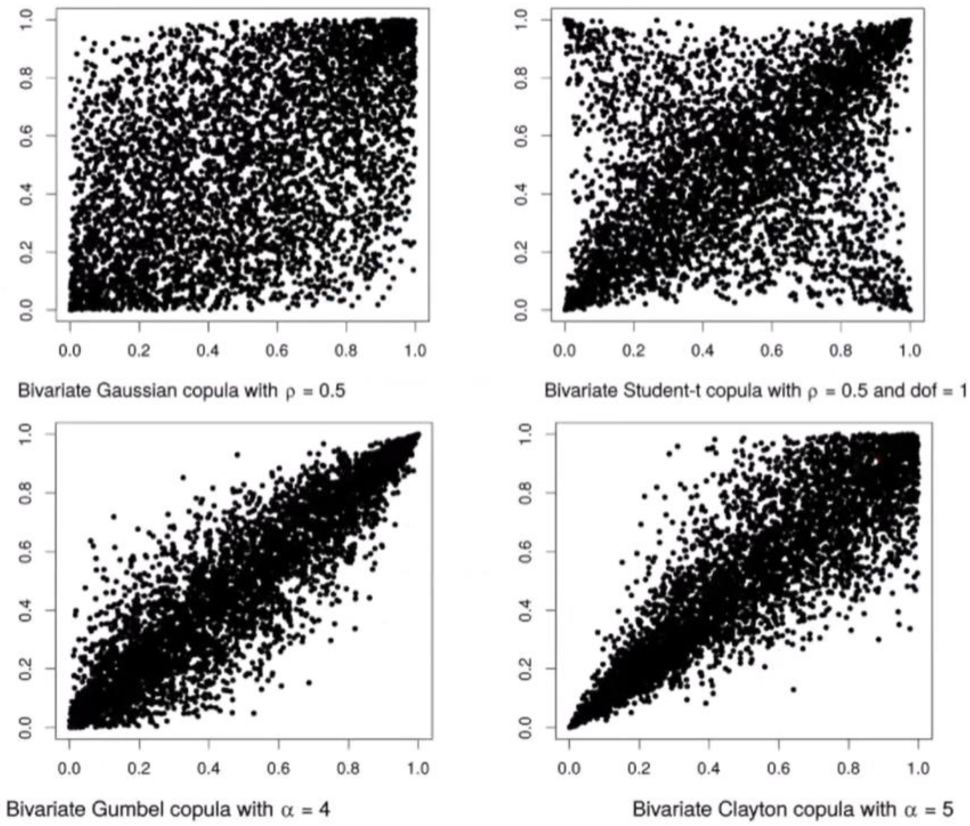

Figure 3.2 Pseudo-observation of four well-known Copulas

3.4 Parameter Estimation

After that, the values for the parameters of the copula models were estimated using both MLE and inversion of Kendall's tau method. In this subsection, the formulation of MLE and Kendall's tau is shown followed by the computation algorithms designed to obtain the values.

3.4.1 Maximum Likelihood Estimation

The Likelihood and Log-likelihood function for the Copula and marginal distribution are as below:

$$L(\mu, \alpha, \theta) = \prod f(x_t) \prod c\big(F(x_t), F(x_{t-1})\big) \tag{3.11}$$

$$\ell(\mu, \alpha, \theta) = \sum_{t=1}^{n} logf(x_t) \ + \sum_{t=2}^{n} log\ c\big(F(x_t), F(x_{t-1})\big) \tag{3.12}$$

PDF of Laplace Distribution:

$$f(x_t) = \frac{1}{2\alpha}\exp\left(-\frac{|x_t - \mu|}{\alpha}\right) \tag{3.13}$$

CDF of Laplace Distribution:

$$F(x_t) = \begin{cases} \frac{1}{2}\exp\left(\frac{x_t - \mu}{\alpha}\right), & \text{and} if\ x_t \leq \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x_t - \mu}{\alpha}\right), & \text{and} if\ x_t > \mu \end{cases} \tag{3.14}$$

The functions 3.13 and 3.14 are the density and cumulative function of Laplace distribution that used to compute the log-likelihood function 3.12. The log-likelihood function is the general form and can be reused by any distribution chosen. After that, the first and second partial derivative of the log-likelihood function were derived for the Newton-Raphson formulation as follows:

Newton-Raphson Formulation:

$$\hat{P}_t = \hat{P}_{t-1} - \ell(\hat{P}_{t-1})^{-1} \cdot \ell'(\hat{P}_{t-1})$$

$$\begin{bmatrix} \hat{\mu}_t \\ \hat{\alpha}_t \\ \hat{\theta}_t \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{t-1} \\ \hat{\alpha}_{t-1} \\ \hat{\theta}_{t-1} \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \alpha} & \frac{\partial^2 \ell}{\partial \mu \partial \theta} \\ \frac{\partial^2 \ell}{\partial \mu \partial \alpha} & \frac{\partial^2 \ell}{\partial \alpha^2} & \frac{\partial^2 \ell}{\partial \theta \partial \alpha} \\ \frac{\partial^2 \ell}{\partial \mu \partial \theta} & \frac{\partial^2 \ell}{\partial \theta \partial \alpha} & \frac{\partial^2 \ell}{\partial \theta^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \ell}{\partial \mu} \\ \frac{\partial \ell}{\partial \alpha} \\ \frac{\partial \ell}{\partial \theta} \end{bmatrix} \tag{3.15}$$

Since MLE derivation is affected by the choice of Copulas, thus, the log-likelihood function and the partial derivatives of both Copulas have to be computed. The equations below show the example of combining the Laplace CDF, Laplace PDF, and Clayton Copula PDF to construct the log-likelihood and one of the partial derivative.

Log-likelihood for Clayton copula is

$$\ell(\mu, \alpha, \theta) = \sum_{t=1}^{n} log f(x_t) + \sum_{t=2}^{n} log\ c\big(F(x_t), F(x_{t-1})\big) \tag{3.16}$$

$$= \sum_{t=1}^{n} log[\frac{1}{2\alpha}\exp\left(-\frac{|x_t-\mu|}{\alpha}\right)]$$

$$+ \sum_{t=2}^{n}\Big[log(\theta+1) \ + log\big(\mathrm{F}(\mathrm{x_t})\mathrm{F}(\mathrm{x_{t-1}})\big)^{-(\theta+1)} + log\big(\mathrm{F}(\mathrm{x_t})^{-\theta}+\mathrm{F}(\mathrm{x_{t-1}})^{-\theta}-1\big)^{-1/\theta-2}\Big]$$

First order partial derivative with respect to μ

$$\frac{\partial\ell}{\partial\mu} = -\frac{1}{\alpha}\sum_{t=1}^{n}\mathrm{sign}(x_t-\mu) + \sum_{t=2}^{n}\Bigg[-(\theta+1)\frac{\partial\log\big(F(x_t)\big)}{\partial\mu} - (\theta+1)\frac{\partial\log\big(F(x_{t-1})\big)}{\partial\mu} - \left(\frac{1}{\theta}+2\right)\frac{\partial}{\partial\mu}log\big(F(x_t)^{-\theta}+F(x_{t-1})^{-\theta}-1\big)\Bigg] \tag{3.17}$$

The implementation of the MLE using Newthon-Raphson method is designed as follows:

Algorithm 1: Estimate parameters using MLE

1. Define Laplace PDF and CDF functions.
2. Define the Copulas-based Markov Model PDF functions.
3. Define the Copulas log likelihood function with marginal distribution follows the functions defined in step 1.
4. Import the gradient and hessian function from '*autograd*' library.
5. Define the Newton Raphson algorithm that iterate until it converge, where epsilon is smaller than $1x\ 10^{-8}$.
6. Print each iteration and the epsilons of all parameters.
7. Print the output of final gradient and estimated parameters of Laplace and copula.

3.4.2 Inversion of Kendall's Tau

The Kendall's Tau relation to a bivariate Copula is as below:

$$\tau = 4\int_{0}^{1}\int_{0}^{1}(u-\theta+v-\theta-1)(-1/\theta)dudv-1 \tag{3.18}$$

Kendall's Tau to Clayton Copula

$$\tau = \frac{\theta}{(\theta+2)} \tag{3.19}$$

Kendall's Tau to Gaussian Copula

$$\tau = \frac{2}{\pi}arcsin(\theta) \tag{3.20}$$

The formula 3.19 and 3.20 is the closed form formula to obtain the parameters of the Clayton and Gaussian copula by using the Tau obtained. Thus, the algorithm 2 below shows the steps of obtaining rank correlation and parameters of Clayton and Gaussion copula.

Algorithm 2: Estimate parameters using Kendall's Tau
1. Pair the data $(x_0, y_0), \ldots, (x_t, y_t)$, where $t = 0, 1, 2, \ldots$
2. Given $m < n$, if $x_m > x_n$ *and* $y_m > y_n$ *or* $x_m < x_n$ *and* $y_m < y_n$, number of concordant pair +1. Else, number of discordant pair +1.
3. Count the total pairs of concordant and discordant.
4. Obtain the Kendall's Tau correlations.
5. Define the Tau to copula parameters according to the equations 3.3 and 3.4.
6. Pass the tau value into the functions and print the result.

3.5 Simulation

Subsequently, the same n of samples of SPX500 index was simulated according to each parameter estimated. The inverse of Copulas are derived to generate dependent random variables from the generated uniform distribution U1 and U2 in the range [0,1]. The formula below are the inverse of Clayton copula.

$$V = \left(U1^{-\theta}\cdot\left(U2^{-\frac{\theta}{1+\theta}}-1\right)+1\right)^{-\frac{1}{\theta}} \tag{3.21}$$

To assess whether the chosen models adequately captured the dependence structure and serial dependence inherent in the data, RMSE is calculated. If the results were close to zero, means the result is satisfactory. However, if the results were unsatisfactory, modifications to the models were made before continuing to the final phase.

Algorithm 3: Simulate the data using the estimated parameters
1. Generate two independent uniform variables with n samples.
2. Generate dependent variable using inverse Copula formula and the estimated parameters.
3. Inverse the generated data using the inverse function of Laplace CDF using the estimated parameters (MLE) or actual parameters (Kendall's).
4. Plot the actual data and simulated data.
5. Compute the RMSE.

3.6 Result Evaluation

The purpose of this final phase was to address the second objective of the paper, which was to evaluate the results by comparing the RMSE of each combination of estimator and copula-based Markov model. RMSE is used to measure the dispersion of the simulated data from the actual data. The difference of them is squared to make sure it is positive, avoiding cancelation of positive and negative values. The formula of the RMSE is as follows:

$$RMSE = \sqrt{MSE} = \sqrt{\sum_{i=1}^{n} \frac{(s_t - a_t)^2}{n}} \tag{3.22}$$

3.7 Flow chart

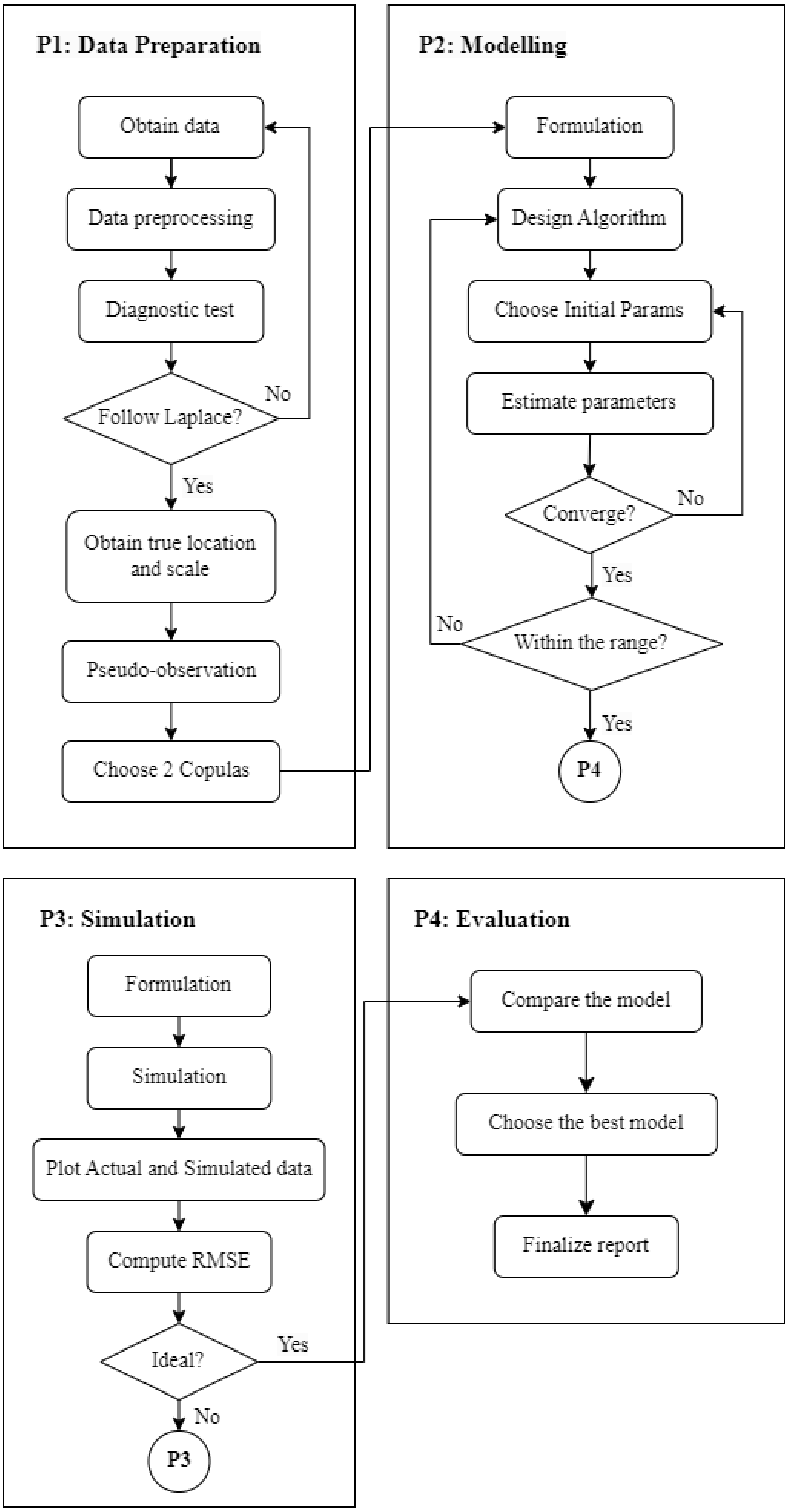

Figure 3.3 The flow chart of the research methodology

CHAPTER 4
RESULTS AND DISCUSSIONS

4.1 Introduction

The findings discussed in this section covered all the phases as outlined in the previous section. This section will delve into the significant findings of the model diagnostic tests, estimated parameters, and performance of each Copula. These findings will be explained and discussed.

4.2 Obtain data and Data Preprocessing

The SPX500 index data retrieved from various sources covers the period from 1992 to 2024. To reduce the volatility of the data, the dataset is trimmed, retaining only 10 years, starting from January 4th, 2010. After this process, the data is split, and the logarithmic return of the closing price is calculated using the training data.

Upon examination, it was observed that only the logarithmic return of the first row displayed a 'NaN' value due to the absence of the previous value. Consequently, to maintain data integrity, the first row corresponding to the date is removed. Subsequently, the refined data is stored in the data frame presented below.

	Date	Open	High	Low	Close	Volume	Log_return
1	03/01/1994	466.510010	466.940002	464.359985	465.440002	270140000	-0.002168
2	04/01/1994	465.440002	466.890015	464.440002	466.890015	326600000	0.003111
3	05/01/1994	466.890015	467.820007	465.920013	467.549988	400030000	0.001413
4	06/01/1994	467.549988	469.000000	467.019989	467.119995	365960000	-0.000920
5	07/01/1994	467.089996	470.260010	467.029999	469.899994	324920000	0.005934
...	...	...	...	...	...	...	...
5323	23/02/2015	2109.830078	2110.050049	2103.000000	2109.659912	3093680000	-0.000303
5324	24/02/2015	2109.100098	2117.939941	2105.870117	2115.479980	3199840000	0.002755
5325	25/02/2015	2115.300049	2119.590088	2109.889893	2113.860107	3312340000	-0.000766
5326	26/02/2015	2113.909912	2113.909912	2103.760010	2110.739990	3408690000	-0.001477
5327	27/02/2015	2110.879883	2112.739990	2103.750000	2104.500000	3547380000	-0.002961

Figure 4.1 Preprocessed data

4.3 Diagnostic test

Some visualization and statistical tests are carried out as the normality and stationarity assumptions are stringent. All of these assumptions play a vital role in the model selection and formulation afterwards.

4.3.1 Normality test

Although the copulas do not require the marginal distribution to be normally distributed, its normality will ease the work afterwards as many predefined packages, such as '*Copula.Markov*' in R, are restricted for normal marginal distribution functions. (Huang and Emura, 2019) Thus, the QQ plot of normal distribution against the SPX500 is plotted.

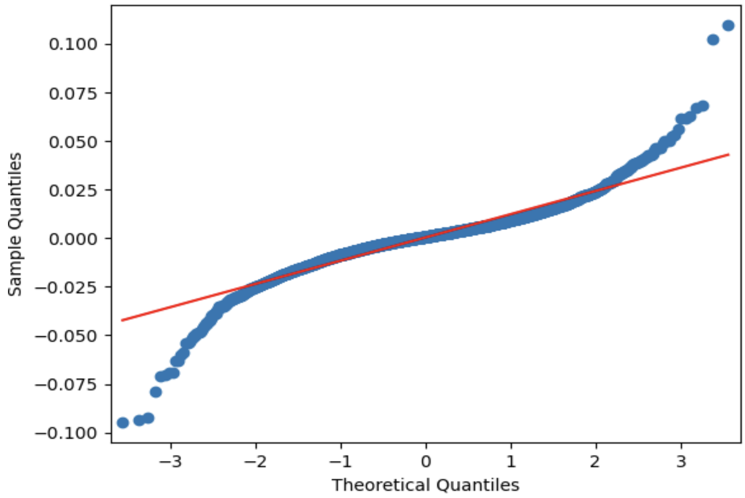

Figure 4.2 Normal QQ-plot of log return.

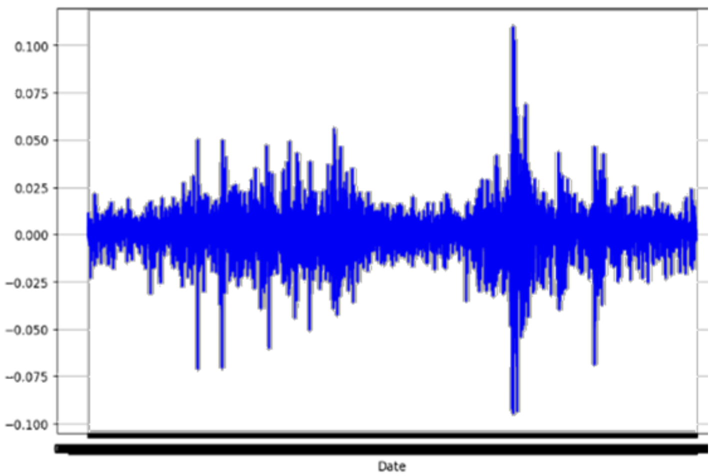

Figure 4.3 Line plot of log return.

Figure 4.2 shows that the data is over-dispersed relative to a normal distribution, thus the normality is violated. This aligned with the claims of many researchers that the financial data are not normally distributed and the assumption of normality without any diagnostic test is a poor practice that will lead to further problems. Besides, characteristics of fatter tails and a large number of outliers can be concluded from the QQ plot as well. Thus, normal distribution might be a good choice for the data and the packages in R will not be suitable for this case.

4.3.2 Stationary test

The stationary assumption for time series data is vital as this ensures that the model is robust over time. To test the stationary of the data, both the ADF Test and KPSS Test are carried out.

```
ADF Test Results for Log_return:
ADF Statistic: -15.40293082252661
p-value: 3.197387432384522e-28
Critical Values: {'1%': -3.431784550351163, '5%': -2.8621739027033426, '10%': -2.5671074294833836}

KPSS Test Results for Log_return:
KPSS Statistic: 0.2799236141656677
p-value: 0.1
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
```

Figure 4.4 ADF and KPSS test results

As the *p*-value for ADF is smaller than 0.05 and the test statistic is smaller than the critical value, the null hypothesis of non-stationary is rejected. Besides, the *p*-value for the KPSS test is 0.1, which is larger than 0.05. Thus, we can conclude that the logarithmic return of the time series data is stationary.

4.3.3 Distribution test

Student's *t*-distributions and Laplace distribution are two widely used distributions while modelling financial data. (Nadaf *et al*., 2022; Sun *et al*., 2018) Thus, choices of distributions were limited down to these two and various tests were carried out to test the goodness-of-fit of these distributions with the data.

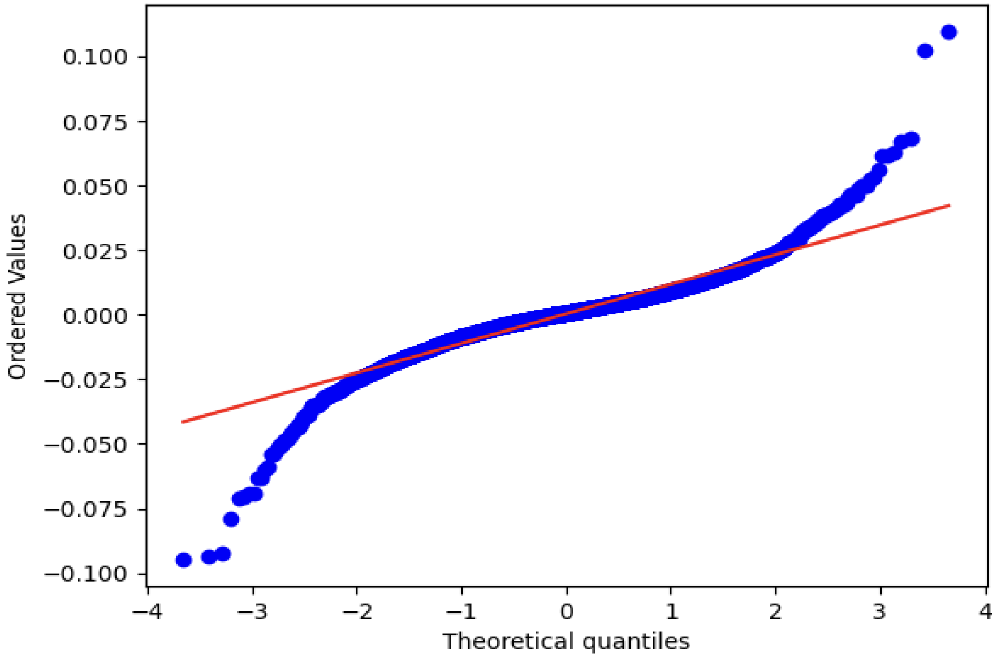

Figure 4.5 *t*-distribution QQ-plot.

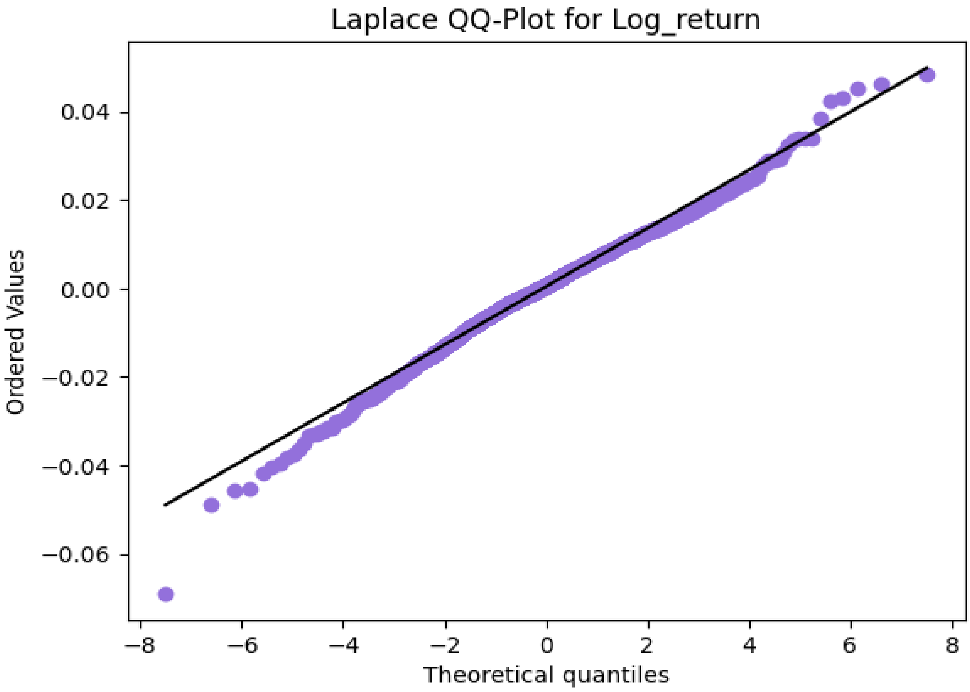

Figure 4.6 Laplace distribution QQ-plot.

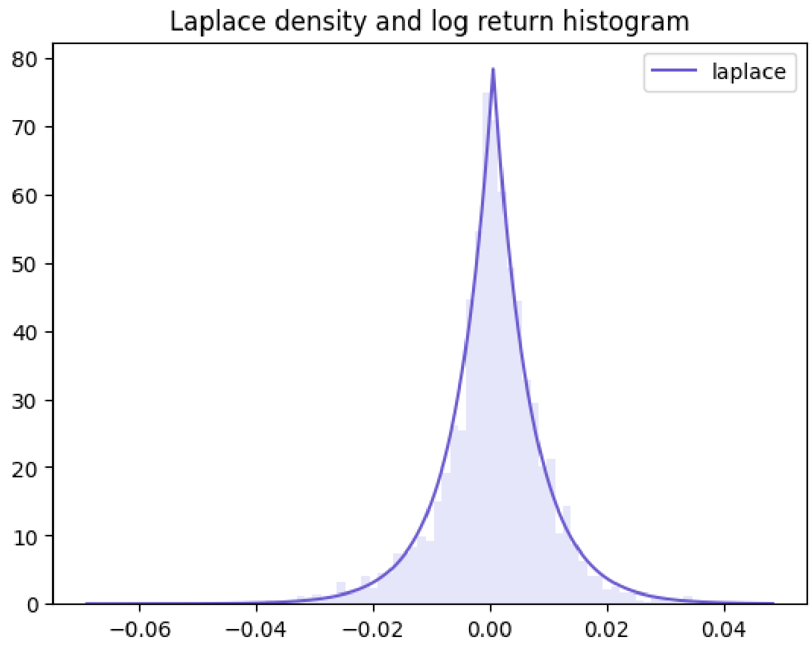

Figure 4.7 Density histogram compared with Laplace distribution.

Both Figure 4.4 and Figure 4.5 showed that the Laplace distribution might fit the data better than *t*-distributions. To prove this, further investigation is performed using the Kolmogorov-Smirnov (KS) test. The results are shown in Figure 4.5 below.

```
Student's t:
KS statistic: 0.4977675609327342, p-value: 0.0
Laplace:
KS statistic: 0.022902396612338338, p-value: 0.14073459801031385
```

Figure 4.8 KS Test results for *t* and Laplace distributions.

A higher KS statistic of student's *t*-distribution indicates a greater difference with the data. The extremely low *p*-value (close to zero) suggests strong evidence against the null hypothesis that the sample comes from a Student's *t*-distribution. Therefore, it's likely that the sample does not follow a Student's *t*-distribution. Besides, the *p*-value of 0.1407 suggests that there is evidence that the sample comes from a Laplace distribution at a significance level of 0.05 (or 5%).

4.4 Estimated parameters

The Kendalls tau obtained is around -0.02818, showing a slight negative correlation of the data. Thus the estimated parameters for Clayton copula and Gaussian copula are around -0.05428 and -0.04426 respectively.

The location and scale of the Kendall's Tau is the actual parameters calculated from the datasets. By comparing the table below it shows that the MLE of Clayton copula obtained the closest scale compare to the MLE of the Gaussian. However, both obtained the similar locations.

Table 4.1: Estimated parameters of Clayton and Gaussian using MLE and Kendall's Tau.

Copula		MLE	Kendall's Tau
Clayton	μ	2.08E-03	5.93E-04
	α	6.83E-03	6.37E-03
	θ	4.29E+00	-5.48E-02
	RMSE	**0.013322**	**0.013379**
Gaussian	μ	2.06E-03	5.93E-04
	α	8.96E-03	6.37E-03
	θ	0.4273	-4.43E-02
	RMSE	**0.015414**	**0.019091**

4.5 Evaluation

After that, data is simulated using the formulated algorithm mentioned in section 3.5. The simulated data and actual data are then plotted as below.

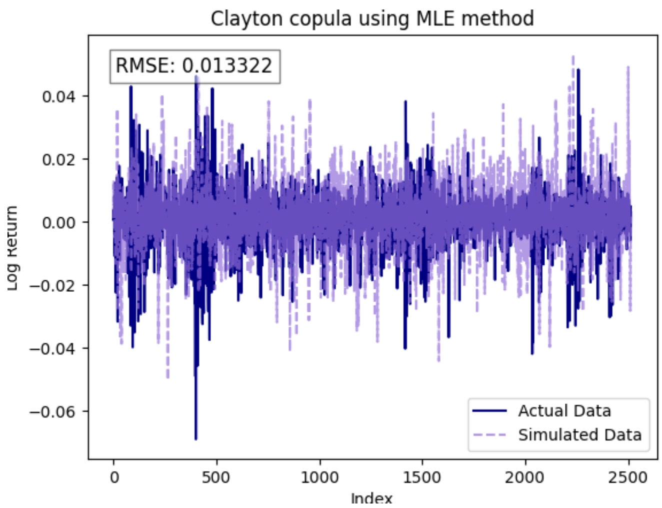

Figure 4.9 Clayton copula and MLE method.

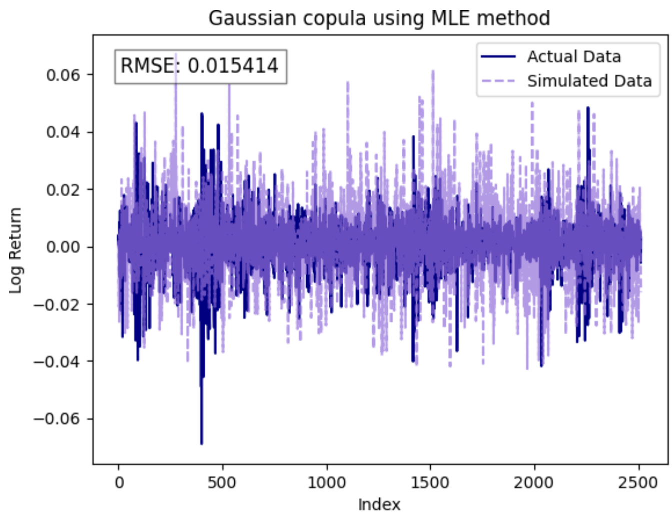

Figure 4.10 Gaussian copula and MLE method.

Figure 4.8 and 4.9 shows the line plots of simulated and actual data using Clayton and Gaussian copulas, where both of their parameters estimated using MLE method. The light purple and dark blue line indicating the line plot of simulated data and actual data respectively. The darker purple in the middle of line is where the simulated data overlay the actual data. It shows very clearly that both the line in Clayton is more similar compare to the Gaussian copula. Conclusion can be made that the Clayton perform better than Gaussian when the MLE is used due to the lower RMSE.

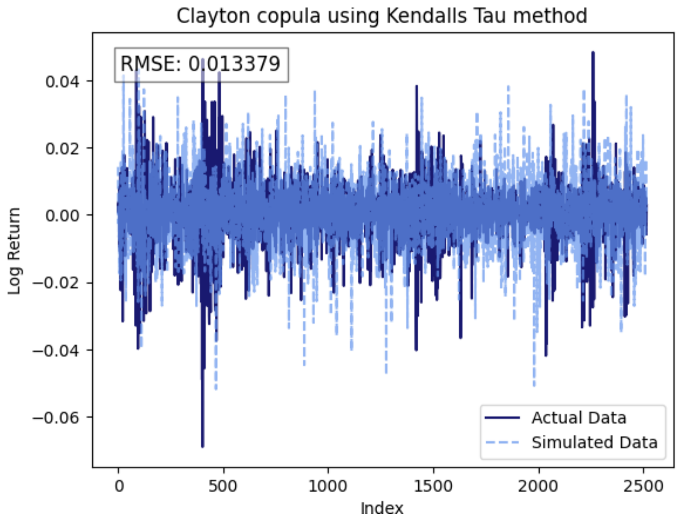

Figure 4.11 Clayton copula and Kendall’s Tau method.

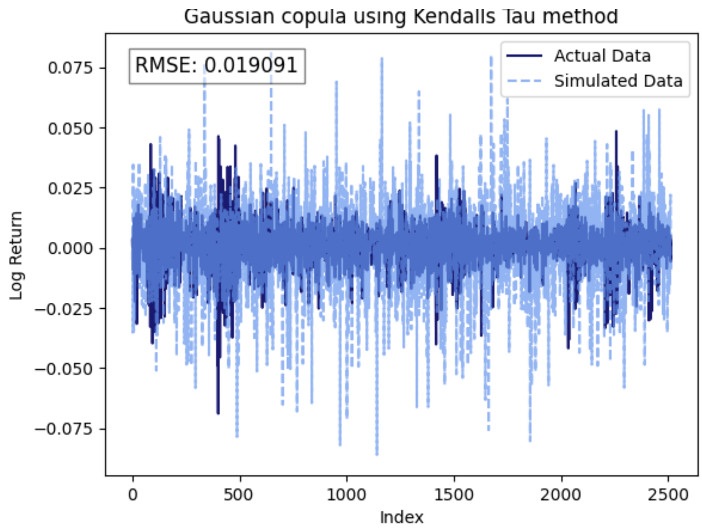

Figure 4.12 Gaussian copula and Kendall's Tau method.

Similar to the MLE method, the simulation of both Copula where their parameter were estimated using Kendall's Tau are shown in Figure 4.10 and 4.11. The Figure 4.11 showing the worst case were the simulated data that is too dispersed. From all the figures above, great differences between the best and worst performed models are shown. Thus, the Clayton is once again outperformed the Gaussian copula. On top of that, it is obvious that the MLE has better performance than Kendall's Tau. In short, the combination of Copula and MLE is the best model for modelling Laplace marginal distributed data.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

In conclusion, this study aimed to, firstly, identify the marginal distribution of SPX500's log returns and make sure it is following the Laplace distribution. Secondly, estimate the parameters of copula-based Markov models using both the Maximum Likelihood Estimation (MLE) method and the inversion of Kendall's Tau method and lastly assess the performance of the Clayton and Gaussian copulas.

The findings reveal that the Clayton copula provides a better fit for the ten-year SPX500 daily log returns, outperforming the Gaussian copula. Both MLE and Kendall's Tau yielded comparable results for the Clayton copula, with MLE having a slight advantage. However, for the Gaussian copula, MLE performed significantly better than the inversion of Kendall's Tau method. In short, this study had succesfully explored a combination of models that has been under-examined, providing insights into their performance for Laplace distributed time series data.

5.2 Recommendations

This study employs the Clayton copula, Gaussian copula, and Laplace distribution to model SPX500 log returns. However, future research could explore a broader range of methods and models to gain deeper insights. For instance, comparing the performance of other copulas, such as the Gumbel, Student-*t* or Frank copulas, along with different marginal distributions which could offer valuable comparisons. Additionally, employing alternative parameter estimation techniques, such as the Method of Moments or Bayesian inference, might enhance model accuracy. Finally, integrating forecasting models such as GARCH or its variants could provide a dynamic view of volatility and further enrich the analysis.

REFERENCES

Alqawba, M., and Diawara, N. (2020). Copula-based Markov zero-inflated count time series models with application. *Journal of Applied Statistics*, *48*(5), 786–803. https://doi.org/10.1080/02664763.2020.1748581

Chen, X., Zhang, X., and Bo, W. (2022). Copula-based time series with filtered nonstationarity. *Journal of Econometrics*, *228*(1), 127–155. https://doi.org/10.1016/j.jeconom.2020.10.008

D’Amico, G., Petroni, F., Regnault, P., Scocchera, S., and Storchi, L. (2019). A copula-based Markov reward approach to the credit spread in the European Union. *Applied Mathematical Finance*, *26*(4), 359–386. https://doi.org/10.1080/1350486x.2019.1702068

Darsow, W., Nguyen, B., and Olsen, E.T. (1992). Copulas and Markov processes. *Illinois Journal of Mathematics*, *36*, 600-642. doi:10.1215/ijm/1255987328.

Dewick, P. R., and Liu, S. (2022). Copula modelling to analyze financial data. *Journal of Risk and Financial Management*, *15*(3), 104. https://doi.org/10.3390/jrfm15030104

Durante, F., and Sempi, C. (2015). Principles of copula theory. *CRC Press*. https://doi.org/10.1201/b18674

Emura, T., Lai, C., and Sun, L. (2023). Change point estimation under a copula-based Markov chain model for binomial time series. *Econometrics and Statistics*, *28*, 120–137. https://doi.org/10.1016/j.ecosta.2021.07.007

Emura, T., Long, T. H., and Sun, L. H. (2016). R routines for performing estimation and statistical process control under copula-based time series models. *Communications in Statistics: Simulation and Computation*, *46*(4), 3067–3087. https://doi.org/10.1080/03610918.2015.1073303

Gagniuc, P. (2017). Markov Chains: From Theory to Implementation and Experimentation. *USA, NJ: John Wiley* and *Sons. pp*. 1–235. ISBN 978-1-119-38755-8.

Huang, X., and Emura, T. (2019). Model diagnostic procedures for copula-based Markov chain models for statistical process control. *Communications in Statistics. Simulation and Computation*, *50*(8), 2345–2367. https://doi.org/10.1080/03610918.2019.1602647

Huang, X., and Emura, T. (2022). Computational methods for a copula-based Markov chain model with a binomial time series. *Communications in Statistics: Simulation and Computation*, 1–18. https://doi.org/10.1080/03610918.2022.2061514

Huang, X., Wang, W., and Emura, T. (2020). A copula-based Markov chain model for serially dependent event times with a dependent terminal event. *Japanese Journal of Statistics and Data Science*, *4*(2), 917–951. https://doi.org/10.1007/s42081-020-00087-8

Kim, J. M., Baik, J., and Reller, M. (2019). Control charts of mean and variance using copula Markov SPC and conditional distribution by copula. *Communications in Statistics: Simulation and Computation*, *50*(1), 85–102. https://doi.org/10.1080/03610918.2018.1547404

Long, T., and Emura, T. (2014). A control chart using copula-based Markov chain models. *MPRA.* https://econpapers.repec.org/paper/pramprapa/57419.htm

Nadaf, T., Lotfi, T., and Shateyi, S. (2022). Revisiting the Copula-Based trading method using the Laplace marginal Distribution function. *Mathematics*, *10*(5), 783. https://doi.org/10.3390/math10050783

Pepi, N., Gunardi, and Rosadi, D. (2024). Copula-Based Markov Chain Logistic Regression modeling on binomial time series data. *MethodsX*, *12*, 102509. https://doi.org/10.1016/j.mex.2023.102509

Sun, L. H., Lee, C. S., and Emura, T. (2018). A Bayesian inference for time series via copula-based Markov chain models. *Communications in Statistics: Simulation and Computation*, *49*(11), 2897–2913. https://doi.org/10.1080/03610918.2018.1529241

Sun, L., Huang, X., Alqawba, M., Kim, J. M., and Emura, T. (2020). Copula-Based Markov models for time series. *Springer Singapore.* https://doi.org/10.1007/978-981-15-4998-4