

Application of 4-dimensional Copulas In Calculating Value-At-Risk for the Portfolio of 4 SP500 companies

Kairzhan Bolatbekov

May 11, 2023

Supervisor: Dr. Dongming Wei
Second Reader: Dr. Rustem Takhanov

Abstract

Portfolio risk management is a process aimed at maintaining profit streams and reducing uncertainties in investment decisions. Value-at-Risk (VaR) is a widely used metric to quantify the potential loss of profits. Although historical simulations and Gaussian distribution are common methods for estimating VaR, modelling the joint multivariate distribution of portfolio investments can be challenging. Copula models offer a solution to these challenges for joint distributions. In this study, we calculated VaR and Conditional Value-at-Risk (CVaR) for a portfolio consisting of the four least correlated stocks among the 15 largest companies in the SP 500 using historical simulations and copula models. We evaluated portfolio based on equal weighting. The optimal ARIMA-GARCH model was selected using Akaike Information Criteria (AIC) values. Furthermore, the performance of the VaR estimations was compared and analyzed using goodness-of-fit tests.

Keywords: VaR, Copula, ARIMA-GARCH, Risk Management, Data Series

1 Introduction

In the risk management domain, risk signifies the numerical probability of incurring a loss or experiencing returns that are lower than anticipated (12). A key responsibility of financial risk managers is to curtail client losses while adhering to specified risk levels. Ascertaining the extent of risk is a vital component of risk mitigation.

This project represents the value of a portfolio consisting of d investments, each with weights w_1, \dots, w_d , at a given time t as $V_t = \sum_{i=1}^d w_i V_{(i,t)}$, where $V_{(i,t)}$ denotes the value of the i -th investment in the portfolio at time t (12). Additionally, all weights are non-negative and collectively sum up to 1 ($w_i \geq 0$, $\sum_{i=1}^d w_i = 1$). Each investment can be influenced by multiple risk factors. Examples of non-mutually exclusive risks affecting an investment are interest-rate risks, management risks, marketability risks, political risks, and unsystematic and systematic risks, among others (10). While some risks can be readily quantified, others cannot. A portfolio investment risk analyst's role is to evaluate the impact of various risk factors on all investments within the portfolio. Consequently, the portfolio's value hinges on both time and risk factors that might concurrently influence all d investments. Therefore, V_t is rewritten as $V_t = f(t, \mathbf{Z}_t)$, where f represents a function, t denotes the time period, and \mathbf{Z}_t signifies a d -dimensional random vector of risk factors that is equal to $\mathbf{Z}_t = [Z(1,t), \dots, Z(d,t)]$ (12). In this project, the 4-dimensional case is considered as an example, with d equating to 4 in all calculations.

Moreover, the portfolio's loss between day t and $t + 1$ is defined as the difference in portfolio values across these two days ($L_{t+1} := L_{[t\Delta, (t+1)\Delta]} = -(V_{t+1} - V_t)$). In this case, Δ typically corresponds to $1/365$, representing one day in a calendar year, and $t \in \mathbb{N}$ (12). To assess a portfolio's risk, it is necessary to identify the loss distribution function: $F_L(y, t\Delta) := P(L_{[t\Delta, (t+1)\Delta]} \leq y)$, where y represents the potential portfolio loss value (12). Crucially, understanding the joint distribution function for all d investments in the portfolio is essential for estimating the loss distribution function. However, this requires knowledge of the dependence structure of the investments. The utilization of time-series models and copulas as a means to estimate risks and discern the dependence structure of investments will be presented, and this theory will subsequently serve as the foundation for analyzing 4 stocks listed in SP500 corporations.

2 Theory

2.1 Value-at-Risk

One of the most common methods for assessing market risks is VaR. VaR calculates the maximum expected loss at a specific confidence level (typically 95% or 99%) over a particular time horizon (usually a year for commodities) under normal market conditions (12). A positive VaR value indicates the worst potential loss, while a negative VaR value implies a higher likelihood of increased profits (4). Absolute VaR is defined as the value of the absolute losses of the portfolio in dollars. VaR can also be employed to estimate each investment's marginal contribution to the portfolio risks and to determine the expected loss if any investment is removed from the portfolio (7).

Definition. Given a confidence level $\alpha \in (0, 1)$, the *VaR* of our portfolio at the confidence level α is represented by the smallest number y such that the probability that the loss L surpasses y is at most $1 - \alpha$.

$$VaR_\alpha = \inf y \in \mathbb{R} : P(L_{[t\Delta, (t+1)\Delta]} > y) \leq 1 - \alpha = \inf y \in \mathbb{R} : F_L(y, t\Delta) \geq \alpha \quad (1)$$

VaR is easily interpretable, as it encapsulates risks in a single number. For instance, if our model suggests a yearly VaR at a 95% confidence level of \$1 million, it means there is a 5% chance of experiencing a loss greater than \$1 million in a year (12). In this project, we will consider confidence levels of 90%, 95%, and 99.5%.

Various methods exist for estimating VaR, with the following three being the most common: analytical estimation based on return distribution assumptions, empirical computations using past returns histograms, and Monte Carlo simulations-based computations (12). Each of these methods has its pros and cons due to the assumptions they make when modelling.

For instance, VaR can be modelled using normality assumptions (known as parametric or normal VaR) or t-student distribution assumptions (referred to as t-series VaR) of returns. To be accurate, the financial loss distribution functions should follow normal or t-student distributions. However, financial losses typically exhibit high kurtosis values, indicating fatter tails (4). This implies a higher probability of extreme positives and negatives, which is called a leptokurtic distribution. As a result, normal VaR may not be the most appropriate way to model VaR for financial data, as it might significantly underestimate or overestimate VaR values (7). This has been demonstrated by various empirical evidence (16). Due to these frequent errors, there is substantial criticism against using this model in such cases. Hence, it is crucial to focus on tail dependence, as a good model should capture tail behaviour, be easily implemented, and be extendable for larger portfolios (16).

Furthermore, VaR can be calculated using non-parametric predictions based on historical returns, which do not make any assumptions about risk-factor distributions (7). This method

relies on the past financial performance of investments. The main critique of this approach is that the present cannot always accurately predict the future with high certainty (10). The success of this method depends on the relevance of the data used for various risk factors (12). If new risk factors emerge, this approach will quickly fail. Additionally, the assumption that all days have equal weight in making predictions is not always accurate. As a result, estimates derived from historical returns might be higher or lower than the actual risk and will be influenced by volatility trends. To mitigate the impact of volatility, various volatility models are employed, with the GARCH model being a frequently used example that will be utilized in this paper (10), (2).

As previously mentioned, parametric VaR estimation for a portfolio can become extremely complex due to the need to model the joint multivariate distribution function. This complexity stems from the joint distribution function for a random vector of risk factors, which implicitly describes both the individual marginal behaviour of risk factors and the dependence structure between them (10). The dependence structure between risk factors can take any form. While assuming mutual independence between risk factors is the simplest approach, it is often unrealistic due to factors such as macroeconomic systematic risk factors (e.g., the state of the economy, coronavirus pandemics) that may impact all investments in the portfolio simultaneously. Consequently, in most situations, there will be dependence between different investments in the portfolio, and assuming independence might be an unrealistic idea.

2.2 Copulas

The concept of copula can prove to be useful in situations like the one mentioned above where we should not simply assume independence but derive the dependence structure of various investments. Copulas are such dependence functions that can isolate the dependencies and provide us with models of univariate distributions separated from their dependence (4). Additionally, the dependence by copulas is expressed on a quantile scale, which is similar to VaR and is also quite interpretable (12). Copula-based VaR models might also provide better and more accurate VaR forecasts compared to parametric and historical simulation methods, according to various researches in the field(1),(16).

Definition. A d -dimensional copula is a distribution function on $[0, 1]^d$ with standard uniform marginal distributions. We denote it $C(\mathbf{u}) = C(u_1, \dots, u_d)$.

Copula function C is a mapping of the form $C: [0, 1]^d \rightarrow [0, 1]$. This means that it is a mapping of a d -dimensional unit cube into the unit interval. The following properties must hold.

- $C(u_1, \dots, u_d)$ is d -increasing (increasing in each component).
- $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ for all $i \in 1, \dots, d, u_i \in [0, 1]$.

Here, "d-increasing" stands for a d -dimensional analog of a non-decreasing function of one variable (14).

Theorem. For all copulas $C(u_1, \dots, u_d)$ we have the Fréchet bounds

$$\max\left\{\sum_{i=1}^d u_i + 1 - d, 0\right\} \leq C(\mathbf{u}) \leq \min\{u_1, \dots, u_d\} \quad (2)$$

The lower bound is denoted as $W(\mathbf{u})$ and the upper bound is $M(\mathbf{u})$ (12).

In 1959 Sklar highlighted the applications of copulas in the studies of multivariate distribution functions. This theorem suggests that copulas are contained in all multivariate distribution

functions and that they can be used with univariate distribution functions to construct multivariate cdfs (12).

Sklar's Theorem. Let F be a joint distribution function with marginal distribution functions F_1, \dots, F_n . Then, there exists a copula $C: [0, 1]^d \rightarrow [0, 1]$, such that, for all x_1, \dots, x_n in $(-\infty, \infty)$,

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)) \quad (3)$$

If the marginals are continuous, then C is unique. If that's not the case, C is uniquely determined on $\text{Ran}F_1 \times \dots \times \text{Ran}F_n$, where $\text{Ran}F_i$ is the range of F_i . Also, if C is a copula and F_1, \dots, F_n are univariate dfs, then F , defined above is a joint distribution function with margins F_1, \dots, F_n (14).

From Sklar's theorem we get that copula links marginal distributions to a joint distribution. Thus, for continuous distribution functions, the question of obtaining the joint distribution function is changed to the selection of the appropriate copula based on goodness-of-fit tests (4).

2.3 Copula types and application to the 4 dimensional model

There exist various forms of copulas, including fundamental, explicit, and implicit types. Fundamental copulas serve as representations for specific dependence structures, while explicit copulas possess closed-form expressions, and implicit copulas employ Sklar's theorem, deriving from well-known multivariate distribution functions (12). Some fundamental copulas include independence ($\prod_{i=1}^d u_i$) and comonotonicity ($M(\mathbf{u}) = \min u_1, \dots, u_d$) copulas. Random variables associated with such copulas are referred to as independent and comonotonic, respectively (14). The lower Fréchet bound ($W(\mathbf{u}) = \max \sum_{i=1}^d u_i + 1 - d, 0$) functions as a copula only in 2-dimensions, but not in $d \geq 3$. This specific copula is known as the counter monotonicity copula (12).

Regarding explicit copulas, Archimedean, Elliptical, and Extreme-Value copulas are prevalent (16). In the present study, two Archimedean copulas will be employed, specifically Gumbel and Clayton copulas, which belong to the well-researched Archimedean copula family and have been selected for this paper. The Archimedean model's general formula is as follows:

$$C(u_1, u_2, \dots, u_d) = \phi^{[-1]}(\phi(u_1) + \phi(u_2) + \dots + \phi(u_d)) \quad (4)$$

here ϕ is a strict generator, with ϕ^{-1} being completely monotonic on $[0, \infty)$ (5). As in this project, we consider the 4-dimensional case, the formula for Archimedean models will have a form of

$$C(u_1, u_2, u_3, u_4) = \phi^{[-1]}(\phi(u_1) + \phi(u_2) + \phi(u_3) + \phi(u_4)), \quad (5)$$

where ϕ is a generator function of the copula (14). Overall, not all copulas can be extended to $d \geq 3$ and this depends on the following theorem.

Theorem. Let ϕ be a continuous strictly decreasing function to $[0, \infty)$ such that $\phi(0) = \infty$ and $\phi(1) = 0$, and let ϕ^{-1} denote the inverse of ϕ . If C is the function from $[0, 1]^d$ to $[0, 1]$, then C is an d -dimensional copula for all $d \geq 2$ if and only if ϕ^{-1} is completely monotonic on $[0, \infty)$ (14).

Additionally, the Gaussian copula is going to be used in this paper after checking individual marginal distribution functions and transforming them into normal distribution functions (4). It belongs to the Elliptical copula family and has a symmetric distribution of multivariate data (12). The gaussian copula is an implicit formula that does not have a simple closed form. Its formula for d -dimensions can be represented as follows.

$$C_P^{Ga}(u_1, u_2, \dots, u_d) = P(\Phi(X_1) \leq u_1, \dots, \Phi(X_d) \leq u_d) = \Phi_{\mathbf{P}}(\Phi^{-1}(u_1) + \Phi^{-1}(u_2) + \dots + \Phi^{-1}(u_d)) \quad (6)$$

Here $\mathbf{X} \sim N_d(\mathbf{0}, P)$, where P is the correlation matrix of \mathbf{Y} , where $\mathbf{Y} \sim N_d(\mu, \Sigma)$ is a Gaussian random vector. Φ is the standard univariate normal distribution function and $\Phi_{\mathbf{P}}$ denotes the joint distribution function of \mathbf{X} . Also, C_P^{Ga} , Gaussian copula, highlights the parametrization of copula by $\frac{1}{2}d(d-1)$ parameters of the correlation matrix (12).

The generator functions and some important characteristics of Archimedean 4-dimensional copulas selected for this capstone project are summarized in the Table I. Domain for Clayton copula is $[-1, \infty]$ and for Gumbel copula is $[1, \infty]$

Table 1: First three columns of selected Archimedean four-variate copulas

Copula	Generator	4-dimensional Copula
Clayton	$\phi(t; \theta) = \frac{t^{-\theta} - 1}{\theta}$	$(u_1^{-\theta} + u_2^{-\theta} + u_3^{-\theta} + u_4^{-\theta} - 3)^{-1/\theta}$
Gumbel	$\phi(t; \theta) = (-\ln(t))^\theta$	$exp\{-[(-\ln(u_1))^\theta + (-\ln(u_2))^\theta + (-\ln(u_3))^\theta + (-\ln(u_4))^\theta]^{1/\theta}\}$

It is essential to mention the various applications of 4-dimensional copulas in different fields. Copulas have found widespread use in finance, insurance, and risk management due to their ability to model complex dependence structures among multiple random variables (15; 6). In particular, 4-dimensional copulas can be applied to study the joint behaviour of four distinct assets or risk factors.

In insurance and actuarial science, 4-dimensional copulas are used to model the joint distribution of claim frequencies and severities for different lines of business, enabling a comprehensive assessment of the insurer's risk exposure. Moreover, in portfolio management and asset allocation, 4-dimensional copulas can offer insights into the dependence structure among various asset classes or investment strategies, facilitating more informed diversification and risk management decisions. Recent advancements in copula-based models have also explored the applications of deep learning techniques for more accurate estimation and forecasting, especially in the field of finance (18).

To generate 4 dimensional copulas we used Pair Copula Construction, which allows for free specification of $n(n-1)/2$ bivariate copulas. The modeling scheme is based on a decomposition of a multivariate density into $n(n-1)/2$ bivariate copula densities, of which the first $n-1$ are dependency structures of unconditional bivariate distributions, and the rest are dependency structures of conditional bivariate distributions.(11). For the four-variate case, this method will look like in equation 7 and Figure 1.

$$\begin{aligned}
C(u_1, u_2, u_3, u_4) = & C_{12}(u_1, u_2) \cdot C_{23}(u_2, u_3) \cdot C_{34}(u_3, u_4) \\
& \cdot C_{13|2}((F(u_1|u_2), F(u_3|u_2))) \cdot C_{24|3}(F(u_2|u_3), F(u_4|u_3)) \\
& \cdot C_{14|23}(F(u_1|u_2, u_3), F(u_4|u_2, u_3))
\end{aligned}$$

Where

$$\begin{aligned}
F(u_1|u_2) &= \frac{\partial C_{12}(u_1, u_2)}{\partial u_2} \\
F(u_3|u_2) &= \frac{\partial C_{23}(u_2, u_3)}{\partial u_2} \\
F(u_2|u_3) &= \frac{\partial C_{23}(u_2, u_3)}{\partial u_3} \\
F(u_4|u_3) &= \frac{\partial C_{34}(u_3, u_4)}{\partial u_3} \\
F(u_1|u_2|u_3) &= \frac{\partial C_{13|2}(F(u_1|u_2), F(u_3|u_2))}{\partial F(u_3|u_2)} \\
F(u_4|u_2|u_3) &= \frac{\partial C_{24|3}(F(u_4|u_3), F(u_2|u_3))}{\partial F(u_2|u_3)}
\end{aligned} \tag{7}$$

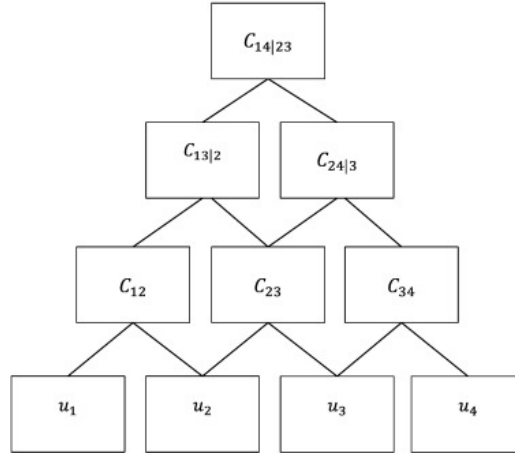


Figure 1: Four-dimensional Pair Copula Construction.

2.4 Dependence measures

There are various dependence measures, but the ones that will be discussed in this section are the most commonly used ones: Pearson's linear correlation, rank correlation (Kendall's tau and Spearman's rho) and coefficients of tail dependence.

Pearson's linear correlation measures the *linear* dependence between variables and can take values from -1 to 1. If the value of ρ is positive, then random variables are positively correlated, if it is negative, then - negatively correlated. For random variables X and Y with $Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$, it is defined as following:

$$Corr = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{8}$$

Correlation takes an important part in the financial and portfolio management theory but there are some cautions. This concept is natural only in the context of multivariate normal dfs.

Moreover, as it can be seen from the formula above, correlation makes sense only if variances of X and Y are finite. This might create issues for heavy-tailed distributions prevalent in finances, as it might not be the most suitable model to describe dependence between random variables (12).

Rank correlations are copula-based measures of dependence. This means that they only depend on the copula and not on the marginal distribution functions, compared to Pearson's linear correlation which considers both of those values (12). Overall, they are easily calculated only from the ranks (orders) of data in the sample of each random variable. Two common rank correlation coefficients are Kendall's tau and Spearman's rho and they will be discussed in this section.

Kendall's tau is a measure of concordance. For random vectors (X_1, X_2) and $(\tilde{X}_1, \tilde{X}_2)$, concordant pairs have the following relation: $(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0$. On the other hand, discordant pairs are such pairs which have $(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0$. If we try to interpret this, with the increase of one random variable, another one tends to increase, then the probability of them being concordant rather than discordant is higher (12).

Using this definition of concordance, we can denote Kendall's tau ($\rho_\tau(X_1, X_2)$) in such a way that

$$\rho_\tau(X_1, X_2) = P((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0) - P((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0) \quad (9)$$

For the trivariate case, Joe's extension of Kendall's tau using bivariate copulas is the following:

$$\rho_\tau(X_1, X_2, X_3) = \frac{1}{3} \{ \rho_\tau(X_1, X_2) + \rho_\tau(X_1, X_3) + \rho_\tau(X_2, X_3) \} \quad (10)$$

As for the relation of Kendall's tau with bivariate copulas, Nelsen suggests that

$$\rho_\tau(X_1, X_2) = -1 + 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) \quad (11)$$

Spearman's rho is another measure of dependence. It is defined as linear correlation of unique copula associated with the general multivariate random vector \mathbf{X} .

$$\rho_S(\mathbf{X}) = \rho(F_1(X_1), F_2(X_2), \dots, F_d(X_d)) \quad (12)$$

Here, F_1, \dots, F_d are marginal distribution functions (14). Both of those rank correlation dependence measures also take values from -1 to 1, similar to Pearson's correlation. Value of -1 is associated with countermonotonicity and value of 1 - with comonotonicity (12).

Spearman's rho also has the formula suggested by Nelsen:

$$\rho_S(X_1, X_2) = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3 \quad (13)$$

It is stated that for rank correlations for continuous marginal dfs depend only on the unique copulas (5). Therefore, Kendall's tau inverse will be used to infer the parameters of copulas based on the ranks associated with our sample. Also, because the structure of dependence changes over time, time series models will be used for analysis of the financial data.

Moreover, as identified above, financial data frequently departs from normality and has a leptokurtic distribution with skewness, also known as fat-tail problem. This problem is present in both univariate and multivariate settings, where large market movements tend to occur together. This concept is *tail dependence* (5). This concept is necessary for risk managers to know to work against unwanted events (1). Also, in finances high losses are of interest for risk managers, so upper and lower tail dependence coefficients are usually computed to analyze

the joint behavior of random variables. The formulas for lower and upper tail dependence are written below.

$$\lambda_L = \lim_{\alpha \rightarrow 0^+} \frac{C(\alpha, \alpha)}{\alpha} \quad (14)$$

$$\lambda_U = \lim_{\alpha \rightarrow 1^-} \frac{1 - 2\alpha + C(\alpha, \alpha)}{1 - \alpha} = \lim_{\alpha \rightarrow 1^-} 2 - \frac{1 - C(\alpha, \alpha)}{1 - \alpha} \quad (15)$$

Moreover, tail dependence coefficients for the chosen trivariate copulas are listed in the Table II (1).

Table 2: Tail dependence of copula families

Copula name	Parameter range	Lower tail	Upper tail
Gaussian	$(-1, 1)$	0	0
Clayton	$\alpha \in [-1, 0) \cup (0, \infty)$	$2^{-\frac{1}{\alpha}}$	0
Gumbel	$\alpha \in [1, \infty)$	0	$2 - 2^{-\frac{1}{\alpha}}$

2.5 Time series basics

For financial return series there are some empirical observations that apply to most of series of changes in the risk-factors (for example, log returns in this project). Those facts are as follows:

- (1) Return series are not independent and identically distributed;
- (2) Series of absolute or squared returns have significant serial correlation;
- (3) Conditional expected returns approach zero;
- (4) Volatility is not fixed and varies over time;
- (5) Return series possess leptokurtic distribution (there are narrower centers and higher chances of extreme values compared to normal distribution);
- (6) Extreme returns tend to appear in clusters (if there are extreme results, they are usually followed by other extreme ones) (12).

In this project we will be able to empirically check if those facts apply to the data series of the assets in the portfolio.

Definition. *Time series model* for a single risk factor denoted by $(X_t)_{t \in \mathbb{Z}}$ is a family of rvs, indexed by integers and defined on the probability space (Ω, \mathcal{F}, P) .

The definition of stationarity is important for time series. So, the following two definitions should provide some explanation of strictly and weakly stationary time series.

Definition. Time series $(X_t)_{t \in \mathbb{Z}}$ is called *strictly stationary* if $X_{t_1}, \dots, X_{t_n} = X_{t_1+k}, \dots, X_{t_n+k}$, for all $t_1, \dots, t_n, k \in \mathbb{Z}$ and for all $n \in \mathbb{N}$.

Definition. Time series $(X_t)_{t \in \mathbb{Z}}$ is called *covariance (weakly) stationary* if the first two moments exists and satisfy the following:

$$\mu(t) = \mu, t \in \mathbb{Z} \quad (16)$$

$$\gamma(t, s) = \gamma(t + k, s + k), t, s, k \in \mathbb{Z} \quad (17)$$

From the definition of covariance stationary time series, we see that covariance between X_t and X_s only depends on their separation $|s - t|$, known as *lag* (12). Thus, this means that stationary time series do not have any trends or effects of seasons. They are not dependent on time. Non-stationary series, on the other hand, show some dependence on seasons or other similar trends and depend on the particular time index. Time series analysis is concerned with changing non-stationary time series into stationary time-series. This is done after the

identification and removal of trends (3). There are various tests to check the trends and correlation of data. One of the possible solutions is checking the ACF and PACF plots.

Definition. *Autocorrelation Function*, also denoted as ACF - $\rho(h)$ of covariance stationary time series $(X_t)_{t \in \mathbb{Z}}$. It is the following:

$$\rho(h) = \rho(X_h, X_0) = \gamma(h)/\gamma(0), \forall h \in \mathbb{Z}, \quad (18)$$

where h is the *lag*.

ACF summarizes how observation values are correlated with lag values. On the y-axis, correlation coefficient can take values between -1 and 1 (3). On the other hand, we can also look at Partial Autocorrelation Function (PACF). It summarizes how an observation is correlated with lag values given that is not accounted for by previous observations with lags (3). After calculations, the results of autocorrelation and partial autocorrelations of residuals of time series can be plotted. Such plots are called autocorrelation plots or correlograms and they play an important role in analyzing the stationarity of data.

Augmented Dickey-Fuller test is a formal test to check the stationarity of data. This is a unit root test that checks the existence of trends behind the time series. In this test we have two hypotheses:

H_0 : Time series has a unit root, which means it is non-stationary. That is there is a time dependent structure.

H_1 : Time series doesn't have a unit root, which means it is stationary. That is there is no time dependent structure (3).

2.6 Time series models

As dependence structure between assets change with time, the time series should be used to analyze our return series. Additionally, volatility can be higher or lower than average values in some periods. Models like ARIMA and GARCH have been used in most latest researches to model financial data (4),(16).

The Autoregressive Integrated Moving Average Model (ARIMA) is a standard class of a statistical model that is used for time series forecast and analysis (3). This model creates covariance stationary time series and uses dependency between residual errors from a moving average model with a number of lagged observations and observation values. ARIMA(p,d,q) has three parameters, for which integer values are substituted to indicate what kind of ARIMA model is used in the forecasting. The parameters show the next factors:

p - how many lag observations are included in the model(lag order);

d - how many times raw observations are differenced (degree of differencing);

q - how large is the moving average window (order of moving average).

ARIMA (p,d,q) model is ARMA (p,q) model differenced d times (3).

The formula for ARMA(p,q) model is as follows:

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} - \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} + \dots - \theta_q \epsilon_{t-q}, \quad (19)$$

where y_t - data on which the ARMA model is applied, C - constant, ϵ_t - white noise. The parameters ϕ_i are AR coefficients and parameters θ_i are MA coefficients. ARMA(p,q) can also be rewritten in terms of lags:

$$(1 - \sum_{i=1}^p \phi_i L^i) y_t = (1 - \sum_{j=1}^q \theta_j L^j) \epsilon_t, \quad (20)$$

where L - lag operator(2). This is often reduced to:

$$\phi_p(L) y_t = \theta_q(L) \epsilon_t, \quad (21)$$

where $\phi_p(L) = (1 - \sum_{i=1}^p \phi_i L^i)$ and $\theta_q(L) = (1 - \sum_{j=1}^q \theta_j L^j)$

The general formula for ARIMA(p,d,q) model in terms of lag operator is as follows:

$$\phi_p(L)(1 - L)^d y_t = \theta_q(L)\epsilon_t, \quad (22)$$

Both ACF and PACF plots are used as graphical diagnostic tests to decide on the values of p,d,q for ARIMA model. If ACF trails off and PACF has a hard cut-off after a particular lag, the appropriate model is AR and lag is a value of p. If PACF trails off and ACF has a hard cut-off after a particular lag, the appropriate model is MA and lag is a value of q. If, both of those functions trail off, the appropriate model is ARIMA (3).

*The Generalized Autoregressive Conditional Heteroscedasticity, also denoted as **GARCH***, is an exponential decay model. This model is designed to capture volatility clustering of data returns. Volatility, being a measure of intensity of random changes in returns of investments, can affect assets vastly. GARCH model using past returns and historical volatilities models the variance (4).

Suppose, r_t is a stochastic process of daily log returns. Then a univariate normal GARCH(m,n) process is defined in the following way:

$$r_t = \mu_t + \epsilon_t \quad (23)$$

$$\epsilon_t = \sigma_t z_t \quad (24)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^n \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^m \beta_j \sigma_{t-j}^2 \quad (25)$$

such that r_t - log return at time t, μ_t - expected value of the conditional log return at time t, ϵ_t - mean corrected return at time t (4).

The simplest GARCH model - GARCH (1,1). It is quite frequently used in finances and can capture the volatility of the market pretty well (4). It is defined in the following way:

$$r_t = \mu_t + \epsilon_t \quad (26)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (27)$$

Moreover, GARCH model's parameters can be found and checked using historical data. Sample data with its volatility clusters is used to estimate parameters and fit the most appropriate GARCH model (4). It is usually done in the software due to the complexity of calculations.

2.7 Criteria of checking

To evaluate the goodness of a model, information-theoretic approach considers the information discrepancy compared to true distribution generating the data as the most important criterion (8). In this project *Akaike and Bayesian information criteria* (AIC and BIC respectively) for choosing a suitable model, quite commonly used information criteria, are used.

AIC takes an important role as a model selection criterion in different fields (8). It is identified as following:

$$AIC = -2\ln(\hat{L}) + 2p, \quad (28)$$

where p - number of estimated parameters of model, \hat{L} - maximum value of the likelihood function.

This evaluation criterion analyzes the model, parameters of which are estimated using MLE method. As we can see from the formula, it penalizes the models with high number of estimated

parameters and rewards the models that have high log-likelihood values. Overall, it is a fairly flexible method to capture the "badness" of the model.

BIC is also used very widely. This criterion evaluates the models from the Bayesian standpoint.

$$BIC = -2\ln(\hat{L}) + p \times \log(N), \quad (29)$$

where p - number of estimated parameters of model, \hat{L} - maximum value of the likelihood function, N - sample size.

Similarly to AIC, BIC is used as an information criterion to evaluate models estimated by the MLE method, given the sample size is sufficiently large (8). For sample sizes greater than 8, BIC penalizes the model more compared to AIC. However, the overall idea of penalizing the models with high number of estimated parameters and rewarding the ones with high log-likelihood values stays the same.

When comparing different models, the model with the smallest values of AIC and BIC is considered to be better than others.

3 Application

First, we will clean the data by removing dates with incomplete closing price information. Next, we will transform the series of daily closing prices using log differences—a common transformation for financial data. As a result, we will have $Y_t = \log P_t - \log P_{t-1} = \log \frac{P_t}{P_{t-1}}$, where P_t represents the closing price on day t (16), (1), (4). Then, we will perform basic descriptive statistics for the log-returns of the investments, including measures of central tendency, dispersion, kurtosis, and skewness. As previously mentioned, financial data often exhibit skewed distributions and fat tails (1). Additionally, we will compute Pearson's linear correlation and Kendall's tau values for the four investments in the portfolio. We will plot the returns data for all four investments and examine potential volatility clusters, investigating the causes of these clusters. We will also conduct normality and unit-root tests, such as the Shapiro-Wilk and Augmented Dickey Fuller tests, to determine if the return series follows a normal distribution and are stationary. Furthermore, we will examine ACF and PACF plots for all investments in the portfolio and perform the Ljung-Box test to assess data autocorrelation, which helps detect volatility clusters and shows the potential impact of today's movement on future movements (4). The null hypothesis for this test is that the data exhibits no serial autocorrelation. As is customary with formal tests, we will base our judgments on the p-values of this test.

If the data is not independently and identically distributed, we will fit it into a time series model. Here, we will fit the return series to ARIMA models if there is a serial correlation in the returns and insignificant serial correlation in the squared returns. If a significant serial correlation exists in the squared observations, we will apply GARCH models to capture data volatility. We will select the most suitable model based on the Akaike Information Criterion and Bayesian Information Criterion, with the smallest value representing the most appropriate model. We will use ACF and PACF correlograms along with the Ljung-Box test once more to evaluate the model's appropriateness and assess whether the time-series models have improved serial correlation in residuals and square residuals. Next, we will compute the 4-dimensional copula parameters using the rank-based Maximum Pseudo-Likelihood (MPL) estimation procedure (9), employing pseudo-observations of standardized residuals to estimate the marginals and copula parameters. This procedure is also widely used in other research (16). We will then fit copulas to the time series and choose the most suitable one based on the goodness-of-fit test of Cramer Von Mises (S_n) and tail dependence coefficients (9). The Cramer Von Mises test

measures how closely the empirical copula matches the fitted copula. Models with smaller S_n values or higher p-values are considered more appropriate.

3.1 Data Descripton

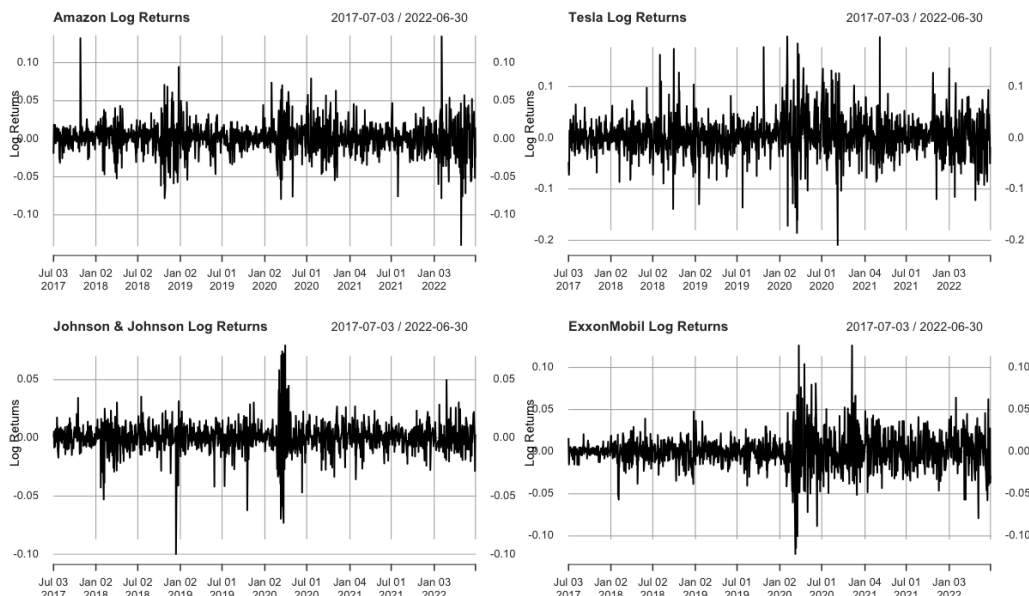


Figure 2: Log-returns

Yahoo Finance data for Closing Prices of Amazon, Tesla, Johnson&Johnson, and Exxon-Mobil for the period between 07-01-2017 and 07-01-2022 is used. Data was cleared and log returns of the closing prices were calculated. Log Returns can be seen in Figure 2.

Table 1 contains some basic statistical information about log-returns of each company.

Company	Mean	Variance	Skewness	Kurtosis
Amazon	0.000173031	0.000487055	-0.363357	5.42749
Tesla	0.00357789	0.00196852	-0.272662	3.81645
JNJ	0.000334513	0.000195966	0.0173411	7.66999
XOM	0.000177193	0.000613054	-0.16025	4.04234

One can observe that the log returns for the companies exhibit a positive mean. The log returns of JNJ are right-tailed, whereas those of the other companies are left-tailed. All the companies display substantial kurtosis, which strongly suggests the non-normality of the log returns. The small p-values obtained from the Jarque-Bera test for the normality of the log returns corroborate this observation. The Kolmogorov-Smirnov test also yields similar findings. According to the Kolmogorov-Smirnov test, if the p-value is below 0.05, the distribution is likely not Gaussian. The p-values for Amazon, Tesla, Johnson&Johnson, and ExxonMobil are 0.0002449, 0.0001795, 0.0000006, and 0.00406235, respectively.

Amazon, Tesla, Johnson&Johnson, and ExxonMobil have the following correlation coefficients:

In this project, an equally weighted portfolio will be constructed, assigning a weight of 0.25 to each asset. The correlation coefficients are relatively low since the portfolio has been designed to select a combination of the four smallest correlation coefficients among the top 15 S&P companies.

Table 3: Correlation coefficients of companies

Company 1	Company 2	Correlation coefficient
AMZN	TSLA	0.4128
AMZN	JNJ	0.2701
AMZN	XOM	0.2120
TSLA	JNJ	0.1387
TSLA	XOM	0.2054
JNJ	XOM	0.3425

3.2 Calculating VAR using historical data

Value at risk is one of the most widely used risk measures in finance. VaR was popularized by J.P. Morgan in the 1990s. The executives at J.P. Morgan wanted their risk managers to generate one statistic that summarized the risk of the firm’s entire portfolio, at the end of each day. What they came up with was VaR.

In order to formally define VaR, we begin by defining a random variable L , which represents the loss to our portfolio. L is simply the opposite of the return to our portfolio. If the return of our portfolio is \$600, then the loss, L , is +\$600 (12).

Monte Carlo simulations are widely used throughout finance, and they can be a very powerful tool for calculating VaR. Imagine that we have N random variables, X_1, X_2, \dots, X_N , representing the returns of different stocks. In order to describe the relationships between each of the variables, we could form an $N \times N$ covariance matrix, where each element, $\sigma_{i,j}$ corresponds to the covariance between the i th and j th random variables (13):

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \dots & \sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \sigma_{n,1} & \dots & \sigma_{n,n} \end{bmatrix} \quad (30)$$

If the covariance matrix satisfies certain minimum requirements, we can decompose the covariance matrix, rewriting it in terms of a lower triangular matrix, L , and its transpose, L' , which is an upper triangular matrix. This is what is known as a Cholesky decomposition [3].

$$\Sigma = LL' \quad (31)$$

Monte Carlo simulation happens by taking L part of the Cholesky decomposition and multiplying it by a vector of i.i.d standard normal variables. By comparison with historical and parametric Value at Risks, Monte Carlo simulations var are generally considered to be the most flexible (12).

90% 99% and 99.5% confidence level VARs were calculated using different methods for portfolios using data from 2017-07-01 to 2022-07-01. Predictions were made with 100000 simulations.

Table 4: Historical VARs using different methods

Portfolio	Historical Var	Normal Var	t-dist Var	MC Var
90%	1.172986	1.313834	1.314509	1.218557
99%	2.494242	2.359622	2.362600	2.38972
99.5%	2.881441	2.609340	2.613267	2.430796

3.3 Determining Marginal Distributions

To continue observing the nature of the series, Augmented Dickey-Fuller (ADF) test also needs to be conducted. As usual, the results are analyzed using the p-value, where a p-value below 0.05 suggests us to reject H_0 , and a p-value greater than 0.05 suggests we fail to reject H_0 . We fail to reject the Null hypothesis for all investments in the portfolio, since all of the p-values are greater than 0.05. This means that all time series seem to have a unit root and appear to be non-stationary.

Additionally, autocorrelation and partial autocorrelation functions are used to check the correlation of series. According to ACF plots, the returns have some serial autocorrelation. Additionally, high serial correlation in the squared returns signifies the existence of volatility clusters. This is indeed true and agrees with our analysis of volatility clusters above. Ljung-Box test with a lag of 20 is also conducted to check the autocorrelation of time series. The p-values are 0.005914 for AMZN, 0.08386 for TSLA, 4.411e-16 for JNJ, and 7.083e-6 for XOM.

So, returns seem to be non-normal, and non-stationary and have serial autocorrelation in both ordinary returns and squared returns. As a result, ARIMA-GARCH models are used to ensure that the volatility of data is captured. After fitting the most appropriate ARIMA-GARCH model, we should plot ACF and PACF correlograms to visually inspect the usefulness of ARIMA-GARCH models in making residuals serially independent.

Autocorrelation seems to be largely reduced for all investments after the use of ARIMA-GARCH models. The chosen ARIMA-GARCH models for every investment in the portfolio based on the lowest values of AIC and the Ljung-Box test's p-values are summarized in Table IV. Using those correlograms and Ljung-Box p-values, it seems that at 95% confidence level, the residuals for all four investments are serially independent.

Table 5: Companies and chosen ARIMA(p,d,q)-GARCH(m,n) models based on AIC criterion

Company	ARIMA-GARCH model type	Ljung-Box p-values
AMZN	ARIMA(3,0,3)-GARCH(1,1)	0.7382
TSLA	ARIMA(2,0,2)-GARCH(1,1)	0.4967
JNJ	ARIMA(3,0,4)-GARCH(1,1)	0.8632
XOM	ARIMA(2,0,3)-GARCH(1,1)	0.07914

3.4 Choosing and fitting copula

Table 6: Copulas and parameter estimates

Copula	Parameter	Log-likelihood	Goodness of fit p-values	AIC	BIC
Gumbel	1.063	24.8369	0.19508	-47.67381	-42.53812
Normal	0.1217	47.23343	0.16723	-92.46686	-87.33117
Clayton	0.1493	56.53083	0.22935	-111.0617	-105.926

From the Table V it looks as if all copulas are quite good fits for the transformed data series. Clayton copula shows the best values of AIC, BIC and goodness of fit test's p-values, so it is used further to calculate VaR, as the best-fit copula. Based on researches, Clayton copula is the most suitable choice for dependence structure in cases of any financial turmoils (1) and the fact that it has shown to be the most appropriate in this time-period corresponds to the economy's condition in 2020 quite well.

3.5 Estimating VaR

Monte-Carlo simulations, where losses from the adjusted copula are randomly generated, are used in this section. Additionally, marginal distribution functions are estimated. Here, the normal and t-student distribution functions are fitted. VaR is estimated with both of those marginals and then, the results are compared.

As a result, Clayton Copula with 0.1493 parameters is used with both t-student and normal marginal distributions. The estimations of VaR for the equally weighted portfolio using Clayton copula with different marginals and historical simulations are summarized in Table VI.

Table 7: Value-at-Risk

Model	Marginal	90%	99%	99.5%
Clayton Copula	t-student	0.9124409	1.6913028	1.8667663
Clayton Copula	normal	1.227861	2.251884	2.482578
Historical		1.172986	2.494242	2.881441

Overall, it seems that Clayton copula with normal distribution marginals shows the highest VaR values and VaR calculated from the normal marginals has the lowest VaR values. Due to both of those VaR estimations having very strong assumptions, it seems that Clayton copula with normal marginals is the best fit. It has a higher VAR for 90% confidence level but for 99% and 99.5% confidence levels VARs are slightly lower. In general, there is little difference between Historical and Clayton Copula VARS.

Conclusion

This study aimed to estimate the Value at Risk (VaR) for an equally weighted portfolio consisting of Amazon, Tesla, Johnson&Johnson, and ExxonMobil stocks. The data analysis revealed that the log returns for the companies exhibit positive means, non-normality, and non-stationarity. ARIMA-GARCH models were employed using AIC criteria to choose the best fit to capture the volatility of the data and to make the residuals serially independent.

In order to model the dependence structure among the portfolio assets, copulas were utilized. The Clayton copula, with the best AIC, BIC, and goodness of fit test p-values, was chosen as the most appropriate copula for this case. VaR estimation was conducted using Monte Carlo simulations, and both t-student and normal marginal distributions were fitted. The results showed that the Clayton copula with normal marginal distributions provided the closest to historical VaR values.

In conclusion, the Clayton copula with normal marginals was considered the best fit for estimating VaR in the given portfolio. The VaR values at 90%, 99%, and 99.5% confidence levels were 1.227861, 2.251884, and 2.482578, respectively. The study provides valuable insights into the risk assessment of the selected portfolio, and the methodology can be extended to other portfolios and assets for risk management purposes.

Further Investigations

In order to improve the understanding of the risk characteristics of the chosen equally weighted portfolio, several further investigations can be conducted. This section outlines potential extensions to the current research project.

Extending Copula Dimensions: The current project uses a four-dimensional copula to model the dependence structure of the assets in the portfolio. It is possible to extend

the dimensions of the copula to include more assets in the analysis, which would provide a better understanding of the portfolio's overall risk and dependence structure. Including a larger number of assets would enable a more comprehensive analysis of the diversification benefits and potential tail risks.

Exploring Other Copula Types: In this project, Clayton, Gumbel, and Normal copulas were tested, and the Clayton copula was chosen as the best fit. However, other copula types, such as the Student's t copula, Frank copula, and Joe copula, could also be considered to better understand the dependence structure of the assets in the portfolio. Different copulas may better capture the dependence structure, particularly during periods of market stress.

Additional Value-at-Risk (VaR) Calculation Methods: The current project uses several methods for calculating VaR, including historical, normal, t-distribution, and Monte Carlo simulations. Additional methods, such as the Filtered Historical Simulation, Extreme Value Theory (EVT), and Conditional Value-at-Risk (CVaR), could also be employed to provide a more comprehensive understanding of the portfolio's risk profile. Also, as it was performed in the recent article about Copula ARMA-GARCH by Siroos Shahriari, S.A. Sisson, and Taha Rashidi, fitted ARMA-GARCH models can be compared to the mean absolute percentage error (MAPE) and root mean square error (RMSE) values of model predictions on synthesised data to demonstrate the goodness of fit (17)

Incorporating Conditional Value-at-Risk (CVaR): CVaR, also known as Expected Shortfall (ES), measures the expected loss in the worst-case scenarios beyond the VaR threshold. Incorporating CVaR analysis would provide a more accurate estimation of potential losses during extreme events and help to better understand the tail risks of the portfolio.

Extending the Time Period: The current project examines a 5-year time period for the selected assets. To better understand the risk and dependence structure of the portfolio, it is possible to extend the time period under analysis. Investigating a longer time period could provide valuable insights into how the relationships between the assets and the overall risk of the portfolio change over time. This would also help to capture any structural changes in the market, evolving market conditions, and the impact of various economic events on the portfolio's risk profile.

By extending the dimensions of the copula, exploring other copula types, employing additional VaR calculation methods, and incorporating CVaR, the project could provide a more comprehensive understanding of the portfolio's risk characteristics and aid in the development of more robust risk management strategies.

Acknowledgements

I cannot express enough appreciation for my capstone supervisor, Professor Dongming Wei, who not only granted me the opportunity to work on this exciting project but also offered priceless advice and guidance. My heartfelt gratitude goes to Professor Rustem Takhanov, who generously accepted the role of the second reader for my capstone. I must also acknowledge Professor Tolga Etgu for the encouragement and support he provided, attending our weekly presentations and motivating us to push ourselves.

References

- [1] R. Ab Razak, "Portfolio risks of bivariate financial returns using copula-var approach: A case study on malaysia and u.s. stock markets," *Global Journal of Pure and Applied Mathematics*, 2016.

- [2] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th ed. Wiley, 2008.
- [3] J. Brownlee, *Introduction to time series forecasting with Python: how to prepare data and develop models to predict the future*. Jason Brownlee, 2017.
- [4] N. H. Chan and H. Y. Wong, *Handbook of financial risk management simulations and case studies*. Wiley, 2013.
- [5] U. Cherubini, E. Luciano, and W. Vecchiato, *Copula methods in finance*. John Wiley & Sons, 2004.
- [6] H. Joe, *Dependence modeling with copulas*. CRC Press, 2014.
- [7] P. Jorion and P. Jorion, *Value at risk: the new benchmark for managing financial risk*. McGraw-Hill, 2009.
- [8] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer, 2008.
- [9] S. Latif and F. Mustafa, “Trivariate distribution modelling of flood characteristics using copula function—a case study for kelantan river basin in malaysia,” *AIMS Geosciences*, 2020.
- [10] C. F. Lee, A. C. Lee, and J. C. Lee, *Handbook of quantitative finance and risk management*. Springer, 2010.
- [11] P. S. C. Marcelo Brutti Righi, “Analyzing the dependence structure of various sectors in the brazilian market: A pair copula construction approach,” *Economic Modelling*, vol. 35, pp. 199–206, 2013.
- [12] A. J. McNeil, F. Rüdiger, and P. Embrechts, *Quantitative risk management: concepts, techniques and tools*. Princeton University Press, 2015.
- [13] M. B. Miller, *Mathematics and Statistics for Financial Risk Management*. John Wiley & Sons, 2012.
- [14] R. B. Nelsen, *An introduction to copulas*, 2nd ed. Springer, 2006.
- [15] ———, *An introduction to copulas*. Springer Science & Business Media, 2006.
- [16] M. Sampid and H. Hasim, “Estimating value-at-risk using a multivariate copula-based volatility model: Evidence from european banks,” *International Economics*, 2018.
- [17] T. R. Siroos Shahriari, S.A. Sisson, “Copula arma-garch modelling of spatially and temporally correlated time series data for transportation planning use,” *Transportation Research Part C: Emerging Technologies*, vol. 146, 2023.
- [18] Y. Zhu, X. Liu, and W. Zhang, “Deep learning in copula-based models for financial forecasting,” *Neural Networks*, vol. 141, pp. 251–267, 2021.

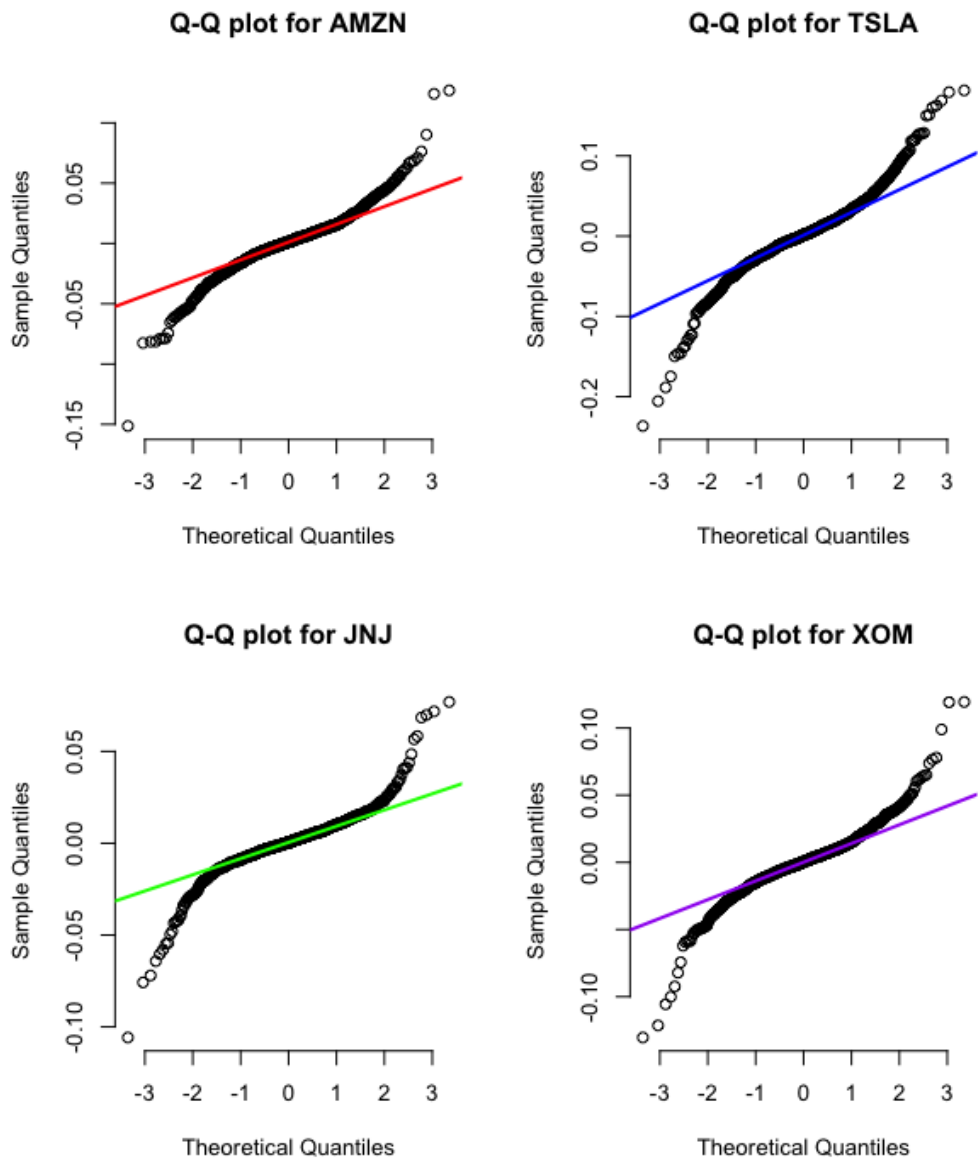


Figure 3: Log-returns

Appendix

Listing 1: My R code

```

library("readxl")
library("e1071")
library("car")
library("tseries")
library("copula")
library("VineCopula")
library("aTSA")
library("forecast")
library("rugarch")
library("gofCopula")
library("corrplot")
library("RColorBrewer")

```

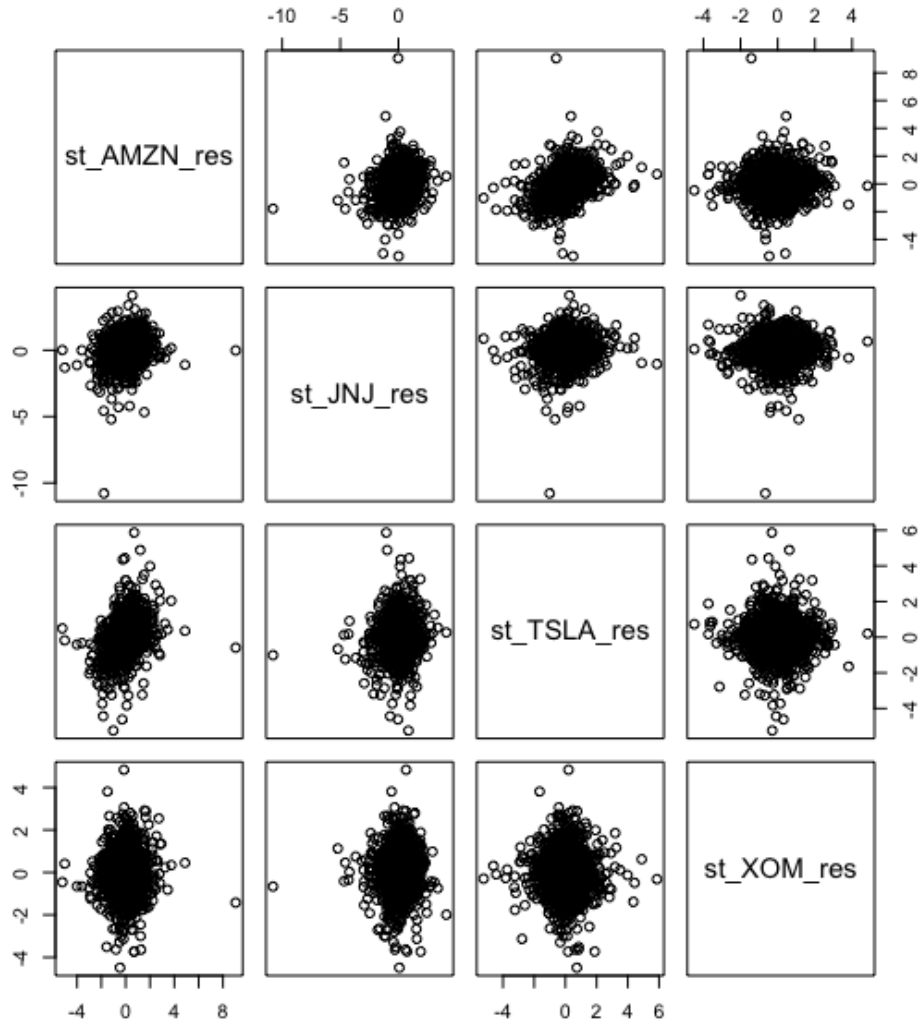


Figure 4: Log-returns

```

library("univariateML")
library("MASS")

data = read_excel("data2.xlsx")

sapply(data, mean, na.rm=TRUE)
logAMZN = diff(log(data$ClosingPrice_AMZN), lag=1)
logTSLA = diff(log(data$ClosingPrice_TSLA), lag=1)
logJNJ = diff(log(data$ClosingPrice_JNJ), lag=1)
logXOM = diff(log(data$ClosingPrice_XOM), lag=1)
summary(logAMZN)

#Shapiro Wilk formal test of normality
shapiro.test(logAMZN)
shapiro.test(logJNJ)
shapiro.test(logTSLA)
shapiro.test(logXOM)

```

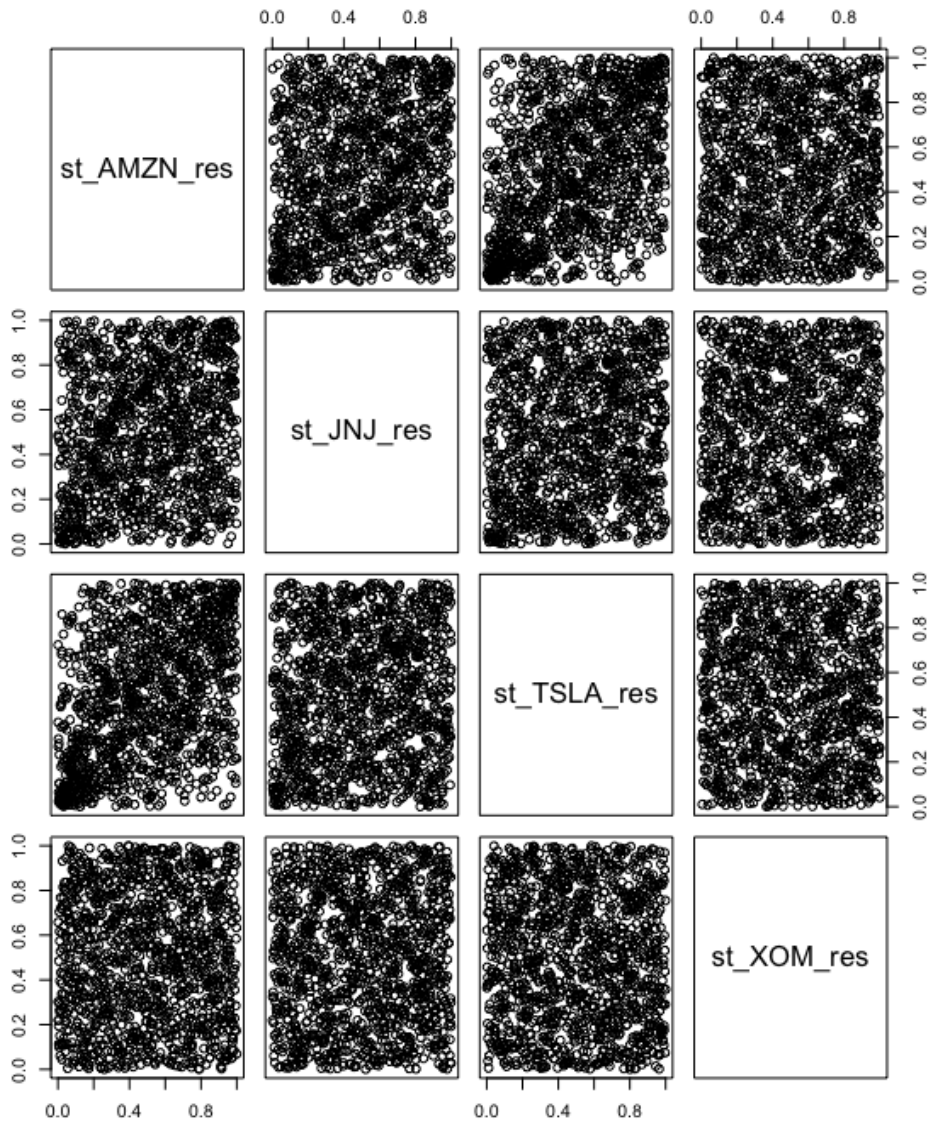


Figure 5: Log-returns

```

#Autocorrelation functions of returns and returns squared
par(mfrow=c(2,2))
acf(logAMZN, main="AMZN_returns")
acf(logAMZN^2, main="AMZN_returns_squared")
pacf(logAMZN, main="AMZN_returns")
pacf(logAMZN^2, main="AMZN_returns_squared")
par(mfrow=c(2,2))
acf(logTSLA, main="TSLA_returns")
acf(logTSLA^2, main="TSLA_returns_squared")
pacf(logTSLA, main="TSLA_returns")
pacf(logTSLA^2, main="TSLA_returns_squared")
par(mfrow=c(2,2))
acf(logJNJ, main="JNJ_returns")
acf(logJNJ^2, main="JNJ_returns_squared")

```

```

pacf(logJNJ, main="JNJ_returns")
pacf(logJNJ^2, main="JNJ_returns_squared")
par(mfrow=c(2,2))
acf(logXOM, main="XOM_returns")
acf(logXOM^2, main="XOM_returns_squared")
pacf(logXOM, main="XOM_returns")
pacf(logXOM^2, main="XOM_returns_squared")

#Box-Ljung test with 20 lags
Box.test(logAMZN, lag=20, type=c("Ljung-Box"))
Box.test(logTSLA, lag=20, type=c("Ljung-Box"))
Box.test(logJNJ, lag=20, type=c("Ljung-Box"))
Box.test(logXOM, lag=20, type=c("Ljung-Box"))

#Augmented-Dickey-Fuller
adf.test(logAMZN)
adf.test(logTSLA)
adf.test(logJNJ)
adf.test(logXOM)

#ARIMA model
AMZN = as.numeric(logAMZN)
TSLA = as.numeric(logTSLA)
JNJ = as.numeric(logJNJ)
XOM = as.numeric(logXOM)
AMZN = AMZN [!is.na(AMZN)]
TSLA = TSLA [!is.na(TSLA)]
JNJ = JNJ [!is.na(JNJ)]
XOM = XOM [!is.na(XOM)]
AMZN_f.aic = Inf
AMZN_f.order = c(0,0,0)
TSLA_f.aic = Inf
TSLA_f.order = c(0,0,0)
JNJ_f.aic = Inf
JNJ_f.order = c(0,0,0)
XOM_f.aic = Inf
XOM_f.order = c(0,0,0)
for (p in 1:4) for (d in 0:2) for (q in 1:4) {
  AMZN_curr.aic = AIC(arima(AMZN, order=c(p, d, q),
    optim.control=list(maxit=1000)))
  TSLA_curr.aic = AIC(arima(TSLA, order=c(p, d, q),
    optim.control=list(maxit=1000)))
  JNJ_curr.aic = AIC(arima(JNJ, order=c(p, d, q),
    optim.control=list(maxit=1000)))
  XOM_curr.aic = AIC(arima(XOM, order=c(p, d, q),
    optim.control=list(maxit=1000)))
  if (AMZN_curr.aic < AMZN_f.aic) {
    AMZN_f.aic = AMZN_curr.aic
    AMZN_f.order = c(p, d, q)
    AMZN_f.arima = arima(AMZN, order=AMZN_f.order)
  }
}

```

```

}
if (TSLA_curr.aic < TSLA_f.aic) {
  TSLA_f.aic = TSLA_curr.aic
  TSLA_f.order = c(p, d, q)
  TSLA_f.arima = arima(TSLA, order=TSLA_f.order)
}
if (JNJ_curr.aic < JNJ_f.aic) {
  JNJ_f.aic = JNJ_curr.aic
  JNJ_f.order = c(p, d, q)
  JNJ_f.arima = arima(JNJ, order=JNJ_f.order)
}
if (XOM_curr.aic < XOM_f.aic) {
  XOM_f.aic = XOM_curr.aic
  XOM_f.order = c(p, d, q)
  XOM_f.arima = arima(XOM, order=XOM_f.order)
} }
cat ("ARIMA_order_for_AMZN:", AMZN_f.order, "\n")
cat ("ARIMA_order_for_TSLA:", TSLA_f.order, "\n")
cat ("ARIMA_order_for_JNJ:", JNJ_f.order, "\n")
cat ("ARIMA_order_for_XOM:", XOM_f.order, "\n")

```

```

par(mfrow=c(2, 2))
acf(resid(AMZN_f.arima))
acf(resid(AMZN_f.arima)^2)
pacf(resid(AMZN_f.arima))
pacf(resid(AMZN_f.arima)^2)
par(mfrow=c(2, 2))
acf(resid(TSLA_f.arima))
acf(resid(TSLA_f.arima)^2)
pacf(resid(TSLA_f.arima))
pacf(resid(TSLA_f.arima)^2)
par(mfrow=c(2, 2))
acf(resid(JNJ_f.arima))
acf(resid(JNJ_f.arima)^2)
pacf(resid(JNJ_f.arima))
pacf(resid(JNJ_f.arima)^2)
par(mfrow=c(2, 2))
acf(resid(XOM_f.arima))
acf(resid(XOM_f.arima)^2)
pacf(resid(XOM_f.arima))
pacf(resid(XOM_f.arima)^2)

```

#GARCH

```

AMZN_garch = garch(AMZN, trace=FALSE)
AMZN_res = AMZN_garch$res[-1]
AMZN_res = AMZN_res[!is.na(AMZN_res)]
TSLA_garch = garch(TSLA, trace=FALSE)
TSLA_res = TSLA_garch$res[-1]
TSLA_res = TSLA_res[!is.na(TSLA_res)]
JNJ_garch = garch(JNJ, trace=FALSE)

```

```

JNJ_res = JNJ_garch$res[-1]
JNJ_res = JNJ_res[!is.na(JNJ_res)]
XOM_garch = garch(XOM, trace=FALSE)
XOM_res = XOM_garch$res[-1]
XOM_res = XOM_res[!is.na(XOM_res)]
summary(AMZN_garch)
summary(TSLA_garch)
summary(JNJ_garch)
summary(XOM_garch)

```

#ACF PACF after garch

```

par(mfrow=c(2, 2))
acf(AMZN_res)
acf(AMZN_res^2)
pacf(AMZN_res)
pacf(AMZN_res^2)
par(mfrow=c(2, 2))
acf(TSLA_res)
acf(TSLA_res^2)
pacf(TSLA_res)
pacf(TSLA_res^2)
par(mfrow=c(2, 2))
acf(JNJ_res)
acf(JNJ_res^2)
pacf(JNJ_res)
pacf(JNJ_res^2)
par(mfrow=c(2, 2))
acf(XOM_res)
acf(XOM_res^2)
pacf(XOM_res)
pacf(XOM_res^2)

```

#Standardized residuals with mean of 0, standard deviation of 1

```

sA = sd(AMZN_res)
sT = sd(TSLA_res)
sJ = sd(JNJ_res)
sX = sd(XOM_res)
st_AMZN_res = AMZN_res/sA
mean(st_AMZN_res)
sd(st_AMZN_res)
st_JNJ_res = JNJ_res/sJ
mean(st_JNJ_res)
sd(st_JNJ_res)
st_TSLA_res = TSLA_res/sT
mean(st_TSLA_res)
sd(st_TSLA_res)
st_XOM_res = XOM_res/sX
mean(st_XOM_res)
sd(st_XOM_res)
#Pseudo-observations

```

```

mg = cbind(st_AMZN_res , st_JNJ_res , st_TSLA_res , st_XOM_res )
set.seed(13)
pseudo = pobs(as.matrix(mg))
pairs(mg) #to plot and see the correlations of pairwise returns of assets
pairs(pseudo)

fitcop_norm = fitCopula(normalCopula(dim=4), data=pseudo, method="mpl")
summary(fitcop_norm)
tau(normalCopula(param=0.1207))
fitcop_g = fitCopula(gumbelCopula(dim=4), data=pseudo, method="mpl")
summary(fitcop_g)
tau(gumbelCopula(param=1.063))
fitcop_cl = fitCopula(claytonCopula(dim=4), data=pseudo, method="mpl")
summary(fitcop_cl)
tau(claytonCopula(param=0.1293))

cat("Log-likelihood_of_Gaussian_copula:", logLik(fitcop_norm), "\n")
cat("AIC_of_Gaussian_copula:", AIC(fitcop_norm), "\n")
cat("BIC_of_Gaussian_copula:", BIC(fitcop_norm), "\n")
cat("Log-likelihood_of_Gumbel_copula:", logLik(fitcop_g), "\n")
cat("AIC_of_Gumbel_copula:", AIC(fitcop_g), "\n")
cat("BIC_of_Gumbel_copula:", BIC(fitcop_g), "\n")
cat("Log-likelihood_of_Clayton_copula:", logLik(fitcop_cl), "\n")
cat("AIC_of_Clayton_copula:", AIC(fitcop_cl), "\n")
cat("BIC_of_Clayton_copula:", BIC(fitcop_cl), "\n")

#Greatness-of-fit test for copulas using parameters from above
gofCopula(normalCopula(param=0.1207, dim=4), as.matrix(mg), N=100, method="Sn")
gofCopula(claytonCopula(param=0.1293, dim=4), as.matrix(mg), N=100, method="Sn")
gofCopula(gumbelCopula(param=1.063, dim=4), as.matrix(mg), N=100, method="Sn")

#Marginal Normal Dists
mod_Gauss_AMZN = fitdistr(st_AMZN_res, "normal")
mod_Gauss_AMZN
cat("AIC_of_normal_marginal_distribution_of_AMZN:", AIC(mod_Gauss_AMZN), "\n")
cat("BIC_of_normal_marginal_distribution_of_AMZN:", BIC(mod_Gauss_AMZN), "\n")
mod_Gauss_TSLA = fitdistr(st_TSLA_res, "normal")
mod_Gauss_TSLA
cat("AIC_of_normal_marginal_distribution_of_TSLA:", AIC(mod_Gauss_TSLA), "\n")
cat("BIC_of_normal_marginal_distribution_of_TSLA:", BIC(mod_Gauss_TSLA), "\n")
mod_Gauss_JNJ = fitdistr(st_JNJ_res, "normal")
mod_Gauss_JNJ
cat("AIC_of_normal_marginal_distribution_of_JNJ:", AIC(mod_Gauss_JNJ), "\n")
cat("BIC_of_normal_marginal_distribution_of_JNJ:", BIC(mod_Gauss_JNJ), "\n")
mod_Gauss_XOM = fitdistr(st_XOM_res, "normal")
mod_Gauss_XOM
cat("AIC_of_normal_marginal_distribution_of_XOM:", AIC(mod_Gauss_XOM), "\n")
cat("BIC_of_normal_marginal_distribution_of_XOM:", BIC(mod_Gauss_XOM), "\n")

#Marginal t-student distributions

```



```

mod_t_AMZN = fitdistr(st_AMZN_res, "t")
mod_t_AMZN
cat("AIC_of_t-student_marginal_distribution_of_AMZN:", AIC(mod_t_AMZN) ,"\n")
cat("BIC_of_t-student_marginal_distribution_of_AMZN:", BIC(mod_t_AMZN) ,"\n")
mod_t_TSLA = fitdistr(st_TSLA_res, "t")
mod_t_TSLA
cat("AIC_of_t-student_marginal_distribution_of_TSLA:", AIC(mod_t_TSLA) ,"\n")
cat("BIC_of_t-student_marginal_distribution_of_TSLA:", BIC(mod_t_TSLA) ,"\n")
mod_t_JNJ = fitdistr(st_JNJ_res, "t")
mod_t_JNJ
cat("AIC_of_t-student_marginal_distribution_of_JNJ:", AIC(mod_t_JNJ) ,"\n")
cat("BIC_of_t-student_marginal_distribution_of_JNJ:", BIC(mod_t_JNJ) ,"\n")
mod_t_XOM = fitdistr(st_XOM_res, "t")
mod_t_XOM
cat("AIC_of_t-student_marginal_distribution_of_XOM:", AIC(mod_t_XOM) ,"\n")
cat("BIC_of_t-student_marginal_distribution_of_XOM:", BIC(mod_t_XOM) ,"\n")

#simulation of copulas
r = 10000
norm_cop_est = normalCopula(0.1207, dim=4)
sim_val_normc = rCopula(r, norm_cop_est)
gumb_cop_est = gumbelCopula(1.063, dim=4)
sim_val_gumbelc = rCopula(r, gumb_cop_est)
clay_cop_est = claytonCopula(0.1293, dim=4)
sim_val_claytonc = rCopula(r, clay_cop_est)

#with normal marginals
sim_val_AMZN = qnorm(sim_val_claytonc[,1], mean=mod_Gauss_AMZN$estimate[[1]],
                    sd=mod_Gauss_AMZN$estimate[[2]])
sim_val_TSLA = qnorm(sim_val_claytonc[,1], mean=mod_Gauss_TSLA$estimate[[1]],
                    sd=mod_Gauss_TSLA$estimate[[2]])
sim_val_JNJ = qnorm(sim_val_claytonc[,1], mean=mod_Gauss_JNJ$estimate[[1]],
                    sd=mod_Gauss_JNJ$estimate[[2]])
sim_val_XOM = qnorm(sim_val_claytonc[,1], mean=mod_Gauss_XOM$estimate[[1]],
                    sd=mod_Gauss_XOM$estimate[[2]])

#with t margins
sim_val_AMZN_t = qnorm(sim_val_claytonc[,1], mean=mod_t_AMZN$estimate[[1]],
                      sd=mod_t_AMZN$estimate[[2]])
sim_val_TSLA_t = qnorm(sim_val_claytonc[,1], mean=mod_t_TSLA$estimate[[1]],
                      sd=mod_t_TSLA$estimate[[2]])
sim_val_JNJ_t = qnorm(sim_val_claytonc[,1], mean=mod_t_JNJ$estimate[[1]],
                      sd=mod_t_JNJ$estimate[[2]])
sim_val_XOM_t = qnorm(sim_val_claytonc[,1], mean=mod_t_XOM$estimate[[1]],
                      sd=mod_t_XOM$estimate[[2]])

w = cbind(0.25,0.25,0.25,0.25)
alpha = c(0.9, 0.99,0.995)
MC_data = cbind(sim_val_AMZN, sim_val_TSLA, sim_val_JNJ, sim_val_XOM)

```

```

MC_Lsim = -(MC_data%*%t(w))
MC_data_t = cbind(sim_val_AMZN_t, sim_val_TSLA_t, sim_val_JNJ_t, sim_val_XOM_t)
MC_Lsim_t = -(MC_data_t%*%t(w))
quantile(MC_Lsim, alpha)
quantile(MC_Lsim_t, alpha)
#historical simulations
quantile(mg, alpha, method="historical")
normal_var <- mean(mg) - sd(mg) * qnorm(1 - alpha)
t_degrees_freedom <- df.residual(lm(mg ~ 1))
t_var <- mean(mg) - sd(mg) * qt(1 - alpha, df = t_degrees_freedom)

#Visualizations
par(mfrow=c(2, 2))
plot(st_AMZN_res, type="l", main="Standardized_Residuals_of_AMZN_GARCH_Model")
plot(st_TSLA_res, type="l", main="Standardized_Residuals_of_TSLA_GARCH_Model")
plot(st_JNJ_res, type="l", main="Standardized_Residuals_of_JNJ_GARCH_Model",
plot(st_XOM_res, type="l", main="Standardized_Residuals_of_XOM_GARCH_Model",

cor_matrix <- cor(mg)
corrplot(cor_matrix, method="circle", tl.col="black", title="Correlation_Mat

# Load required libraries
library(ggplot2)

# Create a new data frame with dates and closing prices
closing_prices <- data.frame(
  Date = data$Date,
  AMZN = data$ClosingPrice_AMZN,
  TSLA = data$ClosingPrice_TSLA,
  JNJ = data$ClosingPrice_JNJ,
  XOM = data$ClosingPrice_XOM
)

# Reshape the data frame into long format for ggplot2
library(tidyr)
closing_prices_long <- gather(closing_prices, key = "Stock", value = "Closing

# Create the plot with ggplot2
p <- ggplot(data = closing_prices_long, aes(x = Date, y = ClosingPrice, color
  geom_line() +
  labs(title = "Stock_Closing_Prices", x = "Date", y = "Closing_Price") +
  theme_minimal()

# Display the plot
print(p)

```