Statistical Arbitrage using Pairs Trading

Abstract

Statistical arbitrage has played a pivotal role in the evolution of modern financial markets, particularly with the advent of electronic trading. Originally pioneered by quantitative hedge funds in the late 1980s, statistical arbitrage relies on complex mathematical models and computational algorithms to exploit short-term mispricing between related securities. The technique saw widespread success during the 1990s, with firms utilizing high-frequency trading to capture small profits at scale. Over time, it has evolved into a key strategy for hedge funds and institutional investors, particularly through approaches such as pairs trading, which seeks to profit from the mean-reverting behavior of two historically correlated assets.

This project builds on that foundation by applying pairs trading as a form of statistical arbitrage. In the initial phase of the project, stock data from multiple equities was imported and cleaned to ensure its suitability for analysis. The data was then analyzed using the Spearman correlation coefficient, a non-parametric measure that identifies pairs of stocks in the same sectors with a historical tendency to move in tandem. This method was selected for its ability to capture monotonic relationships between assets, providing a robust foundation for identifying potential pairs suitable for mean-reversion strategies.

The next stage of the project involves fitting the marginal distributions of these pairs to identify the most appropriate models for each stock's behavior. By fitting these distributions, the goal is to extract key parameters that can then be used to select appropriate copulas, which will model the joint dependency structure of the pairs. Copulas allow for a more flexible and sophisticated understanding of how pairs move together compared to linear correlation alone. By applying conditions to the probabilities generated by these copulas, the strategy aims to take advantage of statistical inefficiencies in the market. Ultimately, the project seeks to exploit these inefficiencies through the mean-reverting nature of stock pairs, generating profit by taking positions when the price divergence exceeds historical norms and converges back to equilibrium.

Table of Contents

Contents

1. Introduction **1**

2. Literature Review **3**

2.1 Distance method……...4

2.2 Cointegration method……...5

2.3 Time series method……..8

2.4 Other approaches……10

2.5 Copula approach……13

3. Methodology **15**

3.1 Data Extraction…….....15

3.2 Data Cleaning……15

3.3 Conversion to log returns……...16

3.4 Finding Pairs……...16

3.5 Choosing Pairs……....17

3.6 Fitting marginal distributions to pair……..18

3.7 Testing for statistical significance of fitted distributions……...19

3.8 Fitting copulas to distributions……...20

3.9 Trading strategy……..22

4. Results and Discussion **24**

4.1 Financial Services (ABCAPITAL and CANFINHOME)……...24

4.2 Information Technology (TCS and INFY)…….....27

4.3 Oil, Natural Gas and other consumable fuels (OIL and ONGC).....29

4.4 Fast moving consumer goods (NESTLEIND and DABUR).....32

4.5 Automobile and Auto components (CEATLTD and MRF).....35

4.6 Healthcare (CIPLA and GLENAMRK).....37

5. Conclusion **41**

6. References **42**

List of figures

Figure 1: Q-Q Plot of ABCAPITAL and CANFINHOME.....24
Figure 2: Distribution fit of CANFINHOME.....25
Figure 3: Distribution of ABCAPITAL.....25
Figure 4: Returns of the algorithm on finance sector.....26
Figure 5: Q-Q Plot of TCS and INFY.....27
Figure 6: Distribution fit of INFY.....28
Figure 7: Distribution fit of TCS.....28
Figure 8: Results of the algorithm on IT sector.....29
Figure 9: Q-Q Plot of OIL and ONGC.....29
Figure 10: Distribution fit of OIL.....30
Figure 11: Distribution fit of ONGC.....31
Figure 12: Results of the algorithm on the Oil sector.....31
Figure 13: Q-Q Plot of NESTLEIND and DABUR.....32
Figure 14: Distribution fit of NESTLEIND.....33
Figure 15: Distribution fit of DABUR.....33
Figure 16: Results of the algorithm on the FMCG sector.....34
Figure 17: Q-Q Plot of CEATLTD and MRF.....35
Figure 18: Distribution fit of CEATLTD.....36
Figure 19: Distribution fit of MRF.....36
Figure 20: Results of the algorithm on the Auto sector.....37
Figure 21: Q-Q Plot of CIPLA and GLENMARK.....37
Figure 22: Distribution fit of CIPLA.....38
Figure 23: Distribution fit of GLENMARK.....39
Figure 24: Results of the algorithm on the healthcare sector.....39

List of tables

Table 1: Weights of industries in the NIFTY 500 index.....17
Table 2: Distribution fit parameters of the finance sector.....25
Table 3: Copula goodness-of-fit parameters for finance sector.....26
Table 4: Detailed result of the finance sector.....26
Table 5: Distribution fit parameters of the IT sector.....27
Table 6: Copula goodness-of-fit parameters for IT sector.....28
Table 7: Detailed result of the IT sector.....29
Table 8: Distribution fit parameters of the Oil sector.....30
Table 9: Copula goodness-of-fit parameters for Oil sector.....31
Table 10: Detailed result of the Oil sector.....32
Table 11: Distribution fit parameters of the FMCG sector.....33
Table 12: Copula goodness-of-fit parameters for FMCG sector.....34
Table 13: Detailed result of the FMCG sector.....34
Table 14: Distribution fit parameters of the Auto sector.....35
Table 15: Copula goodness-of-fit parameters for Auto sector.....36
Table 16: Detailed result of the Auto sector.....37
Table 17: Distribution fit parameters of the healthcare sector.....38
Table 18: Copula goodness-of-fit parameters for healthcare sector.....39
Table 19: Detailed result of the healthcare sector.....40

Chapter 1
Introduction

The financial markets, despite their sophistication and the advances in technology, remain imperfect and subject to fundamental inefficiencies. These inefficiencies arise from a variety of factors, such as liquidity imbalances, delayed reactions to information, and psychological biases of investors. Even in an age dominated by high-frequency trading and automated systems, market participants continue to overreact or underreact to news, creating short-term mispricing. Statistical arbitrage, particularly through pairs trading, seeks to capitalize on these inefficiencies by identifying predictable patterns in the relationships between assets. This strategy assumes that while individual stock prices may deviate in the short term, they will revert to a mean over time, offering a profitable opportunity to those who can effectively model and predict these movements.

Traditional technical analysis, which involves the study of price charts, patterns, and historical volume to predict future market behavior, has become less effective in today's complex and highly automated financial environment. Markets have grown faster, more efficient, and are increasingly dominated by institutional players using advanced quantitative techniques. The advent of algorithmic and high-frequency trading has diminished the edge that individual traders or even traditional investors might gain from relying on simple chart patterns or historical price trends. The fast-paced nature of modern trading has led to a situation where technical indicators are often lagging, resulting in missed opportunities or delayed reactions to significant market movements. In contrast, statistical arbitrage leverages mathematical models and real-time data analysis, offering a more sophisticated and adaptive method to profit from short-term market fluctuations.

Pairs trading, as a form of statistical arbitrage, provides an elegant solution to this challenge by focusing on the relative value between two correlated stocks rather than attempting to predict the direction of an individual asset. By identifying pairs of stocks that exhibit strong historical relationships, traders can construct strategies that rely on these stocks reverting to their mean relationship after deviating due to market noise or temporary inefficiencies. This approach is particularly powerful in volatile or uncertain markets, where directional trades become riskier and harder to predict. In such scenarios, the relative pricing of pairs offers a lower-risk alternative, as the expectation is that the spread between the two will narrow, regardless of the broader market movements.

By modeling the joint distribution of asset pairs using copulas, this strategy enables the generation of more robust trading signals that account for asymmetries and non-linear relationships. The incorporation of copulas allows for precise entry and exit thresholds based on conditional probabilities, leading to a better balance between profitability and risk exposure. As a result, this method not only enhances profitability in favorable market conditions but also mitigates drawdowns during adverse scenarios, making it a compelling choice for our analysis.

Chapter 2

Literature Review

The review paper of Krauss [1] classifies the various methods of pairs trading into 5 different approaches as stated below:

1. **Distance approach**: This strategy is the most thoroughly studied framework for trading pairs. Distance metrics are used to find securities that move together throughout the formation period. Simple threshold criteria are utilized to initiate trading signals during the trading period. This strategy's main advantages are its transparency and ease of use, which enable extensive empirical applications. The primary conclusions show that distance pairs trading is lucrative in a variety of markets, assets and time intervals.
2. **Cointegration approach**: Cointegration tests are used in this case to find co-moving securities during the formation phase. During the trading session, trade signals are generated by basic algorithms, most of which are based on the GGR threshold rule. The equilibrium relationship of recognized pairings is econometrically more reliable, which is the main advantage of these tactics.
3. **Time series approach**: Generally, the formation phase is disregarded in the time series method. Every author in this field works under the assumption that earlier investigations have created a set of co-moving securities. Rather, they concentrate on the trading window and how various techniques for time series analysis, such as representing the spread as a mean-reverting process, might produce optimal trading signals.
4. **Stochastic control approach**: The formation period is disregarded, much like in the time series approach. The goal of this method is to determine, in relation to other accessible assets, the ideal portfolio constituents for the 2 assets in a pairs trade. For this portfolio problem, value and optimal policy functions are found using stochastic control theory.

5. **Other approaches**: This bucket includes additional pairs trading frameworks that have little connection to previously discussed methods and a small body of supporting literature. The Principal Components Analysis (PCA), copula, and machine learning and integrated predictions approaches are all included in this area.

2.1 Distance Approach

The earliest reference to the strategy of pairs trading is found in the paper of Gatev et al. [2] and this paper introduces the concept of pairs in a simple 2-step process. Initially, identify two stocks whose prices have historically moved in tandem throughout a formation period. Second, keep an eye on their spread throughout the next trading session. Short the winner and purchase the loser if the prices diverge and the spread expands. The spread will return to its historical mean if there is an equilibrium relationship between the two stocks. After that, a profit is possible when the positions are switched.

To elaborate on the method of distance of Gatev et al. [2], it is performed on the data from the CRSP daily files from 1962 to 2002. First, a cumulative total return index P_{it} is constructed for each stock i and normalized to the first day of a 12 months formation period. Second, with n stocks under consideration, the sum of Euclidean squared distance (SSD) for the price time series of n*(n−1)/2 possible combinations of pairs is calculated. The top 20 pairs with minimum historic distance metric are considered in a subsequent six months trading period. Prices are normalized to the opening day of the trading session. trades are closed at mean reversion, the end of the trading period, or upon delisting of the given asset. Trades are opened when the spread diverges by more than $2 * \sigma$ (standard deviation of the spread).

The reasons for which this method is not used in this project:

- The choice of Euclidean squared distance for identifying pairs is analytically suboptimal.
- **Spread Variance**: Let V(.) represent the sample variance and P_{it} and P_{jt} the normalized price time series of the stocks i and j in a pair. Thus, we can express empirical spread variance $V(P_{it} - P_{jt})$ as follows:

$$V(P_{it} - P_{jt}) = \frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt})^2 - \left(\frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt})\right)^2$$

The paper solves for the average sum of squared distances for the formation period as follows:

$$\overline{SSD_{ijt}} = \frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt})^2 = V(P_{it} - P_{jt}) + \left(\frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt})\right)^2$$

In the above equation, we can see that for the Euclidean distance to be minimum between the 2 stocks the spread needs to be 0. On one hand, we make profit with the help of the spread between the 2 stocks in a pair (the basic definition of pairs trading) and on the other hand we aim to minimize the spread by using the above metric.

2.2 Cointegration Approach

The most cited work for this method belongs to Vidyamurthy [3]. He developed a univariate cointegration approach for trading pairs as a theoretical framework without practical applications.

Three crucial steps form the basis of the framework:

1. Pre-selection of possibly cointegrated couples using basic or statistical data.
2. Tradability testing using a proprietary methodology.
3. Trade rule formulation using non-parametric techniques.

However, the idea behind his method is similar to the idea of cointegrated pairs given by Do et al. [4] and Puspaningrum [5].

Preselection of pairs: This study splits the log price P_{it} of a security i into a stationary, idiosyncratic component ε_{it} and a nonstationary, common trends component n_{it} by utilizing Stock and Watson's common trends model (CTM) [6]. Similarly, its return r_{it} is made up of a particular return r_{it}^{s} and common trends return r_{it}^{c}.

Examine a portfolio that consists of one long unit of the security i and one short unit of security j. The spread m_{ijt} between the two securities represents the portfolio price time series, and the first difference Δm_{ijt} represents the return time series.

$$m_{ijt} = p_{it} - \gamma p_{jt} = n_{it} - \gamma n_{jt} + \epsilon_{it} - \gamma \epsilon_{jt}$$

$$\Delta m_{ijt} = r_{ijt} = r_{it}^{c} - \gamma r_{jt}^{c} + r_{it}^{s} - \gamma r_{jt}^{s}$$

The common return components of this pair must be equal up to the value of γ, the cointegration coefficient, for it to be cointegrated. After that, the spread time series becomes stationary as they cancel each other out. The Arbitrage Pricing Theory (APT) of Ross [7] is utilized in this work to find equities that have comparable common return components. The return on stock i can be stated using APT as an orthogonal statistical factor model as follows (Tsay [8]):

$$r_{it} - \mu_i = \beta_i' f_t + \epsilon_{it}$$

Thus, f_t holds the $k * 1$ factor returns, epsilon(it) is the idiosyncratic (specific to that asset) error of r_{it}, and β_i indicates a $k * 1$ vector of factor loadings for stock i. Either the mean return μ_i is ignored throughout the study, or returns are assumed to be normalized implicitly. The paper states that if stocks i and j have factors β_i and β_j that are equal up to a value of γ, then stocks i and j form a cointegrated system if APT holds true for all time periods. The portfolio returns can therefore be stated as follows:

$$r_{ijt} = r_{it} - \gamma r_{jt} = \beta_i' f_t - \gamma \beta_j' f_t + \epsilon_{it} - \gamma \epsilon_{jt} = \epsilon_{it} - \gamma \epsilon_{jt}$$

According to the paper, the returns of the CTM correspond to the idiosyncratic returns of APT, and the common trend returns of the CTM to the common factor returns of APT. It denotes a perfectly cointegrated pair, as defined by the above equation, in which the common factor returns cancel each other out and are identical up to a scalar. This implies that a common factor return similarity metric may now be used to preselect equities that may be cointegrated.

We have not used this model for our project due to the following reasons:

1. More investigation is necessary into the unusual pairing of CTM and APT, particularly with relation to the presumption that APT is valid across all time periods.
2. The factors for the markets are hard to find, and the paper offers no advice on how to choose a suitable factor model.

But, the cointegration approach is preferred over the distance approach as it more profitable in comparison as found out in the results of Huck and Afawubo [9] for the S&P 500 and Bogomolov [10] for the Australian Stock market.

Other papers like Dunis and Ho [11] use multivariate cointegration framework i.e. they choose a basket of stocks along with another basket and trade them as pairs. Currently, we focus on the univariate framework (i.e. trading one stock versus another stock in a pair) in our project and intent to explore the above method in future.

2.3 Time Series Approach

The most cited paper in this domain is of Elliot et al. [12]. Observed in Gaussian noise, it specifically explains the spread with a mean-reverting Gaussian Markov chain. The latent state variable x_k is thought to go through a mean-reverting process:

$$x_{k+1} - x_k = (a - bx_k)\tau + \sigma\sqrt{\tau}\epsilon_{k+1}$$

Thus, $a \in R_0^+$, b > 0, $\sigma \geq 0$ and $\epsilon_k \sim N(0\,,1)$ are satisfied. For k = {0, 1, 2...,} the time $t_k = k\tau$ is discrete. With mean-reversion strength b, this process reverts to its mean, $\mu = \frac{a}{b}$. Another way to write it is:

$$x_{k+1} = A + Bx_k + C\epsilon_{k+1}$$

where, A = aτ, B = 1 − bτ and C = $\sigma\sqrt{\tau}$. In continuous time, it is possible to explain the state process with the Ornstein-Uhlenbeck process:

$$dx_t = \rho(\mu - x_t)dt + \sigma dW_t$$

where, a standard Brownian motion defined on a probability space is denoted by dW_t. The mean is indicated by the parameter $\mu = \frac{a}{b}$, while the mean-reversion speed is described by $\rho = b$. The measurement equation is the second part of a state space model: In this case, the observed spread is defined as the product of some Gaussian noise $\omega_t \sim N(0,1)$ and the state variable x_k:

$$y_k = x_k + D\omega_k, \qquad D > 0$$

This model states that a trade in pairs is entered when $y_k \geq \mu + c\left(\frac{\sigma}{\sqrt{2p}}\right)$ or $y_k \leq \mu + c\left(\frac{\sigma}{\sqrt{2p}}\right)$. As a result, c represents a fixed value, and Elliott et al. offer no instructions on how to find it. At time T, the position is inverted, signifying the Ornstein-Uhlenbeck process's first passage time outcome.

The above process has three advantages:

1. The state space model and the Kalman Filter can be used to estimate the parameters of the completely tractable model.
2. It is possible to use the continuous time model for predicting. As long as the spread actually adheres to this strict model, important pairs trading concerns like expected time of holding the asset and expected returns can be answered in detail.
3. Mean-reversion is the foundation of the strategy and is essential to pairs trading.

We did not use the Ornstein-Uhlenbeck process in our project due to the following reasons:

1. **Assumption of Constant Mean Reversion Speed**: The OU process assumes a constant pace of mean reversion (denoted by the parameter b in the above formulae). In real financial markets, the speed of mean reversion may not be constant, as market conditions change over time due to factors like volatility, liquidity, or external news. This makes the OU process potentially unrealistic in dynamically changing environments such as the financial markets.
2. **Gaussian Noise and Constant Volatility:** The OU process assumes that the noise driving the process is normally distributed and that the volatility of the spread remains constant over time. However, financial markets often exhibit volatility clustering and fat-tailed distributions (i.e., extreme movements are more likely than under a normal distribution), which the OU process does not account for. This can lead to an underestimation of risks and missed opportunities during periods of high volatility.

3. **Stationarity Assumption:** The OU process assumes that the spread between the two securities is stationary, meaning it fluctuates around a constant mean over time. However, many asset pairs may only exhibit short-term mean reversion, and their relationship could break down over longer periods due to fundamental changes in the underlying assets. This assumption of stationarity might lead to pairs being selected that are not truly mean-reverting over the long run.

The next approach i.e. **Stochastic Control Approach** uses methods similar to the Ornstein-Uhlenbeck processes and thus due to the above reasons we move forward with the other approaches as discussed in Krauss's [1] review paper.

2.4 Other Approaches

- **Machine Learning Approach:** A major use of ML techniques was done by Huck [13] and Huck [14]. Three steps made up the methodology the paper used: trading, outranking, and forecasting. During the forecasting phase, n*(n − 1)/2 possibilities of pairs can be created by considering a universe of n stocks. Huck employs Elman neural networks to produce return forecasts for each security i that are one week ahead of time, depending on the historical returns of stocks i and j, where i, j ∈ {1..., n}. Therefore, for every security i, a total of (n – 1) return projections are generated per month. Huck employs an ELECTRE III, a Multi-Criteria Decision Method (MCDM), in the outranking stage. Using a set of criteria, a set of alternatives are ranked using this method. The following formula can be used to determine any stock i's performance in relation to criterion j:

$$x_{ij} = \hat{x}^{i,t+1}_{|X_{i,t},X_{j,t}} - \hat{x}^{j,t+1}_{|X_{i,t},X_{j,t}}$$

Accordingly, the expected spread, or the difference between the predicted returns of securities i and j, depending on their historical performance, represents the stock's performance. Next, to establish the ranking procedure, the preference, indifference, and veto thresholds are determined. Undervalued stocks are ranked highest and overpriced stocks are ranked lowest using ELECTRE III, which creates an outranking of the pairs. The top m stocks in the ranking are purchased and the bottom m stocks are shorted during the trading stage. The positions are closed, a new rating is made, and the process is repeated following a week of trading. Even though the results of the strategy were impressive with a 54% prediction accuracy and more than 0.8% excessive weekly returns, we did use the above strategy in our project because of the following reasons:

1. The paper did not eliminate survivorship bias from the database used to conduct the test. This meant that the companies which went bankrupt during the said timeline were not taken into account which might have inflated the result.
2. **Overfitting**: In Huck's [15] paper machine learning models like Elman neural networks were used, which can easily overfit, especially when applied to small datasets or datasets that are not sufficiently diverse.
3. **Small and Specific Dataset**: The paper applied machine learning models to the S&P 100 stocks over the period of 1992 to 2006, which could be considered relatively narrow in terms of both the number of assets and the time frame.

- **Principal Component Analysis:** For the U.S. equities market, Avellaneda and Lee [15] created a statistical arbitrage approach that they used to stocks with a market value of more than $1 billion USD at the time of the trading. They employ two different methods during their development stage to break down stock returns into their idiosyncratic and systematic components. These methods are comparable to the Common Trends Model (CTM) that was previously covered in the literature study. The first method involves regressing a stock's R_i returns on the relevant sector ETF in the paper: $\boldsymbol{R_i = \beta_i F + \epsilon_i}$

Thus, β_{ij}F represents the systematic part of the portfolio (where F is the factor return and β_i is the factor loading) and F stands for the returns of the relevant sector ETF. On the other hand, the idiosyncratic component is represented by ε_i. The second method takes into consideration a multi-factor model consisting of m factors:

$$R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \epsilon$$

The paper constructs m eigen-portfolios in accordance with this statistical factor model using Principal Component Analysis (PCA). They also create an equity valuation relative-value model. It is supposed that stock returns fulfill the following differential equation based on the multi-factor model mentioned above:

$$\frac{dP_{it}}{P_{it}} = \mu_i dt + \sum_{j=1} \beta_{it} \frac{dI_{jt}}{I_{jt}} + dX_i$$

Therefore, the residual X_{it} is supposed to follow an Ornstein-Uhlenbeck (OU)-process, and μ_i indicates stock price drift. These two elements line up with stock i's peculiar returns. The systematic returns are represented by the remaining summand, which might originate from either the statistical factor model (m > 1) or the matching sector ETF (m = 1).

This method has impressive returns with a Sharpe ratio (risk adjusted return) of 1.44 from the period of 1997 to 2007 for Principal Component Analysis (PCA) and 1.1 for ETF (exchange traded fund) based strategies.

This strategy was not chosen for our project due to the following reasons:

1. The outcomes of the method are not robust to data mining.
2. **Linear Relationships:** PCA assumes linear relationships between variables (assets). Pairs trading, however, often relies on finding co-movement between assets that may not be strictly linear.

3. **Inability to Capture Mean-Reverting Behavior:** Pairs trading strategies typically rely on the assumption that the spread between two assets exhibits mean-reverting behavior. PCA focuses on explaining variance and does not explicitly model mean reversion.

2.5 Copula Approach

The most cited paper in this method is that of Liew et al. [16]. In a formation period, pairs are built based on previously discussed correlation or cointegration criteria.

Next, for the two components i and j of a pair, the log returns r_{it} and r_{jt} are computed. Subsequently, the return time series' marginal distribution functions F_i and F_j are computed. In this paper, parametric distribution function fitting was chosen. Two uniform variables, U= $F_i(r_{it})$ and V = $F_j(r_{jt})$, are produced by applying the probability integral transform by putting the returns, r_{it} and r_{jt}, into their respective distribution functions. We can now determine a suitable copula function. Starting with five copulas that are frequently used in financial applications, the study assesses several information criteria to decide which one fits the best. They calculate the conditional marginal distribution functions as first partial derivatives of the copula function C (u, v) by utilizing the best-fitted copula:

$$P(U \leq u \mid V = v) = \frac{\partial C(u,v)}{\partial v};\ \ P(V \leq v \mid U = u) = \frac{\partial C(u,v)}{\partial u}$$

One may regard a stock to be comparatively overvalued (undervalued) if the conditional probability is less than or equal to 0.5. When the conditional probabilities are within the tail regions of their conditional distribution functions—that is, below their 5 percent and above their 95 percent confidence levels—the author advises to enter the market. Specifically, whenthe transformed returns of stocks i and j fall beyond the confidence intervals determined by $P(U \leq u \mid V = v) = 0.05$ and $P(V \leq v \mid U = u) = 0.95$, the stocks are purchased and sold short, respectively. When the conditional probabilities

return to 0.5, they switch places. Copulas can identify better trading opportunities because they are a suitable model for such intricate dependency structures.

We have started implementing the last method that is stated above i.e. using copulas for pairs trading, the results of which are stated in the next chapter.

Chapter 3

Methodology

3.1 Data Extraction

For this project, we have initially chosen the NIFTY500 stocks OHLCV (open-high-low- close) data. This data is imported via the yahoo finance library in python. The procedure for data extraction is explained below:

1. The symbols of all the NIFTY500 stocks are imported from the NSE website.
2. Using the yfinance library of python, the past 4 years OHLCV data of the stocks is imported.

3.2 Data Cleaning

After extracting the data in the previous step, we start cleaning the data in the following steps:

1. The empty files are removed from the directory (which may have been caused byerrors in the API)
2. Columns other than the “Adjusted Close” are removed from the data as here we consider the daily data and are concerned only with the closing prices of stocks.
3. The listing dates of all the stocks are taken into a list and sorted.
4. The stocks that were listed after 4th July 2019 are removed from the directory asthey don’t have enough data for us to perform the analysis.
5. The data before 4th July 2019 for all the stocks is removed and thus the period we select for analysis is 4th July 2019 to 4th July 2024.
6. Now that all the stocks have the data for the selected period, we round off the “Adjusted Close” column for all the stocks down to 2 decimal places.

7. As we have to use percentage change rather than price data, we convert the prices to percentage change.
8. The dataframe is then converted to a .csv file and saved.

3.3 Conversion to cumulative log returns

The above saved csv file is opened and all the columns of the dataframe are added by 1 as returns can have a negative value which will give an error when we take the logarithm of the same. After this, we take the logarithm of data and then we take the data's cumulative sum to the exponential. Thus, we get the cumulative logarithmic returns of all the selected stocks.

3.4 Finding Pairs

After finding the cumulative log returns for all the stocks we now find pairs among these which we can them select to implement our strategy in python. To do this, we follow the below steps:

1. We split the data in half into training and testing datasets so that we can find correlated stocks with the training data and trade the pair in the testing period.
2. After this, we take out the correlation between the cumulative log returns of all the stocks via the **spearman correlation** (as we are more focused on the ranking rather than the linear relationship between stock returns).
3. From the above list we select the pairs of stocks that have correlation greater than 0.85 as we need variance between the stocks to trade the pair efficiently.

NIFTY 500 index can majorly be divided into the following sectors based on the composition of the index:

Table 1: Weights of industries in the NIFTY 500 index

Sector	Weight in the index (%)
Financial Services	28.98
Information Technology (IT)	9.64
Oil, Gas and Consumable Fuels	7.82
Fast moving consumer goods (FMCG)	7.09
Automobile and auto components	7.00
Healthcare	6.32

Source: https://www.niftyindices.com/Factsheet/ind_nifty_500.pdf

3.5 Choosing Pairs

We have chosen the following pairs from each sector to proceed:

- **Financial Services**: (Aditya Birla Capital (ABCAPITAL), Canfin Homes (CANFINHOME))
- **Information Technology (IT)**: (Tata Consultancy Services (TCS), Infosys (INFY))
- **Oil, Gas and Consumable Fuels**: (Oil and Natural Gas Corporation (ONGC), Oil India (OIL))
- **Fast Moving Consumer Goods (FMCG)**: (Nestle India (NESTLEIND), Dabur India (DABUR))
- **Automobile and Auto Components**: (CEAT Ltd. (CEATLTD), MRF Ltd. (MRF))
- **Healthcare**: (Cipla Ltd. (CIPLA), Glenmark Pharmaceuticals Ltd. (GLENMARK))

The pairs above are selected keeping in mind that they belong to the same industry and for most cases are direct competitors of each other. This enables us to establish a baseline similarity between the stocks in each pair which justifies the assumption of mean reversion in pairs which is the underlying concept of pairs trading.

3.6 Fitting marginal distributions to pairs

We fit different marginal distributions to the returns series of the pairs like:

- **Normal Distribution:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}, \qquad x \in R$$

- **T-Distribution:**

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}, \quad v \in (1, n-1)$$

- **Logistic Distribution:**

$$f(t) = \frac{e^z}{\sigma(1+e^z)^2}; \;\; z = \frac{t-\mu}{\sigma}, \quad t, \mu \in (-\infty, \infty), \;\; \sigma > 0$$

We mainly fit the above 3 distributions to the returns time series of all the pairs and find out the best fit using AIC (Akaike Information Criteria), BIC (Bayesian Information Criteria) and HQIC (Hannan-Quinn Information Criteria). These values are used to judge the goodness-of-fit of the distributions over the return time series. **AIC** is effective for finding the best predictive model, prioritizing goodness-of-fit but with a moderate penalty for complexity. **BIC**, which imposes a stronger penalty on model complexity, is preferred when the goal is to select the true underlying model, especially in large datasets. **HQIC** balances between AIC and BIC, offering robustness in scenarios where sample sizes are moderate. These are calculated as follows:

- **AIC:**

$$AIC = -2 * (LL) + 2 * K$$

- **BIC:**

$$BIC = -2 * (LL) + \log(N) * K$$

- **HQIC:**

$$HQIC = -2 * (LL) + 2 * \log(\log(N)) * K$$

where,

LL = Log-likelihood of the model (maximum likelihood which depicts the fit of the model) ($LL \in (-\infty, \infty)$)
K = Number of predictors in the model, including the intercept and any additional indicators ($K \in (I^+)$)
N = Sample size ($N \in (I^+)$)

Considering the above values for AIC, BIC, and HQIC, we select the distribution which best fits our time series and proceed to the further process with it.

3.7 Testing for statistical significance of the fitted distributions

We perform the one-sample Kolomogorov-Smirnov test (KS test) on both pairs with their fitted distributions we get from the previous method against their respective return time series of the same period. The p-value that this test outputs, helps us decide whether the fit of the distribution is statistically significant or not. It is calculated by the following formula:

$$D_n = sup_x |F_n(x) - F(x)|$$

$$p \approx 2 * \sum_{k=1}^{\infty} (-1)^{k+1} e^{\left(-2k^2 n {D_n}^2\right)}$$

where:

- F_n(x): Empirical CDF of the sample.
- F(x): Hypothesized CDF.
- D_n: KS statistic, the maximum vertical distance between F_x and F(x).

The distribution fit is statistically significant when the p-value in the above formula exceeds a critical value. In our case we take the critical value to be 0.08, i.e. when the KS value is greater than 0.08, we can effectively reject the null hypothesis with a 99% confidence interval. After rejecting the null hypothesis, we start finding the best copula that fits our previously found distribution.

3.8 Fitting copulas to distributions

A copula fundamentally represents a cumulative distribution function, wherein the individual marginal distributions adhere to the uniform (0, 1) distribution. A copula is characterized by its ability to encapsulate the joint distribution of variables by transforming each variable into its marginal distribution.

A d-dimensional Copula can be written as:

$$C_Y(u_1, u_2, \dots, u_d) = P\{F_{Y_1}(Y_1) \le u_1, \dots, F_{Y_d}(Y_d) \le u_d\}$$

$$= P\{Y_1 \le {F_{Y_1}}^{-1}(u_1), \dots, Y_d \le {F_{Y_d}}^{-1}(u_d)\}$$

$$= F_Y\{{F_{Y_1}}^{-1}(u_1), \dots, {F_{Y_d}}^{-1}(u_d)\}$$

$$Further\ u_j = F_{Y_j}(y_j), \forall j = 1,2,3, \dots, d, hence$$

$$F_Y(y_1, \dots, y_d) = C_y\{F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)\}$$

Here, F_{Y_1}, F_{Y_2},…, F_{Y_d} are the cumulative distribution functions of Y_1, Y_2,…, Y_d respectively.

We have fitted our distributions on different copulas like Gaussian, Gumbel, Frank, Joe and Clayton. The formulae for the same are given below:

- **Gaussian Copula:**

$$C_R^G(u_1, \dots, u_d) = \varphi_R\left(\varphi^{-1}(u_1), \dots, \varphi^{-1}(u_d)\right)$$
$$u_1, u_2, \dots, u_d \in (0,1)$$

Where R is the correlation matrix and $\varphi^{-1}(.)$ *is the inverse of the standard normal cumulative distribution function.*

- **Clayton Copula:**

$$C_{cl}(u_1, u_2, \dots, u_d|\theta) = \left({u_1}^{-\theta} + {u_2}^{-\theta} + \cdots + {u_d}^{-\theta} + 1 - d\right)^{\frac{-1}{\theta}}$$

$$u_1, u_2, \dots, u_d \in (0,1) \text{ and } \theta > 0$$

- **Gumbel Copula:**

$$C(u, v) = e^{-\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{-\frac{1}{\theta}}}, \quad u, v \in (0\,, 1) \text{ and } \theta > 0$$

- **Frank Copula:**

$$C(u, v) = -\frac{\ln\left(\frac{1 + g(u)g(v)}{g(1)}\right)}{\theta}, \quad u, v \in (0\,, 1)$$

$$g(x) = e^{-\theta x} - 1, \quad \theta > 0$$

Where u and v are the respective cumulative distribution functions (CDF) of the pair we are fitting the copula on.

- **Joe Copula:**

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

$$u, v \in (0\ ,1)\ and\ \theta > 0$$

The above copulas are fitted on the marginal distributions we got earlier to get the best fitting copula using the AIC, BIC and HQIC tests and are also tested for statistical significance by using the KS test. The best fitting copula is then selected and the trading logic is applied on it to produce trading signals.

3.9 Trading Strategy

After the previous step, we calculate the conditional probabilities $P(U \leq u \mid V = v)$ and $P(V \leq v \mid U = u)$. Using these conditional probabilities, we get the trading signals by the following conditions:

a = 0.9 ($For\ our\ analysis$)

b = 0.5 ($For\ our\ analysis$)

If we currently have a long position (i.e. long position on 1st stock and short on the 2nd stock):

If ($P(U \leq u \mid V = v)$ > (1-b)) or ($P(V \leq v \mid U = u)$ < b) we **close** the long position

Else we keep our long position intact.

If we currently have a short position (i.e. short position on 1st stock and long on 2nd stock):

If ($P(U \leq u \mid V = v)$ > b) or ($P(V \leq v \mid U = u)$ > (1-b)) we **close** the short position

Else we keep our short position intact.

If we currently do not have any open position:

If $(P(U \leq u \mid V = v) < (1\text{-a}))$ or $(P(V \leq v \mid U = u) > \text{a})$ we **open** a long position

Else if $(P(U \leq u \mid V = v) > \text{a})$ or $(P(V \leq v \mid U = u) < (1\text{-a}))$ we **open** a short position

Else we do not open either position.

Chapter 4

Results and discussion

The results for all the sectors are given below:

4.1 Financial Services (ABCAPITAL and CANFINHOME):

We first made the Q-Q (Quantile-Quantile) graph of the returns of these stocks.

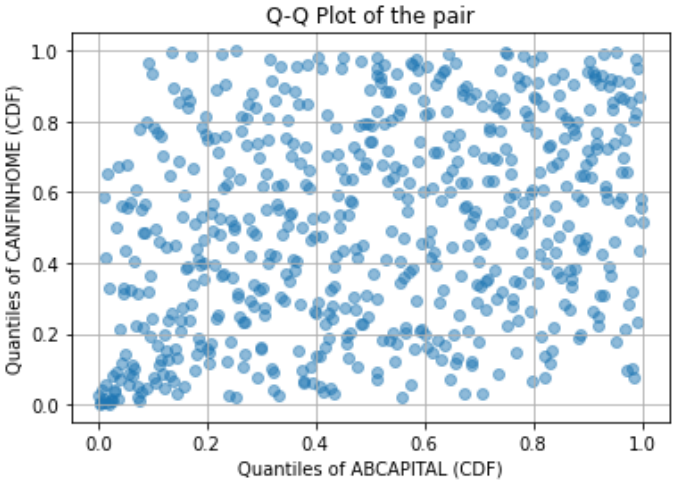

Fig 1: Q-Q Plot of ABCAPITAL and CANFINHOME

This shows that the returns of both the stocks is uniformly distributed across quantiles but there is concentration of returns on the lower side of the graph which suggests that it would fit well on distributions that have fatter tails. We plotted the fit of different distributions on the returns data of both ABCAPITAL and CANFINHOME. We found out that the t-distribution fits this data the best, confirming our assumption. We also tested for statistical significance using the KS test and the results for all these distributions are given below:

Table 2: Distribution fit parameters of finance sector

Stock	Distribution	AIC	BIC	KS Value
ABCAPITAL	Student-t	-2617.75	-2604.47	0.824
ABCAPITAL	Normal	-2556.14	-2564.99	0.035
ABCAPITAL	Logistic	-2611.58	-2598.30	0.637
CANFINHOME	Student-t	-2797.72	-2784.44	0.869
CANFINHOME	Normal	-2716.70	-2707.85	0.024
CANFINHOME	Logistic	-2786.31	-2773.03	0.603

The graphs for the same can be plotted as shown below:

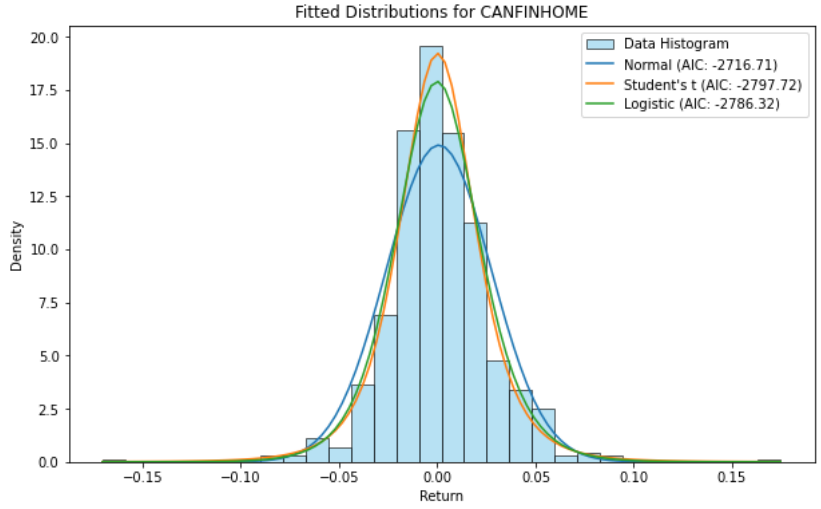

Fig 2: Distribution fit of CANFINHOME

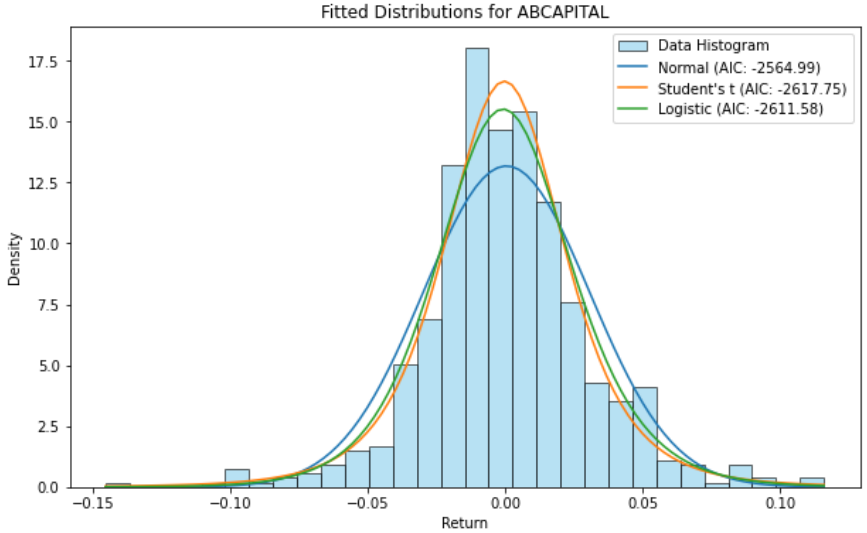

Fig 3: Distribution fit of ABCAPITAL

The result after fitting the copulas to this pair is as follows:

Table 3: Copula goodness-of-fit parameters for finance sector

Copula type	Parameter	AIC	BIC	KS Value
Gaussian	0.32	-307.09	-302.66	0.388
Clayton	0.62	-112.77	-108.34	0.551
Gumbel	1.19	-40.61	-36.18	0.003
Frank	1.94	-59.16	-54.74	0.224
Joe	1.16	-14.98	-10.55	0.102

The returns for the strategy are shown below:

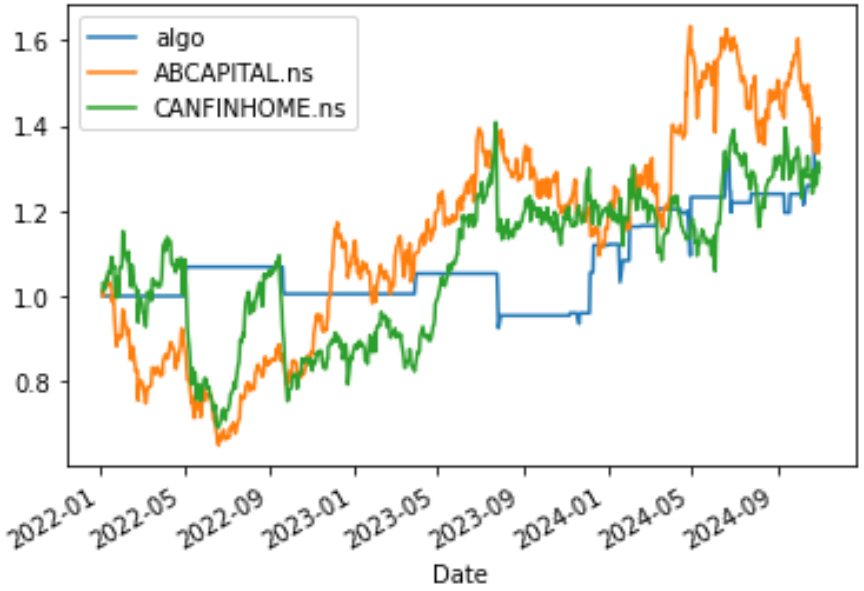

Fig 4: Returns of the algorithm on finance sector

The detailed result is shown in the table below:

Table 4: Detailed result of finance sector

Strategy	Absolute Return	Annual Return	Sharpe Ratio	Max. Drawdown
Algorithm	0.393	0.127	0.745	-0.134
ABCAPITAL	0.341	0.111	0.482	-0.379
CANFINHOME	0.286	0.094	0.435	-0.399

4.2 Information Technology (TCS and INFY):

We first made the Q-Q (Quantile-Quantile) graph of the returns of these stocks.

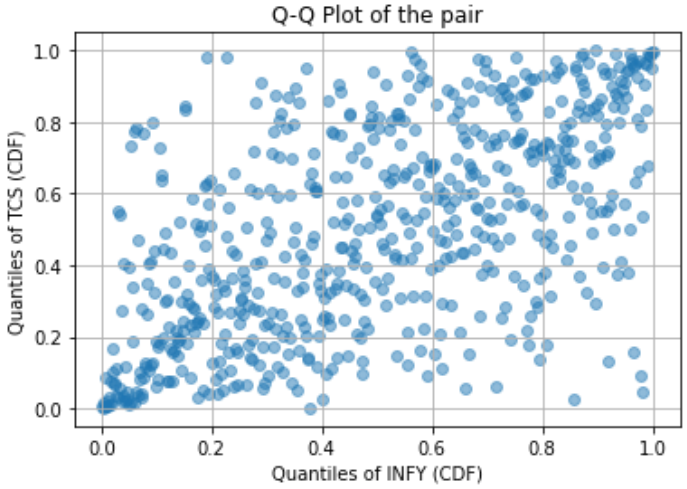

Fig 5: Q-Q Plot of TCS and INFY

This shows that the returns of both the stocks is uniformly distributed across quantiles but there is concentration of returns on the lower side of the graph which suggests that it would fit well on distributions that have fatter tails. We plotted the fit of different distributions on the returns data of both TCS and INFY. We found out that the t-distribution fits this data the best, confirming our assumption. We also tested for statistical significance using the KS test and the results for all these distributions are given below:

Table 5: Distribution fit parameters of IT Sector

Stock	Distribution	AIC	BIC	KS Value
INFY	Normal	-3237.41	-3316.85	0.0033
INFY	Student-t	-3354.82	-3341.54	0.935
INFY	Logistic	-3330.13	-3228.55	0.3
TCS	Normal	-3052.33	-3043.47	0.00000137
TCS	Student-t	-3276.78	-3263.05	0.79
TCS	Logistic	-3216.57	-3203.47	0.120

The graphs for the same can be plotted as shown below:

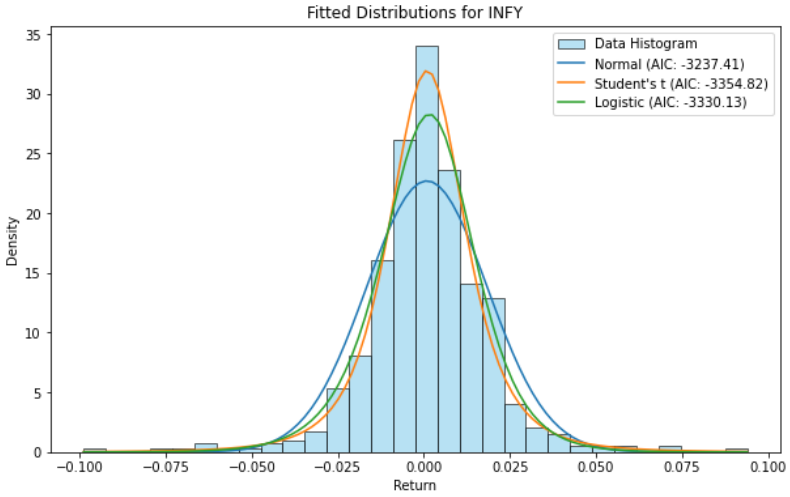

Fig 6: Distribution fit of INFY

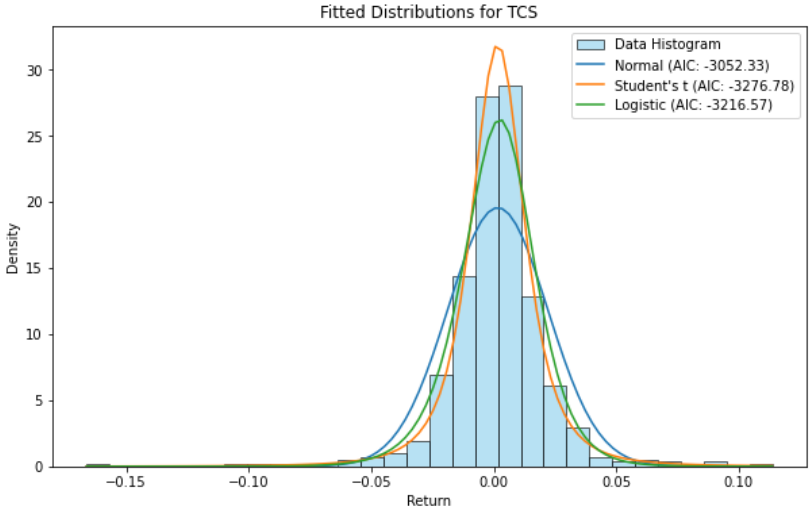

Fig 7: Distribution fit of TCS

The results after fitting the copulas to this pair are:

Table 6: Copula goodness-of-fit parameters for IT sector

Copula type	Parameter	AIC	BIC	KS Value
Gaussian	0.62	-478.80	-474.37	0.576
Clayton	1.19	-28.91	-282.48	0.414
Gumbel	1.67	-281.58	-277.16	0.314
Frank	4.61	-272.37	-267.94	0.281
Joe	1.82	-204.17	-199.74	0.058

The returns for the strategy are shown below:

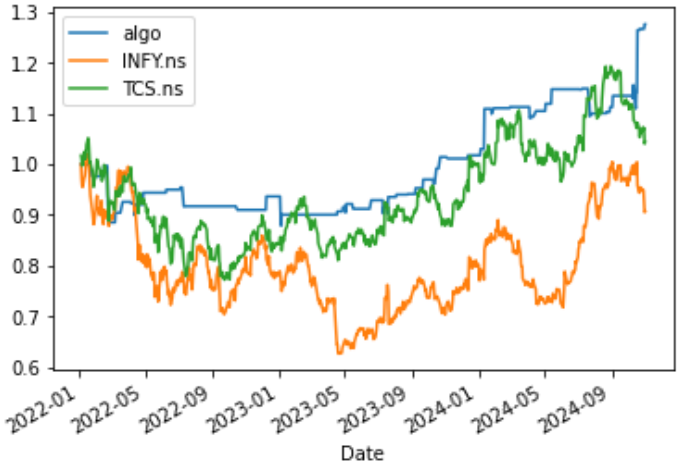

Fig 8: Results of the algorithm on IT Sector

The detailed result is shown in the table below:

Table 7: Detailed result of IT sector

Strategy	Absolute Return	Annual Return	Sharpe Ratio	Max. Drawdown
Algorithm	0.320	0.105	0.851	-0.121
TCS	0.029	0.010	0.153	-0.265
INFY	-0.085	-0.031	-0.003	-0.386

4.3 Oil, Natural Gas and other consumable fuels (OIL and ONGC):

We first made the Q-Q (Quantile-Quantile) graph of the returns of these stocks.

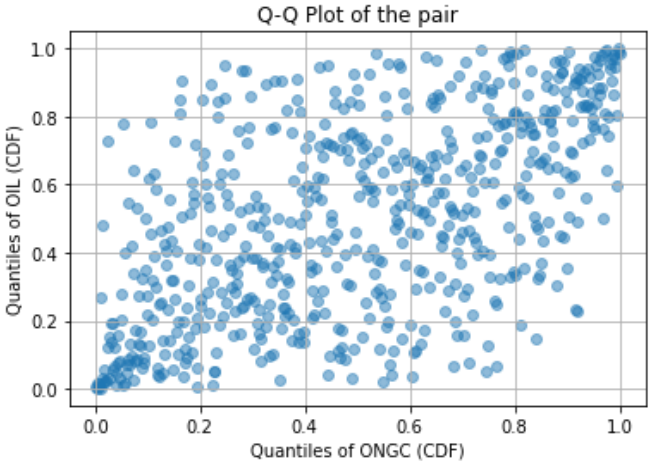

Fig 9: Q-Q Plot of OIL and ONGC

This shows that the returns of both the stocks is uniformly distributed across quantiles but there is concentration of returns on the lower side of the graph which suggests that it would fit well on distributions that have fatter tails. We plotted the fit of different distributions on the returns data of both OIL and ONGC. We found out that the t-distribution fits this data the best, confirming our assumption. We also tested for statistical significance using the KS test and the results for all these distributions are given below:

Table 8: Distribution fit parameters of Oil Sector

Stock	Distribution	AIC	BIC	KS Value
OIL	Normal	-2681.50	-2672.64	0.004
OIL	Student-t	-2790.91	-2777.63	0.998
OIL	Logistic	-2769.07	-2755.79	0.530
ONGC	Normal	-2690.75	-2681.89	0.0005
ONGC	Student-t	-2845.46	-2832.18	0.790
ONGC	Logistic	-2809.82	-2796.54	0.380

The graphs for the same can be plotted as shown below:

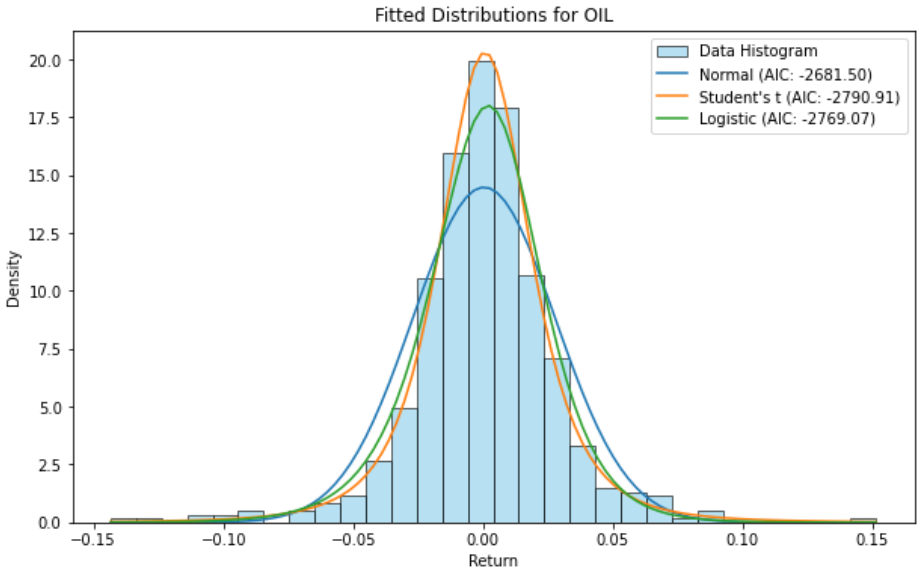

Fig 10: Distribution fit of OIL

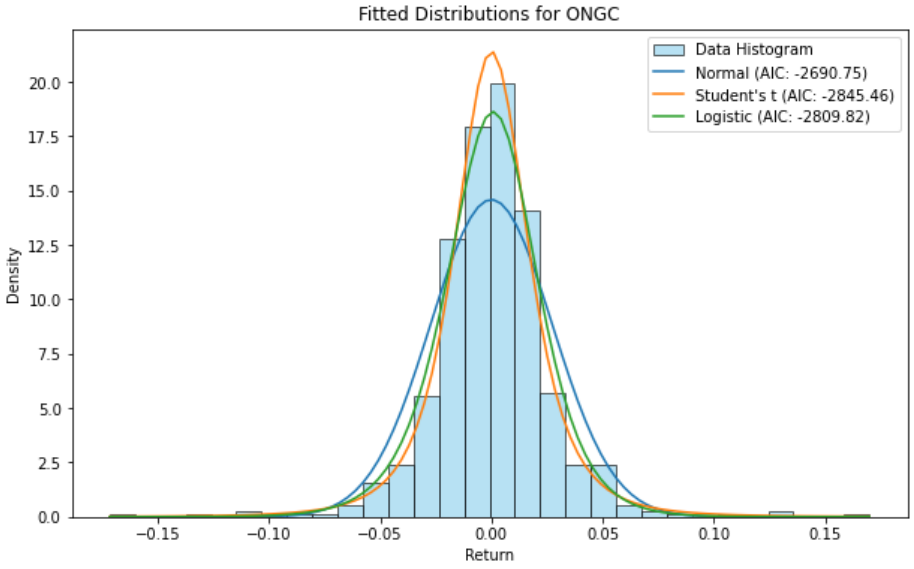

Fig 11: Distribution fit of ONGC

The results after fitting the copulas to this pair are:

Table 9: Copula goodness-of-fit parameters for Oil sector

Copula type	Parameter	AIC	BIC	KS Value
Gaussian	0.59	-459.92	-455.50	0.411
Clayton	1.07	-254.59	-250.17	0.303
Gumbel	1.64	-267.43	-263.00	0.332
Frank	4.26	-244.46	-240.04	0.419
Joe	1.80	-201.18	-196.75	0.019

The returns for the strategy are shown below:

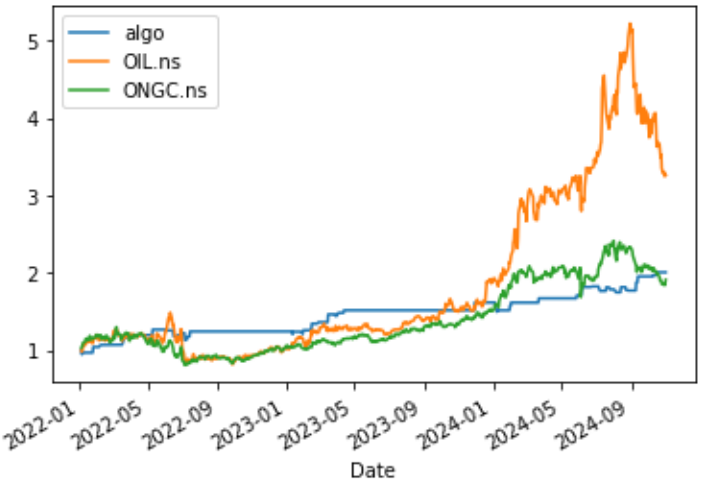

Fig 12: Results of the algorithm on the Oil sector

The detailed result is shown in the table below:

Table 10: Detailed result of Oil sector

Strategy	Absolute Return	Annual Return	Sharpe Ratio	Max. Drawdown
Algorithm	1.084	0.303	1.655	-0.107
OIL	2.291	0.536	1.223	-0.449
ONGC	0.848	0.247	0.832	-0.379

4.4 Fast moving consumer goods (FMCG) (NESTLEIND and DABUR):

We first made the Q-Q (Quantile-Quantile) graph of the returns of these stocks.

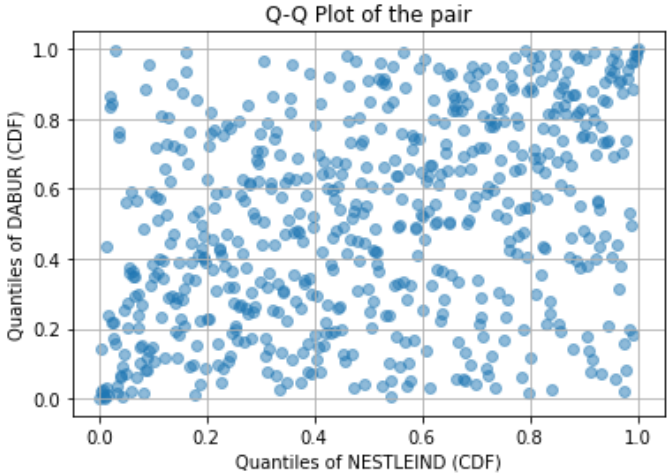

Fig 13: Q-Q Plot of NESTLEIND and DABUR

This shows that the returns of both the stocks is uniformly distributed across quantiles but there is concentration of returns on the lower side of the graph which suggests that it would fit well on distributions that have fatter tails. We plotted the fit of different distributions on the returns data of both NESTLEIND and DABUR. We found out that the t-distribution fits this data the best, confirming our assumption. We also tested for statistical significance using the KS test and the results for all these distributions are given below:

Table 11: Distribution fit parameters of FMCG Sector

Stock	Distribution	AIC	BIC	KS Value
NESTLEIND	Normal	-3402.30	-3393.45	0.002
NESTLEIND	Student-t	-3534.72	-3521.43	0.737
NESTLEIND	Logistic	-3510.58	-3497.58	0.474
DABUR	Normal	-3323.60	-3314.75	0.003
DABUR	Student-t	-3463.08	-3449.80	0.549
DABUR	Logistic	-3436.29	-3423.00	0.548

The graphs for the same can be plotted as shown below:

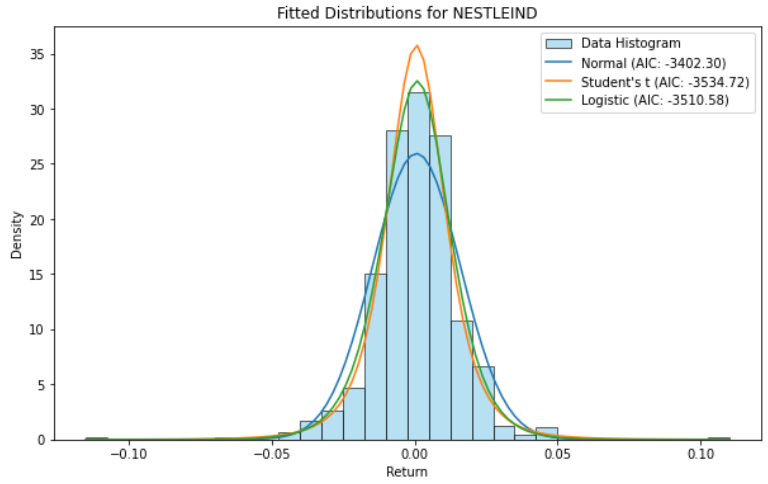

Fig 14: Distribution fit of NESTLEIND

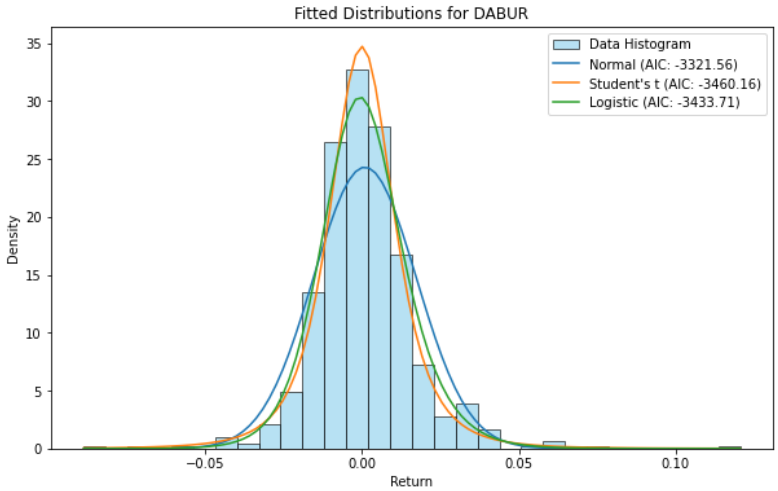

Fig 15: Distribution fit of DABUR

The results after fitting the copulas to this pair are:

Table 12: Copula goodness-of-fit parameters for FMCG sector

Copula type	Parameter	AIC	BIC	KS Value
Gaussian	0.38	-346.33	-341.90	0.490
Clayton	0.59	-93.85	-89.42	0.281
Gumbel	1.31	-109.88	-105.45	0.336
Frank	2.45	-88.64	-84.22	0.189
Joe	1.38	-86.16	-81.73	0.058

The returns for the strategy are shown below:

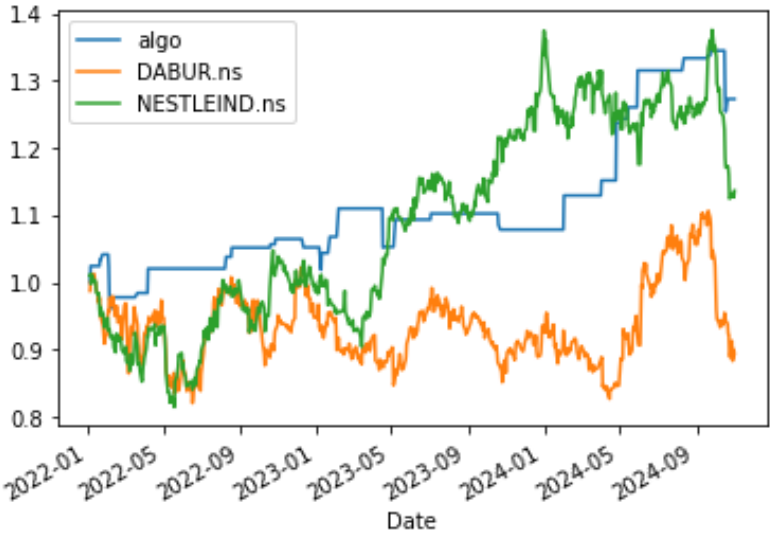

Fig 16: Results of the algorithm on the FMCG sector

The detailed result is shown in the table below:

Table 13: Detailed result of FMCG sector

Strategy	Absolute Return	Annual Return	Sharpe Ratio	Max. Drawdown
Algorithm	0.272	0.090	0.928	-0.077
NESTLEIND	0.124	0.043	0.319	-0.195
DABUR	-0.094	-0.035	-0.054	-0.201

4.5 Automobile and Auto components (CEATLTD and MRF):

We first made the Q-Q (Quantile-Quantile) graph of the returns of these stocks.

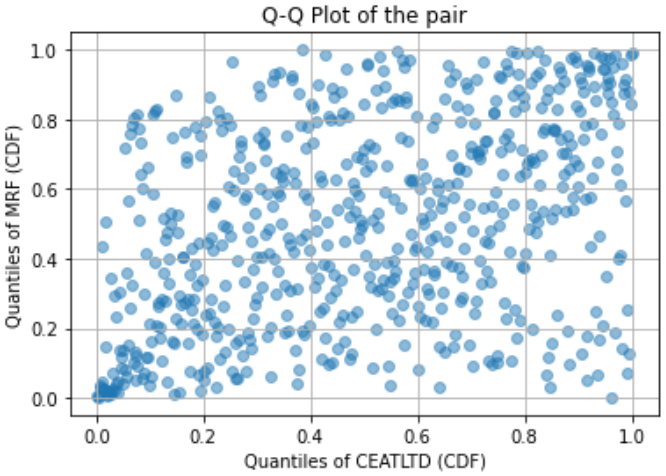

Fig 17: Q-Q Plot of CEATLTD and MRF

This shows that the returns of both the stocks is uniformly distributed across quantiles but there is concentration of returns on the lower side of the graph which suggests that it would fit well on distributions that have fatter tails. We plotted the fit of different distributions on the returns data of both CEATLTD and MRF. We found out that the t-distribution fits this data the best, confirming our assumption. We also tested for statistical significance using the KS test and the results for all these distributions are given below:

Table 14: Distribution fit parameters of Auto Sector

Stock	Distribution	AIC	BIC	KS Value
CEATLTD	Normal	-2942.40	-2933.54	3.755*e-07
CEATLTD	Student-t	-3149.29	-3136.00	0.87
CEATLTD	Logistic	-3081.14	-3067.85	0.032
MRF	Normal	-3155.57	-3146.71	9.589*e-05
MRF	Student-t	-3306.72	-3293.43	0.516
MRF	Logistic	-3267.06	-3523.78	0.070

The graphs for the same can be plotted as shown below:

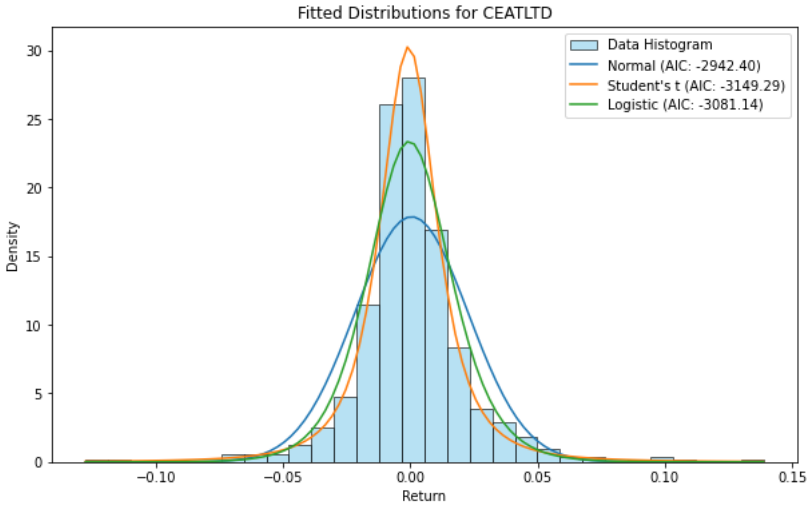

Fig 18: Distribution fit of CEATLTD

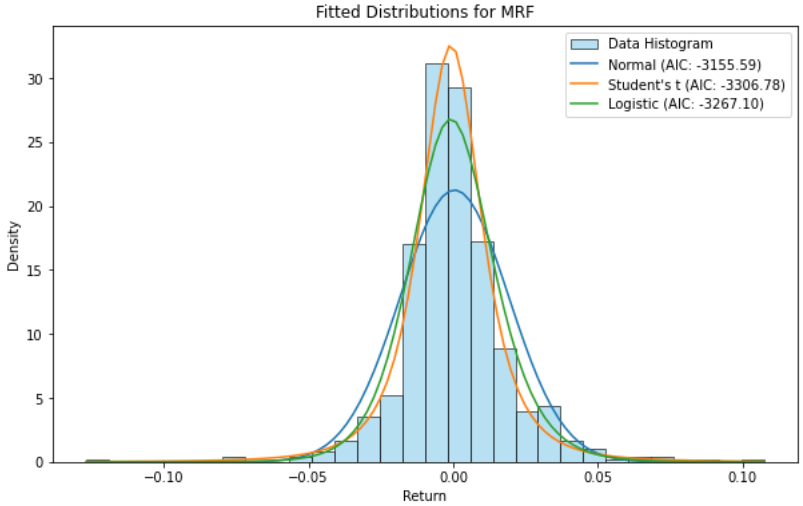

Fig 19: Distribution fit of MRF

The results after fitting the copulas to this pair are:

Table 15: Copula goodness-of-fit parameters for Auto sector

Copula type	Parameter	AIC	BIC	KS Value
Gaussian	0.38	-375.05	-370.62	0.197
Clayton	0.59	-160.63	-156.21	0.411
Gumbel	1.31	-116.76	-112.33	0.025
Frank	2.45	-119.30	-114.87	0.134
Joe	1.38	-78.65	-74.22	0.004

The returns for the strategy are shown below:

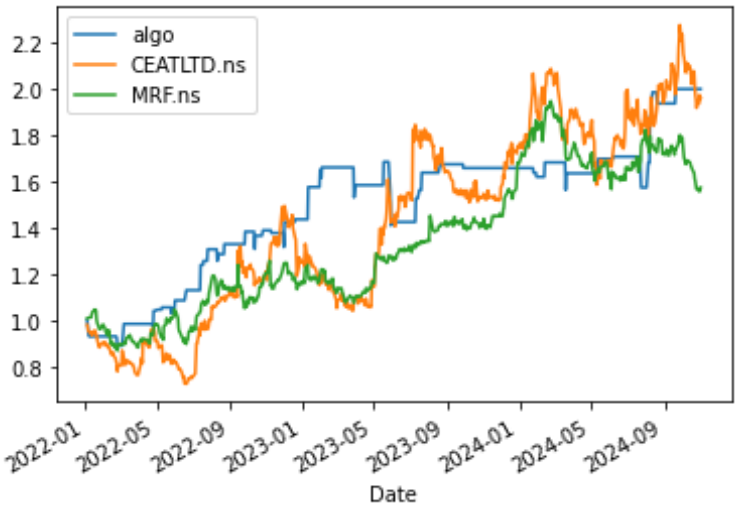

Fig 20: Results of the algorithm on the Auto sector

The detailed result is shown in the table below:

Table 16: Detailed result of Auto sector

Strategy	Absolute Return	Annual Return	Sharpe Ratio	Max. Drawdown
Algorithm	1.001	0.284	1.233	-0.163
CEATLTD	1.001	0.284	0.852	-0.304
MRF	0.5622	0.174	0.827	-0.201

4.6 Healthcare (CIPLA and GLENMARK):

We first made the Q-Q (Quantile-Quantile) graph of the returns of these stocks.

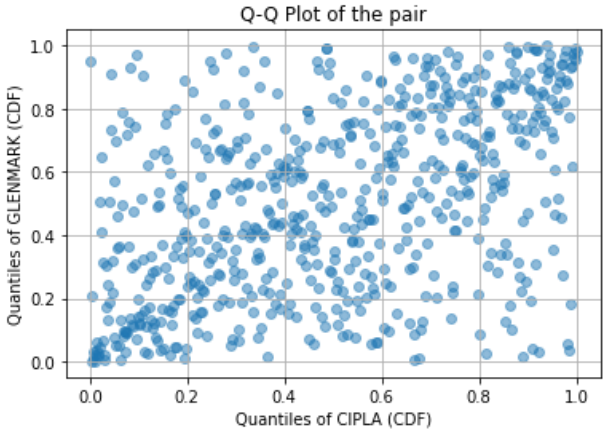

Fig 21: Q-Q Plot of CIPLA and GLENMARK

This shows that the returns of both the stocks is uniformly distributed across quantiles but there is concentration of returns on the lower side of the graph which suggests that it would fit well on distributions that have fatter tails. We plotted the fit of different distributions on the returns data of both CIPLA and GLENMARK. We found out that the t-distribution fits this data the best, confirming our assumption. We also tested for statistical significance using the KS test and the results for all these distributions are given below:

Table 17: Distribution fit parameters of Healthcare Sector

Stock	Distribution	AIC	BIC	KS Value
CIPLA	Normal	-3089.96	-3081.10	4.9*e-06
CIPLA	Student-t	-3206.08	-3192.80	0.47
CIPLA	Logistic	-3182.72	-3169.44	0.059
GLENAMRK	Normal	-2588.59	-2579.73	6.18*e-06
GLENMARK	Student-t	-2850.81	-2837.52	0.92
GLENMARK	Logistic	-2791.74	-2778.45	0.142

The graphs for the same can be plotted as shown below:

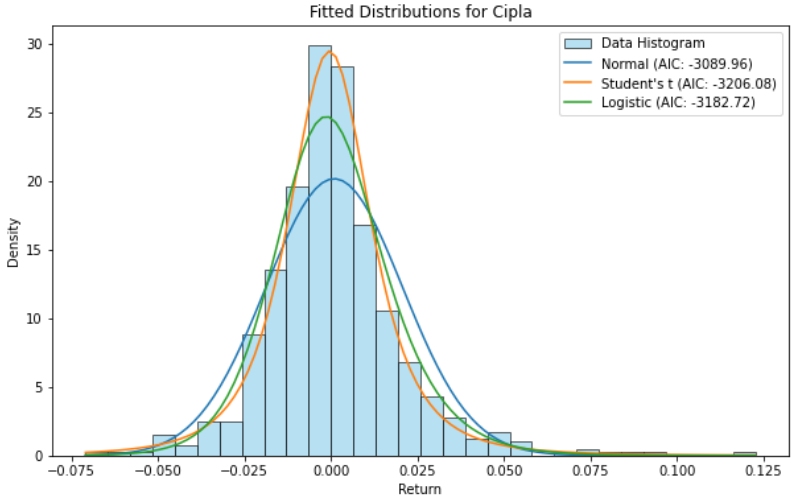

Fig 22: Distribution fit of CIPLA

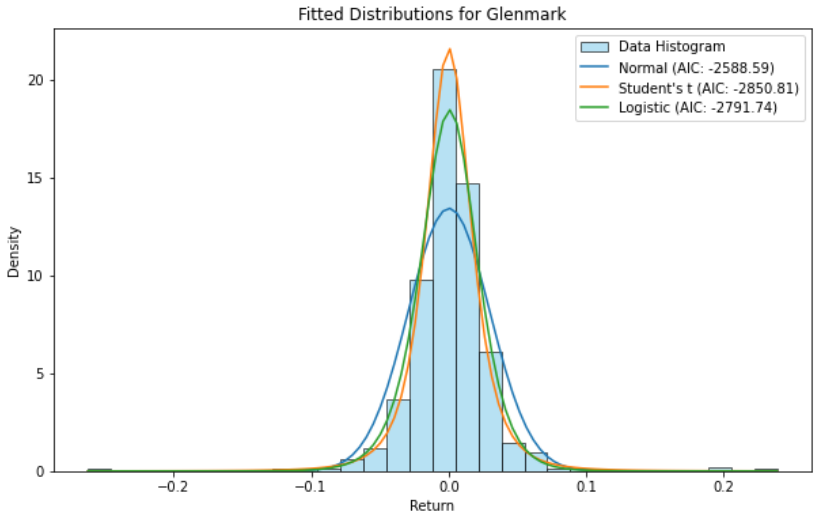

Fig 23: Distribution fit of GLENMARK

The results after fitting the copulas to this pair are:

Table 18: Copula goodness-of-fit parameters for healthcare sector

Copula type	Parameter	AIC	BIC	KS Value
Gaussian	0.49	-405.23	-400.80	0.213
Clayton	0.78	-139.69	-135.26	0.251
Gumbel	1.41	-146.44	-142.02	0.165
Frank	3.40	-158.45	-154.02	0.359
Joe	1.49	-106.55	-102.12	0.015

The returns for the strategy are shown below:

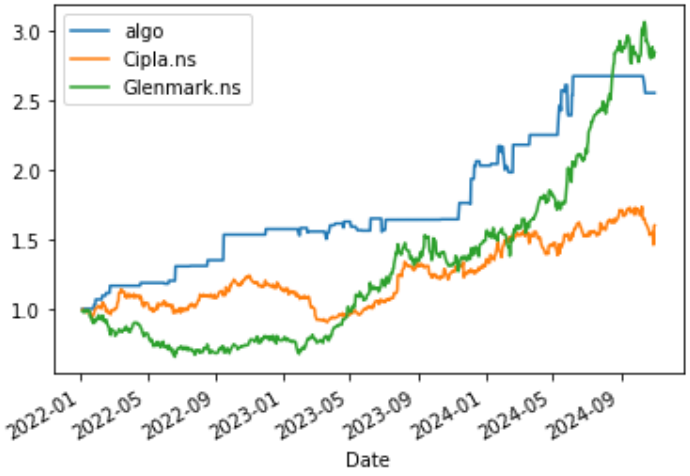

Fig 24: Result of the algorithm on the healthcare sector

The detailed result is shown in the table below:

Table 19: Detailed result of healthcare sector

Strategy	Absolute Return	Annual Return	Sharpe Ratio	Max. Drawdown
Algorithm	1.550	0.401	1.772	-0.086
CIPLA	0.609	0.187	0.847	-0.270
GLENMARK	1.886	0.465	1.381	-0.339

From all the above results, it can be evidently said that the copula-based pairs trading strategy is more superior than holding an individual stock from a return standpoint and also from a risk standpoint. In all cases, we can see that the Sharpe ratio (the measure of risk adjusted returns) of the strategy is far better than the basic buy-and-hold strategy and also provides better returns in most cases even when our trading period was a bull run period in the indian markets.

Chapter 5

Conclusion

The project successfully demonstrated the efficacy of a pairs trading strategy leveraging conditional probability through copulas to generate superior risk-adjusted returns compared to the traditional buy-and-hold strategy. By selecting pairs within the top six sectors of the NIFTY 500 index by weightage, the methodology ensured diversification while focusing on sectors with significant market representation. The use of copulas allowed for a nuanced modeling of dependencies between asset returns, capturing tail dependencies and non-linear relationships that conventional correlation-based approaches might miss. This led to informed trading decisions and improved entry-exit timing, crucial for the profitability of pairs trading.

The results revealed that the pairs trading strategy consistently outperformed the buy-and-hold approach in terms of Sharpe ratio, reutrns and maximum drawdowns across all six sectors. These metrics underscore the strategy's robustness in generating higher returns relative to risk while minimizing exposure to severe losses during market downturns. By dynamically adjusting positions based on the conditional probabilities derived from the copula models, the strategy effectively capitalized on mean-reverting opportunities. This not only highlights the practical utility of copula-based modeling in financial trading but also establishes a compelling case for adopting statistical arbitrage techniques in equity markets. The findings are particularly relevant for institutional investors and quantitative traders aiming to optimize returns with controlled risk exposure.

Chapter 6

References

1. Krauss, Christopher (2015): Statistical arbitrage pairs trading strategies: Review and outlook, IWQW Discussion Papers, No. 09/2015, Friedrich-Alexander-Universität Erlangen- Nürnberg, Institut für Wirtschaftspolitik und Quantitative Wirtschaftsforschung (IWQW), Nürnberg
2. Gatev, E., Goetzmann, W. N., and Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. Review of Financial Studies, 19(3):797–827.
3. Vidyamurthy, G. (2004). Pairs trading: Quantitative methods and analysis. John Wiley & Sons, Hoboken, N.J.
4. Do, B., Faff, R., and Hamza, K. (2006). A new approach to modeling and estimation for pairs trading. In Proceedings of 2006 Financial Management Association European Conference.
5. Puspaningrum, H. (2012). Pairs trading using cointegration approach. PhD thesis,University of Wollongong.
6. Stock, J. H. and Watson, M. W. (1988). Testing for common trends. Journal of theAmerican Statistical Association, 83(404):1097.
7. Ross, S. A. (1976). The arbitrage theory of capital asset pricing. Journal of EconomicTheory, 13(3):341–360.
8. Tsay, R. S. (2010). Analysis of financial time series. Wiley series in probability and statistics. John Wiley & Sons, Hoboken, N.J., 3rd edition.
9. Huck, N. and Afawubo, K. (2015). Pairs trading and selection methods: is cointegration superior? Applied Economics, 47(6):599–613.
10. Bogomolov, T. (2011). Pairs trading in the land down under. In Finance and CorporateGovernance Conference.
11. Dunis, C. L. and Ho, R. (2005). Cointegration portfolios of European equities for index tracking and market neutral strategies. Journal of Asset Management, 6(1):33–52.
12. Elliott, R. J., Van Der Hoek*, John, and Malcolm, W. P. (2005). Pairs trading. QuantitativeFinance, 5(3):271–276.
13. Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index.European Journal of Operational Research, 196(2):819–825.

14. Huck, N. (2010). Pairs trading and outranking: The multi-step-ahead forecasting case. European Journal of Operational Research, 207(3):1702–1716.

15. Avellaneda, M. and Lee, J.-H. (2010). Statistical arbitrage in the US equities market. Quantitative Finance, 10(7):761–782.

16. Liew, R. Q. and Wu, Y. (2013). Pairs trading: A copula approach. Journal of Derivatives & Hedge Funds, 19(1):12–30.