International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 6, June-2014, pp: (344-355), Impact Factor: 1.252, Available online at: www.erpublications.com

Multiday Pair Trading Strategy using Copula

Apoorva Uday Nayak[1], Megha Ugalmugale[2], Vishwanjali Gaikwad[3], Vahida Z. Attar[4]
[1,2,3,4]Department of Computer Engineering College of Engineering, Pune, India

Abstract: Pair trading is a market neutral strategy meant to generate profit regardless of whether equities rise or fall. The strategy of matching a long position (buy) with a short position (sell) in two historically correlated stocks is known as pair trading. Some traditional pair trading techniques use correlation or cointegration as a dependence measure and assume symmetric distribution of data along the mean zero. However, the occurrence of non-linear distributions is quite frequent in financial assets and thus the use of linear correlation coefficient as a dependence measure is erroneous and may lead to misleading results. This may trigger wrong trading signals and may fail to recognize profit opportunities. This paper presents an overview of applications of different copulas to develop a model to suggest profitable strategy by analyzing data and providing trading signals using machine learning and other statistical analysis techniques. Copula is a relatively new pair trading technique. Copulas are much more realistic. They can be applied regardless of the form of marginal distribution thus providing much more robustness and flexibility in practical applications.

I. INTRODUCTION

The stock market is one of the most important sources for companies to raise money. It is also considered the primary indicator of a country's economic strength and development. However, a great risk is involved on the part of investors due to the unpredictable nature of stock markets. The prediction of stock prices has always been a challenging task. Investors are always in search of profitable trading strategies which provide accurate signals of when to enter and exit the trade and how to invest i.e. long or short. Pair trading is one such investment strategy. It is a market neutral strategy meant to generate profit regardless of whether equities rise or fall. The strategy of matching a long position (buy) with a short position(sell) in two historically correlated stocks is known as pair trading. This paper is a study of a model developed using the copula approach of pair trading.

II. LITERATURE SURVEY

A. Pair Trading

Pair trading was developed to generate significant and consistent returns while controlling risk. It is a market neutral strategy meant to generate profit regardless of whether equities rise or fall. The strategy of matching a long position with a short position in two historically correlated stocks is known as pair trading. The buying of a security such as a stock with the expectation that the asset will rise in value is known as long position whereas the sale of a borrowed security with the expectation that the asset will fall in value is known as short position. The temporary weakening of correlation between the two paired stocks is the point of interest in pair trading. This can be caused by temporary changes in demand and supply or large buy and sell orders for one of the paired stocks. The divergence between paired stocks can also be caused due to one company being in the news for some important reason. Pair trading takes advantage of this temporary mispricing of the assets and hence, categorized as statistical arbitrage trading strategy. In this scenario, one stock is observed to go up while the other stock is observed to fall. Pair trading involves going short on the outperforming stock and going long on the underperforming one with the expectation that the paired stocks will converge over time. Hence, pair trading is a convergence trading strategy. [1]

For example, consider two companies A and B belonging to the same sector. Ideally, both A and B should observe similar rise or fall depending on their demand in the market. However, if the price of stock A were to go up significantly while B observed no change, the pair trading strategy would suggest buying stock B and selling stock A. If the price of stock B rose to close that gap in price, the trader would make money on stock B, while if the price stock A fell, he would make money on having shorted the stock A. Pair trading is thus a mean-reverting strategy with the expectation that prices will eventually return to their historical trends. It is a market-neutral strategy which helps reduce the market risk. Even if the market plummets on a particular day and the two stocks move down with it,

the trade results in a gain on the short position and a loss on the long position which is negated thus leaving the profit close to zero in spite of the large move.

B. Advantages of Pair Trading

1) Market Neutrality: - Pair trading is independent of market movements and hence it is a market neutral strategy. It has the ability to make profit whether the market is going up, down or sideways. This is because the strategy does not depend on market direction, but on the relationship between the two instruments.

2) Self - Funding: - Pair trading is self-funding since the short sale returns can be used to buy the long position.

3) Risk Control: - Pair trading is basically the matching of a long position with a short position of correlated instruments. Suppose a trader is long on stock A and short on stock B. Since the two stocks are correlated, if the entire sector suffer from losses the price may fall for both stocks. In this example, any losses sustained through the long position will be diminished by the gain in the short position. Therefore, even though the entire sector is down on the day, the traders' net position may remain neutral because of the low correlation to the market averages.

C. Traditional Pair Trading Strategies

1) Distance Approach: For paired securities/stocks, the co-movement in pair is determined by squared difference between two price series, knows as distance. If this distance is greater than certain threshold value i.e. if they diverge then the overpriced security is shorted and long position is taken on underpriced security. The positions are closed when they converge.

2) Cointegration Approach: In cointegration, two time series are linearly combined to a single time series. If the time-series deviates from its long standing mean then the long and short positions are taken assuming that the time-series will revert to mean.[2]

D. Copula

Copula is a statistical measure that represents a multivariate uniform distribution. It examines the association or dependence between random variables.[As defined by Investopedia, wikipedia]. Copula is a mathematical tool that can be used to solve many financial problems. However this concept wasn't used in finance till the late 90s. The notion of copula was first introduced in Sklar. Sklar's Theorem states that any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the variables.[3] For example, let H(x, y) be a joint distribution of two random variables X and Y with marginal distribution functions F and G. Then there exists a copula C such that

$$H(x, y) = C(F(x), G(y))$$

The converse also holds true. For any univariate distribution functions F and G and any copula C, the function H is a two-dimensional distribution function with marginals F and G. If F and G are continuous, then C is unique. Practical advantage of copula-based approach to modelling is that appropriate marginal distributions for the components of a multivariate system can be selected freely, and then linked through a suitable copula. That is, copula functions allow one to model the dependence structure independently of the marginal distributions. In pairs trading, the assumption of and the use of correlation or cointegration as a dependence measure is central. Some traditional techniques also assume symmetric distribution of data along the mean zero. However, the occurrence of non-linear distributions is quite frequent in financial assets. Negative skewness and/or excess kurtosis are also observed. This results in lower and upper tail dependencies. Outside the world of elliptical distributions, the use of linear correlation coefficient as a dependence measure is erroneous and may lead to misleading results. This may trigger wrong trading signals and may fail to recognize profit opportunities. Copulas are much more realistic. A significant shortcoming which they overcome is the non-assumption of normal distribution for all types of data and the use of different dependence measures other than the linear correlation coefficient. Copula can be applied regardless of the form of marginal distribution thus providing much more robustness and flexibility in practical applications. Modeling asset returns is one of the most important problems in finance. Copula provides a powerful framework for describing dependence structures without rigid assumptions. Copula measures lower and upper tail dependencies, considers linear and non-linear relationship and results in far richer set of information with the shape and nature of dependency between the stock pairs. Copula is invariant under strictly monotonic transformations and hence the same copula is obtained regardless of the form of data i.e. Price series and price series when converted to return series have the same copula.

The two main steps involved in the pair trading with copula technique is the selection of an appropriate copula for the stock pairs and identification of trading opportunities and relative positions between stock pairs. This separation of procedure is important as it provides greater flexibility and the analyst can use different marginal distributions to resolve the diversity in financial risks (Ane and Kharoubi, 2003). Decomposing the joint distribution into individual marginal distributions and a copula helps as it allows for the construction of better models of the individual variables than would be possible if only explicit multivariate distributions were considered. Applying the best-fitting marginal distribution ensures that all information regarding the dependence structure between random variables are accurately captured before estimating its joint distribution without rigid assumptions.[4][5]

III. METHODOLOGY

A. Data Downloading

Backtesting is one of the most commonly used approaches for testing a trading strategy. In order to confirm its effectiveness, simulations are to be done of the developed model/strategy on relevant past data that can be downloaded from websites providing financial data. Data can be downloaded using the Python language using modules such as webbrowser, urllib etc. The urllib module provides a high-level interface for fetching data across the World Wide Web [5]. Fetching data from a url involves calling the urllib.urlopen(url) function and reading from the returned stream object [6]. For example,

fetch_data = urllib.urlopen(“www.abc.com’’)
data = fetch_data.read()
The strings in data can be matched by using regular expressions by importing the ‘re’ module. The webbrowser module provides a high-level interface to allow displaying Web-based documents to users. The data for each constituent of the chosen index is to be downloaded in the following format
The data only at the end of the day is taken into consideration. Hence, this is a multiday trading strategy.

B. Compilation of required data

The downloaded data can now converted to a form which is easier to use and to work upon. One way of doing it is to create the following matrices.

1) Closing Price Matrix: Closing Price Matrix is the matrix of close prices of all the constituents of an index over a period of time. The rows correspond to the days whereas the columns correspond to the closing prices of different constituents of the index. Closing price may not be available for some dates. This may be because of non-existence of the stock. Insert -1 for such cases in the matrix.

2) Volume Matrix: Volume depicts the number of shares traded on a given day. Volume matrix is the matrix of trading volume of all the constituents of an index over a period of time. The rows correspond to the days whereas the columns correspond to the trading volume of different constituents of the index. As index cannot be traded, it is assigned a common value for all days. If volume is not available, insert -1 in the matrix.

3) Index Membership Matrix: Index Membership matrix is a matrix indicating the presence (or absence) of a constituent in the index at a given point of time. Indexes undergo changes in their constituents. Market condition of the stock is one of the reasons for inclusion or exclusion of a stock from its index. Stocks which were a part of the index few years back may not be present in the index anymore and vice versa. The rows correspond to the days whereas the columns correspond to different constituents of the index (present and past). One in the index membership matrix indicates that the stock was present in the index on the given day whereas zero indicates that the stock was not a part of index on that particular day.

C. Data checking

This step involves checking data for its errors, rectifying them and finally creating rectified matrices for close price and volume.

1) Data Missing Cases: Closing price and volume when not available have been marked with -1 in their respective matrices. Consider the data from the most reliable website (Website1) as the reference data. If closing price (or volume) on day t is not present and marked as -1 in the main matrix but the data is available for that day on another website (Website2) then it can be rectified in the main matrix by

$x = \text{Website2}(t) - \text{W ebsite2}(t-1)$
$\text{Website1}(t) = \text{Website1}(t-1) + x$

2) Checking Splits and Bonus: Sometimes a company divides its existing shares into multiple shares. This action is called stock split and it increases the number of the company's outstanding shares by dividing each share, which in turn reduces its price. Sometimes a company issues additional shares to its current shareholders in proportion to the number of shares that each shareholder owns, without him having to pay any additional cost. When a company declares such splits and bonuses the stock charts show dramatic movement(if the data is not adjusted). The data can be checked for such splits and bonuses in two steps:

1. Visual Checks - Data can be checked visually for drastic movement.
2. Checking with a Script - A script to find percentage change between closing prices for consecutive days can be used to identify splits and bonuses. A change of 10% or more is a considerable change and may be due to splits or bonuses which are to be adjusted using adjusting factor.

D. Pairs formation

The fundamental task involved in pair trading is the selection of pairs. Stock pairs with similar characteristics help resolve the issue of the true value of stocks being frequently unknown. It has been observed that highly correlated pairs come from the same sector.

Date	Open	High	Low	Close	Volume Traded

E. Model Development

In any approach used for pair trading there are two different periods: the formation period and the trading period. Pairs are formed during the formation period by observing the price behavior from the historical data. In the copula approach for pair trading, appropriate distributions and relevant parameters are determined during the formation period. These estimates from the formation period are used to identify profitable trades during the trading period.

1) Calculating returns: Price returns are the rate of returns on an investment portfolio where only the capital appreciation is considered while the income generated due to dividend and interests is not taken into account. Benefit of using returns over the normal price series is that returns provide normalization. Statistical analysis and machine learning require evaluation of analytic relationships amongst two or more variables. Normalization

measures all variables in a comparable metric despite originating from price series of unequal values thus making it conducive to statistical analysis. There are two types of returns

1. Simple Returns - Simple Returns are given by

$$\text{Simple Returns} = \frac{P(t) - P(t-1)}{P(t-1)}$$

Where, P(t) is the closing price of the stock on day t.

2. Logarithmic Returns - Logarithmic Returns are given by

$$\text{Logarithmic Returns} = \frac{P(t) - P(t-1)}{P(t-1)}$$

Where, P(t) is the closing price of the stock on day t.

The price series is to be converted to logarithmic returns for further computations.

2) Rolling Window: The concept of rolling window requires one to set the window size. Window size specifies the number of days to be considered at a time. If the window size is set to 120, start with day 121. If i = 121, consider (i−120) to i days each time and then increment i.

I.e. on day 121, consider the data from day 1 to day 121 for analysis
on day 122, consider the data from day 2 to day 122 and so on.

3) Finding Uniform distribution: Copula is a multivariate probability distribution which requires the marginal probability distribution of each variable to be uniform. Uniform distribution is the probability distribution where the random variable assumes each of its values with an equal probability. For example when a fair die is tossed, each element of the sample space 1, 2, 3, 4, 5, 6 occurs with probability 1/6. Empirical distribution function can be used as it steps up by 1/n at each of the n data points.[6]

4) Model Selection: This step involves finding the distribution which adequately fits the data. We consider the two parametric families of copula: Elliptical copula, Archimedean copula

1. Elliptical Copula

Elliptical copula are those relating to elliptical distributions. They are also known as implicit copulas. These copulas follow linear dependence structure. The linear correlation coefficient is used to measure their linear dependence. They do not have simple closed forms but are extracted from multivariate distribution functions using Sklar's theorem. The asset returns which are well represented as elliptical distributions belong to elliptical copulas. Most common elliptical copulas are Gaussian and Student-t copulas. Normal Copula or the Gaussian copula is the copula of the multivariate normal distribution. The distribution of a normal random variable with mean 0 and variance 1 is called standard normal distribution. It is a bell shaped curve. In contrast with the normal copula, student-t copula can model tail dependence. It allows for joint fat tails and an increased probability of joint extreme events compared with the Gaussian copula. Two parameters namely linear correlation and degree of freedom are used to capture the correlation between the risk classes. Increasing the value of decreases the tendency to exhibit extreme co-movements. Hence, student-t copula is generally considered a better solution to model the dependence structure for operational risk data.

The standard bivariate Students t-distribution has a disadvantage that both the marginal distributions must have the same tail heaviness which may not be true for stock returns. It does not account for asymmetries in data. Hence, we need to study another class of parametric copulas which is the Archimedean copulas.

2. Archimedean Copula

One of the key disadvantages of elliptical copulas is that they do not have closed form expressions. Archimedean copulas, on the other hand, are not derived from multivariate distribution functions but admit an explicit formula. They are also known as explicit copulas. Archimedean copulas have non - elliptical distribution and follow non - linear dependence structure. They describe stronger dependence between extreme data. They can model dependence with only one parameter monitoring the strength of dependence even in high dimensional cases. They make use of Kendall Tau for calculating dependency. Although restrictive for higher dimensional cases, they fit bivariate return distributions extremely well. As the project demands analyzing stocks in pairs, only bivariate distributions are generated and the use of Archimedean copulas is completely justified. Clayton, Gumbel and Frank are asymmetric

Archimedean copulas. Clayton copula exhibits greater dependence in the negative tail than in the positive while Gumbel copula exhibits greater dependence in the positive tail than in the negative. Clayton copula produces a tight correlation at the low end of each variable. It interpolates between perfect negative and perfect positive dependence.[7] Gumbel Copula produces more correlation at the two extremes of the correlated distribution but has its highest correlation in the maxima tails.[8] Choice of copula forms an important step. Selection of the best copula can be made by maximizing the likelihood function value. Consider the five copulas explained above - Normal, Student-T, Clayton, Gumbel, Frank. Calculate maximum log likelihood and subsequently the value of information criterion for all five copulas. Copula with the lowest value of information criterion is selected for the current window.

5) Conditional probabilities: This step involves calculation of conditional probabilities using for the copula selected in the above step. As we consider pairs of stock each time, one stock can be taken as U and the other stock as V. Conditional Probabilities P ($U \leq u|V = v$) and P ($V \leq v|U = u$) can be calculated by taking the last element of current window for stock 1 as u and the last element of current window for stock 2 as v.

1. **Normal Copula**

The formula for $P(U \leq u|V = v)$ is given as

$$\Phi\left(\frac{\Phi^{-1}(u) - \theta\Phi^{-1}(v)}{\sqrt{1-\theta^2}}\right)$$

The formula for $P(V \leq v|U = u)$ is given as

$$\Phi\left(\frac{\Phi^{-1}(v) - \theta\Phi^{-1}(u)}{\sqrt{1-\theta^2}}\right)$$

Here,
φ is cumulative distribution function and θ is the correlation coefficient.
$\theta = 1$ indicates perfect positive correlation
$\theta = 0$ indicates no correlation
$\theta = -1$ indicates perfect negative correlation
For a random variable X cumulative distribution function is given by $F(x) = P(X \leq x)$ which is the probability that X takes a value less than or equal to x. For continuous distributions, like in the case of normal distribution, it gives the area under the probability density function from $-\infty$ to x. All observations of any random variable X can be converted to observations of a normal random variable Z with mean 0 and variance 1. This can be done by
$z = (x - \mu)/\sigma$

When X takes the value x, Z will take the value z given by the above formula. This z when looked up in the normal probability table gives the area under the curve to the left of z which is the probability of finding random variable X less than or equal to x[9].
$\Phi^{-1}(u)$ is the inverse of the standard univariate Gaussian distribution function. It gives the z value corresponding to which lies an area of u to the left in the normal probability curve.

2. **Student -T Copula**

The formula for $P(U \leq u|V = v)$ is given as

$$t_{\theta 1} + 1\left(\sqrt{\frac{\theta_1 + 1}{\theta_1 + t_{\theta 1}{}^{-1}(v)^2}} * \frac{t_{\theta_1}^{-1}(u) - \theta_2 t_{\theta 1}{}^{-1}(v)}{\sqrt{1-\theta_2^2}}\right)$$

The formula for $P(V \leq v|U = u)$ is given as

$$t_{\theta 1} + 1\left(\sqrt{\frac{\theta_1 + 1}{\theta_1 + t_{\theta 1}{}^{-1}(u)^2}} * \frac{t_{\theta_1}^{-1}(v) - \theta_2 t_{\theta 1}{}^{-1}(u)}{\sqrt{1-\theta_2^2}}\right)$$

Here,
θ_1 is the degree of freedom which is given as sample size - 1
θ_2 is the linear correlation coefficient
$t_{\theta 1}(T)$ - gives the area α under the T distribution curve to the left of the T value for θ_1 degrees of freedom
$T = \bar{x}-\mu/(S/\sqrt{n})$
$t^{-1}_{\theta 1}(u)$ is inverse of the standard univariate Student-T distribution with θ_1 degrees of freedom, expectation 0 and variance θ_1.
It gives the T value that leaves area α to the left where $\alpha = t_{\theta 1}(u)$

3. Clayton Copula

The formula for $P(U \leq u|V = v)$ is given as

$$v^{-(\theta+1)}\left(u^{-\theta}+v^{-\theta}-1\right)^{-\frac{1}{\theta}-1}$$

The formula for $P(V \leq v|U = u)$ is given as

$$u^{-(\theta+1)}\left(u^{-\theta}+v^{-\theta}-1\right)^{-\frac{1}{\theta}-1}$$

Here, the parameter θ defines the degree of dependence between the variables using Kendall's Tau and can be given as $\boldsymbol{\theta} = 2 * \tau/1-\tau$
$\theta \in [-1, +\infty)\setminus\{0\}$
$\theta = \infty$ indicates perfect dependence[10].

4. Gumbel Copula

The formula for $P(U \leq u|V = v)$ is given as

$$C_\theta(u,v) * \left[(-lnu)^\theta+(-lnv)^\theta\right]^{\left(\frac{1-\theta}{\theta}\right)} * (-lnv)^{\theta-1} * \frac{1}{v}$$

The formula for $P(V \leq v|U = u)$ is given as

$$C_\theta(u,v) * \left[(-lnu)^\theta+(-lnv)^\theta\right]^{\left(\frac{1-\theta}{\theta}\right)} * (-lnu)^{\theta-1} * \frac{1}{u}$$

Here, the parameter θ defines the degree of dependence between the variables using Kendall's Tau and can be given as
$\theta = 1/1-\tau$
$\theta \in [1, +\infty)$
$\theta = \infty$ indicates perfect dependence
$\theta = 1$ indicates independence

5. Frank Copula

The formula for $P(U \leq u|V = v)$ is given as

$$\frac{e^{-\theta v}\left(e^{-\theta u}-1\right)}{\left(e^{-\theta}-1\right)+\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}$$

The formula for $P(V \leq v|U = u)$ is given as

$$\frac{e^{-\theta u}\left(e^{-\theta v}-1\right)}{\left(e^{-\theta}-1\right)+\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}$$

Parameter θ is given by

$$\frac{[D_1(\alpha)-1]}{\alpha} = \frac{1-\tau}{4}$$

D1(θ) is Debye function with n = 1.
Debye function is given as

$$D_n(x) = \frac{n}{x^n}\int_0^x \frac{t^n}{e^t-1}\,dt.$$

$\theta \in [-\infty, \infty)\setminus\{0\}$

6) Thresholds that trigger the trades: The conditional probabilities calculated in the previous step act as primary thresholds for trade selection. For example, if we consider stock - index pairs and decide 0.25 and 0.75 as the thresholds, buy pair will opened when

Stock Conditional Probability given Index < 0.25 or
Index Conditional Probability given Stock > 0.75

and will be closed when

Stock Conditional Probability given Index > 0.5 if trigger due to stock or
Index Conditional Probability given Stock < 0.5 if trigger due to index

Similarly, sell pair will be opened when

Stock Conditional Probability given Index > 0.75 or
Index Conditional Probability given Stock < 0.25

and will be closed when

Stock Conditional Probability given Index < 0.5 if trigger due to stock or
Index Conditional Probability given Stock > 0.5 if trigger due to index

7) Conditions to be satisfied for trade selection:

1. Number of trades - Count of maximum number of trades allowed should not be exceeded.
2. Index membership - The stock should be present in index membership matrix as otherwise all the parameters calculated with respect to index are erroneous.
3. Extreme values of conditional probability – Extreme values of correlation (nearer to 0 or nearer to 1) indicate the weakening of correlation among pairs which is a conducive property for the start of trade. Absolute value of (0.5 - conditional probability value) is calculated and trades with highest values are selected.
4. Maximum trades per stock - A given stock may form pairs with a number of stocks. However, we need to restrict the number of trades that it can be involved in on a particular day by specifying a threshold.
5. Volume case - If a company is in news, the trade may not be mean reverting. Such cases can be identified by checking volume. If the volume is more than double of the average of the last 10 days of the stock then it is advisable to not trade such stocks or square off the trade if the position is already holding.

8) Formula to calculate Profit and Loss:

Buy Pair Open

a. $\text{stock\_pnl\_return} = \frac{\text{close\_price\_stock}}{\text{previous\_close\_price\_stock}} - 1$

stock_pnl_return gives the absolute return. It is the percentage measure of profit or loss of capital invested on stock. For example if the previous_close_price_stock is 100 and the current close_price_stock is 105 then by above formula we get stock_pnl_return as 5% which indicates 5% profit. If the current_close_price_stock is 98 then by above formula we get stock_pnl_return as −2% which indicates 2% loss.

b. stock_pnl_$ = previous_stock_pnl_$ * (1 + stock_pnl_return)

stock_pnl_$ gives the the progressive appreciation or depreciation of the capital invested in stock since the start of trade till the current day. 0.5 is usually assumed to be invested in the stock at the beginning of the trade.

c. $\text{index\_pnl\_return} = -1 * \left(\frac{\text{price\_index}}{\text{previous\_price\_index}} - 1\right)$

index_pnl_return gives the absolute return. It is the percentage measure of profit or loss of capital invested on index. If the previous_price_index is 100 and current_price_index is 102 then it is a 2% profit. However, as the buy pair is open, the investor is buying stock and selling index which is a 2% loss to the
investor. Hence, the expression is multiplied by -1 to negate the effect of absolute return.

d. index_pnl_$ = previous_index_pnl_$ * (1 + index_pnl_return)

index_pnl_$ gives the the progressive appreciation or depreciation of the capital invested in index since the start of trade till the current day. 0.5 is usually assumed to be invested in index at the beginning of the trade.

e. total_money = stock_pnl_$ + index_pnl_$

total_money gives the progressive appreciation or depreciation of the capital invested in the stock-index pair as a whole since the start of trade till the current day. As 0.5 is assumed to be invested in stock and 0.5 in index the total money invested in pair at the beginning of the trade is 1.

f. pair_pnl_return = $\left(\frac{\text{previous\_stock\_pnl\_\$}}{\text{previous\_total\_money}} * \text{stock\_pnl\_return}\right) + \left(\frac{\text{previous\_index\_pnl\_\$}}{\text{previous\_total\_money}} * \text{index\_pnl\_return}\right)$

pair_pnl_return gives the absolute return of pair.

g. pair_pnl_$ = (stock_pnl_$ − previous_stock_pnl_$) + (index_pnl_$ − previous_index_pnl_$)

pair_pnl_$ = total_money – prev_total_money

pair_pnl_$ gives the profit / loss made in a day by trading the stock - index pair.

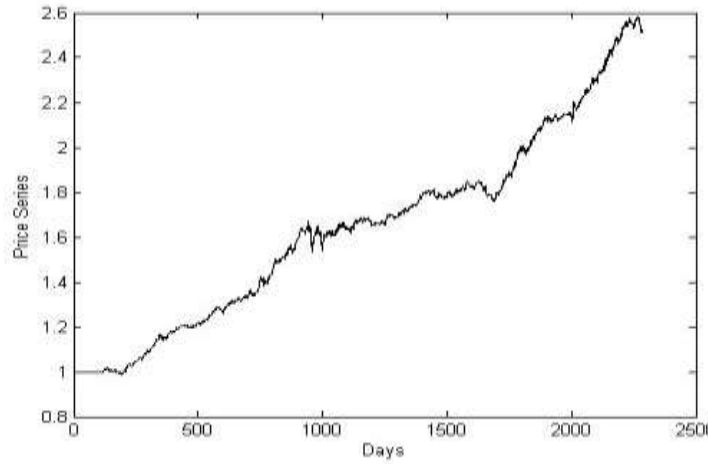
Sell pair open - Following are the changes if sell pair is open

a. stock_pnl_return $= -1 * \left(\frac{\text{close\_price\_stock}}{\text{previous\_close\_price\_stock}} - 1\right)$

b. index_pnl_return = $\frac{\text{price\_index}}{\text{previous\_price\_index}} - 1$

IV. RESULTS AND OBSERVATIONS

A 9 year period data starting from 1 January 2005 for the CNX NIFTY index was downloaded and used for analysis. The above methodology was implemented by forming stock-stock pairs and stock-index pairs and an appropriate copula was selected each time which best satisfied the dependence features of the data. Results were calculated by varying the parameters such as size of window, conditional probability thresholds, maximum number of active pairs allowed on a given day and maximum number of trades allowed per stock on a given day.
Given below is the graph of returns for the stock – stock pairs of CNX NIFTY. Window size was set to 120, 0.25 and 0.75 were the conditional probability thresholds, at most 12 pairs could be active on a day and a stock could be involved in at most two trades.



The above combination of parameters gives 10.7% annual return for CNX NIFTY. Annual return can be calculated as

PriceSeries = LogReturnToPrice(PortfolioReturn)

$$\text{AnnualReturn} = \left(\frac{\text{PriceSeries(Last Element)}}{\text{PriceSeries(1)}}\right)^{\left(\frac{252}{\text{length (Pirce Series)}} - 1\right)}$$

CONCLUSION

A model to suggest profitable strategy by analyzing data and providing trading signals has been developed using machine learning, copula and other statistical analysis techniques. Pairs trading can seek profits from the difference in price change between two related instruments when the pairs are appropriately formed. The Copula function instead of assuming normality of financial data, describes the dependency between the two instruments by relating the joint distribution with their most appropriate marginal distributions. Thus use of copula for implementing pairs trading exhibits high profits.

ACKNOWLEDGEMENT

The authors wish to express their heartfelt thanks to Dr. Aniruddha Pant and Mr. Rohit Walimbe of Algo Analytics Financial Consultancy Pvt. Ltd. for their immense contribution in improving the quality of this paper.

REFERENCES

[1]. Trading Strategies With Copulas; Authors: Yolanda Stander, Danil Marais, Ilse Botha
[2]. Algorithmic Pairs Trading: Empirical Investigation of Exchange Traded Funds; Author: Miika Sipil
[3]. Using Gaussian Copulas in Supervised Probabilistic Classification; Authors: Rogelio Salinas-Guti rrez, Arturo Hern ndez- Aguirre, Mariano J. J., Rivera-Meraz, and Enrique R. Villa-Diharce
[4]. Modelling the dependence structure of financial assets: A survey of four copulas; Author: Kjersti Aas
[5]. Pairs Trading: A Copula Approach; Authors: Rong Qi Liew, Yuan Wu
[6]. Asymptotic Statistics; Authors: A. W. van der Vaart
[7]. Quantitative Risk Management: Concepts, Techniques, and Tools; Authors: Alexander J. McNeil, Rdiger Frey, Paul Embrechts
[8]. Measuring Operational and Reputational Risk: A Practitioner's Approach; Authors: Aldo Soprano, Bert Crielaard, Fabio Piacenza, Daniele Ruspantini
[9]. Probability and Statistics for Engineers and Scientists; Authors: Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers and Keying
[10]. Vose SoftwareT M Reference Number: M-M0788-A

About Authors

Apoorva Uday Nayak is a final yearstudent pursuing Bachelor degree at the College of Engineering, Pune. Her current areas of interest include machine learning, data mining, finance, probability and statistics. She was an intern at the Barclays PLC.

Megha Ugalmugale is a final year student pursuing Bachelor degree at the College of Engineering, Pune. Her current areas of interest include machine learning, data mining, UNIX, finance, probability and statistics. She was an intern at the Barclays PLC.

Vishwanjali Gaikwad is a final year student pursuing Bachelor degree at the College of Engineering, Pune. Her current areas of interest include machine learning, data mining.

Dr. Vahida Z. Attar is an Assistant Professor in Department of Computer Engineering and IT at the College of Engineering, Pune. Her current areas of interest include machine learning, data mining, information retrieval.

APPENDIX

Information Criteria:

As mentioned earlier, the two main steps involved in the pair trading with copula technique is the selection of an appropriate copula for the stock pairs and identification of trading opportunities and relative positions between stock pairs. Visually, the type of distribution (normal, student-t, clayton, gumbel, frank) into which the data fits can be determined by plotting the graphs and selecting the corresponding copula of the distribution which resembles the graph. However doing this computationally is not possible and some other methods other than visual aids are required. Information Criteria is a test which helps in selection of a model which that adequately fits data. We look at the following three commonly used information criterion.

Akaike Information Criterion (AIC)

$$AIC = 2 * k - 2\ln[Lmax]$$

Bayesian information criterion (BIC)

$$BIC = \ln[n]k - 2\ln[Lmax]$$

HannanQuinn information criterion(HQIC)

$$HQIC = 2\ln[\ln[n]]k - 2\ln[Lmax]$$

where,
n = number of observations (e.g. data values, frequencies)
k = number of parameters to be estimated (e.g. the
Normal distribution has 1: linear correlation, Student-t distribution has 2: linear correlation and degree of freedom)

Lmax = the maximized value of the Likelihood (fit the parameters by MLE and record the natural log of the Likelihood) for the estimated model.

Any one of the above criterion can be used to find the model which adequately fits the data. It can be observed that the above three formulae are closely related and all three of them employ the usage of maximum log likelihood.

Calculate the values of the selected criterion for all candidate models are found using the above formulae. The model with the lowest value is selected as the model which fits the data to the nearest. Lmax can be increased by increasing the number of parameters (k). An increase in k increases the value of criterion. Hence, Information Criteria not only value goodness of fit but discourage over fitting i.e. having too many parameters relative to the size of data.

Beta condition:

Equal weights have been assigned to stock and index while calculating profit and loss. A more refined way is to assign β-weights. β-weights measure the volatility of a stock as compared to the index. B is calculated using regression analysis. Regression analysis is a statistical process of determining the relationship between two variables. Linear regression uses one independent variable to predict the outcome of the dependent variable. The relationship between index and stock is given by
$Y= \alpha+ \beta*X$

Considering stock - index pairs
$Returnstock = \alpha+ \beta*$ Returnindex

If β is 1 then it means that if the index goes up by 1% then stock also goes up by 1%. If β is 2 then it means that if the index goes up by 1% then stock goes up by 2%. This means that the stock is more volatile and should be assigned less weight. The larger the absolute value of the β weight, the more influence the stock has on the market index. β-weights for a stock-index pair are calculated as:
Weight of Stock= $1/(1+ \beta)$; Weight of Index= $\beta/(1+ \beta)$

Kendall Tau
Kendall's Tau coefficient is a statistic used to measure of the strength of dependence between two quantities. It reflects the association between the two quantities when ranked by each of the quantities. Consider two samples X and Y each of size n and each containing distinct elements.
Let X = {x1, x2, x3, ...,xn} and Y = {y1, y2, y3, ..., yn}
(x1, y1),(x2, y2),(x3, y3), ...,(xn, yn) are the joint set of observations for X and Y .
The total number of possible pairings between the observations of X and Y is $n*(n-1)/2$.

The Kendall's Tau coefficient is defined as

where,
a pair (xi, yi),(xj, yj) is concordant if (xi>xj and yi>yj) or (xi<xj and yi<yj) ;
a pair (xi, yi),(xj, yj) is discordant if(xi>xj and yi<yj) or (xi<xj and yi>yj)
As the denominator represents the total number of possible pairs, the value of Kendall's Tau should lie in between $-1 \leq \tau \geq 1$.