A Study of Non-Linear Pair Trading Strategy using Mixed Copula and Machine Learning

IMMEDIATE

Compiled February 28, 2023

Over the past years, researchers have found numerous limitations of traditional linear pair trading methods, mainly because it only applies to a small number of assets. In this paper, our team presents a non-linear pair trading strategy that generates a positive alpha based on historical data. In this paper, we explain the mixed-copula and machine learning techniques used to select two index pairs MID-NDX and RUT-N225, and develop a new strategy that outperforms canonical traditional linear pair trading methods. In the two-year data backtest, our mixed-copula strategy significantly outperforms benchmark strategies.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

Pairs trading is a well-known trading strategy developed in the 1980s and a fundamental equity investment tool used by many hedge funds. The traditional strategy consists of two fundamental steps. First, the strategist needs to identify asset pairs or clusters, usually involving equities or market indices, that display similar return patterns. Second, the strategist needs to statistically summarize the observed relationship and deploy automated trading strategies based on concepts such as mean reversion, relative value, and spread trading. In the 1980s, this technical method was extremely popular. However, an essential drawback of this strategy is it heavily relies on the correlation coefficient as a form of dependency. Though seemingly useful in applications, this can give inaccurate trading signals for data that are not normally distributed. Moreover, as the statistical arbitrage depletes, individual and institutional investors are discouraged by the high commission involved in the process and have gradually phased out the traditional correlation- and co-integration-based method because in reality most of the cross-price interactions are non-linear.

In this paper, we examined the possibility of applying the combination of copula and machine learning methods. We first utilize supervised machine learning on optimizing multivariate regression and then construct a copula mix to derive a model that generates trading signals. This allows us to accurately summarizes the non-linear interactions between the indices' empirical multivariate distributions. In addition, we incorporated FastDTW, a linear-model-based supervised machine learning model to evaluate potential index pairs and expedite the pair selection process prior to mix-copula model training. We hereby demonstrate the potency of mix-copula pairs trading and enhance the preceding pair-selection process with dynamic time wrapping and time series clustering.

Furthermore, we tested and tuned various parameters and hyper-parameters, including thresholds for signal generation and frequencies of model updates, in an effort to stress test our model's robustness and optimize the return. The yielded strategy is proved to be superior compared to both traditional and advanced linearity-based pairs trading methods including distance-based rolling regression pairs trading, and Kalman Filter-based pairs trading.

Section 2 summarizes the past research conducted on the subject. Section 3 explains our methodology, while we explain our data source and pre-processing in Section 4. Section 5 discusses the results obtained using the Copula method and Section 6 both concludes and elaborates on possible further research.

2. LITERATURE REVIEW

Traditional pair trading strategies are usually based on linear models. In most cases, traditional pairs trading strategies are market neutral and are good for risk diversification. They are also used as an additional tool in the process of investment portfolio asset allocation. Some of the most popular statistical tools used are distance and co-integration methods and the most popular process used in the field of stochastic processes is the co-integration[3]. It was shown in past research that pair trading strategy generated by the distance method and co-integration yield residual series with better properties compared to other traditional pair trading methods. However, to avoid the over-dependence on correlation coefficient, the investors should program the strategy to incorporate different time frames to include the high volatility of the market during economic crisis [2]. More, the distance and co-integration approach only

capture the dependencies between stock prices with elliptically distributed random variables, and situations like that are rare in practice [1]. This motivates us to use copula-based models to address the current need for non-linear-model-based strategies in pair trading.

The object of pair trading strategy using copula is to apply optimal copula between two stock returns and use the relative positions between stock pairs to obtaining a winning position. In most cases, the copula approach requires the investors to identify the marginal distributions, the relevant copula function, and the conditional probability distribution functions. The marginal distribution functions and the related parameters can be obtained by the values of the cumulative log returns. Using the cumulative distribution function values obtained by each stock, the investors can then select the copula function to use. Finally, the investors need conditional probability functions. Stocks are identified as being relatively undervalued if the conditional probability is less than 0.5 and relatively overvalued if the conditional probability is greater than 0.5. In addition, the values of the conditional probabilities are also an indication of certainty or confidence regarding the position of stocks. Therefore, the execution of trade should be done when one of the conditional probabilities is close to 1.

3. METHODOLOGY

A. Copula

Copula has been widely used to model bivariate distributions, specifically to capture co-movements of asset prices or returns in finance. Before the financial crisis, David X. Li found a way to model the correlation of events, such as the default correlation among a pool of mortgage-backed securities, by the Gaussian copula function. The Gaussian copula was popular for its capacity of security valuation. Yet unfortunately, it largely underestimated the actual correlation among MBS during the financial crisis and indirectly caused the collapse of the financial market in 2008[9]. Later more copula models such as Clayton copula, Joe copula, and Frank copula are used in the modeling of correlation among two or even more variables. The nature of the copula function provides an alternative way to explain the non-linear relationship among assets and therefore this approach has been studied and applied by researchers in pair trading.

A.1. Definition

A copula function maps univariate marginals to multivariate distributions which helps to model the relationship among variables Fig 1. For bivariate copula, it is defined as $C : [0,1] \times [0,1] \rightarrow [0,1]$. Based on Sklar's Theorem, which states that all the multivariate distributions can be written as the function of univariate distributions, and a copula function can be used to describe the dependence structure among those variables. Given two uniform random variables U_1 and U_2, their joint distribution can be written as a bivariate copula function:

$$C(u_1, u_2; \theta) = P(U_1 \leq u_1, U_1 \leq u_2) \tag{1}$$

Thus, for any given two univariate distributions and corresponding cumulative density functions $F_1(x_1)$ and $F_2(x_2)$, the copula function can be written as:

$$C(F_1(x_1), F_2(x_2); \theta) = F(x_1, x_2) \tag{2}$$

As a result, the bivariate distribution can be specified in terms of two arbitrary univariate marginal distributions.

The conditional probabilities generated based on copula functions are crucial for pairs trading and are defined as:

$$P(U_1 \leq u_1 | U_2 = u_2) = \frac{\partial C(u_1, u_2)}{\partial u_2} \tag{3}$$

$$P(U_2 \leq u_2 | U_1 = u_1) = \frac{\partial C(u_1, u_2)}{\partial u_1} \tag{4}$$

and the copula density $c(u_1, u_2)$ is defined as:

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2} \tag{5}$$

Copulas are also defined by two categories: Archimedean Copulas and Elliptical Copulas. The most commonly used copulas, such as Clayton, Gumbel, Joe, and Frank, are Archimedean Copulas. Gaussian and Student-t copulas are two famous Elliptical Copulas. A bivariate copula falls into the category of Archimedean Copulas if it is generated as:

$$C(u_1, u_2; \theta) = \phi^{-1}(\phi(u_1; \theta) + \phi(u_2; \theta)) \tag{6}$$

where $\phi : [0,1] \times \theta \rightarrow [0, \infty)$. ϕ is called the generator function for the copula and ϕ^{-1} is its inverse function. The generator function is used to distinguish one copula from others. The generator functions for common copulas are listed below:

1. Gumbel: $\phi(u, \theta) = (-\ln u)^{\theta}, \theta \in [1, +\infty)$
2. Frank: $\phi(u, \theta) = -\ln \frac{e^{u\theta}-1}{e^{-\theta}-1}, \theta \in (-\infty, +\infty)$
3. Clayton: $\phi(u, \theta) = \frac{u^{-\theta}-1}{\theta}, \theta \in [-1, +\infty)$
4. Joe:$\phi(u, \theta) = \ln 1 - (1-u)^{\theta}, \theta \in [-1, +\infty)$

The Elliptical copulas, mainly Gaussian copula and Student-t copula, can be represented in a similar way. The bivariate Gaussian copula with correlation matrix $R \in [-1,1]^{d \times d}$ can be defined as:

$$C_R(u) = \Phi_R(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \tag{7}$$

where Φ_R is the joint Gaussian distribution with R as its correlation matrix and Φ^{-1} is the inverse of the cumulative density function of a standard normal distribution. The Student-t, therefore, can be defined as:

$$C_{R,v}(u) = \Phi_{R,v}(\Phi_v^{-1}(u_1), \Phi_v^{-1}(u_2)) \tag{8}$$

where v is the degrees of freedom in the Student-t distribution.

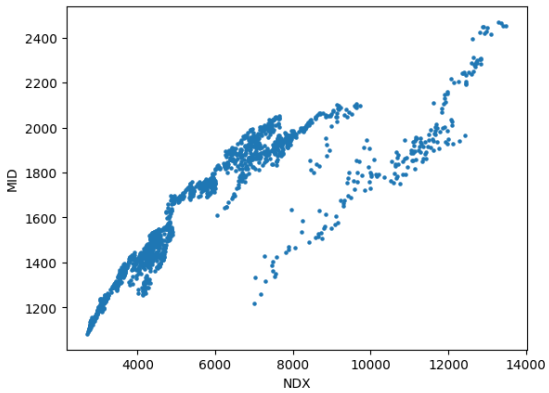

Fig. 1. Scatter Plot of MID-NDX Performance After Uniform Transformation: X: NDX Closing; Y: MID Closing; data range: 2013/02/07 - 2021/01/25

A.2. Mixed Copula

In this paper, a mixture of Archimedean Copulas will be used to model the relationship between two indices. The conditional probability will be employed to develop the trading strategies. A mixed copula is simply a linear combination of a group of copulas with respective positive weights summing up to one. In the mixed copula, θs need to be specified for each component. Therefore, the mixed copula can be written as:

$$C_{mix}(u_1,u_2,\theta,w) = w_1C_1(u_1,u_2,\theta_1)+...+w_nC_n(u_1,u_2,\theta_n) \quad \textbf{(9)}$$

where $\sum_{n=1}^{\infty} w_n = 1$ is required.

Clayton-Gumbel-Joe(CGJ) mixed copula is adopted to develop the model. These three copulas are chosen since Clayton copula mainly captures the left tail dependence between two variables, while Gumbel and Joe copulas are focusing more on the right tail dependence. Therefore, by combining these three copulas, it may effectively capture the tail dependence between two indices that are ignored by the traditional Gaussian copula. The CGJ model can be defined as:

$$w_{CGJ} = w_cC_c(u_1,u_2,\theta_c)+w_gC_g(u_1,u_2,\theta_g)+w_jC_j(u_1,u_2,\theta_j) \quad \textbf{(10)}$$

A.3. Training Algorithm

All the data of the pairs studied in this paper are ranging from February 7, 2013, to January 20, 2023. The data in the training data set ranges from February 7, 2013, to January 25, 2021, and the data in the testing data set ranges from January 27, 2021, to January 20, 2023. To find the parameters and weights in a mixed copula, the maximum likelihood estimator method is used to search for the optimal value of parameters and weights. The maximum likelihood estimator method aims to maximize the following equation:

$$Q(\phi) = \sum_{t=1}^{T} log(\sum_{n=1}^{3} w_nC_n(u_{1t},u_{2t},\theta_n)) + \delta(\sum_{n=1}^{3} w_n - 1) \quad \textbf{(11)}$$

where $\phi = (w^T,\theta^T)^T$ is a vector of all parameters and weights involved. The Pycop package is employed here to identify the optimal parameters and weights for the mixed copula model.

However, when the U_1 and U_2 are nearly linearly related to each other, the normal maximum likelihood method may not yield satisfying results. Pycop package in our strategy utilizes Scipy optimizers that assume the twice differentiability of expressions. Yet these bivariate distributions are nearly linear after transformation. This will incapacitate model convergence.

Therefore, a numerical method using the Expectation-Maximization(EM) algorithm is adopted to find the optimal parameters and weights. The EM algorithm needs to maximize the likelihood function defined as:

$$Q(\phi) = \sum_{t=1}^{T} log(\sum_{n=1}^{3} w_nC_n(u_{1t},u_{2t},\theta_n))$$

$$-T\sum_{n=1}^{3} p_{\lambda,\alpha}(w_n) + \delta(\sum_{n=1}^{3} w_n - 1) \quad \textbf{(12)}$$

where $p_{\lambda,\alpha}$ is is the smoothly clipped absolute deviation (SCAD) penalty function proposed by Fan and Li in 2001 [4]. λ and α are the tuning parameters. The penalty function is able to drive the under-performed copula components to 0. The SCAD penalty function component is usually defined by its derivative:

$$p'_{\lambda,\alpha}(w_n) = \lambda\{I(\beta \le \lambda) + \frac{(\alpha\lambda-\beta)_+}{(\alpha-1)\lambda}I(\beta > \lambda) \quad \textbf{(13)}$$

Under the E-step, we iteratively calculate the following equation using the old θ and w to find the new weights:

$$w_{new,k} = (w_kp'_{\lambda,\alpha}(w_k) - \frac{1}{T}\sum_{t=1}^{T}\frac{w_kC_k(u_{1,t},u_{2,t},\theta_k)}{\sum_{n=1}^{3} w_nC_n(u_{1,t},u_{2,t},\theta_n)}) \quad \textbf{(14)}$$

After calculating the weights for each copula component in the mixed copula model, θ for each copula component should be generated under the M-step to maximize the likelihood function. The Newton-Raphson method is used to find the optimal θ for each copula component. Then iteratively repeat the E-step and M-step until weights and copula parameters converge.

A.4. Trading Strategy and Backtesting

In the belief that low probability events (i.e. events outside Neutral Band) would eventually drop back to below thresholds, we developed a neutral position arbitrage strategy to conduct the pairs trading. Our underlying assumptions are infinite divisibility of assets, zero tradings commissions, and no slippage.

The trading strategy is a simple price threshold strategy based on conditional probability generated by copulas. The conditional probabilities are calculated by differentiating the cumulative copula function $C(u_1,u_2,\theta)$ in terms of U_1 and U_2 as equations (3) and (4).

Given two indices U_1 and U_2, the prices observed from the market at end of the day are u_1 and u_2 respectively. Therefore, the conditional probability will indicate whether one index is overvalued compared to the other one. The judgment of undervalue and overvalue will be based on the threshold parameter preset in the model. The trading signal generation process and the neutral band are as follows:

1. If $P(U_1 \leq u_1 | U_2 = u_2) <$ lower threshold **and** $P(U_2 \leq u_2 | U_1 = u_1) >$ upper threshold, signals of long U_1 and short U_2 will be generated for the subsequent trading day.
2. If $P(U_2 \leq u_2 | U_1 = u_1) <$ lower threshold **and** $P(U_1 \leq u_1 | U_2 = u_2) >$ upper threshold, signals of short U_1 and long U_2 will be generated for the subsequent trading day.
3. Neutral Band $= 0.5 \pm parameter$, when two conditional probabilities are running within the neutral band, no signal will be generated, and thus no position opened.
4. Positions are held until the two conditional probabilities cross through the middle of the Neutral Band (Exit Threshold$= 0.5$). Positions will then be closed.
5. When the signals are reversed, the positions are closed and reversed.
6. All trading actions including opening, closing, and reversing positions are conducted at market opening.

Notice here that the "**and**" relationship when generating signals can also be changed to "**or**" relationship. Each time when long and short positions are created, cash-neutral will be achieved. This ensures the value of the long position is equal to the value of the short position.

For example, if $LowerThreshold = 0.35$ and $P(U_1 \leq u_1 | U_2 = u_2) < LowerThreshold$, which means the probability of U_1's stock price running below u_1 given U_2 is trading at u_2 is below 0.35, under the "**or**" relationship, signals of long U_1 and short U_2 with equal dollar value will be generated.

Despite a cash-natural strategy, we post a cash outflow restriction of $10,000 at the beginning of the test period. Technically speaking, this is to facilitate the calculation of percentage return; realistically, this is often the margin requirement for short-selling activities. At each trade signal, we invest all $10,000 in long one asset and short the same amount for the other in the pair in order to establish the arbitrage. No net cash outflow is needed. Both positions are not closed until neutral or reversed signals. Upon closing, the net gain/loss from the pairs trading is recognized and accumulated in the initial cash restriction. This allows the strategy to continue with rolling profits. No new positions are opened on the last day of the backtest period, and all open positions are closed.

Percentage returns are calculated as the ratio of cumulative cash outflow allowance and $10,000. Although our approach is purely arbitrage-focused and thus should yield infinite percentage returns under our assumptions, this pseudo-return allows the comparison of performance among pairs and exploration of different model hyper-parameters.

Besides the return calculation, two backtesting methods are used to generate the trading signals. First of all, the whole data set will be split into training and testing data set with the weight of 0.8 and 0.2 respectively. On the other hand, the rolling window method will be employed to generate signals. The size of the rolling window will be three years and each year the model will be recalibrated. The second method will be more practical in terms of real-life trading. The results of these two backtesting methods will be demonstrated later in this paper.

B. Pair Evaluation with Dynamic Time Wrapping

We employed fast dynamic time wrapping (FDTW) for pair selection for three purposes. First, this machine learning method is used in substitution of traditional judgment-based selection. This ensures a wider range of index pairs could be evaluated fairly, and no alpha would be missed or misinterpreted. Second, this process enables the quantification of pair-wise time series relationships, which further supports a pair-similarity vs. return analysis. Last but not least, pair selection prior to trading is also a regularization process for our mix-copula trading model training. This saved training resources and time immensely.

A linear method originally developed for speech recognition by Sakoe and Chiba in 1978 [8], DTW is capable of comparing similar features across time series pairs at different timing **illustration here**. Fundamentally, the algorithm looks for the optimal alignment between a time series pair by cross-comparing observations from one series with time-adjacent observations in the other series. The two-time series X and Y of length m and n discrete time steps can be simplified as:

$$X = x_1, x_2, ..., x_m$$

$$Y = y_1, y_2, ..., y_n$$

where x_i and y_j are observations at each time step.

The distances $c_{i,j}$, usually Euclidean, between each pair of observations are cost increments for each time step increment. DTW algorithm aims to identify the optimal path that accumulates the minimum incremental costs across all time increments. This is easily achieved with dynamic programming **figure**. The path can be visualized as a single-directional walk across the cost matrix, where incremental costs for each time step are recorded in each cell:

$$C_{i,j} = c_{i,j} + min(c_{i-1,j}, c_{i,j-1}, c_{i-1,j-1}) \quad \textbf{(15)}$$

After row-wise incremental cost computation, the minimum cost can be identified at the terminal cell, and the minimum cost path can be retraced.

To reduce computational complexity and accelerate algorithm deployment, we employed the FDTW algorithm developed by Salvador and Chan [10], where DTW is recursively executed to refine a coarse-scale path constructed with enlarged time steps into the minimum cost path.

The minimum FDTW costs are inversely proportional to the similarity of time series pairs. Evaluating all pairs across the 18 selected major stock indices, we obtain the heat map that highlights the best and worst matching index pairs on a linear scale Fig 2.

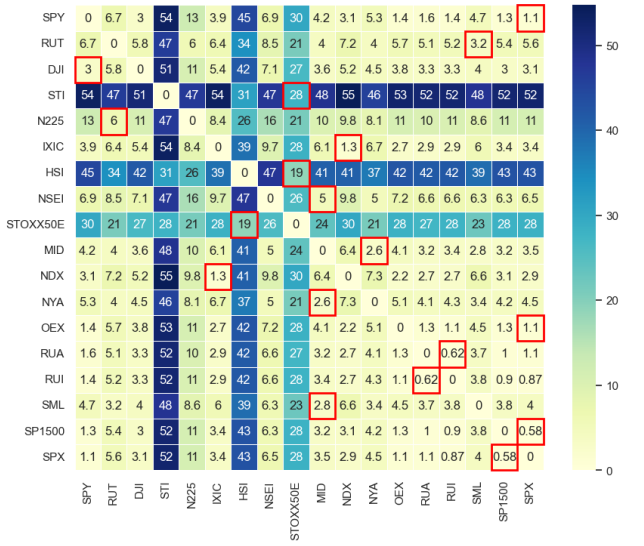

Fig. 2. Heat Map of FDTW Scores for All Pairs : small scores correspond to closer match of time series pairs; best match with each index is circled

C. Baseline Methods

C.1. Distance Method

One of the most popular approaches in pairs trading (especially statistical trading) is the distance method. It was developed and revised by Gatev, Goetzmann, and Rouwenhorst [5] and was the best-known in the academic community [11]. The mean excess return over the first year in their backtest period is 11 percent. However, the profitability decreases over their research period. This method requires the investors to compare the calculated price difference of two closely correlated securities to a predetermined threshold.

For our benchmark, we employed rolling regression with a window size of 120 days (6 months) for forecasting the correlation coefficient $\hat{\beta}_t$ between index prices $S_{Y,t}$ and $S_{X,t}$ in the pair for each day. The residue ϵ_t is then compared with its rolling standard deviation to generate long/short signals. If the observed residual is greater than the threshold value, it means the prices will converge and the trade following such expectation should be profitable.

The trading rules are as follows:

1. If $\epsilon_t > k\sigma_t$, long $S_{Y,t}$, and short $S_{X,t}$
2. If $\epsilon_t < k\sigma_t$, short $S_{Y,t}$, and long $S_{X,t}$
3. Positions are opened, closed, and reversed following the same rules as the mixed-copula model.

C.2. Kalman Filter

Rolling regression-based pairs trading considers the short-term correlation nature of equity indices. Yet, given that the noisy market signals are largely dependent on investors' short-term imbalance of perceptions or actions, the latent noise could impose serious challenges on the accuracy of measurement of market signals, and thus the performance or even validity of the generated trading signals. The Kalman Filter [6] is a mathematical technique that has been developed to estimate and forecast parameters of an evolving system by combining a series of observations with uncertain measurements.

In our paper, Kalman Filter-based benchmark aims to derive the true asset value (unobservable) from an assumption of constant co-integration between stock indices, using the market prices as noisy measurements [7].

The correlation factor between two asset prices $S_{Y,t}$ and $S_{X,t}$ can be determined by estimating β_t, as represented by the Measurement Equation:

$$S_{Y,t} = \beta_t S_{X,t} + v_t, \text{where } v_t \sim N(0, \sigma_v^2) \tag{16}$$

We can further assume that β is not a constant but varies with time, which facilitates that we can capture the dynamics of the system while ensuring that the residuals v_t are stationary. For the sake of simplicity, it can be assumed that the evolution of beta follows a random walk (the correlation factor $F_t = 1$), represented by the State Equation:

$$S_{Y,t} = \beta_t = F_t\beta_{t-1} + w_t, \text{where } w_t \sim N(0, \sigma_w^2) \tag{17}$$

The Kalman Filter can thus be implemented on the dynamic linear regression system. For each time step, a state prediction is first conducted to obtain *a priori* predictions for β_t and σ_w^2 using the state equation. *A posteriori* estimates of β_t and σ_w^2 are obtained after adjusting the *a priori* prediction with the Kalman gain K_t (for simplicity, only estimation related to β_t is shown here):

$$\check{\beta}_t = F_t\beta_{t-1} \tag{18}$$

$$\hat{\beta}_t = \check{\beta}_t + K_t(S_{Y,t} - \check{\beta}_t S_{X,t}) \tag{19}$$

where

$$K_t = \frac{S_{t,CA}}{S_{t,CA}^2 + \gamma^{-1}} \tag{20}$$

and

$$\gamma = \frac{\hat{P}_{t|t-1}}{\sigma_v^2} \tag{21}$$

is the Signal-to-Noise Ratio (SNR), referring to the relationship between the state variance and measurement error. In instances where SNR is low, the measurements are deemed to be noisy and uninformative, leading to greater reliance on prior knowledge such as $\hat{\beta}_{t|t-1}$. On the other hand, in situations where SNR is high, greater weight is assigned to the observation as it is perceived to be more reliable. By assuming that $\hat{\beta}$ is time-varying and follows a random walk, the filter captures the dynamics of the system and produces stationary residuals.

We conducted a comparison of the estimates of the beta coefficient, utilizing both the Kalman Filter method and the rolling linear regression method with a 6-month look-back period. The Kalman Filter model is more responsive in nature due to its ability to assign greater weight to the most recent observations and adjust the weightings based on the level of noise present in the measurements.

Our pair trading strategy's signal is dependent on the residuals v_t, which is assumed to have mean-reversion properties with mean at 0. At the close of each trading day, we update our estimate of β_t, and subsequently calculate the residuals.

Meanwhile, we monitor the residuals' rolling volatility σ_t. A trading signal is generated if the closing residual is significantly large (i.e. outside the Bollinger Band); otherwise, no position is taken. Specifically:

1. If $v_t > k\sigma_t$, long $S_{Y,t}$, and short $S_{X,t}$
2. If $v_t < k\sigma_t$, short $S_{Y,t}$, and long $S_{X,t}$
3. Positions are opened, closed, and reversed following the same rules as the mixed-copula model.

For our analysis, we consider k values of 0.5, 1, and 2. Although the scaling parameter can be optimized by modeling the mean-reversion process, we do not explore this further in this study.

4. DATA AND EDA

In this study, we aimed to investigate the performance of our trading strategy on various stock indices from January 20, 2013, to February 7, 2023. To achieve this, we collected data from Yahoo Finance API in Python for the following stock indices: SPY (SPDR S&P 500 ETF Trust) [1], RUT (Russell 2000 Index), DJI (Dow Jones Industrial Average), STI (Straits Times Index), N225 (Nikkei 225 Index), IXIC (Nasdaq Composite Index), HSI (Hang Seng Index), NSEI (Nifty 50 Index), STOXX50E (EURO STOXX 50 Index), MID (S&P 400 Mid Cap Index), NDX (Nasdaq 100 Index), NYA (NYSE Composite Index), OEX (S&P 100 Index), RUA (Russell 3000 Index), RUI (Russell 1000 Index), SML (S&P Small Cap 600 Index), SP1500 (S&P Composite 1500 Index), and SPX (S&P 500 Index). We chose these stock indices because they represent a diverse set of economies and industries across the globe.

To accurately reflect the returns and mitigate foreign currency risks, we converted all stock indices into US dollars. We collected foreign exchange rates of various currencies against the US dollar over the past 10 years from Yahoo Finance, and converted the adjusted close price of each stock index as the final price for each trading day. Since stock index prices may fluctuate during the trading day, we decided to use only the adjusted close price of each stock index because our strategy may involve stock indices from different time zones.

Due to differences in holidays or other market closures in each country, the integration of data inevitably leads to inconsistencies in some dates, which we must exclude before conducting our analysis. We collected approximately 2500 data points for each stock index from January 20, 2013, to February 7, 2023. We converted the data points into US dollars based on the exchange rates of the corresponding day. We then merged the data based on the trading date and aligned it to Eastern Time in the United States. This allowed us to create a consistent and comparable time series of stock index prices. The final available data set consists of 2061 data points, with a data coverage ratio of approximately 82.44%.

Our data collection and processing methods were designed to minimize biases and ensure the accuracy of our results. By collecting data for a diverse set of stock indices and adjusting for differences in currency, we have created a robust dataset that can be used to analyze the performance of global stock markets over the past decade.

All possible combinations of index pairs are tested, and MID-NDX and RUT-N225 are chosen as the characteristic pairs for further analysis. For further analysis of the two pairs, please refer to the conclusion section. Here we first want to perform exploratory data analysis(EDA) on the index price of these two special pairs, divided into five directions for testing, namely: Correlation, Jarque-Beta test for normality, Augmented Dickey-Fuller test for stationarity, Ljung-Box test for autocorrelation and Engle's ARCH test for ARCH effect. Here we only show the results, please see our EDA folder for the specific test code.

A. Correlation

We used the correlation function that comes with the python pandas dataframe for the linear correlation test. The tested correlation for index price is about 0.952 for MID-NDX and about 0.916 for RUT-N225, as both correlations are greater than 0.5 and can be considered to represent strong or large correlations.

B. Jarque-Bera Test for Normality

The Jarque-Bera test is used to check whether a given data set is normally distributed. In our test, if the difference between two stock indexes follows a normal distribution, then it will have a p-value of less than 0.05. The p-values of MID-NDX and RUT-N225 were tested to be close to 0, and their differences can be considered to follow a normal distribution.

C. Augmented Dickey-Fuller Test for Stationarity

Augmented Dickey-Fuller test can be used to do the unit root test for stationarity, and if the unit root exists, it will cause unpredictable results in time series analysis. In our test, if the p-value is less than 0.05 then the data has no unit root and is stationary. The tested p-values are about 0.63 for MID-NDX and 0.79 for RUT-N225, which are both greater than 0.05 and can be considered as they have unit roots and non-stationary.

D. Ljung-Box Test for Autocorrelation

The Ljung-Box test can be used to check whether there is autocorrelation in the time series. In our test, two data sets are autocorrelated if the p-value is less than 0.05. After testing, the p-values of both MID-NDX and RUT-N225 are close to 0 and can be considered as autocorrelated for both special pairs.

E. Engle's ARCH Test for ARCH Effect

Engle's ARCH test is a Lagrange multiplier test to assess the significance of the ARCH effect (Autoregressive Conditional Heteroskedasticity). In our test, if the p-value is less than 0.05 then there is an ARCH effect present in both data sets. after testing, the p-values of both MID-NDX and RUT-N225 are close to 0 and can be considered as having an ARCH effect present in both special pairs.

[1] This is not a stock index but we use it because it is a widely known and used ETF and it is designed to track the S&P 500 stock market index

F. EDA conclusion

To conclude, we have enough data to show that the index prices of each of the two special pairs, MID-NDX and RUT-N225, are strongly correlated and they are non-stationary; the two indexes of each pair are autocorrelated, there is an arch effect present, and their differences follow a normal distribution. Therefore, these two special pairs are suitable for pair trading.

5. MIXED COPULA PAIRS TRADING STRATEGY PERFORMANCE

A. Performance against Benchmarks

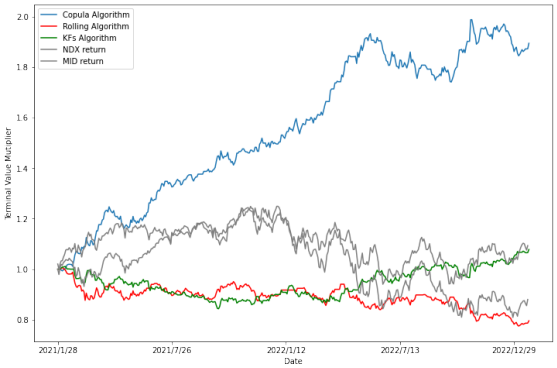

Fig. 3. Cumulative Return from MID-NDX Pairs Trading by Different Algorithms: X: 2-year backtest time horizon; Y: cumulative return as a multiple of initial position limit; 3 thresholds tested and 2 initial stock return

To compare the performance of the three different algorithms, we generated and plotted the results of three different algorithms: mixed Copula (Copula)[2], Rolling distance method (Rolling), and Kalman Filters strategy (KFs).

We chose the NDX (NASDAQ 100 Index) and MID (S&P MidCap) pair as an example to evaluate the algorithms. The detailed pair selection and comparison process are described in subsection DD. The training data set consisted of the stock price for NDX and MID from 2013/02/07 to 2021/01/25, and the test data set was the NDX and MID price from 2021/01/25 to 2023/1/20.

We ran back-testing to evaluate the performance of each algorithm and compared the results in a table. The table compared the number of trades, maximum percentage drawdown, win rates for executed trades, and final cumulative P/L. As the table shows, the mixed copula strategy outperformed the other algorithms, with a higher winning rate, a smaller maximum percentage drawdown, and a significantly higher final cumulative profit while maintaining a relatively low total trading number. The detailed result could be find on Fig. 7. Table of Pairs Trading Results vs. Different Thresholds.

As indicated in Fig. 4., the signals generated by mixed copula algorithm are evenly distributed in the test time of two years, and the proportion of long and short of a single stock is balanced.

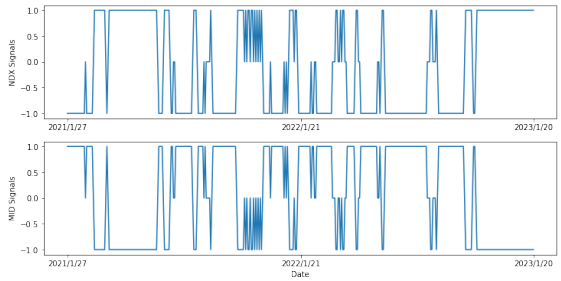

Fig. 4. Signals Generated by Mixed Copula Algorithm: X: 2-year backtest time horizon; Y: Trading Signals; +1 stands for Long position and -1 stands for Short

These results indicate that the mixed copula algorithm is a robust pair trading strategy that can generate profitable trades while minimizing risk. Furthermore, we found that the mixed copula algorithm is applicable to many other pair combinations, making it a valuable tool for traders looking to diversify their portfolios and maximize returns. In conclusion, the mix copula method is a powerful and versatile tool that can help traders identify profitable pair trading opportunities and improve their trading performance.

B. Different Thresholds vs. Returns

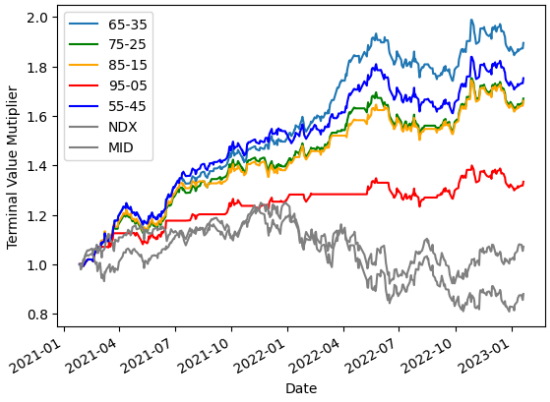

Fig. 5. Cumulative Return from MID-NDX Pairs Trading at Different Thresholds: X: 2-year backtest time horizon; Y: cumulative return as a multiple of initial position limit; 5 thresholds tested

[2]Standard hyperparameters (copula mix = 'CGJ', Neutral Band width = 0.3, 8 years training and 2 years backtest, no model update) are used in this most subsequent sections unless otherwise specified.

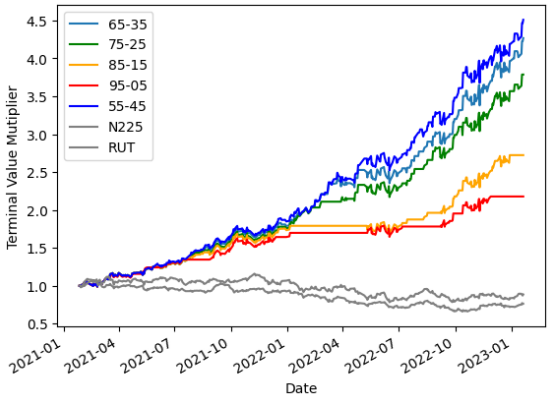

Fig. 6. Cumulative Return from RUT-N225 Pairs Trading at Different Thresholds: X: 2-year backtest time horizon; Y: cumulative return as a multiple of initial position limit; 5 thresholds tested

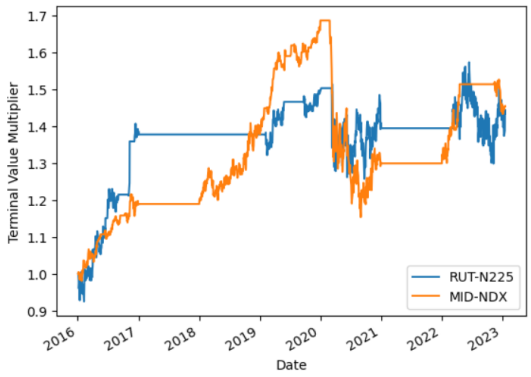

Fig. 8. Cumulative Return from RUT-N225 and MID-NDX Pairs Trading Using Rolling Model Update: X: 7 year backtest time horizon; Y: Terminal Value multiplier with initial fund of $10000

For the two most profitable pairs, MID-NDX (US-US) and RUT-N225 (US-non-US), we tested to optimize the strategy's return by altering thresholds 7. Cumulative returns peaked for Neutral Bands of widths around 0.30-0.10.

The mix-copula model generated signals performing extremely consistently across different thresholds. This indicates the model's prediction is almost equally precise across different probability regions. Win rates seem to be most strongly related to cumulative returns.

The model and strategy's results in terms of maximum drawdown are also comparable across the range of thresholds. As a result, investors should expect our strategy to carry a VaR of less than 10% over 2 year period.

C. Mixed Copula Model Update

Rolling model update method is also applied to the data. The rolling window size is set to be 3 years and recalibrate the model once a year. In general, the trading signals generated each year are based on the model trained using the previous three-year data. The rolling model update method is applied to both RUT-N225 and MID-NDX pairs trading. As a result, the value multiplier for the rolling model update method does not move as smoothly as that in the traditional train-test method. What's more, as shown in Fig 8, the overall value multiplier using the rolling model method is around 1.4, which means the total return from 2016 to 2023 is only 40 percent which is far lower than the performance in the test sample under the traditional train-test method. There are two reasons behind this finding.

First of all, the train-test model feeds more training data points to the mixed copula model and this will help to generate the cumulative distribution function of two indices that are close to their actual cumulative distribution function. Therefore, the prediction made using the train-test method in the test sample will be more accurate and signals will be generated more frequently compared to that in the rolling model update method. The lack of data points feed for each rolling model may generate less effective trading signals and the frequency of generation is also low. This explains the non-trading period in 2017 and 2021 for both RUT-N225 and MID-NDX pairs.

Also, during the 2020 pandemic period, the market was volatile and the copula model may not be able to generate promising trading signals since the model generated based on data ranging from 2017 to 2019 may not be able to capture the distribution of the indices in 2020 and therefore, the return during 2020 is unsatisfying for both pairs selected here.

D. Yields from Different Pairs vs. FDTW distances

Of 153 unique index pairs tested, the mixed copula model converged in the training of 117 pairs. Over 2 year backtest period, the strategy yielded an average cumulative return of 46.98% (CAGR = 21.24%), and a 57% standard deviation. The top performer pair is RUT-N225, which yielded a stunning 326.66% return over two years (CAGR = 106.56%). The median return is 26.29% (CAGR = 12.38%), beating the same-period S&P 500 average yearly return of 4.68%, and on par with its 10-year average of 12.35%. Only signals from 14 pairs resulted in a loss, with the greatest loss of 20.13% occurring in the RUT-STI pair.

It is notable that nearly all of the above-average returning pairs are US-non-US index pairs. Given the leading few pairs at the same time generated remarkably high yields, we propose two reasons that may contribute to this observation. First, Us-non-US pairs typically possess multivariate distribution features that are advantageous to the model training. The lack of long-term co-movement prevents the time series pairs to become perfect matches, which troubles model convergence; yet the pairs are still close enough on the FDTW scale to avoid underfitting. Second, there might be a potential data mismatch on time. As

Pair	Thresholds	Trades	Win Rate (%)	Maximum Drawdown (%)	Terminal Value Mutiplier (%)
MID-NDX	0.55-0.45	360	55.28	7.61	175.19
	0.65-0.35	**354**	**55.93**	**6.98**	**189.45**
	0.75-0.25	345	55.36	7.18	166.97
	0.85-0.15	320	55.94	5.73	166.02
	0.95-0.05	222	54.95	5.15	133.36
	Rolling	372	49.46	13.07	80.00
	KF	311	51.13	9.06	108.00
RUT-N225	**0.55-0.45**	**354**	**60.17**	**6.46**	**450.81**
	0.65-0.35	338	59.47	6.33	426.66
	0.75-0.25	315	60.00	6.18	378.64
	0.85-0.15	261	59.00	5.67	272.33
	0.95-0.05	202	59.90	5.61	217.80
	Rolling	221	58.37	48.58	44.00
	KF	282	56.74	36.73	51.00

Fig. 7. Table of Pairs Trading Results vs. Different Thresholds: 5 thresholds tested, optimal thresholds are highlighted

Yahoo Finance does not specify the timing of data collected for foreign market indices, the model may have matched US closing prices with Asian closing prices, which could be future data not foreseeable hours before actual Asian market closing. This extra information could therefore be factored into the return as additional profit.

Excluding all pairs involving non-US indices, the remaining 45 pairs yielded an average return of (CAGR = 10.17%). MID-NDX now yields the highest return of 89.45% (CAGR = 37.64%). It was originally ranked 20th.

No linear relationship is observed between FDTW score and mixed copula pairs trading returns Fig 9. However, FDTW analysis provided several interesting insights and may provide critical references for investors' pair selection process.

Profitable pairs are clustered around the FDTW score of 5 to 10 and above 40. The sometimes disjoint trends for poorly paired time series may introduce impedance for market price auto-correction, which may explain the latter range.

In addition, FDTW scores match the lackluster outcome for US-US pairs. Low FDTW scores corroborate the common knowledge of strong co-movement patterns of several major US equity indices. Such a pattern is advantageous for traditional correlation/co-integration methods, thus making them popular targets for canonical pairs trading strategies. Yet the narrow domains of sample space trouble empirical distribution estimation in copula estimation, thus making them unattractive to our strategy. In some extreme cases, the extreme correlation/co-integration of time series could lead to an inability of our algorithm to converge during training. All pairs that failed to converge during training had FDTW scores < 6 Fig 10.

Moreover, most of the arbitrage opportunities for closely related US-US pairs could be already exploited. This finding illustrates our model's capacity of capitalizing on unexplored and latent non-linear time series relations.

Another reason for investors to avoid too-closely related pairs is that too much similarity could also diminish the chances of successful model training.

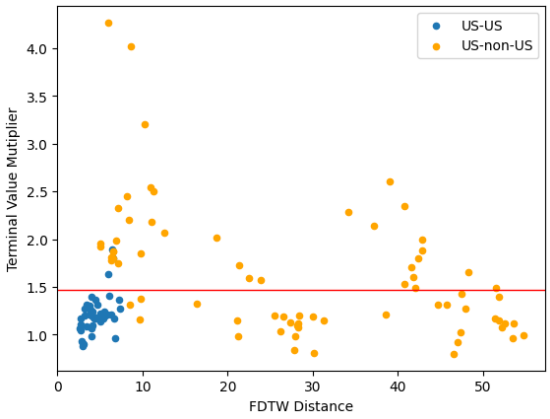

Fig. 9. Scatter Plot of Return Profiles Against FDTW Distance: X: FDTW distance between asset pairs; Y: pairs trading terminal results as multiples of initial investment; Purple: US-US pairs, Yellow: US-non-US pairs; horizontal cutoff: average returns as multiples of initial investment; number of pairs = 117

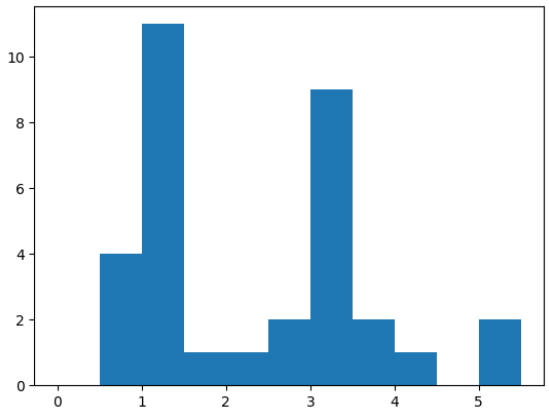

Fig. 10. Histogram of FDTW Distance of Pairs for Which Mix Copula Optimization Failed: X: FDTW distance between asset pairs; Y: number of asset pairs in each bin; number of pairs = 36

6. CONCLUSION

Mixed Copula Clayton-Gumbel-Joe is used to model the bivariate distribution between two indices. The normal maximum likelihood method and Expectation-Maximization (EM) algorithm are employed to find the optimal weights and copula parameters. In this paper, we verified a strategy composed of mixed copula and machine learning techniques. The traditional train-test method and rolling model update method are used to backtest the performance of our trading strategy on MID-NDX and RUT-N225 pairs. Significant alpha was found in dozens of potential index pairs, the most distinguished pairs being RUT-N225 and MID-NDX, with CAGRs being 106.56% and 37.64% respectively.

Besides validating the strategy, this report also provided a rudimentary exploration of model regularization and optimization. Although no indicative relationship is identified, FDTW provides a good first-step filtering tool for pair selection. It also acts as a regularization tool for individual investors with limited computational abilities. Optimization on hyperparameter is tested, and profit-maximizing thresholds and model update schemes are identified.

However, we recognize pairs trading using copula is still a relatively new approach in the trading area after all. There is still much discovery and further improvements to be made in the field to overcome the limitations of our current work. Several standard market assumptions are used in our methodology. We assume infinite divisibility of assets, no trade commissions, no slippage, and no commission. The last one may contribute to the most inaccuracy in our back-testing results. Another potential concern is that our strategy assumes overnight holding of positions, which might post higher margin requirements. This might reduce our projected earnings as more of the initial cash restriction would be used in satisfying these restrictions rather than earning alpha. The last major concern is the potential data-time mismatch. The foreign stock index closing might be future data that incorporates unobtainable alpha into our backtest results. This issue should be prioritized in future exploration.

Furthermore, future researchers can use our approach in analysis in other asset spaces. With a much larger sample space, the inclusion of the pairs selection method will be the most essential step in pairs trading.

REFERENCES

1. Fernando B. Sabino da Silva, Flávio Ziegelman, and João Caldeira. A pairs trading strategy based on mixed copulas, Nov 2017.
2. Mario Carrasco Blázquez, Carmen De la Orden De la Cruz, and Camilo Prado Román. Pairs trading techniques: An empirical contrast. *European Research on Management and Business Economics*, 24(3):160–167, 2018.
3. Binh Do and Robert Faff. Does simple pairs trading still work? *Financial Analysts Journal*, 66(4):83–95, 2010.
4. Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
5. Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006.
6. R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.
7. Marko Kolanovic. Dynamic system: Kalman filtering, 2017.
8. H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.
9. Felix Salmon. Recipe for disaster: The formula that killed wall street, Feb 2009.
10. Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.
11. R. Todd Smith and Xun Xu. A good pair: Alternative pairs-trading strategies. *Financial Markets and Portfolio Management*, 31(1):1–26, 2017.