**Machine Learning based approach for Pairs Selection in Financial Institutions**

Basharat Shafi Masoodi1 ,Peerzada Zubair Ul Islam2 andSitara Karim³

1Department of Finance and Human Resources, Genetikode, India

2Department of Computer Sciences and Engineering, Islamic University Of Science and technology Awantipora, India

³Department of Finance and Banking, Sunway University, Malaysia

Corresponding Author:

Basharat Shafi Masoodi

Department of Finance and Human Resources,

Genetikode, India

Email: bsmasoodi@gmail.com

**Abstract**

Pairs trading is the commonly recognized statistical arbitrage strategy that involves identifying pairs of securities whose prices tend to move mutually. We used a rolling-window machine learning approach to identify cointegrated pairs from the financial companies and found 13 potential trading pairs. Different time windows were used for training, testing and validation of the data to identify consistent pairs and observed that 3 pairs were consistent in all three rolling windows. We also backtested a mean reversion strategy designed to profit from these co-integrated pairs and obtained performance statistics such as total return, daily Sharpe, and CAGR. Additionally, we measured the Hurst exponent of the trading pairs to confirm that the strategy was mean reverting. The results showed that the strategy was successful in generating positive returns for all three pairs, with an average total return per trade of 0.34, 0.39, and 0.23 for the financial trading companies FMNB-FNLC, GFED-RBCAA, and MBWM-AMNB, respectively. The daily Sharpe for per trade was 0.4, 0.41, and 0.34 for FMNB-FNLC, GFED-RBCAA, and MBWM-AMNB, respectively.Hurst exponent results showed that the H value of less than 0.5 for trading pairs which were consistent in the training and all the testing datasets which clearly indicates the strategy is mean reverting. Overall, the results suggest that the mean reversion strategy based on cointegrated pairs was successful in generating positive returns and that the Hurst exponent provided further confirmation that the strategy was mean reverting

**Introduction**

One of the most well-known strategies among different algorithmic trading methods is the statistical arbitrage strategy[1, 2]. Statistical arbitrage is a profitable situation stemming from pricing inefficiencies among financial markets. Statistical arbitrage is not a real arbitrage opportunity, but it is merely possible to obtain profit applying past statistics [3, 4]. Statistical arbitrage strategies are often deployed based on mean reversion property, but they can also be designed using other factors. Pairs trading is the commonly recognized statistical arbitrage strategy that involves identifying pairs of securities whose prices tend to move mutually. Whenever the relationship between financial securities behaves abnormally, the pair would be traded. Then the open positions will be closed when the unusual behaviour of pairs reverts to their normal mode [5].

According to Krauss (2017), pairs trading strategy is a two-step process. The first step, which is called the formation period, attempts to find two or more securities whose prices move together historically. In the second step, which is the trading period, we seek abnormalities with their price movement to profit from statistical arbitrage opportunities [6]. There are two general approaches to find appropriate pairs of assets in the formation period: the heuristic approach and the statistical approach. The heuristic approach, which is regularly called the distance approach, is more straightforward than the statistical approach, which the latter is based on the cointegration concept. A pairs trading strategy is a mean reversion strategy on the spread, or price difference, of two financial assets. When the spread increases or decreases away from the mean, this strategy predicts the spread of the cointegrated pair of stocks will revert back to the mean. In the trading period, we can combine our strategy with different mathematical tools such as stochastic processes, stochastic control, machine learning, and other methods to improve the results [7, 8].

In machine learning, the algorithm learns the spread mean reversion using training data and then correctly predicts it using testing data. When the spread increases to a given threshold, the stock pair is traded by simultaneously entering into a short position (sell) for the higher price stock and a long position (buy) for the lower price stock. If the spread decreases to a given threshold the stock pair is traded by simultaneously entering into a short position for the lower price stock and a long position for the higher price stock. When the spread of the cointegrated pair, reverts back to the mean, one or both positions will be profitable. Each position is held until the opposite position is entered into, or no position and finally the position is exited.

**Methodology**

**Data Mining and Pre-processing:**

The data used in this study was downloaded from Yahoo! finance, It provides financial news, and commentary including stock quotes, press releases, financial reports, in original content and  provides high-quality intra-day data. The data used for analysis in this work is in 1 day frequency. We download data of 4500 stocks from Yahoo! Finance which was filtered and only data related to financial companies was selected. After filtering the data we got 231 companies. Out of 231 companies we finally selected 188 financial companies because there was missing data.

**Co-integration Analysis**

The idea of the pairs trading strategy comes from the identification of stationary price series. Stock price series are found to be non-stationary due to their stochastic behaviour [9]. Therefore, we aim to explore the effectiveness of cointegration-based pairs trading, which are used for selecting the potential pairs. From these datasets, pairs of stocks were selected using the cointegration analytical. The Engle Granger test was adopted to test cointegration. The test uses the residuals to see if unit roots were present, using Augmented Dickey-Fuller test. The residuals would be practically stationary if the time series is cointegrated. We selected top pairs of stocks that have long-run equilibria for downstream analysis. The cointegration equation is determined by formula:

$$y\_{t}=∝\_{0}+∝\_{1}x\_{t}+\in \_{t}$$

where $∝\_{1}$is the cointegration coefficient.

**Kalman Filter and Spread**:

Once we obtain cointegrated pairs to focus on, we apply the Kalman filter scheme to data to produce the estimate of the time-varying regression coefficients for each pair in order to get a dynamic hedge ratio without giving a specific regression window length and with a presence of statistical noise. The Kalman filter is a recursive algorithm which allows one to upgrade model estimates using new information. Basically, the Kalman filter is based on the representation of the dynamic system with a state space regression model (Eq 1-6). It formulates the beta dynamics through an autoregressive process driven by a white noise process [10].

$x\left(k+1\right)=F\left(k\right)x\left(k\right)+G\left(k\right)u\left(k\right)+v\left(k\right)$ (1)

$y\left(k\right)=H\left(k\right)x\left(k\right)+w\left(k\right)$ (2)

$\hat{x}\left(k\right)=F\left(k\right)\hat{x}\left(k\right)+G\left(k\right)u\left(k\right)$ (3)

$\hat{y}\left(k\right)=H(k)\hat{x}(k+1|k)$ (4)

$∆x=f(y\left(k+1\right),\hat{x}\left(k\right))$ (5)

$\hat{x}\left(k+1\right)=\hat{x}\left(k\right)+∆x$ (6)

*Where*

*x(k) is the n – dimensional state vector (unknown)*

*u(k) is the n – dimensional input vector (unknown)*

*y(k) is the p – dimensional output vector (unknown, measured)*

*F(k), G(k), H(k) are appropriately dimensioned system matrices (known)*

*v(k), w(k) are zero – mean, white Gaussian noise with noise (known) covariance matrices Q(k), R(k)*

Then spread between stocks was computed for each pair with the coefficients obtained. For a stationary and cointegrated series, a spread between the pair of assets, which is essentially a difference between their prices, is assumed to fluctuate around a mean value. The spread can be calculated by finding slope, usually called hedge ratio, of a linear regression between two time series.

To endorse the competence of the proposed strategies, we examine the performance out of sample. The procedure is as follows: The period of the training dataset is from January, 11, 2010, to December 31, 20014, comprising 1050 data points; the test dataset -1 covers the period from 01-01-2014, to 31/12/2016, comprising 715 data points; the test dataset -2 covers the period from 03-01-2017, to 31-12-2018, comprising 503 data points and the test dataset -3 covers the period from 03/01/2000, to 30/10/2020, comprising 715 data points. The training data consists of daily prices for 5 years from 2010-2014, and the testing data consists of daily prices for 2015-16;2017-18 and 2019-2021. The daily prices data for the training set is pre-processed, and used to generate the 10 input features for each day which include risk - free rate, Total Return, Daily Sharpe, Daily Sortino, CAGR, Maximum Drawdown, Calmar Ratio. These 10 features represent the state of the environment for that day.

The selectedyear-window as described above was used to identify dynamic cointegrated pairs and, for these chosen cointegrated pairs, we estimated the cointegration relationship. Then, the resulting estimates are applied to trades during a forward-ahead period. This procedure is repeated through the remaining sample period. It can be argued that this procedure can be improved by increasing the frequency of re-estimation of the portfolio pairs, and many other alternatives for the out of-sample exercise could be implemented. In this way, we have adopted a conservative perspective on presenting here the out-of-sample strategy, and other different variations to our strategy design have been left to the robustness check section.

Mean-reverting strategy was implemented which is designed to profit from the co-integrated trading pair. The hypothesis is that the spread of co-integrated pair is mean reverting in nature thus providing arbitrage opportunities if the spread deviates significantly from the mean. Mean reverting can be measured with respect to the Hurst exponent (H) of the pair with H< 0.5 showing time series is mean reverting, H=0.5 showing time series is a Geometric Brownian Motion and H>0.5 for trending time series [11].

**Results:**

We started by performing an Augmented Dickey-Fuller test as unit roots are a necessary condition for cointegration using a rolling-window approach within a given industry at the 5% significant level. Specifically, we use a five-year window from t to t + 5 (2010-2014) to identify cointegrated pairs. The process was repeated using two year window from t to t + 2 (2015-2016), followed by another two years window from t to t + 2 (2017-2018) and further three year window from t to t + 3 (2019-2021). Table 1shows cointegrated pairs in the portfolio.



Table 1: Cointegrated pairs in the portfolio

**Potential trading pairs**

Out of the 13 cointegrated pairs found across all rolling windows, we observed 3 cointegrated pairs consistent in the all four rolling windows. Figure 1,2,3 shows the price movements of the cointegrated stocks to validate and confirm the better performance of our approach.



Figure 1: Line Graph demonstrates the relationship in price data between FMNB and FNLC trading pair with time with Training data, Test data I, Test data II and Test data III.



Figure 2: Line Graph demonstrates the relationship in price data between GFED and RBCAA trading pair with time with Training data, Test data I, Test data II and Test data III



Figure 3: Line Graph demonstrates the relationship in price data between MBWM and AMNB trading pair with time with Training data, Test data I, Test data II and Test data III

**Performance Summary**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cointegrated Pairs** | **Risk-free rate** | **Total Return** | **Daily Sharpe** | **Daily Sortino** | **CAGR** | **Max Drawdown** | **Calmar Ratio** |
| FMNB-FNLC | 0.00% | 0.34  | 0.4 | 0.51 | 6.16% | -12.69% | 0.49 |
| GFED-RBCAA | 0.00% | 0.39  | 0.41 | 0.55 | 6.94% | -15.27% | 0.45 |
| MBWM-AMNB | 0.00% | 0.23  | 0.34 | 0.47 | 4.35% | -17.81% | 0.24 |

**Table 2:** The summary of performance statistics obtained from backtesting of the strategy. The statistics include Risk-free rate, Total Return, Daily Sharpe, Daily Sortino, CAGR, Max Drawdown and Calmar Ratio. Briefly, average rate of total return per trade is 0.34, 0.39 and 0.23 for FMNB-FNLC, GFED-RBCAA and MBWM-AMNB respectively. Similarly, Daily Sharpe for per trade is 0.4, 0.41 and 0.34 for FMNB-FNLC, GFED-RBCAA and MBWM-AMNB respectively. In addition, we also observed CAGR per trade is 6.16%, 6.94% and 4.35% for FMNB-FNLC, GFED-RBCAA and MBWM-AMNB respectively.

**Mean Reversion Strategy and Hurst Exponent**

The idea of mean reversion is rooted in a well-known concept called regression to the mean. This is a theory first observed by statistician Francis Galton and it explains how extreme events are usually followed by more normal events. A mean reversion trading strategy involves betting that prices will revert back towards the mean or average. Markets are forever moving in and out of phases of mean reversion and momentum. Therefore, it is possible to develop strategies for both types of market. A simplistic example of a mean reversion strategy is to buy a stock after it has had a large fall in price. When a stock has seen a big drop, there’s usually a good chance that it will bounce back to a more normal level. For example, if the Dow Jones Industrial Average drops 20% this month, it will probably fall less than that next month [12, 13].

This Mean-reverting strategy was implemented which is designed to profit from the co-integrated trading pairs. Mean reverting is measured with respect to the Hurst exponent (H) of the trading pair. The results showed the H value of less than 0.5 for trading pairs which were consistent in the training and all the testing datasets which clearly indicates the strategy is mean reverting (Table 3).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **In-Sample** |  |  | **Out-Sample I** | **Out-Sample II** | **Out-Sample III** |
| Stock1 | Stock1 | Hurst | Hurst | Hurst | Hurst |
| GFED | RBCAA | 0.34 | 0.21 | 0.29 | 0.3 |
| FMNB | FNLC | 0.39 | 0.3 | 0.33 | 0.27 |
| MBWM | AMNB | 0.38 | 0.31 | 0.3 | 0.22 |

**Table 3:** Hurst exponent values for the three trading pairs in training and test datasets.

**Conclusion**

This study identified cointegrated trading signal by estimating the volatility of data using the Kalman filter and calculating time adaptive regression coefficient based on it. As for underlying universe for the strategy, we confine ourselves to consider the most cointegrated pairs as a basket for our experiment. A common concern about the performance of pairs portfolio strategies is that its implementation is not realistic: closing prices are used for liquidating trading positions, and some risks related to the strategy are, for instance, close positions with next day prices, used to be ignored. To examine the impact of this question in our strategy, we explore the out-of-sample performance using daily open prices for trades. This new dataset is applied to the dynamic and constant methods. Therefore, the current study identified 3 high ranking pairs from trading portfolios based on a set of in-sample statistical learning. This verified that the strategy has better profitability and stability compared to traditional strategy [14].

**Discussion**

The results of the study indicate that a mean reversion strategy based on cointegrated pairs can be profitable, and the performance of the strategy is better compared to a traditional strategy. The study used a rolling-window approach to identify cointegrated pairs within a given industry, and then used a mean reversion strategy on these pairs [15]. The results showed that out of the 13 cointegrated pairs found across all rolling windows, three pairs were consistent in all four rolling windows, and these pairs were used for the mean reversion strategy. The performance summary of the strategy showed that the average rate of total return per trade was 0.34, 0.39, and 0.23 for FMNB-FNLC, GFED-RBCAA, and MBWM-AMNB, respectively. The daily Sharpe for per trade was 0.4, 0.41, and 0.34 for FMNB-FNLC, GFED-RBCAA, and MBWM-AMNB, respectively. The CAGR per trade was observed to be 6.16%, 6.94%, and 4.35% for FMNB-FNLC, GFED-RBCAA, and MBWM-AMNB, respectively.

Furthermore, the study also evaluated the mean reversion nature of the strategy by calculating the Hurst exponent (H) of the trading pairs. The H values of less than 0.5 for trading pairs consistent in the training and all the testing datasets indicated that the strategy was mean-reverting. One interesting aspect of the study is the use of the mean reversion strategy, which is based on the concept of regression to the mean. Mean reversion has been a popular approach to trading for many years, and it involves betting that prices will revert back to their average or mean over time. This approach is based on the assumption that extreme events or price movements are likely to be followed by more normal movements. In the context of this study, mean reversion is measured using the Hurst exponent, which is a measure of the long-term memory of a time series. A Hurst exponent of less than 0.5 indicates that a time series is mean reverting, while a value greater than 0.5 indicates that it is trending [16]. The results of the study showed that the Hurst exponent for the three trading pairs that were consistent across all four rolling windows was less than 0.5, indicating that the strategy was indeed mean reverting. Another interesting aspect of the study is the use of a rolling-window approach to identify cointegrated pairs. Cointegration is a statistical property of time series that indicates a long-run equilibrium relationship between them [17]. In this study, cointegration was identified using the Augmented Dickey-Fuller test, which is a widely used test for unit roots in time series. The use of a rolling window approach allowed the study to capture changes in the cointegration relationships between the stocks over time, which is important in a dynamic market environment.

These are several studies that investigate pairs trading strategies using various machine learning techniques and evaluation metrics. In one of the study, the most important feature for determining the future direction of stock return is the most recent lag variable (return difference). The backtest of the alternative model showed favorable total returns of 10.62% with small volatility over a five-year period (11/30/2014 - 11/28/2019). The backtest of the baseline model showed a loss and a more severe maximum drawdown than the loss from total returns. The classification model is effective and makes the pair trading strategy profitable. Sang-Ho Kim et al compare state-of-the-art pairs trading methods on two datasets using return rate, Sharpe ratio, and maximum drawdown as evaluation metrics. They introduce HDRL-Trader, which integrates multiple novel techniques, and show its effectiveness in generating profits [18]. Another study by Georgios Sermpinis et al evaluates five pairs trading strategies using different methods to execute trading actions on commodity markets and proposes a novel CA-DRL structure for decision-making. They find that the traditional methods are ineffective, while the CA-DRL strategy obtains high returns and low risks [19]. Jiayu Wu describes a pairs trading strategy that uses price and technical indicators and trains a binary classification SVM model with a linear kernel. They introduce new metrics that focus on directional prediction and identify an "Opportunity Window" for freely entering trades. The authors suggest future work could include adding more features to the model and updating the spread model over time [20]. These studies provide insights into the effectiveness of various machine learning techniques and evaluation metrics in pairs trading strategies and offer suggestions for future research directions.

However, it is important to note that the results of the study are based on historical data and backtesting, and there is no guarantee that the strategy will perform as well in the future. Market conditions can change quickly and unpredictably, and past performance is not necessarily indicative of future results. Additionally, the study only considers a limited set of trading pairs, and there may be other pairs that are cointegrated and could be profitable to trade.

In conclusion, the study provides valuable insights into the potential of a mean reversion strategy based on cointegrated trading pairs. The use of a rolling-window approach to identify cointegrated pairs and the measurement of mean reversion using the Hurst exponent provide a robust and adaptable framework for the strategy. While the results of the study are promising, it is important to exercise caution and conduct further research before implementing the strategy in a real-world trading environment. The study provides evidence that a mean reversion strategy based on cointegrated pairs can be profitable and has better performance compared to a traditional strategy. However, it is important to note that this study is limited to a specific industry and time period, and the results may not generalize to other industries or time periods. Additionally, the study did not consider transaction costs or slippage, which could impact the profitability of the strategy in a real-world setting.

1. Balladares K, Ramos-Requena JP, Trinidad-Segovia JE, Sánchez-Granero MA: **Statistical Arbitrage in Emerging Markets: A Global Test of Efficiency.** *Mathematics* 2021, **9**.

2. Carta S, Consoli S, Podda AS, Recupero DR, Stanciu MM: **Statistical arbitrage powered by Explainable Artificial Intelligence.** *Expert Systems with Applications* 2022, **206**.

3. Hossain S: **High-Frequency Trading (HFT) and Market Quality Research: An Evaluation of the Alternative HFT Proxies.** *Journal of Risk and Financial Management* 2022, **15**.

4. Zaharudin KZ, Young MR, Hsu WH: **High‐frequency trading: Definition, implications, and controversies.** *Journal of Economic Surveys* 2021, **36:**75-107.

5. Clegg M, Krauss C: **Pairs trading with partial cointegration.** *Quantitative Finance* 2017, **18:**121-138.

6. Krauss C: **Statistical Arbitrage Pairs Trading Strategies: Review and Outlook.** *Journal of Economic Surveys* 2017, **31:**513-545.

7. Tadi M, Kortchemski I: **Evaluation of dynamic cointegration-based pairs trading strategy in the cryptocurrency market.** *Studies in Economics and Finance* 2021, **38:**1054-1075.

8. Vidal T, Crainic TG, Gendreau M, Prins C: **Heuristics for multi-attribute vehicle routing problems: A survey and synthesis.** *European Journal of Operational Research* 2013, **231:**1-21.

9. Tokat E, Hayrullahoğlu AC: **Pairs trading: is it applicable to exchange-traded funds?** *Borsa Istanbul Review* 2022, **22:**743-751.

10. de Moura CE, Pizzinga A, Zubelli J: **A pairs trading strategy based on linear state space models and the Kalman filter.** *Quantitative Finance* 2016, **16:**1559-1573.

11. Huck N: **Pairs trading and outranking: The multi-step-ahead forecasting case.** *European Journal of Operational Research* 2010, **207:**1702-1716.

12. Ramos-Requena JP, Trinidad-Segovia JE, Sánchez-Granero MA: **Introducing Hurst exponent in pair trading.** *Physica A: Statistical Mechanics and its Applications* 2017, **488:**39-45.

13. Bui Q, Ślepaczuk R: **Applying Hurst Exponent in pair trading strategies on Nasdaq 100 index.** *Physica A: Statistical Mechanics and its Applications* 2022, **592**.

14. Kim K: **Performance Analysis of Pairs Trading Strategy Utilizing High Frequency Data with an Application to KOSPI 100 Equities.** *SSRN Electronic Journal* 2011.

15. Miao GJ: **High Frequency and Dynamic Pairs Trading Based on Statistical Arbitrage Using a Two-Stage Correlation and Cointegration Approach.** *International Journal of Economics and Finance* 2014, **6**.

16. Kulish V, HorÁK V: **Forecasting the Behavior of Fractal Time Series: Hurst Exponent as a Measure of Predictability.** *Review of the Air Force Academy* 2016, **14:**61-68.

17. Nyakabawo W, Miller SM, Balcilar M, Das S, Gupta R: **Temporal causality between house prices and output in the US: A bootstrap rolling-window approach.** *The North American Journal of Economics and Finance* 2015, **33:**55-73.

18. Kim S-H, Park D-Y, Lee K-H: **Hybrid Deep Reinforcement Learning for Pairs Trading.** *Applied Sciences* 2022, **12**.

19. Sermpinis G, Stasinakis C, Zong X: **Deep Reinforcement Learning and Genetic Algorithm for a Pairs Trading Task on commodities.** *SSRN Electronic Journal* 2020.

20. Wu J: **A Pairs Trading Strategy for GOOG/GOOGL Using Machine Learning** 2015.