

Comparison Analysis of Arima and Z-test for Forecasting the Maximum Return Based on the Pair Trading Strategy

Hao Fu^{1, *, †}, Yaxuan Yang^{2, *, †}, Liangzi Zhu^{3, *, †}

¹Woods College of Advancing Studies, Boston College, Boston, MA, USA

²School of Finance (DBS sino-foreign educational programme), Xiamen University, Xiamen, Fujian, China

³Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA

*Corresponding author: fuhh@bc.edu., 32920182201274@stu.xmu.edu.cn, lzhu41@jh.edu

[†]These authors contributed equally.

Abstract. The stock market trends are of great challenge. Even some of the mutual market's strategies, such as traditional pair trading strategy may still be risky and are likely to suffer an unbearable loss due to an unreasonable stop mechanism. In this paper, a novel approach of pair trading strategy – using the logarithmic return to generate pairs and those deviations to construct trading signals. In addition, two methods that ARIMA and Z-TEST used to construct trading signals are compared according to their corresponding cumulative returns. In part one, the ARIMA model has been used to identify whether there is a constant predictive power and how trading signals should be generated to obtain the highest returns. Therefore, a stationary test has been done, and an optimal threshold has been selected. In part two, the Z-test has been used to distribute the preceding data, and the best threshold to construct trading signals has been found in the same mechanism with that process of ARIMA. As a result, the Z-test method shows a better return than that of the ARIMA model, which could be caused by a more stable forecasting accuracy from the distribution. To conclude, this novel approach is efficacious to lower the risks from unanticipated extreme cases and unwise designs on stopping signals. Furthermore, the Z-test method for signals construction is more stable and profitable than the ARIMA. The research greatly maximizes profits in stock markets by selecting the best model and an optimal threshold that models should be applied.

Keywords: component; Maximum Return Prediction; Pairs Trading Strategy; Time Series Analysis; Cointegration; ARIMA model; Z test

1. Introduction

The various factors and individual behavior factors make the stock price variation very complicated [1]. The stock price (or trend) prediction is a great challenge due to its characteristics of noise and volatility [2]. Therefore, maximizing the return based on keeping low risk is a very interesting and important problem [3]. Obviously, risk resistance (or management) for quantity investment is the basic issue [4].

An uncertain stock market makes trading in a risky environment, which is why many investors tend to make profits in this stock market using different investment strategies. Market-neutral strategies are especially effective regardless of the direction of market assets movement, such as pair trading [5-7]. By building both long and short directions of shares of a pair of highly correlated stocks in a portfolio, pair trading is one of the effective neutralized approaches. Since the early stage, pairs-trading has been very popular in the application of obtaining arbitrage profit [8-12]. Trading on temporary abnormal deviations of those paired prices compared to their long-run equilibrium relationship required a presumption that the pair of stocks are driven by virtually the same economic forces. Therefore, their price variations are expected to be proportionated, and their direction of trend should be the same [13].

The ARIMA model has been proposed and developed by Box and Jenkins [14]. It is also known as the Box-Jenkins methodology composed of a series of activities for identifying the range of data that could still be effective as a factor to influence the present, estimating the magnitude of predictive

power from the past on the present, and diagnosing possible errors happened in the process of prediction [15-18]. In addition, the Z test approach takes advantage of the idea of classical normal distribution, which helps determine the dispersion degree according to the relative frequency of data within a given standard deviation. Be more specifically, among the two methods of building trading signals, the ARIMA models used past data to forecast that of the present. In contrast, the Z test approach used a long-term distribution to determine how deviated from the center the present data are [19].

However, the research on applying these two methods, ARIMA and Z test, for pair trading in different markets to maximize return with relatively low risk and illuminate corresponding advantages and disadvantages is still insufficient. Therefore, this paper's major interest is to construct two distinctive methods and the more profound part is to identify the merits and drawbacks respectively and explain the possible reasons for that outcomes.

The contributions in this paper include 1) Different from the traditional pair trading strategies, where the stock price is used to determine the correlation between two stocks, the logarithmic value is used during the process primarily because using logarithmic return instead of absolute price can avoid some bias from a too wide range of stocks prices, like less than \$1 dollar value for a share of one stock and \$10000 per share for another stock, which might cause an unstable correlation between these two stocks. 2) On the basis on the traditional pair trading strategies, the methods introduced in this paper constructed daily-time-range trading signals, signals only for a single day, so the expected risk is further reduced with a relatively shorter trading period. 3) An attempt to seek out an optimal threshold for both ARIMA and Z test approaches and corresponding validation test on that threshold with out-of-sample data.

Section 2 introduces the details of methods; Section 3 introduces the comparison results of Arima and Z-test; Section 4 gives the conclusion.

2. Methods

In this paper, the Arima and Z-test are used to forecast the best return by pair-trading strategy separately. And several stocks selected from high-tech corporations are used to evaluate their performances.

2.1 Data Preparation

In this paper, 10 stocks are derived from high-tech companies, such as "Apple, Facebook, Microsoft, Amazon, Tencent, Sony, Alibaba, HP, Google, and Intel". The data source is obtained from Yahoo Finance website [20]. Since pairs trading usually concentrates on pairs of stocks from the same sector, then the selection rule is as follows: "collect stocks that share similar price trends" and "take a long-short position when they diverge and unwind on convergence".

Each company's daily stock price data is selected from June 6, 2016, to June 3, 2021. The data are concentrated in the recent five years, and the study has strong timeliness, generality, and representativeness. Besides, considering that the epidemic greatly impacted the stock price in this period, we took before and after the outbreak of the epidemic as the boundary to divide the test set and the training set. Therefore, the test set can effectively represent the risk resistance ability of our model, which is of guiding significance for the study of the market shortly.

Before using data sets to train machine learning models, it is necessary to preprocess the data. The commonly used data preprocessing methods include standardization and normalization. Standardization, also known as Z-score normalization, first calculates the mean and standard deviation of the data and then maps the raw data to the standard normal distribution.

2.2 Pair Trading Strategy Design

Cointegration is a robust measure of the connection between two financial quantities, and mean reversion between financial quantities is the key concept to be included in cointegration tests. More

specifically, if any two stocks' prices are cointegrated, it represents there is an enduring balanced relationship between them. Many different selection methods can be used to achieve pairs trading; however, all of them use common nonstationary factors to find deviations from a balanced asset-pricing framework. Sharing common stochastic trends among a series of nonstationary stock prices in level form is the foundation of the cointegration tests to analyze stock prices. The reason for using 10 high-tech companies' stock prices to do cointegration tests is that this research mainly focuses on the trend between stocks' mean prices over time rather than the individual price changes. Therefore, based on Engle and Granger's most familiar cointegration test, 4 stock pairs meet the requirements (Alibaba and Facebook; Alibaba and Tencent; Facebook and Tencent; HP and Sony).

An alternative statistical measure for cointegration is a correlation, but they focus on different concepts. Assets with high cointegration do not necessarily indicate a high correlation in prices, and correlation tests rather than cointegration tests reflect co-movements in prices. Almost all multivariate financial problems use correlation to do analysis. However, the results of correlation tests may not be stable over time. Due to the instability of correlation tests and stock prices, log-returns of 10 high-tech companies' stocks from June 7, 2016, to June 2, 2021, are calculated to test for correlation. In addition, correlation is a second-moment calculation. Using log-returns is better, so the appropriate results can be ensured by neglecting the higher moments calculation. Since 4 pairs are selected in the cointegration test, it is simple for each pair of stocks to find periods of divergence, figure out why two stock prices are separating and try to take profit through convergence, which is a mean-reversion process. Hence, one highly correlated pair can be selected among the 4 pairs by applying correlation tests on their logarithmic returns, which are Alibaba and Tencent.

2.3 Forecasting Maximum Return Method

2.3.1 Arima-based method

The analysis of time series is based on its stationarity. Here, we chose the most commonly used Augmented Dickey-Fuller unit root test (ADF test). ARIMA (p,d,q) may be more appropriate for much economic time series than simple linear models. In econometrics, to verify the existence of unit roots in the AR(p) process, the null hypothesis $H_0: \beta = 1$ can be tested by formula (1)

$$x_t = c_t + \beta x_{t-1} + \sum_{i=0}^{p-1} \phi_i \Delta x_t - i + e_t \quad (1)$$

where the ADF test statistic is $\frac{\beta-1}{\sqrt{\text{Var}(\beta)}}$. The order of the model is determined by observing ACF and PACF graphs based on the stable sequence of logarithmic returns. Since the results of ACF and PACF charts are not very significant, the Akaike Information Criterion (AIC) is used to measure the good fit of the statistical model. It is assumed that K is the number of parameters, L is the likelihood function, and N is the number of observations, SSR is the sum of squares of residual errors, then $AIC = 2K + N \cdot \ln(SSR/N)$. The purpose is to have good data fitting and avoid overfitting. Therefore, the smallest AIC value model, namely ARIMA (2, 0, 1), was finally selected.

When building trading signals, the predicted values obtained from the ARIMA model above will be used. The main idea is to compare predicted and true values. When the predicted return difference is greater than the true value, it indicates that the difference between the two stocks is underestimated. Therefore, it is necessary to long the stocks with high returns and short the stocks with low returns, to return to the normal value. Similarly, when the predicted return difference is less than the true value, it represents that the difference between the two stocks is overestimated. Consequently, long the stocks with low returns and short the stocks with high returns should be implemented.

An optimal threshold is needed. The next step is to reduce the transaction frequency to reduce unnecessary costs. Therefore, only when the difference reaches a certain level transactions are executed. The value of the difference level is determined by the multiple i of the standard deviation of the sequence. After searching through 0 to 3 (spaced by 0.1), we get the corresponding value where $i=1.1$ is the maximum cumulative return.

2.3.2 Z-test-based method

In the paper, a method for selecting the optimal threshold will be introduced. First, data is split into two parts, one for training and another for testing. Then, cumulative returns from various thresholds from 0.1 to 3.0 are demonstrated in the training data. Additionally, the threshold for the highest cumulative return is selected, which is 0.4 in this paper. Finally, the chosen threshold of 0.4 is used in the testing data to see whether this threshold is still valid in out-of-sample data. The Z value can be calculated by the formula (2)

$$Z = \frac{x - \mu}{\sigma} \quad (2)$$

As the most correlated pair of stocks, BABA and Tencent, have been selected among the most integrated pairs, it is reasonable to conclude that the chosen pair of stocks have pretty similar trend and moving direction. However, a distribution test for their divergence on daily return is still requisite because only if their daily return deviations are normally distributed, it is safe to build trading signals when the absolute value of those deviations is too big. More specifically, extremely big absolute values indicated that they are very distant from the mean of the series of data, so they are likely to be considered as outliers with a relatively low possibility, and that low possibility is the foundation of how the trading signals are generated because of the expectation of that those outliers would eventually revert back to the mean.

For the trading signal generation, the process is clear and direct. When the daily divergence of two stocks' returns is higher than or lower than a given threshold, the signals will be triggered so that the outperforming stock will be sold and the underperforming stock will be bought. Then it would make a profit after the reversions of two stocks. However, an optimal selection of thresholds is always challenging because different thresholds would have disparate trading signals and, correspondingly, a very large gap in cumulative returns.

3. Results and Discussion

In this paper, two tests are given to evaluate the performance of the proposed method. Firstly, the ADF test is used to evaluate the stationary of return ratio of target stocks. Secondly, the test for pair trading is based on different methods, such as Arima and Z-test.

3.1 Results of ADF test

The results of the ADF test can be found in table 1. The scale of value is in the logarithmic domain.

Table 1. ADF test for logarithmic return

Dickey-Fuller	Lag order	p-value
-20.20	1.00	0.0

Table 1 shows that the p-value equals 0. According to the results in the above table, the p-value is less than 0.01, so the existence of the unit root of the null hypothesis is rejected, which means at the significance level of 5%, the unit root phenomenon is not considered to exist in the sequence. This confirms our preliminary conjecture that the sequence of logarithmic returns of stock prices is stable, which is also the basis for our subsequent construction of the mean value model and volatility model.

3.2 Results of Arima and Z-test

The results of Arima for pair trading are shown in table 2.

Table 2. ARMA (p=2, q=1)

Name	ar1	ar2	ma1
Coefficients	0.787	-0.0248	-0.8537
s.e.	0.091	0.041	0.083
p-value	0.000	0.000	0.000

Especially, the AIC equals -3991.957, and the Prob(JB) equals 0.00.

According to the results in Table 2, most of the model coefficients' p-values are close to 0, indicating that the model coefficient is statistically significant basically. The AIC value of the model is -3991.957. The AIC value is very small, indicates that the model has a good fitting effect. After the Ljung-box test, the p-value is 0.39. It can be considered that the residual difference sequence is classified as a white noise sequence, which conforms to the hypothesis of model construction.

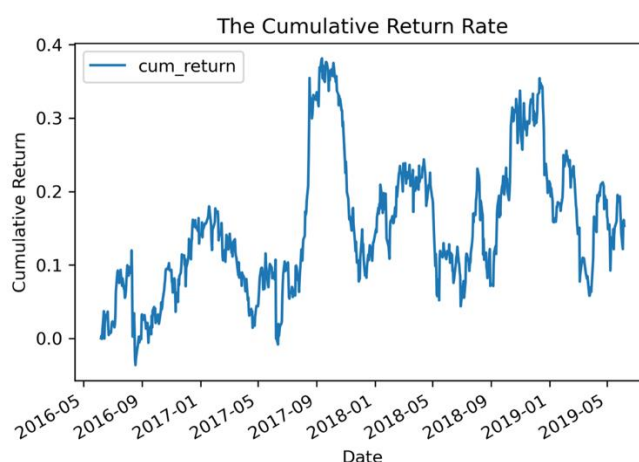

Figure 1. The cumulative return rate.

Figure 1 shows the output result of the training set data, namely the graph of the cumulative return rate from May 2016 to May 2019. As can be seen from the figure, the rate of return fluctuates significantly and stays around 10% for most of the time. Even though it peaked at nearly 40% at the end of 2017, it soon fell sharply to less than 10%, with a cumulative return of 15.31%. This is not ideal when factors such as transaction costs are not taken into account.

Table 3. Different threshold (i) for Maximum Expected Return

i value	cumulative return	i value	cumulative return
0.1	6.84%	1.6	38.18%
0.2	8.62%	1.7	39.94%
0.3	19.15%	1.8	30.13%
0.4	20.92%	1.9	33.70%
0.5	34.23%	2.0	32.82%
0.6	25.78%	2.1	28.04%
0.7	24.95%	2.2	20.80%
0.8	28.12%	2.3	16.60%
0.9	30.40%	2.4	25.66%
1.0	31.93%	2.5	24.71%
1.1	46.12%	2.6	17.16%
1.2	38.89%	2.7	17.16%
1.3	40.16%	2.8	17.16%
1.4	33.39%	2.9	11.24%
1.5	44.65%	3.0	11.24%

Table 3 shows the different threshold values we have a test for exploring the maximum expected return. The best threshold value is 1.1. After obtaining this value, the results of cumulative return can be found in figure 2.

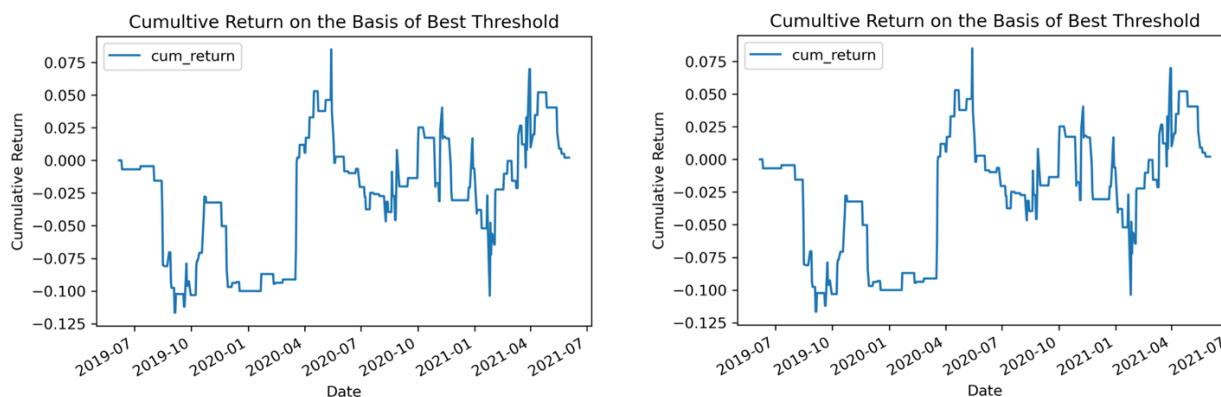


Figure 2. Cumulative return curves based on the best threshold.

Figure 2 shows different cumulative return curves when i changes different values. It can be seen that eventually, accumulated return as a whole has obviously improved, and ultimately the curve of cumulative return is higher as the frequency of transactions diminishing. Therefore, it can be sure that there is an optimal value i lower the increase total return and return volatility. According to the output results, when $i = 1.1$, the cumulative return rate reaches the maximum of 46.12%. Figure 2 shows the performance of the test set data when the i -value takes the optimal threshold in the training set, that is, $i = 1.1$. Since the test set data experienced a special period of COVID-19, the cumulative return from December 2019 to April 2020 was inevitably affected by COVID-19. However, after April 2020, the cumulative return began to turn back, and finally offset the losses caused by COVID-19 and generated positive returns.

The distribution and cumulative return of data are shown in figure 3.

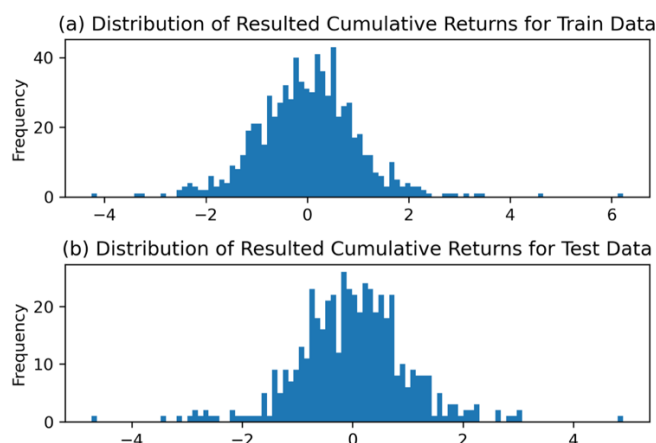


Figure 3. The distribution results of training data and testing data.

In figure 3, subfigure (a) shows the distribution for training data from 2016.6 to 2019.6, and subfigure (b) shows the distribution for the testing data from 2019.7 to 2021.6. The plots for either training or testing data represent that the logarithmic return differences between those two stocks are both normally distributed with a mean of 0. Additionally, the numbers on the x-axis represent the Z score similar to the standard deviation; a Z score of 2 means it is two standard deviations distant from the mean, which is the 0. This zero-mean is reasonable because those two stocks, BABA and Tencent, have a high correlation on daily logarithmic returns, so their daily variations have a plausible expectation to be very close to each other as well as a consequent very close to zero daily deviation.

According to the fundamental property of a normal distribution, data that is more distant from the center would be less frequent and thus less probable. For example, a Z score of 3 is much less likely to happen than that of 1. From the plot above, in the training data, there are only two days of Z score higher than 4 and only one day lower than negative 4, so being higher than 4 or lower than negative 4 is extremely rare and even can be considered as unlikely. Therefore, when such an extreme case happens, a trading signal will be triggered: The overvalued stock will be sold, and the undervalued stock will be bought. However, extreme cases being higher than 4 or lower than negative 4 only happened no more than 5 times within 5 years, so the training outcome, though profitable, would be insecure and unstable to apply in the future data because of this very small sample size.

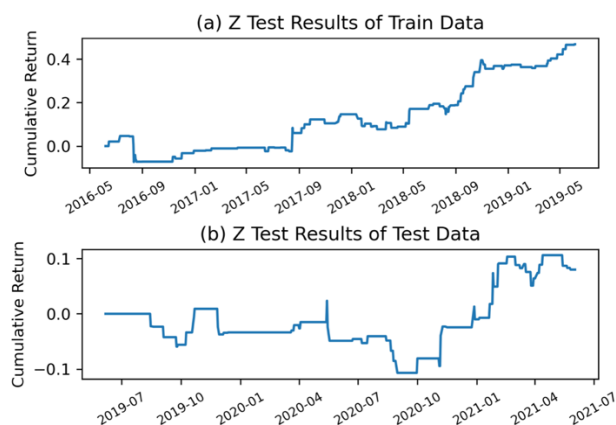


Figure 4. The results of training data and testing data are based on Z-test.

Consequently, the criteria to choose an optimal Z score threshold for triggering trading signals should consider the rarity, exactitude, if more specifically, and the trading times because the latter directly determine whether the sample size is sufficient to render an outcome valid and stable. In this paper, a For-Loop method is used to seek out that optimal threshold from a Z score ranged from 0.1 to 3. The Z score of 0.4 presented the highest cumulative return for the training data as the first plot above shown, which is more than 40 percent cumulative return for 3 years and thus the annual return is higher than 10 percent. In the second plot, the threshold of 0.4 Z score is applied in the testing data, and the result shows a 10 percent total return at the end of June 2021.

As we can see from the return plot in the testing data, although the cumulative return at the end is positive, there was an approximately 10 percent loss at the end of 2020. The insufficiency that caused such a big loss could be that considering the relative relationship of two stock absolute prices is lacking. For example, suppose that BABA is generally higher than Tencent in a relatively constant price, such as \$200 above Tencent. This \$200 deviation is a relative relationship between those two stocks' absolute prices. The consideration of this cannot be neglected, and otherwise, the trading signals based on judging the abnormality would be pointless. Specifically, suppose the stock price of BABA is generally \$200 higher than that of Tencent and both stocks most often have the same daily variation, according to our model. In that case, trading signals will be triggered when daily variation deviation is higher than a given threshold, such as 2 percent: BABA rises 3 percent whereas Tencent only rises 1 percent. Hence, our signals will be selling BABA and buying Tencent, but what if the absolute values of those two stocks have been in a relatively extreme situation, supposing BABA is now \$200 lower instead of higher than Tencent, which is very rare based on the relative relationship between those stocks, a high daily return of BABA thus cannot be considered as abnormal because it is actually a quick reversion on the absolute stock price. Thus, the trading signals based on the daily return to even sell the BABA is becoming absurd.

4. Conclusion

In this paper, 10 high-tech companies' stocks are examined to apply pairs trading strategies. One reasonable stock pair (Alibaba and Tencent) is selected according to the cointegration test and correlation test notion since they are highly cointegrated and highly correlated. Instead of traditional pairs trading strategy that using linear regression method and considering price spread as trade signal, this paper supplements the research on predicting maximum returns based on the pairs trading strategy of ARIMA and Z test. It also plays a certain role in reducing risks, especially in the face of major unexpected shocks. In addition, the introduction of the time series analysis method into pairs trading can provide a new idea to extend the application of pairs trading in financial markets. Forecasting financial data with the method of time series analysis can be regarded as a new way to maximize the rate of return. The setting of stop loss signal can provide a reference for other research on maximizing return rate, that is, to find a reasonable stop loss signal to reduce the volatility of return rate caused by the change of market environment or reduce trading cost by limiting unnecessary trades. Similarly, the positive cumulative return of the Z-test proves the feasibility of this approach. Since the returns are even higher in the Z-test methods, it is reasonable to construct a Z-test method in the future to contrast it with other models before model selection. Therefore, a pairs trading strategy based on ARIMA and Z test can be used to not only help investors to match appropriate stocks pairs to avoid risks but also help investors choose appropriate stop-loss signals and optimal models to reduce possible losses in adverse market conditions. To further reinforce this research, using one model's trading signals to confirm or at least enhance the accuracy of another trading signal may be future works for deeper studies. Or perhaps, combining traditional pair trading strategy, using absolute price comparison to generate trading signals, with this novel attempt in this paper to further improve the stability and liability could also be our future topic.

References

- [1] Boyacioglu M A, Avci D. An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock market return: the case of the Istanbul stock exchange. *Expert Systems with Applications*, 2010, 37(12): 7908-7912.
- [2] Wang B, Huang H, Wang X. A novel text mining approach to financial time series forecasting. *Neurocomputing*, 2012, 83: 136-145.
- [3] Roondiwala M, Patel H, Varma S. Predicting stock prices using LSTM. *International Journal of Science and Research (IJSR)*, 2017, 6(4): 1754-1756.
- [4] Ibrahim M H. Financial market risk and gold investment in an emerging market: the case of Malaysia. *International Journal of Islamic and Middle Eastern Finance and Management*, 2012.
- [5] Cheremushkin S V. Why D-CAPM is a big mistake? The incorrectness of the cosemivariance statistics. *The Incorrectness of the Cosemivariance Statistics* (February 1, 2009), 2009.
- [6] French C W. The Treynor capital asset pricing model. *Journal of Investment Management*, 2003, 1(2): 60-72.
- [7] Vidyamurthy G. *Pairs Trading: quantitative methods and analysis*. John Wiley & Sons, 2004.
- [8] Gatev E, Goetzmann W N, Rouwenhorst K G. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 2006, 19(3): 797-827.
- [9] Elliott R J, Van Der Hoek* J, Malcolm W P. Pairs trading. *Quantitative Finance*, 2005, 5(3): 271-276.
- [10] Andrade S, Di Pietro V, Seasholes M. Understanding the profitability of pairs trading. Unpublished working paper, UC Berkeley, Northwestern University, 2005.
- [11] Andrade S, Di Pietro V, Seasholes M. Understanding the profitability of pairs trading. Unpublished working paper, UC Berkeley, Northwestern University, 2005.
- [12] Hong G, Susmel R. Pairs-trading in the Asian ADR market. University of Houston, Unpublished Manuscript, 2003.

- [13] Gillespie T, Ulph C. Pair trades methodology: a question of mean reversion. Proceedings of International Conference on Statistics, Combinatorics and Related Areas and the 8th International Conference of Forum for Interdisciplinary Mathematics, NSW. 2001.
- [14] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model, 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE, 2014: 106-112.
- [15] Pai P F, Lin C S. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 2005, 33(6): 497-505.
- [16] Merh N, Saxena V P, Pardasani K R. A comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting. Business Intelligence Journal, 2010, 3(2): 23-43.
- [17] Nochai R, Nochai T. ARIMA model for forecasting oil palm price. Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and applications. 2006: 13-15.
- [18] Meyler A, Kenny G, Quinn T. Forecasting Irish inflation using ARIMA models. 1998.
- [19] Stouffer S A, Suchman E A, DeVinney L C, et al. The american soldier: Adjustment during army life. (studies in social psychology in world war ii), vol. 1. 1949.
- [20] Yahoo Finance Website: <https://finance.yahoo.com/>