

Generalization of Cointegration Arbitrage Based on GARCH Model, Improved by Marginal Probability Distribution

Zixu Hou *

Jinan University - University of Birmingham Joint Institute, Jinan University

* Corresponding author. Email: houzixu2019051724@stu2019.jnu.edu.cn

Abstract. As the Russia Ukraine War broke out in February, 2022, a series of international sanctions led to a panic among international hot moneys and slumps in global stock markets. For such a volatile market, hedging strategy is in a position to avoid risks. In finance, the correlation among financial derivative is always a topical issue and investors are concerned about the statistical methods to model the volatility. Recent years, the rapid development in data science makes it possible to record and manipulate massive data of financial derivative series, of which activates statistical arbitrage, a hedging strategy by pairs trading. The most widespread method in practice is cointegration arbitrage and many experiments show that it can always generate positive profits. This paper introduces the backgrounds of statistical arbitrage and relative definitions in cointegration arbitrage like time series, stationary, autocorrelation process, integration and cointegration. Based on the features of financial derivative series, this paper also illustrates error correction model, ARCH model and GARCH model. Finally, this paper shows the current strategy of cointegration arbitrage and raises some improvement by marginal probability distribution.

Keywords: Statistical arbitrage, Trading strategy, Spread series, Cointegration arbitrage, GARCH model, Marginal probability distribution.

1. Introduction

As the Russia Ukraine War broke out in February, 2022, international hot money was caught in panic and slumps occurred in global stock markets. After National Association of Securities Dealers Automated Quotations (NASDAQ) had slumped by near 4% in one market day, Cotation Assistée en Continu 40 (CAC40) and Der DAX slumped by more than 3%. According to the data from Straight Flush, the net outflow of South Bound had reached 13.336 billion Chinese yuan on March 15 and Shanghai Securities Composite Index slumped to 3061.39, breaking the lowest record since August, 2020.

In these days, financial business gets considerable development and quantitative trading should be the most well-known. Statistical arbitrage is a hedging strategy of quantitative short-swing trading based on statistical methods. It generates stable profits from the difference of two specific financial derivative series instead of the individual trends of them.

The earliest statistical arbitrage strategy can be traced back to 1920, raised by Livermore, a legend in Wall Street. For some two stock price series sharing similar trends, his trading strategy is to long the unusual lower one and short the unusual higher one [1].

Until the end of the 20th century, statistical arbitrage strategy had not faded back into financial business. Board and Sutcliffe use cointegration to research on the Nikkei 225 Stock Index among markets, and prove that there exist cross-market arbitrage opportunities on Nikkei 225 Stock Index [2]. After that, Gatev et al. defined the distance between different stock prices to select the trading pairs, and tested the American stock data from 1962 to 1997, achieving 12% annualized rate of return. Later in 2006, Gatev et al. tested the data from 1998 to 2002 again, which still made 11% annualized rate, performing the strong stability of statistical arbitrage strategy [3].

In the 21st century, a number of scholars summarized the system of statistical arbitrage. Hogan et al. gave the strict definition of statistical arbitrage [4]. They indicated that statistical arbitrage comes from the weak-form of market efficiency and also proved that it still works on variable outlays under the condition of liquidity limitation. It went a step further studying the innate character of statistical

arbitrage. Do et al. summarized the methods and strategies for statistical arbitrage, structuring an initial statistical arbitrage system [5]. Later, Elliott et al. used Markov chain under the assumption of mean reversion to maximize the expectation maximum algorithm in the fitted models [6]. Thomaidis et al. added neural network to the GARCH model, providing an example for the later innovations [7].

In these days, statistical arbitrage is an outstanding strategy to generate stable profits. Statistical arbitrage is a kind of neutral strategy by inter-period, inter-market or inter-commodity hedging. Hedging not only provides an alternative way for investors, but also contributes to the balance in the financial market. Statistical arbitrage always performs in the form of automatic programmed trading, because it contains complicated and dynamic computing processes, and requires high response speed when an arbitrage opportunity turns up. Generally speaking, statistical arbitrage has not become widespread in the public but in the professional investment field. The most widespread method in practice is cointegration arbitrage and many experiments show that it can always generate positive profits. However, researches and results will not open to the public as business secrets.

This paper introduces the backgrounds of statistical arbitrage and relative definitions in cointegration arbitrage like time series, stationary, autocorrelation process, integration and cointegration. Based on the features of financial derivative price series, this paper also illustrates error correction model, ARCH model and GARCH model. Then, this paper shows the current strategy of cointegration arbitrage and makes some improvements by marginal probability distribution. Finally, conclusion and expectation are given in the last section.

2. Principles of Cointegration Arbitrage

2.1. Definition of Statistical Arbitrage

Hogan et al. (2003) gave a strict definition of statistical arbitrage as follows [4].

Definition. (Statistical Arbitrage) A strategy, which value at time t is $v(t)$, is called a statistical arbitrage when it satisfies the following conditions:

$$\text{initial wealth } v(0) = 0 \quad (1)$$

$$\lim_{t \rightarrow \infty} E[v(t)] > 0 \quad (2)$$

$$\lim_{t \rightarrow \infty} P[v(t) < 0] = 0 \quad (3)$$

$$\text{if } \forall t < \infty, P[v(t) < 0] > 0, \text{ then } \lim_{t \rightarrow \infty} \frac{\text{Var}[v(t)]}{t} = 0 \quad (4)$$

Remark. According to (3), it allows that $v(t) < 0$ for some finite t , which means that losses may turn up in statistical arbitrage, but its probability converges to 0 as t goes to infinity. Combining condition (2) and (4), we know that as t increases, $v(t)$ has larger probability to be positive. Therefore, it can be regarded as a strategy that generates positive profits with no risks when t is large enough.

2.2. Stationary Time Series and Autoregressive Model

The price series in financial market are time series. There are strictly stationary and weakly (covariance) stationary time series. The former one has a fixed probability distribution all the time and the latter one just requires some constant statistical quantities as follows:

Definition. (Stationarity time series) A time series y_t is weakly (covariance) stationary if it satisfies the following conditions:

$$E(y_t) = \mu \quad (5)$$

$$Var(y_t) = E[(y_t - \mu)^2] = \sigma^2 \quad (6)$$

$$Cov(y_t, y_{t-k}) = E[(y_t - \mu)(y_{t-k} - \mu)] = \gamma_k, \quad k = 1, 2, \dots \quad (7)$$

Remark. A time series is stationary if its mean, variance, and autocovariance are independent of time. Informally, it is one that is not explosive, nor trending, and nor wandering aimlessly without returning to its mean.

To model time series, we should determine the number of dependent lags for eliminating serial correlation in errors. Similar to the sample correlative coefficient, the sample autocorrelation coefficient between observations that are k -lag apart is given by:

$$\rho_k = \frac{Cov(y_t, y_{t-k})}{\sqrt{Var(y_t)Var(y_{t-k})}} = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \quad (8)$$

For testing an autocorrelation is significantly different from zero, a suitable test statistic is:

$$Z_k = \frac{\rho_k - 0}{\sqrt{1/T}} \sim N(0,1) \quad (9)$$

By the hypothesis test with significant level, we can obtain an autoregressive model, denoted AR(p) of y_t , where p is the number of lags.

In AR model, we can predict the values of y_t if we have known the information before time t . But if the time series is nonstationary, the result will make no sense. There are many tests for determining whether a series is stationary or not. The most popular one is the Dickey-Fuller test. According to time series plot, we select the model fitted and use the corresponding test below.

Firstly, we show the Dickey-Fuller Test 1 (without a constant and without a trend):

An AR(1) process $y_t = \rho y_{t-1} + v_t, v_t \sim N(0, \sigma_v^2)$ is stationary when $|\rho| < 1$. We make the first difference of y_t as follows:

$$\Delta y_t = \gamma y_{t-1} + v_t \quad (10)$$

Where $\Delta y_t = y_t - y_{t-1}$ and $\gamma = \rho - 1$. We estimate γ by least squares principle. Then we have a hypothesis test:

$$H_0 : \gamma = 0 \quad (11)$$

$$H_1 : \gamma < 0 \quad (12)$$

Note that the usual t-statistic table is not suit for the test here, and we need to compare the test value τ with the critical values for the Dickey-Fuller test.

Next, we show the Dickey-Fuller Test 2 (with a constant but without a trend):

For an AR(1) process $y_t = \alpha + \rho y_{t-1} + v_t, v_t \sim N(0, \sigma_v^2)$, we similarly make the first difference of y_t , and find out the significance of $\gamma = 0$.

Finally, we show the Dickey-Fuller Test 3 (with a constant and a trend):

For an AR(1) process $y_t = \alpha + \lambda t + \rho y_{t-1} + v_t, v_t \sim N(0, \sigma_v^2)$, we similarly make the first difference of y_t , and find out the significance of $\gamma = 0$.

2.3. Spurious Regression and Cointegration

The regressions are said to be spurious if the result is from the nonstationary series. It will mislead to a significant relationship when there is none. However, the price series in the financial market are nonstationary in practice. The reasons may be the policy changes, bull news, wars and any other unpredictable events. Therefore, to apply our AR model, we need to find the stable relationships among the price series.

Let's think about the nonstationary price series of two derivatives, which are y_t and x_t respectively. Set $\epsilon_t = y_t - \beta_1 - \lambda t - \beta_2 x_t$, where β_1, β_2 and λ are the parameters to be estimated. There is an important case when ϵ_t is stationary time series, then we say y_t and x_t are cointegrated [8]. Cointegration implies that y_t and x_t share similar stochastic impact from the financial market, and, since the ϵ_t is stationary, they never diverge too far from each other.

The test of cointegration is effectively a test of the stationarity of the residuals. If the residuals are stationary, then y_t and x_t are said to be cointegrated; if not, then that y_t and x_t are not cointegrated, and any apparent regression relationship between them is said to be spurious.

The test for the stationarity of the residual is based on the test equation:

$$\Delta e_t = \gamma e_{t-1} + v_t, \Delta e_t = e_t - e_{t-1} \quad (13)$$

Not like the models in Dickey-Fuller test, there is neither a constant nor a trend in the time series of residuals, since they have been expressed by the parameters β_1 and λ . According to the significance that β_1 and λ equal to zero, we choose different critical values for the cointegration test.

To obtain a stable relationship between two price series, we select y_t and x_t cointegrated of which the fitted model is:

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t \quad (14)$$

Where ϵ_t is independent identical distributed with mean 0 and variance σ^2 . Then we define the spread $r_t = y_t - \beta_2 x_t = \beta_1 + \epsilon_t$, which is a stationary time series with mean β_1 and variance σ^2 .

It is exciting that we find the stationary spread series r_t with mean β_1 and it can eliminate the similar impact on y_t and x_t from unpredictable events. Since the mean reversion theorem, now we can develop an abecedarian statistical arbitrage strategy by pairs trading as follows.

Consider a portfolio with initial wealth value 0 when the riskless interest rate is 0. We borrow wealth $y_s - \beta_2 x_s$ at time s , which equals the spread r_s . Then we can buy a unit of y_s long and β_2 units of x_s short when $r_s < \beta_1$. The difference between r_s and β_1 is the residual, denoted by e_s . As soon as r_t goes back to β_1 , we cover the portfolio with the wealth $r_s - \beta_1 > 0$. Then we can generate positive profits with no risks when time t goes to large enough. Symmetrically, in the case of $r_s > \beta_1$, we can also generate positive profits with no risks when time t goes to large enough.

2.4. The Error Correlation Model

However, it is not certain that when the spread will come back to its mean. The error correlation model can estimate the mean reversion speed by the relationship between the price series and their first differences.

Definition. (Order of integration) If the first difference of a process y_t is stationary, y_t is said to be integrated of order one, denoted by $I(1)$. Stationary series are said to be $I(0)$. A relationship between $I(1)$ variables is referred to as a long-run relationship, while that of $I(0)$ variables is referred to as a short-run relationship.

Assuming y_t and x_t are cointegrated and $I(1)$ variables, equivalently $\Delta y_t = y_t - y_{t-1}$ and $\Delta x_t = x_t - x_{t-1}$ are $I(0)$ variables, we can construct a dynamic relationship between Δy_t and Δx_t which embeds a cointegrating relationship, known as short-run error correction model. Consider an autoregressive distributed lag model of y_t and x_t :

$$y_t = \delta + \theta_1 y_{t-1} + \delta_0 x_t + \delta_1 x_{t-1} + v_t \quad (15)$$

It can be shown that

$$\Delta y_t = -\alpha[y_{t-1} - \beta_1 - \beta_2 x_{t-1}] + \delta_0 \Delta x_t + v_t \quad (16)$$

Where $\alpha = 1 - \theta_1$, $\beta_1 = \frac{\delta}{1-\theta_1}$, $\beta_2 = \frac{\delta_0+\delta_1}{1-\theta_1}$

Remark. In this model, $y_{t-1} - \beta_1 - \beta_2 x_{t-1}$ is the error term from the cointegration at time $t - 1$, which will contribute to the correction of the cointegration of Δy_t and Δx_t , multiplied by the mean reversion speed α . Explicitly we prefer the pairs of price series which have larger absolute α .

2.5. Conditional Heteroskedasticity

So far, we have no ideas about the best time to open a position yet. In experiments, the spread series are always conditional heteroskedastic. Here we illustrate models for conditional heteroskedasticity series to estimate the conditional variance at time t when the information before time t is known. It can be used to determine the best trading time.

Autoregressive Conditional Heteroskedasticity model, abbreviated to ARCH model, is popular in finance. It assumes that the square of error term ϵ_t at time t is a nonnegative linear combination of those in the past q periods, which formed as follows:

$$y_t = \mu_t + \epsilon_t \quad (17)$$

$$\epsilon_t = \sigma_t v_t \quad (18)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_q \epsilon_{t-q}^2 \quad (19)$$

Where $\alpha_0 > 0, \alpha_i \geq 0 \forall i > 0$, $\sum_{i=1}^q \alpha_i < 1$ and v_t is an independent and identical distribution with mean 0 and constant variance, μ_t is the expectation of y_t .

Bollerslev developed ARCH model into Generalized Autoregressive Conditional Heteroskedasticity model, abbreviated to GARCH(p, q):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \dots + \beta_p \sigma_{t-p}^2 \quad (20)$$

Where $\alpha_0 > 0, \alpha_i \geq 0, \beta_i \geq 0 \forall i > 0$, $\sum_{i=1}^q \alpha_i + \sum_{i=1}^p \beta_i < 1$.

Remark. In GARCH model, the parameters are too complicated to estimate by ordinary least squares estimation. We use maximal likelihood estimation, setting an initial value of σ_t^2 and computing σ_t^2 by recurrence. We always assume that v_t is normal distributed or t-distributed.

2.6. Current Strategy

In Section 2.3, we have developed an abecedarian statistical arbitrage strategy. In the last section, we are able to predict the conditional variance of the price series at time t to see how much the price will vary from the conditional expectation at some significance. There is a current strategy in stock markets based on GARCH model following.

At the beginning, we need to select the trading pair, of which price series y_t and x_t in financial market which have strong correlation. We can plot the price series for different lengths of period, such as 5 minutes, 15 minutes, 60 minutes, 1 market day and so on, and compute the correlation coefficients between those that look like sharing a similar trend. Moreover, the error correction model can be applied to compare their mean reversion speed coefficients.

In the case that spread series $r_t = y_t - \beta_2 x_t$ is a stationary time series, r_t can be fitted by GARCH model to predict the conditional variance σ_t^2 at time t if the information before t is known. Set $\{\widehat{\beta}_1 - k_1 \widehat{\sigma}_t, \widehat{\beta}_1 + k_1 \widehat{\sigma}_t\}$ as the opening threshold at time t , where k_1 is a positive constant. Here we define a unit of spread r_t as a unit of y_t and $-\beta_2$ units of x_t . When the spread series r_t reaches the opening threshold $\widehat{\beta}_1 - k_1 \widehat{\sigma}_t$, we buy a unit of r_t . While r_t reaches the opening threshold $\widehat{\beta}_1 + k_1 \widehat{\sigma}_t$, we sell a unit of r_t . Besides, to generate profits and control risk, we set the covering thresholds $\{\widehat{\beta}_1 - k_2 \widehat{\sigma}_t, \widehat{\beta}_1 + k_2 \widehat{\sigma}_t\}$ and stopping thresholds $\{\widehat{\beta}_1 - k_3 \widehat{\sigma}_t, \widehat{\beta}_1 +$

$k_3\hat{\sigma}_t\}$ depending on experience, where $0 < k_2 < k_1 < k_3$. We can cover the portfolio as soon as the value of $r_t - \beta_1$ reaches the covering thresholds or stopping thresholds, no matter it wins or loses. It is a key point to determine the vector $k = (k_1, k_2, k_3)^T$, for instance a small k_1 meaning high trading frequency will cause large cost while a large k_1 will miss some arbitrage opportunities.

For the best k_1 , a common method is to test different values of k_1 by fixing step length, and approximating the best interval of k in MATLAB [9]. Furthermore, in the field of statistical learning, size of training data is also important in the result of profits. Setting a proper window length to update the model as soon as a new data point comes out can make the model dynamic [10].

3. application of Marginal probability distribution

3.1. Improvement in Opening and Covering Thresholds

The current strategy above has a good use on conditional variance, based on the mean reversion theorem, believing that the spread r_t will go back to its mean in the future. But the marginal probability distribution of the spread r_t may engender better arbitrage opportunities.

We first define extreme points of time series as follows:

Definition. For a time series $x_t, t = 1, 2, \dots$, we say x_s is an extreme point when $x_s = \max\{x_{s-1}, x_s, x_{s+1}\}$ or $x_s = \min\{x_{s-1}, x_s, x_{s+1}\}$, called local maximum and local minimum respectively.

For any financial derivative which allows to be at long or short position, undoubtedly, a good investing strategy is buying at low price and selling at high price. Then it comes up with the best strategy that is buying at each local minimum and selling at each local maximum, which could generate the maximal profits.

Theorem. For any continuous and bounded function, there exists a local maximum between any two local minimums.

Corollary. For any time series, a local maximum follows by a local minimum and a local minimum follows by a local maximum.

Remark. The corollary indicates that the local minimum and local maximum will take places by turns, responding to buying long and short in turn.

Certainly, it is impossible to make sure that the price at time t is an extreme point. But we can test this hypothesis at some significance level, via the probability distribution of the spread r_t .

For a stationary spread series, we fit it by GARCH model:

$$r_t = \mu_t + \epsilon_t \quad (21)$$

$$\epsilon_t = \sigma_t \nu_t \quad (22)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \dots + \beta_p \sigma_{t-p}^2 \quad (23)$$

Where μ_t is the expectation of r_t . Assuming that ν_t is normal distributed, we estimate α_i, β_i and the parameters in μ_t can be estimated by maximal likelihood estimation. The fitted model is given by

$$\hat{r}_t = \hat{\mu}_t + \hat{\epsilon}_t \quad (24)$$

$$\hat{\epsilon}_t = \hat{\sigma}_t \nu_t \quad (25)$$

$$\hat{\sigma}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 \epsilon_{t-1}^2 + \dots + \hat{\alpha}_q \epsilon_{t-q}^2 + \hat{\beta}_1 \sigma_{t-1}^2 + \dots + \hat{\beta}_p \sigma_{t-p}^2 \quad (26)$$

In (26), from the probability density function of $\hat{\alpha}_i$ and $\hat{\beta}_i$ (this may be complicated to solve), we can obtain the probability density function of $\hat{\sigma}_t^2$. Then by (24), we can obtain the marginal

probability density function of $\hat{r}_t$, which is denoted by $f_{\hat{r}_t}(\hat{r}|I_{t-1})$, where I_{t-1} represents the information before t .

Denote the conditional expectation of $\hat{r}_t$ as $\mu_t(\widehat{I_{t-1}})$ when the information before time t is known. Define the marginal two-sided confidence interval of $\hat{r}_t$ at significant θ as $[\mu_t(\widehat{I_{t-1}}) - k, \mu_t(\widehat{I_{t-1}}) + k]$, such that k holds the equation:

$$\int_{\mu_t(\widehat{I_{t-1}})-k}^{\mu_t(\widehat{I_{t-1}})+k} f_{\hat{r}_t}(\hat{r}|I_{t-1}) d\hat{r} = \theta. \quad (27)$$

Similarly, at time $t + 1$, the marginal two-sided confidence interval of $\hat{r}_{t+1}$ at significant θ is $[\mu_{t+1}(\widehat{I_t}) - k, \mu_{t+1}(\widehat{I_t}) + k]$ when the information before time $t + 1$ is known, such that k holds the equation:

$$\int_{\mu_{t+1}(\widehat{I_t})-k}^{\mu_{t+1}(\widehat{I_t})+k} f_{\hat{r}_{t+1}}(\hat{r}|I_t) d\hat{r} = \theta. \quad (28)$$

In the case of the spread r_s at time s smaller than r_{s-1} , we accept $H_0: r_s < r_{s+1}$ and reject $H_1: r_s \geq r_{s+1}$ at significance θ , when r_s is smaller than the infimum of the marginal two-sided confidence interval of $\hat{r}_{s+1}$ at significance θ . Then we believe that r_s is a local minimum and buy one unit of spread r_s (maybe cover the portfolio). Similarly, In the case of the spread r_s at time s larger than r_{s-1} , we accept $H_0: r_s > r_{s+1}$ and reject $H_1: r_s \leq r_{s+1}$ at significance θ , when r_s is larger than the supremum of the marginal two-sided confidence interval of $\hat{r}_{s+1}$ at significant θ . Then we believe that r_s is a local maximum and sells one unit of spread r_s (maybe cover the portfolio).

Finally, in order to satisfy the condition of statistical arbitrage that generating positive profits with no risks if time t goes to large enough, we also cover the portfolio as soon as r_t goes back to its mean $\hat{\mu}_t$. The portfolio should be at a long position when $r_t < \hat{\mu}_t$, at a short position when $r_t > \hat{\mu}_t$.

Thus, we have developed a set of opening and covering thresholds $\{\mu_{t+1}(\widehat{I_t}) - k, \hat{\mu}_t, \mu_{t+1}(\widehat{I_t}) + k\}$.

3.2. Informal Stopping Thresholds

In the current strategy, we set our stopping thresholds as $\{\widehat{\beta}_1 - k_3 \widehat{\sigma}_t, \widehat{\beta}_1 + k_3 \widehat{\sigma}_t\}$, where k_3 is a positive constant, depending on the conditional variance of spread $\hat{r}_t$ at time t . It may stop the investment in time as soon as the spread r_t is out of expectation too much. But in stock markets, a rebound always follows by an oversold or overbought of a stock price. Thus, selling our portfolio at once may be not the best choice.

Recall the marginal probability density function $f_{\hat{r}_t}(\hat{r}|I_{t-1})$ in the Section 3.1, which can be used to compute the p-value of the observation of r_t at time t , when the information before t is known. Denote the p-value at time t by p_t , which indicates the accuracy of our prediction at time t .

Instead of selling our portfolio as soon as r_t deviates its expectation too much, we can construct a test (maybe the inter-product of p_t) to stop thresholds based on a series of p_t for some recent t . In this way, more observations of r_t can be considered to determine our stopping thresholds.

4. Conclusion

This paper gives some relative definitions in cointegration arbitrage and introduces GARCH model to fit the spread series. It also shows the current strategy of cointegration arbitrage and raises some improvement by marginal probability distribution. The conclusions are as follows.

Firstly, through the cointegration, a stationary spread series r_t is found in the financial market. Since the variance of stationary time series is a constant, it can be shown the probability of the spread

r_t returning to its mean will converge to 1 as time t goes to infinity, making the current strategy satisfy the condition of statistical arbitrage. Secondly, the improvement of opening threshold in Section 3.1 extends the marginal expectation of σ_t^2 to the marginal probability distribution of r_t when the information before t is known. In words, it does not eliminate the conditional variance σ_t^2 as in current strategy but adds in the conditional expectation $\mu_t(I_{t-1})$ to predict the marginal probability distribution of r_t .

There are some deficiencies in this paper. Firstly, note that the estimation by GARCH model, is under the assumption that v_t is normal distributed or t-distributed, not the true distribution of σ_t^2 . It remains a problem to find more appropriate assumptions about the probability distribution of v_t . Secondly, the vector of k in current strategy plays a key role in the return. Fixing the test data, it can be regarded as a map from $k = (k_1, k_2, k_3)^T$ to the return. Multivariable and vector analysis methods may be available to solve the best k . Thirdly, for the improvement of opening and covering threshold, the significance θ takes the place of k_1 and k_2 in the current strategy so that θ will result in various profits depending on what algorithm to be chosen. Besides, in Section 3.2, researcher only gives an idea without a formal formula while experiments are also absent from Chapter 3. Lastly, researcher believes that there will be more developed methods and models applied in statistical arbitrage and the theoretical system will be more completed not far from the future.

References

- [1] E. Lefèvre, *Reminiscences of a Stock Operator*, Wiley, New Jersey, USA, 1923, DOI: 10.2307/2329337.
- [2] . Board, C. Sutcliffe, the dual listing of stock index futures: Arbitrage, spread arbitrage, and currency risk, *Journal of futures markets*, vol. 16, 1996, pp.29 - 54. DOI: 10.1002/ (SICI) 1096 - 9934 (199602)16: 1<29: AID-FUT2>3.0.CO; 2 - J.
- [3] E. Gatev, W. N. Goetzmann, K.G. Rouwenhorst, *Pairs Trading: Performance of a Relative-Value Arbitrage Rule*, *Review of financial studies*, vol. 19, 2006, pp. 797-827. DOI: 10.1093/rfs/hhj020.
- [4] S. Hogan, R. Jarrow, M. Teo, et al., Testing market efficiency using statistical arbitrage with applications to momentum and value strategies, *Journal of financial economics*, vol. 73, 2004, pp. 525 - 565. DOI: 10.1016/j.jfineco.2003.10.004.
- [5] B. Do, R. Faff, K. Hamza, *A New Approach to Modeling and Estimating for Paris Trading*, in: *Proceeding of 2006 Financial Management Association European Conference*, Stockholm, 2006.
- [6] R. J. Elliott, J. Van, W.P. Malcolm, *Pairs Trading*, *Quantitative Finance*, vol. 5, 2005, pp. 271 - 276. DOI: 10.1080/14697680500149370.
- [7] N.S. Thomaidis, N. Kondakis, G.D. Dounias, *An Intelligent Statistical Arbitrage Trading System*, in: *Hellenic Conference on Advances in Artificial Intelligence*, Springer-Verlag, vol. 3955, pp. 596 - 599, 2006, DOI: 10.1007/11752912_77.
- [8] R. Eagle and C. Granger, *Co-Integration and Error Correlation: Representation, Estimation, and Testing*, *Econometrica*, vol. 55, 1987, pp. 251 - 276. DOI: 10.2307/1913236.
- [9] Y.L. Guo, *Research on the statistical arbitrage strategy based on China Ping An A+H Shares*, Harbin, Harbin Institute of Technology, 2019.
- [10] B. Wang, *the statistical arbitrage strategy and the engineering realization of futures based on cointegration*, *Times finance*, vol. 18, 2021, pp.74 - 79.