LUND UNIVERSITY
School of Economics and Management

The application of Ornstein-Uhlenbeck Process model and ARCH/GARCH model in statistical arbitrage

Jianbo Wang
Jianyang Fang

05-2017

Master's Program in Finance

Supervisor: Hossein Asgharian

Abstract

Statistical arbitrage is widely used in the quantitative based trading strategies. In this paper, we mainly use Ornstein-Uhlenbeck (O-U) process model and the GARCH model to estimate the parameters and verify trading signals for the statistical arbitrage. In addition, a new model is created through the combination of O-U model and GARCH model. To estimate the models, HuangXia Bank and Industrial Bank are selected due to the highest correlation among the banks.

Key words: statistical arbitrage, Ornstein-Uhlenbeck model, GARCH model

Table of Contents

1 Introduction

Arbitrage is a very common concept in finance. It means investors can buy the lower-price product and sell it at a higher price at the situation that the financial products have the different prices in the different financial market. In the viewpoint of Velissaris (2010), the ideal arbitrage is a costless strategy which provides investors with the opportunity to get abnormal revenue but with no risk to lose. Nowadays, many different kinds of arbitrage strategies are used by investors and fund managers to obtain high return and reduce the risk.

Statistical arbitrage is one specific form of the arbitrage trading strategies. In the financial market, statistical arbitrage trading is an investing process based on mathematical models. More specifically, statistical arbitrage is using a mathematical model relying on historical data to guide the investors and fund managers to forecast the future value of portfolios to build an arbitrage trading strategy. The mechanism of this kind of arbitrage trading is to research the financial markets that are out of equilibrium level. In other words, the price of a stock is supposed to a certain equilibrium level, and it fluctuates around this kind of level. If the price series moves away from this equilibrium level, it is expected to move back in a certain period. Thus, the opportunity of arbitrage will happen in this kind of process.

In this paper, the specific trading strategy of statistical arbitrage will be researched and discussed. We use the specific trading rule proposed by Bock and Mestel (2009), and from an arbitrage trading strategy, when the price of the stock moves away from an equilibrium level. Moreover, we use the trading signal indicator suggested by Bertram (2009), based on the Ornstein-Uhlenbeck (O-U) process, to define the particular trading strategy. Since the model is built on the time series and the equilibrium level, some related data tests are necessary for the research. Especially, the stationary test should be considered before the model building. Thus, the Dickey-Fuller (D-F) test, Augmented Dickey-Fuller (ADF) test, and Phillips-Perron test, are used to test the stationary of the sequence. Our model and trading strategies are based on the real stock data. More specifically, the stock prices for stocks in the bank industry of China are used to estimate the statistical arbitrage. The total period consists of three years, where two years of data is used to estimate the parameters and verifies the trading signals, and;

the rest of data is used to calculate the return according to the estimated parameters and trading signals.

Through simulating the arbitrage trading process, the highest return is gained by Ornstein-Uhlenbeck process model and from the data out of sample. The combination of O-U model and GARCH model has better performance than pure GARCH model. Perhaps, the stock prices of Chinese banks are comparably stable. This situation causes the arbitrage interval is narrow, and volatility effect causes the return is low in some periods. We contribute to the literature by using a combination of Ornstein-Uhlenbeck model and GARCH model. Through the previous research of these two models, the parameters (Δ) of trading signal and the effect of volatility clustering in spread series are simultaneously considered in the new model. Although the stocks of Chinese banks have better results of arbitrage trading using Ornstein-Uhlenbeck process model, the stocks which are sharply fluctuated may have better arbitrage performance by the new model, but this conjecture may be verified using other data in future research.

In this paper, the next section will give the information about the background of the data that we choose. The related literature will be introduced and evaluated in the third part. After that, the statistical arbitrage trading strategy in section 4. In for the section 5, three different kinds of approaches of the trading signal will be introduced. Next, three different models will be applied to estimate and forecast for the real data and stocks respectively. In section 7, we will compare the three models with each other and state the contribution and the drawback of the research. Finally, this report will be evaluated and summarized in section 8.

2 Literature Review

Bondarenko (2003) introduced the concept of statistic arbitrage, which is different from other arbitrage strategies that cannot have negative payoffs. Statistic arbitrage allows the payoffs be negative as long as the average payoff in a trading period is nonnegative. He also indicated that this arbitrage strategy is a trading with zero-cost, thus in the process, the expected payoff of statistic arbitrage is positive, and meanwhile, the conditionally expected payoff in each trading period is nonnegative. Hogan et al. (2003) specified the concept of statistic arbitrage of Bondarenko (2003). They showed that the statistical arbitrage follows four conditions:

1. v (0) = 0
2. $\lim_{t\to\infty} E^P[v(t)] > 0$.
3. $\lim_{t\to\infty} P(v(t) = 0)$, and
4. $\lim_{t\to\infty} \frac{Var^P[v(t)]}{t} = 0$ If P (v(t) <0)>0 ∀t < ∞.

Which means that (i) This strategy is a zero initial cost (v(0) = 0) self-financing trading strategy (ii) it has the limitation of positive expected discounted profits, (iii) a probability of a loss approach to zero, and (iv) if the probability of the loss does not become zero in finite time, the time-averaged variance will converge to zero. In addition, they tested for the statistical arbitrage rule and researched the two strategies of momentum and value trading. Then, Jarrow et al. (2005) enlarge the set of statistic arbitrage at the foundation of the research of Hogan et al. (2003), and have the supplement of the consistence and the statistical power in the Bonferroni approach through the statistical methods. Meanwhile, they contributed several statistical arbitrage frameworks.

In 1999, Burgess researched the relationship of the components of FTSE 100 using the statistical arbitrage. He found that the co-integration model can be used in the statistical arbitrage strategy. Several years later, Alexander and Dimitriu (2004) applied the method of co-integration on the research of the tracking portfolio and index. They found that co-integration optimal portfolio, which has low- volatility, low- correlation with the market, is superior to the tracking error variance (TEV). Thomaisdis and Kondakis (2006) combined neural networking approach with statistical arbitrage. In this paper, an autoregressive

GARCH model, based on the neural networking, was to seek for the investment portfolio in the stock market. The results verified the possibility of using statistical arbitrage. Meanwhile, the hybrid computational intelligent system was introduced by Thomaidis and Dounias (2006). In this system, nonlinear neural network autoregressive models and GARCH parameterizations of volatility can be used to test the dynamics of the correction of statistical arbitrage opportunities with the pairs of assets. What's more, the authors applied the NN-GARCH model on the forecast of the dynamics of the statistical mispricing. Despite this model was qualified to forecast the short-term changes in volatile levels, it is still questionable giving the variable trading costs and market "frictions".After that, Meucci (2010) showed the multivariate Ornstein-Uhlenbeck process and summarized the discrete-time and continuous-time multivariate process. Furthermore, he interpreted the concept of co-integration and the relationship between it and statistical arbitrage by the illustration of the geometry of the Ornstein-Uhlenbeck dynamics. In addition, Cummins and Bucca (2011) paid their attention to the research of the quantitative trading in refined products markets. They operated the arbitrage process using the optimal statistical arbitrage trading model. In addition, both of them proposed the multiple hypotheses that detected the data snooping bias, and they showed the step-down procedure and the balanced step-down procedure. In their study, unlike the step-down procedure, the balanced procedure successfully identifies any profitable strategies and unbiased of trading applications. After that, Vidyamurthy (2014) argued that statistical arbitrage is based on the thought of relative pricing. He found that two stocks in the similar characteristics have approximately the same stock prices, and simultaneously the price spread could be regarded as the degree of mispricing, which analyzed that a large distance of the price spread has a higher degree of mispricing and then results in achieving a high probability of potential returns. The author also mentioned three parametric methods used in the trading signals in this article, namely, mixtures of Gaussians, ARMA model and Hidden Markov Model.

Huck (2009) described a selection method in the trading portfolio and as well as discussed the pros and cons of its application. He chose S&P 100 index stocks as a stock pool. And at the same time, Elman, a neural networking approach, was selected to predict different returns respectively, then the author adapted to Electre III method to rank for different profits. The result, Huck analyzed that this approach could obtain a perfect trading portfolio under the condition of highly correlated non-linearity.

In 2007, Chng analyzed how trading could obtain profits from four crucial internal components, namely, “negative serial covariance in idiosyncratic returns”; “positive cross-serial covariance in idiosyncratic returns of collaborative firms”; “discrepancy in the unconditional expected return of component stocks” and “lead-lag effects in component stock price reaction to unexpected common factor realizations”. The author also talked about two applications of the model in practical ways. One is for comparing the economic importance with other profitable components. Another is to set a connection between the profitable components and the restriction appeared during the formation. To apply pairs trading in a practical way, in 2009, Perlin set an example of Brazilian financial market to discuss whether pairs trading strategy could apply to the efficient market or not. He also discovered the influence of the arbitrage that data on the condition of different fluency may have. The consequence shows that it is available to use pairs trading strategy in the Brazilian market and its ability of the profitability on the market has a high relation with the data fluency. When selecting daily statistics as sample data, the strategy could help earn a high profit.

3 Data and Methodology

3.1 Date choice and test

According to the concept of statistical arbitrage, we know that the basic idea of this strategy is using the statistical analysis tools to research stocks which have the stable price relationship. In other words, if the stocks have a certain stable relationship, the portfolio of the stocks has a certain equilibrium level. When the price of the portfolio moves away from this level, the equilibrium relationship will drive the price to go back to the equilibrium level. Therefore, the stationary is an important concept in the statistical arbitrage and the stable relationship of stocks is a necessary condition for the research. The choice of the stocks which have the stationary is the first task in the research.

However, there are numerous stocks in the real market. It is difficult to compare and analyze all of the stocks. Therefore, we decide to choose the stocks which are from the same industry. The stocks in the same industry have similarity, have a low difference of the risk factors, and have high probability of similar price trend. In addition, the correlation can be used to choose the appropriate data. For instance, HuangXia Bank and Industrial Bank, both of them come from China bank area, and their correlation is close to 1. Therefore, the stock prices of them can be used to estimate our model. The correlation formula is:

$$\rho = \left| \frac{cov(P_a, P_b)}{\sqrt{var(P_a)var(P_b)}} \right| \tag{1}$$

Where P_a and P_b are the price of stock A and stock B. The pair of data has a high correlation when the value of ρ close to 1. Thus, this pair of stocks can be appropriate to take arbitrage trading.

The test of stationary is necessary for a time series. If the stationary test is ignored, the spurious regression will appear. Then, the total research will become unmeaning. Therefore, the data should be tested before modeling the statistical arbitrage trading. In the section of

Econometrics, there are three kinds of methods to test the stationary: DF (Dickey-Fuller) test, ADF (Augmented Dickey-Fuller) test, and Phillips-Perron test.

ADF test is derived from the DF (Dickey-Fuller) test, which is applied in the AR(1)model ($y_t = \varphi y_{t-1} + u_t$). It is a common approach to test the property of stationary for a time series. The test model can be shown:

$$Y_t=\rho Y_{t-1}+\sum_{i=1}^{p}\beta_i Y_{t-1} + u_t \qquad (2)$$

Where Y_t is time series, t is Time trend.

The null hypothesis is the time series is not a stationary one (H_0: ρ =0), to the contrary, the alternative hypothesis of ρ <0 means that it is stationary. The Phillips-Perron test is similar to the ADF test, has the same null hypothesis, and usually gives the same conclusion. In the research, we decide to use ADF (Augmented Dickey-Fuller) test to test the stationary of the stocks.

3.2 Co-integration test

The non-stationary series $Y_t$can become stationary after being differenced d^{th} order. This kind of series is integration of d^{th} order , we write as $Y_t \sim I(d)$. In most case, if we linearly combine two variables that are I(d), then the combination will also be I(d), and they are co-integrated.

In the research, the Engle-Granger two-step approach can be used to test the co-integration. This method is to take the ADF test for the residuals of the regression. The variables have a stable relationship and co-integration, and then the residuals of the regression equation combined with the variables should be stable. The step of the co-integration test is:

(1). When the variables are integrated of order 1(I(1)), estimating the co-integration regression using OLS

$$Y_t=\beta_0+\beta_1 X_t+\widehat{e_t} \qquad (3)$$

Where $\widehat{e_t}$ is expressed the residual sequence.

(2). Test the stationary of residuals ($\widehat{e_t}$) by the ADF test; the residual equation can be shown:

$$\Delta\widehat{e_t}=\rho\widehat{e_{t-1}}+\sum_{i=1}^{p}\beta_i\Delta\widehat{e_t}+\varepsilon_t \quad (4)$$

The null hypothesis is the time series is non-stationary (H_0: ρ =0). To the contrary, the alternative hypothesis of ρ <0 means that it is stationary, and the variables have a co-integration relationship.

3.3 Error correction model

For the non-stationary variables, they can become stable after being differenced d[th] order, and then the regression model can be built ($\Delta y_t = \beta_1\Delta x_1 + v_t$). However, in the real world, the dependent variables and independent variables are not static equilibrium. More specifically, the lagged independent variables (x_{t-1}) also affect the dependent variables (y_t). Obviously, a simply approach of difference cannot solve the problem of the non-stationary time series. Thus, the Error correction model should be used to solve this kind of problem.

The variables have a long-term relationship:

$$Y_t=\gamma_0+\gamma_1 X_t+\widehat{e_t} \quad (5)$$

However, the real relationship of the variables is:

$$Y_t=\gamma_0+\gamma_1 X_t+\gamma_2 X_{t-1}+\gamma_3 Y_{t-1}+\widehat{e_t} \quad (6)$$

So, we can write the ECM as:

$$\Delta y_t = \gamma_1\Delta x_1 - \lambda(y_{t-1} - \alpha_0 - \alpha_1 x_{t-1}) + \widehat{e_t} \quad (7)$$

Then, the second step of the Engle-Granger two-step approach can be used in the ECM model:

$$\Delta y_t = \beta_1\Delta x_1 + \lambda(\hat{u}_{t-1}) + \widehat{e_t} \quad (8)$$

Where $\hat{u}_{t-1} = y_{t-1} - \hat{\gamma}_0 - \hat{\gamma}_1 x_{t-1}$, and this model takes into account both long-term and short-term effect. Meanwhile, $\lambda = 1-\gamma_3$, $\alpha_0 = \gamma_0/(1-\gamma_3)$, $\alpha_1 = (\gamma_1+\gamma_2)/(1-\gamma_3)$.

In the research, $\alpha_1$is the proportion of hedge trading. For instance, the stock A and B are chosen to take arbitrage, the portfolio of the stock A and B is $\text{Price}^A - \alpha_1\text{Price}^B$. Specifically, at time t, we buy one unit this kind of portfolio means we buy one unit stock A and sell α_1 stock B.

4 Building trading signal

4.1 The rule of the arbitrage trading

The fundamental concept of the statistical arbitrage is mean-reversion. When the price of stock departs the average level, investors can get the opportunity to take arbitrage, until the price goes back to the equilibrium level. In the research of Bock and Mestel (2009), the trading signal is defined as $\Delta\mathrm{Std}_{\mathrm{spread}}$, where the $\mathrm{Std}_{\mathrm{spread}}$is the standard deviation of the spread of the stocks. More specifically, the short sell trading signal is $\mathrm{mean}_{\mathrm{spread}} + \Delta\mathrm{Std}_{\mathrm{spread}}$, which means the investors should short sell the portfolio when the price of the portfolio rises to $\mathrm{mean}_{\mathrm{spread}} + \Delta\mathrm{Std}_{\mathrm{spread}}$ level. Meanwhile, the long buy trading signal is $\mathrm{mean}_{\mathrm{spread}} - \Delta\mathrm{Std}_{\mathrm{spread}}$.

However, the price may move far away from the average level for a certain period. Although the price will return to the average, we have to consider the time cost and the commission cost. Thus, the investors gain loss actually. In order to avoid this situation, the stop loss point should be used in the model. In the research, we set that the difference between stop loss point and the mean of the spread is $2\mathrm{Std}_{\mathrm{spread}}$,. Therefore, the high stop loss point for the statistical arbitrage is $\mathrm{mean}_{\mathrm{spread}} + 2\mathrm{Std}_{\mathrm{spread}}$, and the low stop loss point is $\mathrm{mean}_{\mathrm{spread}} - 2\mathrm{Std}_{\mathrm{spread}}$.

The trading rule can be shown in the table1:

Table 1 trading rule

Begin of trading sign	operation	End of trading sign	operation
Short sell trading signal	Short sell the portfolio	Average level	Long buy the portfolio
Short sell trading signal	Short sell the portfolio	Stop loss point	Long buy the portfolio
Long buy trading signal	Long buy the portfolio	Average level	Short sell the portfolio
Long buy trading signal	Long buy the portfolio	Stop loss poing	Short sell the portfolio

Specifically, for the first type of trading (the second row in Table 1), investors can short sell the portfolio when the price moves away the mean level and rises to the short sell trading signal. Then, the investors should buy back the portfolio and gain positive return when the price of the portfolio goes back to the average level. However, the price will still rise to the stop loss point rather than decrease to the mean (the third row in the table above), and the investors have to buy back the portfolio. Then, the investors will gain a negative return (gain a loss).

4.2 Estimate the trading signal

4.2.1 Using the Ornstein-Uhlenbeck process

The trading signal is $\Delta \mathrm{Std}_{\mathrm{spread}}$, and the $\mathrm{Std}_{\mathrm{spread}}$ is standard deviation of the spread of the stocks, which can be get easily through the spread series. However, the fraction of the signal is difficult to estimate. In this section, we pay attention to the estimation of the fraction of the trading signal (Δ) based on the Ornstein-Uhlenbeck process.

The price of the portfolio combined with stock A and B is:

$$\mathrm{Spread}_t = \log \mathrm{B}_t - \mathrm{n} \log \mathrm{A}_t \qquad (9)$$

Meanwhile, the residual equation can be gained:

$$\mathrm{e}_t = \mathrm{spread}_t - \mathrm{mean}_{\mathrm{spread}} \qquad (10)$$

In the optimal statistical arbitrage trading model of Bertram (2009), the residual follows the Ornstein-Uhlenbeck process:

$$de_t = -\alpha e_t dt + \eta dW_t \qquad (11)$$

Where α, η>0, and W_t is the Wiener process

The auto-regression of the series e_t can be shown:

$$e_t = b * e_{t-1} + \xi_t \qquad (12)$$

After a series of formula transformations and derivations, the parameter of α and η can be shown:

$$\alpha = -\ln(b)/\Delta t \qquad (13)$$

$$\eta = \sqrt{\frac{var(\xi) * 2\alpha}{1-b^2}} \qquad (14)$$

According to the research of Bertram (2009), we know that investors can enter a trade when $e_t = a$, and end the trade when $e_t = m$. Therefore, the return of this trade is (do not consider the trading cost):

$$m - a = mean_{spread} - (mean_{spread} - \Delta Std_{spread}) = \Delta Std_{spread} \qquad (15)$$

The parameter Δ can be gained:

$$\Delta = (m - a)/ Std_{spread} \qquad (16)$$

Meanwhile, Bertram (2009) stated the expected return and the variance for the trading strategy:

$$\mu(a, m, c) = \frac{\alpha (m-a-c)}{\pi[Erfi\left(\frac{m\sqrt{\alpha}}{\eta}\right) - Erfi\left(\frac{a\sqrt{\alpha}}{\eta}\right)]} \qquad (17)$$

$$\sigma^2(a, m, c) = \alpha(m-a-c)^2 \frac{w_1(m\sqrt{2\alpha}/\eta) - w_1(a\sqrt{2\alpha}/\eta) - w_2(m\sqrt{2\alpha}/\eta) + w_2(a\sqrt{2\alpha}/\eta)}{\pi^3 [Erfi\left(\frac{m\sqrt{\alpha}}{\eta}\right) - Erfi\left(\frac{a\sqrt{\alpha}}{\eta}\right)]^3} \qquad (18)$$

Where c is the trading cost, $Erfi(x)$is imaginary error function, and$Erfi(x) = iErfi(ix)$.

In addition, he also showed us the sharp ratio of the trading:

$$S(a,c,r_f) = \frac{-(2a+c+r_f)\sqrt{\alpha\pi \mathrm{Erfi}\,(\frac{a\sqrt{\alpha}}{\eta})}}{\sqrt{(2a+c)^2[w_1\left(\frac{a\sqrt{2\alpha}}{\eta}\right)+w_2(\frac{a\sqrt{2\alpha}}{\eta})]}} \tag{19}$$

Where r_f is the risk-free rate of return.

In our research, we are using the expected return to estimate the trading signal. In order to get the maximum expected return, the parameter a should satisifies:

$$a = -\frac{c}{4} - \frac{c^2\alpha}{4\left(c^3\alpha^3 + 24c\alpha^2\eta^2 - 4\sqrt{3c^4\alpha^5\eta^2 + 36c^2\alpha^4\eta^4}\right)^{\frac{1}{3}}} - \frac{\left(c^3\alpha^3 + 24c\alpha^2\eta^2 - 4\sqrt{3c^4\alpha^5\eta^2 + 36c^2\alpha^4\eta^4}\right)^{\frac{1}{3}}}{4\alpha} \tag{20}$$

4.2.2 Using the ARCH/GARCH model

The trading signal is mentioned as $\mathrm{mean}_{\mathrm{spread}} + \Delta\mathrm{Std}_{\mathrm{spread}}$, and $\mathrm{Std}_{\mathrm{spread}}$ is the standard deviation which is constant. However, in the real stock market, the standard deviation of spread sequence is fluctuant. The situation of volatility clustering may appear in the time series. To analyses the effect of volatility, the ARCH model and GARCH model can be applied in the trading strategy. The ARCH (q) model and GARCH model are defined as:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_q u_{t-q}^2 \tag{21}$$

$$\sigma_t^2 = w + \alpha u_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{22}$$

However, when the lag q might be very large, and non-negativity constraints might be violated, GARCH model is superior to ARCH model. Therefore, the ARCH effect should be tested before using the ARCH/GRACH model. Square the residuals and regress them on q own lags to test for ARCH order q,

$$\hat{u}_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i\, \hat{u}_{t-i}^2 \tag{23}$$

Where $\hat{u}_t$ is residual.

Then, the test statistic is defined as TR^2 and follows the χ^2 distribution. If the null hypothesis is rejected, then the square errors in the regression model follow an AR model and are not constant over time. Meanwhile, the lag q can be gained, and then, we can know which model should be chosen. Next, the new trading signal can be decided as $\text{mean}_{\text{spread}} \pm \Delta\sigma_t$. In order to compare the two kinds of trading signal easily, the Δ is defined as 1 in ARCH/GRACH model. The stop loss point is $\text{mean}_{\text{spread}} \pm 2\sigma_t$.

4.2.3 Combination of the Ornstein-Uhlenbeck process and the ARCH/GARCH model

The fraction of the Δ estimated by the Ornstein-Uhlenbeck process and the σ_t estimated by ARCH/GRACH model were mentioned in previous setion. In this section, these two kinds of factors will be considered simultaneously. In other words, the fraction of the Δ and the σ_t can be combined to a new trading model and strategy. The short sell trading signal is defined as $\text{mean}_{\text{spread}} + \Delta\ \sigma_t$, and the long buy trading signal is: $\text{mean}_{\text{spread}} - \Delta\ \sigma_t$. Meanwhile, the stop loss points are defined as: $\text{mean}_{\text{spread}} \pm 2\ \sigma_t$

5 Result

5.1 The choice and test of the data

5.1.1 The choice of data

We use Shanghai A shares as our database, and select five banks as our research objectives, namely, HuaXia Bank, China Minsheng Bank, China Merchants Bank, Industrial Bank, and China Construction Bank respectively. After that, we collect these data from 19/05/2014 to 17/05/2017 and make them into a line chart (line chart 1 in appendix). From the chart, it is obviously that the trends of all of the stock prices are similar. In addition, a calculation of correlation is set as we compare one bank with another respectively. The result of each correlation for the stock price is shown in Table 2.

Table 2. Price correlations
The table shows the correlation between time-series of stock prices.

Price correlation	Minsheng Bank	Merchants Bank	Industrial Bank	Construction Bank
HuangXia Bank	0.901	0.889	0.959	0.914
Minsheng Bank		0.809	0.928	0.824
Merchants Bank			0.901	0.736
Industrial Bank				0.844

It is manifest to see that the correlation of the stocks between HuangXia Bank and Industrial Bank is most close to 1, which is 0.959. Therefore, these two price series can be the most appropriate data for the research. Because we need to compare the model through the final return, the logarithm price can be introduced in this paper. Using logarithm price is easier and more convenient to gain price return than using stock price directly. In addition, logarithm price is normality assumption, and the range of the logarithm price is $(-\infty, +\infty)$. Thus, in the next step, we attempt to use logarithm for each bank within three years. The theory of using logarithm is based on the following formula:

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}} \approx \log p_t - \log p_{t-1} \tag{24}$$

Followed by the steps above, a new line chart (line chart 2) of these five Logarithm price series is achieved and shown in the appendix. We then calculate the correlation via logarithm, seeing it in Table 3.

Table3. Logarithm price of correlation

The table shows the correlation between time-series of logarithm stock prices.

Logarithm price correlation	Minsheng Bank	Merchants Bank	Industrial Bank	Construction Bank
HuangXia Bank	0.926	0.924	0.971	0.919
Minsheng Bank		0.860	0.945	0.840
Merchants Bank			0.935	0.780
Industrial Bank				0.858

Undoubtedly, after logarithm, the combination of HuangXia Bank and Industrial Bank still appears to be the largest than the rest data.

According to the statistical research of the two tables above, we discover that the 0.959 is the maximum for the correction of these stock prices while we can get 0.971 as the biggest number of the correction of Logarithm price. In such case, we found that when correlating between HuangXia Bank and Industrial Bank whose correlation and logarithm correlation is largest and close to 1. We take these two banks out and make a third graph instead, shown in the appendix (line chart 3).

In this paper, we have a total number of 733 data for each stock, where we decide to select 493 as the sample data. The rest is used for testing the models to verify their performance.

5.1.2 The ADF test of data

Industrial Bank and HuangXia Bank which we choose have similar price trend and high correlation. In order to record and operate by EXCEL and Eviews, HuangXia Bank is represented by X, and Industrial Bank is named Y. Then, the stationary of the data can be tested in this section. The results can be shown in the table below. Both of the logarithm prices of Industrial Bank and HuangXia Bank have unit roots because the p-values are 0.429 and 0.447, and the null hypothesis should be accepted (the results from Eviews are shown in result 1 in appendix). Therefore, the price series of Industrial Bank and HuangXia Bank are non-stationary. Table 4 displays all the information

Table 4. ADF test
The table shows the information of its statistics, p-value and result of the ADF test.

ADF test	ADF test statistic	p-value	result
X	-1.668	0.447	Non-stationary
Y	-1.704	0.429	Non-stationary
dX	-22.507	0.000	Stationary
dY	-22.318	0.000	Stationary

In addition, we take the first order difference on the price series of HuangXia Bank and Industrial Bank respectively (represented by dX and dY). The results are shown in Table 4. The p-values are zero, and the null hypothesis should be rejected. Thus, the two first order difference series are stationary. The series of Industrial Bank and HuangXia Bank are I(1), and they may potentially be co-integrated.

5.1.3 The co-integration test of data

Due to the I (1) of the two variables, the regression can be estimated through Eviews (Result 2 in appendix)

$$Y_t = 0.199 + 0.966{*}X_t + \widehat{e_t} \quad (25)$$

Where, $\widehat{e_t}$ is the residual sequence, and is named resid01 in Eviews

Then, the ADF test is used to the residual $\widehat{e_t}$(Result 3 in appendix). $\widehat{e_t}$ is the residual of the regression combined between Y_t and X_t, and the sum of the residual series is zero, so the intercept and trend are not included in the test equation. The table below indicates that the p-value is 0.003 and the null hypothesis should be rejected. Therefore, the residual series is stationary, and the sequence of Y_t and X_t are co-integration, which shows in table 5

Table 5. ADF test

ADF test	ADF test statistic	p-value	Result
$\widehat{e_t}$	-2.965	0.003	Stationary

5.1.4 Error correction model

Through the Eviews, the lag (1, 1) equation can be gained (Result 4 is in the appendix):

$$Y_t = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 Y_{t-1} + \widehat{e_t} \tag{26}$$

where $\gamma_0 = 0.008, \gamma_1 = 0.820. \gamma_2 = -0.792$, and $\gamma_3 = 0.969$

The lag (1, 1) equation can be transformed to:

$$\Delta y_t = \gamma_1 \Delta x_1 - \lambda(y_{t-1} - \alpha_0 - \alpha_1 x_{t-1}) + \widehat{e_t} \tag{27}$$

Where, $\lambda = 1 - \gamma_3 = 1 - 0.969 = 0.031$,

$$\alpha_0 = \frac{\gamma_0}{1-\gamma_3} = -\frac{0.008409}{1-0.968634} = 0.268, \tag{28}$$

$$\alpha_1 = \frac{\gamma_1+\gamma_2}{1-\gamma_3} = \frac{0.820357-0.792095}{1-0.968634} = 0.900. \tag{29}$$

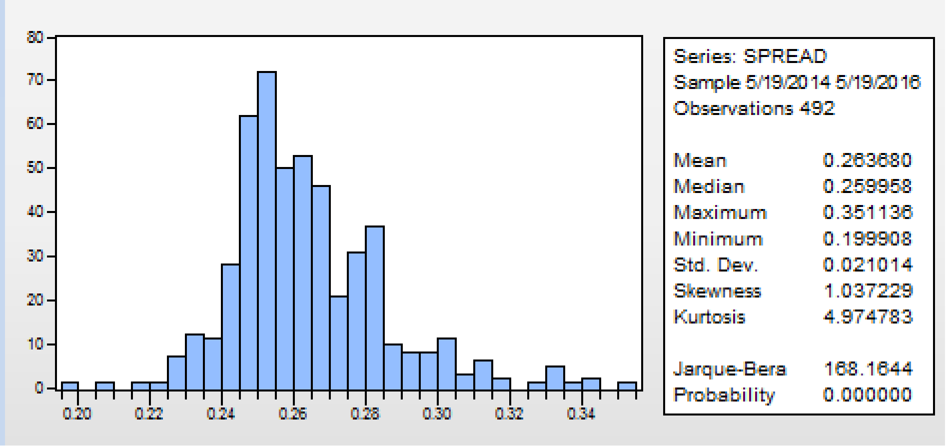
Therefore, the portfolio combined with Y_t and X_t can be presented as:

$$Spread_t = Y_t - 0.9\, X_t \tag{30}$$

Therefore, buy one unit of the portfolio means to buy one unit Y and sell 0.9 fractions of X. Meanwhile, sell one unit of this kind of portfolio is to sell one unit Y and buy 0.9 fractions of X.

The description of the $Spread_t$ can be presented on the chart in figure 1:

Figure 1 the description of the $Spread_t$



The mean of the spread sequence is: $mean_{spread} = 0.284$ and the standard derivation of the series is: $Std_{spread} = 0.021$

5.2 The result estimated by Ornstein-Uhlenbeck process

5.2.1 The estimation of trading signal

The residual of spread series and the auto-regression of the residual can be gained respectively (Result 5 in appendix):

$$e_t = spread_t - mean_{spread} \tag{31}$$

$$\mathrm{e_t} = b * \mathrm{e_{t-1}} + \xi_t \tag{32}$$

Where the coefficient b of the residual equals to 0.979.

And we can gain the residual series ξ_t, which is named resid02 in E-views. The variance of $\xi_t = 0.005605^2$ Assume there are 250 trading days in one year, the certain trading interval is Δt, which equals to $\frac{1}{250} = 0.004$. Meanwhile, the trading cost c can be assumed as 0.002. According to optimal statistical arbitrage trading model of Bertram (2009), the parameters of the trading signal are:

$$\alpha = -\frac{\ln(b)}{\Delta t} = -\frac{\ln(0.979)}{0.004} = 5.300 \tag{33}$$

$$\eta = \sqrt{\frac{var(\xi)*2\alpha}{1-b^2}} = \sqrt{\frac{var(\xi)*2*5,3}{1-0.979^2}} = 0.090 \tag{34}$$

Then, the parameter α and η can be replaced into the equation of parameter a.

$$a = -\frac{c}{4} - \frac{c^2\alpha}{4\left(c^3\alpha^3 + 24c\alpha^2\eta^2 - 4\sqrt{3c^4\alpha^5\eta^2 + 36c^2\alpha^4\eta^4}\right)^{\frac{1}{3}}} - \frac{\left(c^3\alpha^3 + 24c\alpha^2\eta^2 - 4\sqrt{3c^4\alpha^5\eta^2 + 36c^2\alpha^4\eta^4}\right)^{\frac{1}{3}}}{4\alpha} = -0.012 \tag{35}$$

Therefore, the expected return can be maximal when the parameter a equals to -0.012. Then, Δ can be calculated:

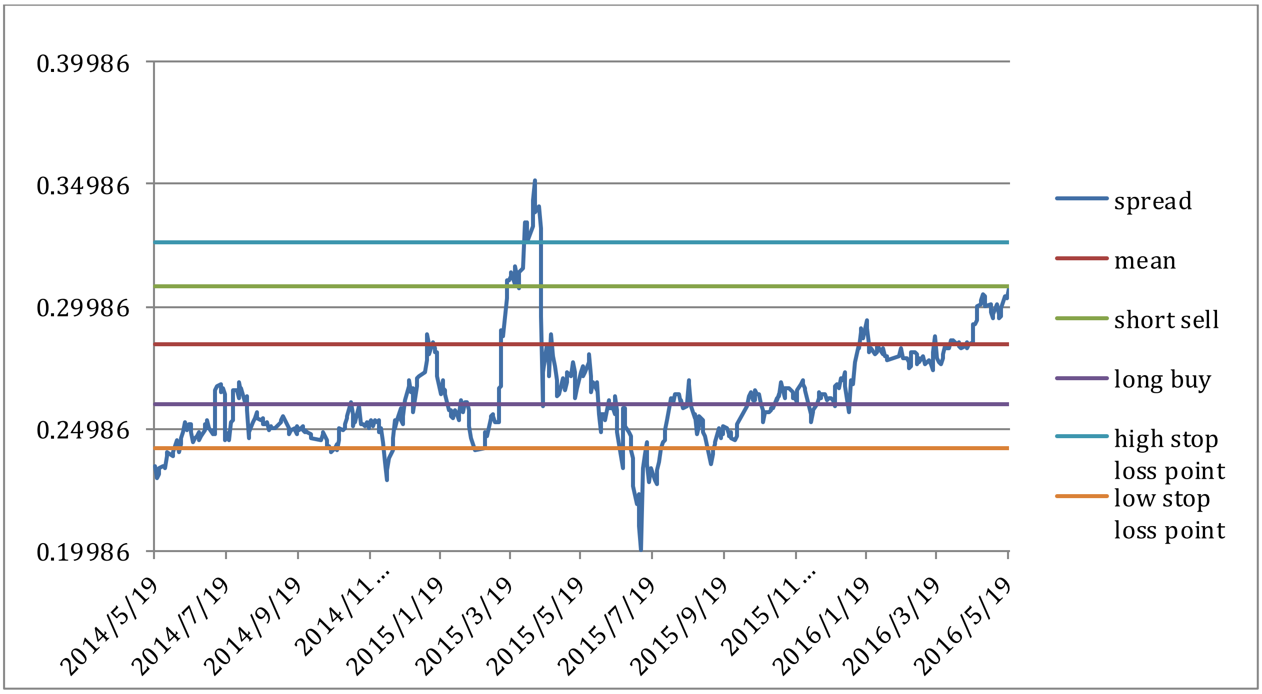
$$\Delta = (\text{m} - \text{a}) / Std_{spread} = 0.024/0.021 = 1.140 \tag{36}$$

According to the research of Bock and Mestel (2009), we can gain the specific trading signal. The short sell trading signal is : $mean_{spread} + \Delta Std_{spread} = 0.284 + 1.140 * 0.021 = 0.308$, and the long buy trading signal is : $mean_{spread} - \Delta Std_{spread} = 0.284 - 1.140 * 0.021 = 0.260$. In addition, the Stop loss point can be expressed as: $mean_{spread} - 2Std_{spread} = 0.284 \pm 2 * 0.021 = (0.326, 0.242)$

5.2.2 The return during the sample period

According to the parameters and trading signals calculated in the previous section, we can simulate the trading strategy from a line chart, which includes the mean level, spread of the portfolio, short sell trading signal, long buy trading signal, high stop loss point and low stop loss point. The graph can be shown in figure 2:

Figure 2 the formation of trading arbitrage with sample data via O-U model



From the line chart, we have 13 operations to complete the statistical arbitrage trading above. When the spread increase (decrease) to the line "long buy", we decide to buy the stock; when the spread goes down (up) to the line "short sell", then we sell the stock instead; when the spread rises or falls to the high(low) stop loss point or goes back to mean level, we do the opposite operations. For example, on 8th, July, 2014, we purchased these two stocks with the spread price of 0.26. Ten days later, on 18th, the price decrease to the low stop loss point. We sold the portfolio with the price of 0.245 to avoid a further loss. In this trading process, 0.0146 was treated as a loss. In another case, when it came to day on 10th, March, 2015, we bought the stocks with the price of 0.26. On the day of 12th, the stocks were sold with the spread price of 0.284 because the spread price increases to the mean level. For the sake of the protection for the current profit, we decided to sell them in order to avoid the potential risk. The return during this trading was positive, means we earned a profit of 0.024. The rest trading results are similar to the operations above. A Two-year investment in these stocks experienced 13 operations, which then was made into a table 5 to display all the details.

Table 5 return of arbitrage trading

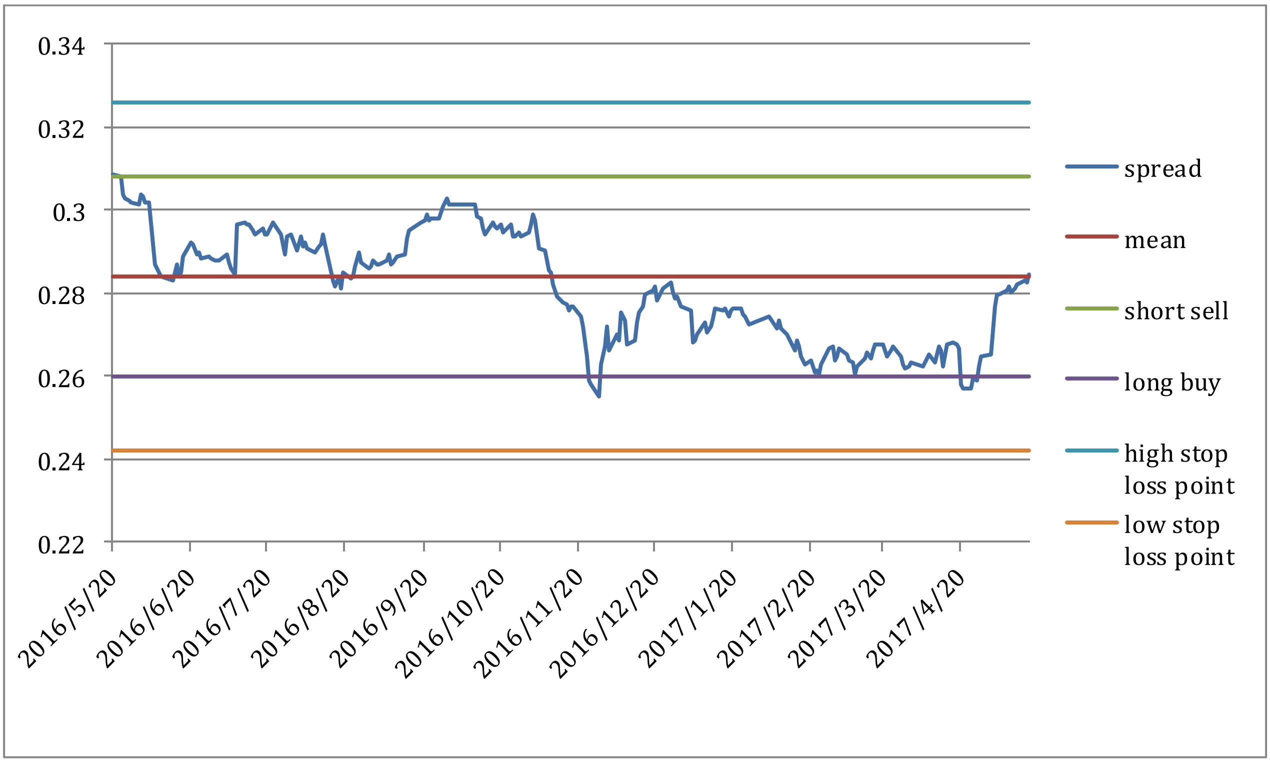
Num of arbitrage	operation	date	Spread price	Price return	Return ratio
1	Long buy	2014-07-08	0.260	-0.260	– 0.015
	Short sell	2014-07-18	0.245	0.245	
2	Long buy	2014-07-26	0.260	-0.260	– 0.011
	Short sell	2014-08-07	0.249	0.249	
3	Long buy	2014-11-10	0.260	– 0.260	– 0.010
	Short sell	2014-12-01	0.250	0.250	
4	Long buy	2014-12-18	0.260	-0.260	0.024
	Short sell	2015-01-05	0.284	0.284	
5	Long buy	2015-02-09	0.260	-0.260	-0.010
	Short sell	2015-02-16	0.250	0.250	
6	Long buy	2015-03-10	0.260	-0.260	0.024
	Short sell	2015-03-12	0.284	0.284	
7	Short sell	2015-03-16	0.308	0.308	-0.018
	Long buy	2015-03-30	0.326	-0.326	
8	Short sell	2015-04-13	0.308	0.308	0.024
	Long buy	2015-04-14	0.284	-0.284	
9	Long buy	2015-04-16	0.260	-0.260	0.024
	Short sell	2015-04-20	0.284	0.284	
10	Long buy	2015-06-02	0.260	-0.260	-0.018
	Short sell	2015-06-19	0.242	0.242	
11	Long buy	2015-06-24	0.258	-0.258	-0.016
	Short sell	2015-06-30	0.242	0.242	
12	Long buy	2015-08-20	0.260	-0.260	-0.018
	Short sell	2015-09-02	0.242	0.242	
13	Long buy	2016-01-04	0.260	-0.260	0.024
	Short sell	2016-01-11	0.284	0.284	
Total return ratio					0.004

It is clear to discover that there are five profitable arbitrages during a two-year trading, and the total return ratio is 0.4 percent. Despite this little profit, it still convinced that with this statistical arbitrage model, investors could achieve return after a 2-year investment. Meanwhile, we find that the negative return exists in the statistical arbitrage; therefore, the arbitrage is not a risk-free strategy. Investors only reduce the investment risk using the statistical arbitrage.

5.2.3 The forecast and estimation of the data out of sample

Then, the parameters and trading signals calculated by the sample data are used to forecast and estimate the spread price out of the same. The simulation is shown in figure 3:

Figure 3 the forecast and estimation of the data out of sample



The procedure for the statistical arbitrage is the same as that shown in the model before. In the past year, only three time periods were suitable for arbitrage, which then were made into a table, showing in Table 6.

Table 6 return of arbitrage trading

Num. of arbitrage	operation	time	Spread price	Price return	Return ratio
1	Short sell	2016-05-23	0.308	0.308	0.024
	Long buy	2016-06-06	0.284	-0.284	
2	Short sell	2016-09-29	0.303	0.303	0.019
	Long buy	2016-11-08	0.284	-0.284	
3	Long buy	2016-11-29	0.260	-0.260	0.024
	Short sell	2016-12-26	0.284	0.284	
Total return ratio					0.067

The table clearly reflects that all the three operations within last year had a positive return. Simultaneously, it describes that the return ratios in the first and third arbitrages have the same value. The total return ratio, 0.067, illustrates that the model application is profitable despite the return of two stocks don't achieve too much return. In such situation, the O-U model is suitable for the investors; particularly, the positive return ratio decreases the expectation of investors who are afraid of the future stock market.

5.3 The result given by the ARCH/GARCH model

5.3.1 Trading signal estimated by ARCH/GARCH model

Through the correlogram of $spread_t$ series (Result 6 in appendix), we can know that $spread_t$ is AR (1) model. Therefore, the auto-regression can be estimated (Result 7):

$$spread_t = 0.009 + 0.966 * spread_{t-1} + u_t \quad (37)$$

Then, the residual u_t can be gained and is named resid03 in Eviews. The line graph of u_t is shown in appendix (Result 8). The ARCH LM test is used to the residual u_t with lag 10 (in Figure 4). From the table, the probability of chi-square distribution is 0.056, and the null hypothesis cannot be rejected. Therefore, the residual u_t does not have the ARCH effect with high order of lags, and the GRACH (1, 1) model can be considered to estimate parameters and trading signals.

Figure 4 the result of ARCH LM test with lag 10

F-statistic	1.817770	Prob. F(10,479)	0.0551
Obs*R-squared	17.95183	Prob. Chi-Square(10)	0.0558

In addition, the GRACH (1, 1) (Result 9 in appendix) model can be shown as:

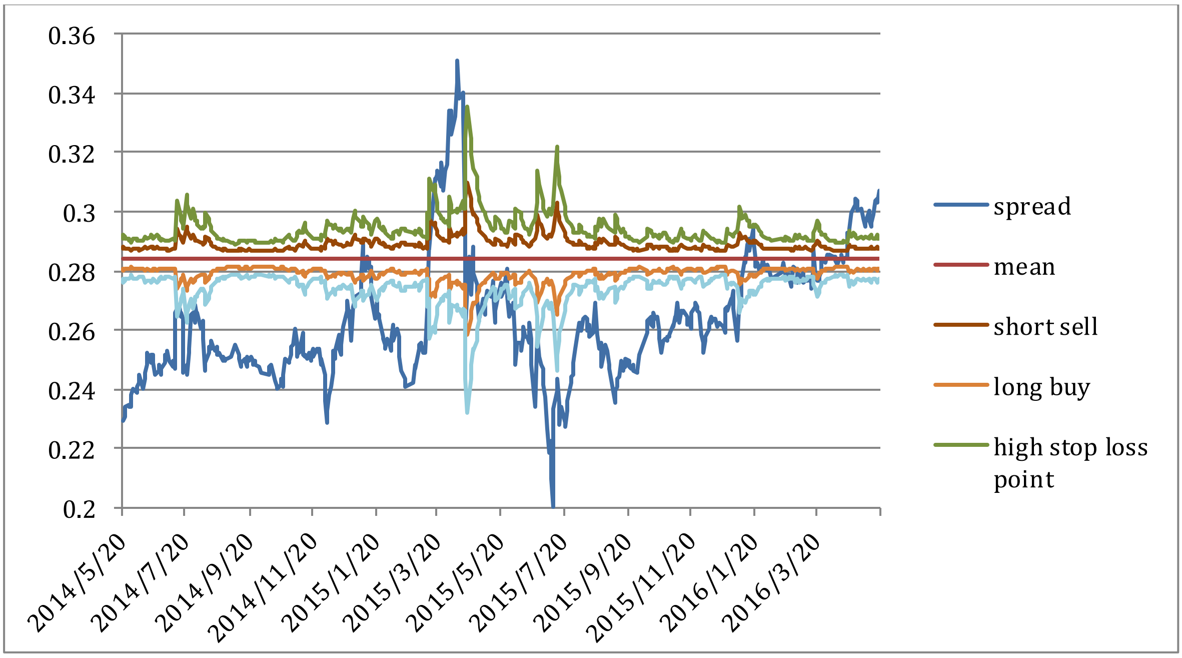
$$\sigma_t^2 = 0.00000166 + 0.270u_{t-1}^2 + 0.707\sigma_{t-1}^2 \quad (38)$$

Therefore, the short sell trading signal is $mean_{spread} + \sigma_t$, and the long buy trading signal is : $mean_{spread} - \sigma_t$. The Stop loss points are: $mean_{spread} \pm 2Std_{spread}$

5.3.2 The return of the sample

Then, we simulate the trading result using the parameters and trading signals calculated by GRACH (1, 1) model, seeing in figure 5

Figure 5 the formation of trading arbitrage with sample data via GARCH model



Because of the volatility, the lines of trading signals and stop loss points are not straight lines, and they are fluctuant around the mean level according to the spread series. It is clearly to notice that two-year period experienced a sharp fluctuation, where the value of spread began with 0.233 and ended with above 0.300. It once climbed to over 0.350 and plunged to less than 0.200. Nearly 12 operations need to be considered and calculated within two years. All the details and information could be checked below via Table 7

Table 7 return of arbitrage trading

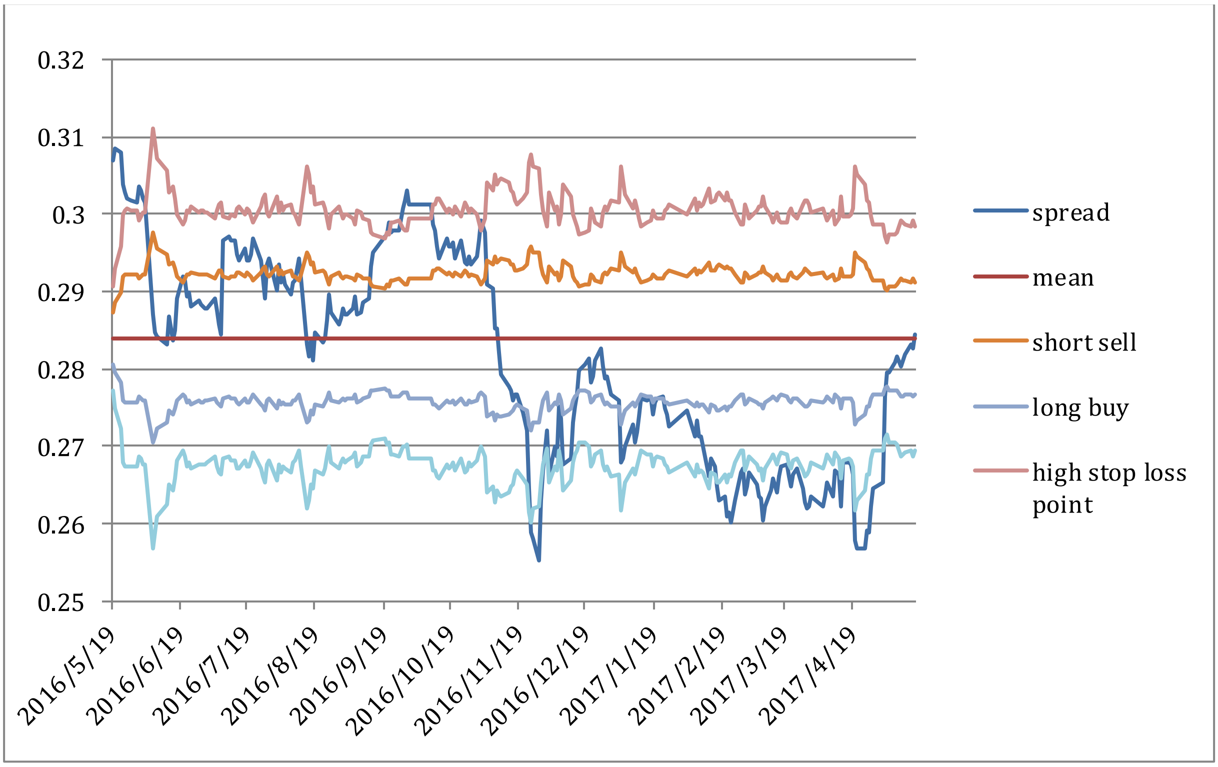
Num of arbitrage	operation	time	Spread price	Price return	Return ratio
1	Long buy	2015-01-05	0.278	-0.278	0.006
	Short sell	2015-01-06	0.284	0.284	
2	Short sell	2015-01-07	0.289	0.289	0.005
	Long buy	2015-01-08	0.284	-0.284	
3	Long buy	2015-01-14	0.279	-0.279	-0.004
	Short sell	2015-01-15	0.275	0.275	
4	Long buy	2015-03-10	0.277	-0.277	0.007
	Short sell	2015-03-11	0.284	0.284	
5	Short sell	2015-03-16	0.295	0.295	-0.011
	Long buy	2015-03-18	0.306	-0.306	
6	Short sell	2015-04-14	0.292	0.292	0.008
	Long buy	2015-04-15	0.284	-0.284	
7	Long buy	2015-04-16	0.243	-0.243	0.041
	Short sell	2015-04-22	0.284	0.284	
8	Long buy	2015-04-27	0.270	-0.270	-0.008
	Short sell	2015-04-29	0.262	0.262	
9	Long buy	2016-01-07	0.276	-0.276	0.008
	Short sell	2016-01-11	0.284	0.284	
10	Short sell	2016-01-15	0.289	0.289	-0.005
	Long buy	2016-01-18	0.294	-0.294	
11	Short sell	2016-01-29	0.290	0.290	0.006
	Long buy	2016-01-21	0.284	-0.284	
12	Short sell	2016-04-16	0.286	0.286	-0.003
	Long buy	2016-04-18	0.289	-0.289	
Total return ratio					0.050

According to the details from the table above, it manifests that more than half of the total operations obtained a positive return. Among the operations, time from 16th April to 22nd April (arbitrage operation 7) experienced the highest return ratio, accounting for approximately 80 percent of the total return ratio. The rest positive value is around 1 percent. The total return ratio is 5 percent, which means that GARCH model is worthwhile to be considered by investors. In addition, the stop loss point is also fluctuant according to the spread series, and it can reduce the investment risk.

5.3.3 Out of sample forecast

Next, the data out of the sample is used to estimate the GARCH (1, 1), and the trading parameters and trading signals are gained from the sample data. The result of the trading is, seeing in figure 6:

Figure 6 the forecast and estimation of the data out of sample



As it can be seen above, over ten operations were suggested to conduct the statistical arbitrage within last year. The fluctuation seems to be sharper and more drastic than that of other tests. The initial point, also regarded as the peak point, fell down since the model started to be tested. The procedure is still the same as those operations introduced before. After a careful and clear calculation, all of the returns of arbitrage trading can be shown in Table 8.

Table 8 return of arbitrage trading

Num of arbitrage	operation	time	Spread price	Price return	Return ratio
1	Short sell	2016-06-06	0.298	0.298	0.013
	Long buy	2016-06-07	0.285	-0.285	
2	Short sell	2016-06-20	0.292	0.292	0.008
	Long buy	2016-07-07	0.284	-0.284	
3	Short sell	2016-07-08	0.292	0.292	-0.005
	Long buy	2016-07-11	0.297	-0.297	
4	Short sell	2016-08-09	0.292	0.292	0.009
	Long buy	2016-08-15	0.283	-0.283	
5	Short sell	2016-09-13	0.291	0.291	-0.006
	Long buy	2016-09-19	0.297	-0.297	
6	Short sell	2016-11-04	0.294	0.294	0.012
	Long buy	2016-11-10	0.282	-0.282	
7	Long buy	2016-11-17	0.276	-0.276	-0.01
	Short sell	2016-11-21	0.266	0.266	
8	Long buy	2016-12-05	0.276	-0.276	-0.007
	Short sell	2016-12-07	0.269	0.269	
9	Long buy	2016-12-14	0.276	-0.276	0.007
	Short sell	2016-12-26	0.283	0.283	
10	Long buy	2017-01-03	0.275	-0.275	-0.007
	Short sell	2017-01-04	0.268	0.268	
11	Long buy	2017-01-23	0.276	-0.276	-0.009
	Short sell	2017-02-10	0.267	0.267	
Total return ratio					0.005

The table illustrates that there are total 11 operations for the test and verification, and despite that there are fewer positive return ratios than the negative ones(5 and 6 respectively); however, the total return ratio, 0.5 percent, represents that when using GARCH model for these stocks, this portfolio is profitable for investors to consider.

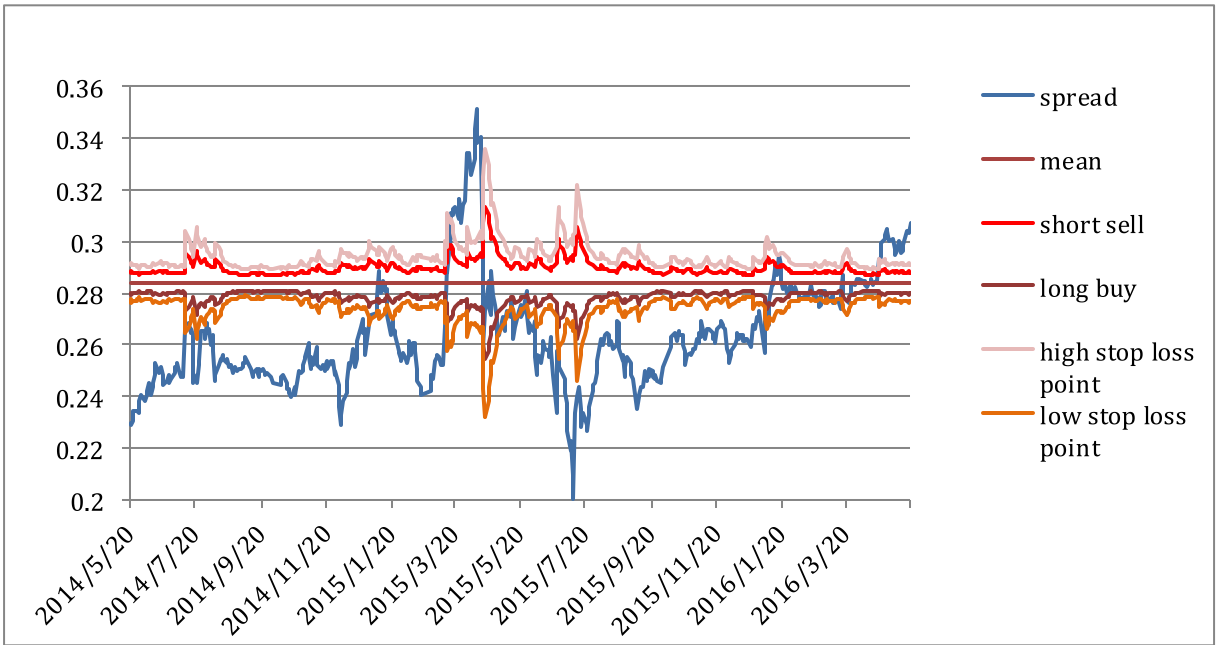
5.4 The result from the combination of Ornstein-Uhlenbeck process and ARCH/GARCH model

In the third part, which is an evolutionary step for us, we device a new model that both combined with O-U model and GARCH model. Obviously, the first two years information as sample data are selected to make the model successfully and the last year data is used to test and verify its correctness and effectiveness as the investors' multiple choices. From the section 5.2.1 and 5.3.1, the short sell trading signal can be expressed as $mean_{spread} + \Delta\sigma_t$, and the long buy trading signal is $mean_{spread} - \Delta\sigma_t$.

5.4.1 The return of the sample

Graph 7 shows the results of the trading strategy during a period of two years.

Figure7 the formation of trading arbitrage with sample data via combined model



As the graph depicts, there are total 12 operations that could be treated as the statistical arbitrages within a two-year observation. Because the parameter Δ= 1,14, which is greater than 1, is added to the trading signal, the arbitrage interval (the range between mean level and short sell/long buy signal) is greater than that of GARCH (1, 1) model. Meanwhile, the range between trading signal and loss stop point is smaller. The operation method is preferred to do in the same way before. A further table is made to note all the operations with their return ratios, which could be seen in Table 9.

Table 9 return of arbitrage trading

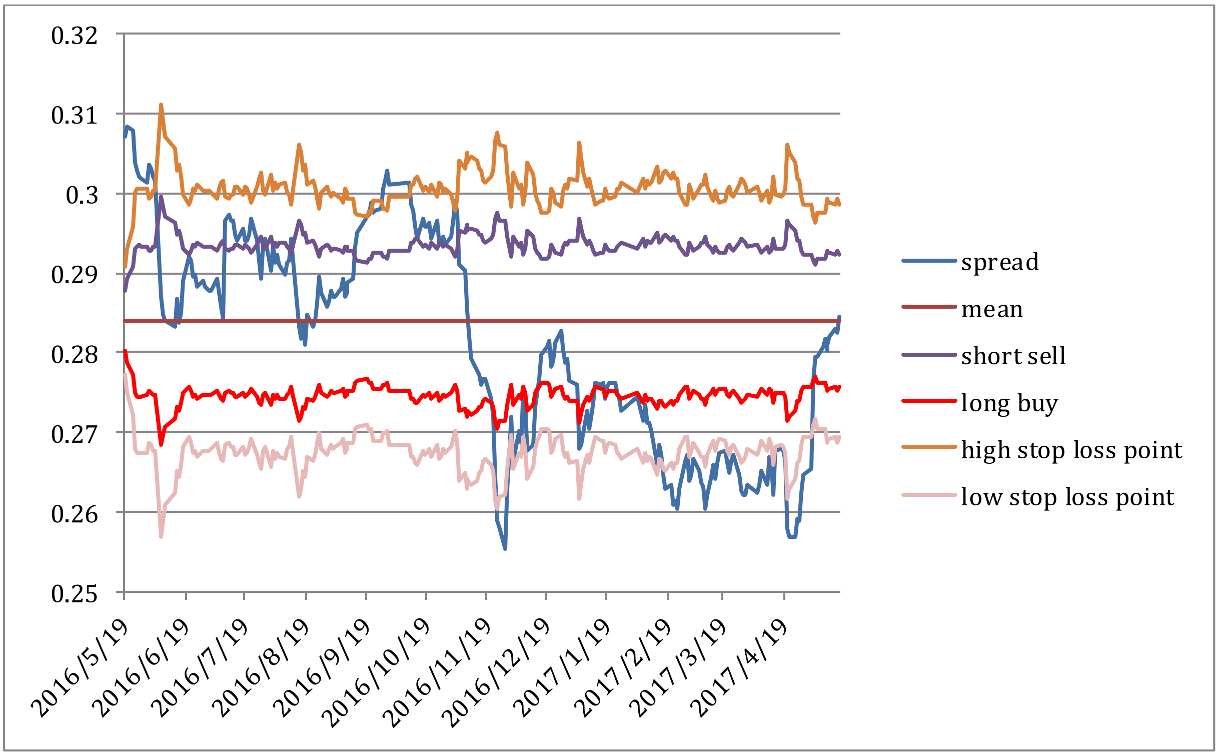
Num of arbitrage	operation	time	Spread price	Price return	Return ratio
1	Long buy	2015-01-05	0.278	-0.278	0.006
	Short sell	2015-01-06	0.284	0.284	
2	Short sell	2015-01-07	0.29	0.29	0.006
	Long buy	2015-01-09	0.284	-0.284	
3	Short sell	2015-03-12	0.298	0.298	-0.008
	Long buy	2015-03-18	0.306	-0.306	
4	Short sell	2015-04-15	0.294	0.294	0.01
	Long buy	2015-04-16	0.284	-0.284	
5	Long buy	2015-04-28	0.270	-0.270	-0.007
	Short sell	2015-04-29	0.263	0.263	
6	Long buy	2016-01-05	0.276	-0.276	0.009
	Short sell	2016-01-11	0.285	0.285	
7	Short sell	2016-01-15	0.291	0.291	-0.004
	Long buy	2016-01-18	0.295	-0.295	
8	Short sell	2016-01-19	0.290	0.290	0.006
	Long buy	2016-01-22	0.284	-0.284	
9	Long buy	2016-01-25	0.279	-0.279	0.005
	Short sell	2016-01-28	0.284	0.284	
10	Long buy	2016-03-11	0.280	-0.280	0.004
	Short sell	2016-03-13	0.284	0.284	
11	Short sell	2016-03-14	0.287	0.287	0.009
	Long buy	2016-03-15	0.278	-0.278	
12	Long buy	2016-04-12	0.284	-0.284	0.003
	Short sell	2016-04-13	0.287	0.287	
Total return ratio					0.040

As the table shows above, the 2-year operations in two stocks had nine records of positive return ratio, where the positive are triple than the negative records. After two-year investment in stocks, investors could achieve a total return ratio as high as 4 percent. The combined model performed to be one choice for investors to consider in the further.

5.4.2 The forecast and return of the data out of sample

According to the two-year historical data, we use the last year and make it into a Figure 8.

Figure 8 the forecast and estimation of the data out of sample



It described the trend of the spread and analyzed that there were 12 opportunities that investors could have the statistical arbitrages. In this combined model, the initial point gradually decreased to the lowest as time went on in spite of some periods when the spread slightly increased. After it plunged to the bottom, the trend then slowly went up until the end. The procedure omits as we have already mentioned the methods before. The table 10 illustrated all the information, as well as each return ratio of each operation.

Table 10 return of arbitrage trading

Num of arbitrage	operation	time	Spread price	Price return	Return ratio
1	Short sell	2016-06-06	0.299	0.299	0.014
	Long buy	2016-06-07	0.285	-0.285	
2	Short sell	2016-06-17	0.293	0.293	0.009
	Long buy	2016-07-04	0.284	-0.284	
3	Short sell	2016-07-05	0.293	0.293	-0.004
	Long buy	2016-07-12	0.297	-0.297	
4	Short sell	2016-08-09	0.293	0.293	0.01
	Long buy	2016-08-15	0.283	-0.283	
5	Short sell	2016-09-12	0.293	0.293	-0.005
	Long buy	2016-09-20	0.298	-0.298	
6	Short sell	2016-10-26	0.294	0.294	-0.005
	Long buy	2016-11-03	0.299	-0.299	
7	Short sell	2016-11-04	0.295	0.295	0.011
	Long buy	2016-11-07	-0.284	-0.284	
8	Long buy	2016-11-18	0.274	-0.274	-0.008
	Short sell	2016-11-21	0.266	0.266	
9	Long buy	2016-12-05	0.275	-0.275	-0.006
	Short sell	2016-12-07	0.269	0.269	
10	Long buy	2016-12-13	0.275	-0.275	0.008
	Short sell	2016-12-26	0.283	0.283	
11	Long buy	2017-01-03	0.274	-0.274	-0.006
	Short sell	2017-01-04	0.268	0.268	
12	Long buy	2017-02-03	0.275	-0.275	-0.008
	Short sell	2017-02-14	0.267	0.267	
Total return ratio					0.010

Form the table, it is clear to notice that in the last year, 12 operations could be used for the statistical arbitrages. The table reminded the investors that half of the operations in the recent one year brought them the positive return ratio. The total return ratio, 1%, expresses this combined model is worthwhile for the investors to consider, despite of the low profitable return. All in all, there are 6 positive returns from two data series and three different kinds of models. The highest return is 6.7% gained from the model of Ornstein-Uhlenbeck process and data series out of sample. Therefore, these three kinds of models can reduce the risk of investment; however, the returns of these models are low. The reason of this situation may be that the stock price is comparatively stable to reduce the opportunities of arbitrage. In addition, the low value of volatility causes that the interval between mean level and trading signal level ($mean_{spread} \pm \Delta\sigma_t$) is small.

6 Contribution and Conclusion

Statistical arbitrage is a common trading strategy. In this paper, we introduce two trading models, i.e. Ornstein-Uhlenbeck process model and GARCH model. We perform a number of related tests to verify the property of the data-used in this analysis. The stock pries HuangXia Bank and Industrial Bank from Chinese stock market are used to test these two models. The paper contributes by using a combination of Ornstein-Uhlenbeck model and GARCH model. The highest return, which is 6.7%, is gained by Ornstein-Uhlenbeck process model using the data out of sample. Therefore, the O-U model is the best model, among these three alternative models, for investors taking statistical arbitrage trading. Because of the stable stock prices of Chinese bank industry and the low standard deviation, the arbitrage interval of O-U model is wider than that of the GARCH model and the combined model. This situation causes that the O-U model can gain higher return than the other two models. Through forecasting the data out of sample, the new combined model has better preference than GARCH model. Because the parameter that is greater than one is added in the trading signal to expend the arbitrage interval, the arbitrage return of combined model becomes higher than that of GARCH model. For the data which is not as stable as the stock prices of Chinese bank industry, because of the effect of volatility clustering, the arbitrage interval is wider when using combined model than O-U model, thus, the return of data using new combined model is higher. However, this should be tested and verified by new data in the future research.

Reference

Alexander, C., & Dimitriu. A. (2002). The Cointegration Alpha: Enhanced Index Traching and Long-Short Equity Market Neutral Strategies, *ISMA Centre Discussion Paper Series in Finance 2002–08.*

Bondarenko, O. (2003). Statistical Arbitrage and Securities Prices, *Review of Financial Studies*, vol. 16, no. 3, pp. 875-919

Burgess, A.N. (1999). Statiscal arbitrage models of the FTSE 100. *International Journal of Forecasting,* vol. 17, no.1, pp. 31-44

Chng, M.T. (2007). Understanding the risks in and rewards for pairs-trading, *June, ICBME.*

Cummins, M., & Bucca, A. (2011). Quantitative Spread Trading on Crude Oil and Refined Products Markets, *Quantitative Finance,* vol. 12, no.12, pp.1857-1875

Hogan, S., Jarrow, R., Teo, M. & Warachka, M. (2003). Testing Market Efficiency using Statistical Arbitrage with Applications to Momentum and Value Strategies, *Journal of Financial economics,* vol. 73, no. 3, pp. 525-565

Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index, *European Journal of Operational Research*, vol. 196, no. 2, pp. 819-825

Jarrow, R., Teo, M., Tse, YK. & Warachka, M. (2005). Statistical arbitrage and market efficieny: enhance theory, robust theory and further applications, *Available Online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=659941, [Accessed 3 Feb 2005]*

Meucci, A. (2010). Review of Statistical Arbitrage, Cointegration, and Multivariate Ornstein-Uhlenbeck. *Available Online: http://ssrn.com/abstract=1404905, [Accessed 15 May 2009]*

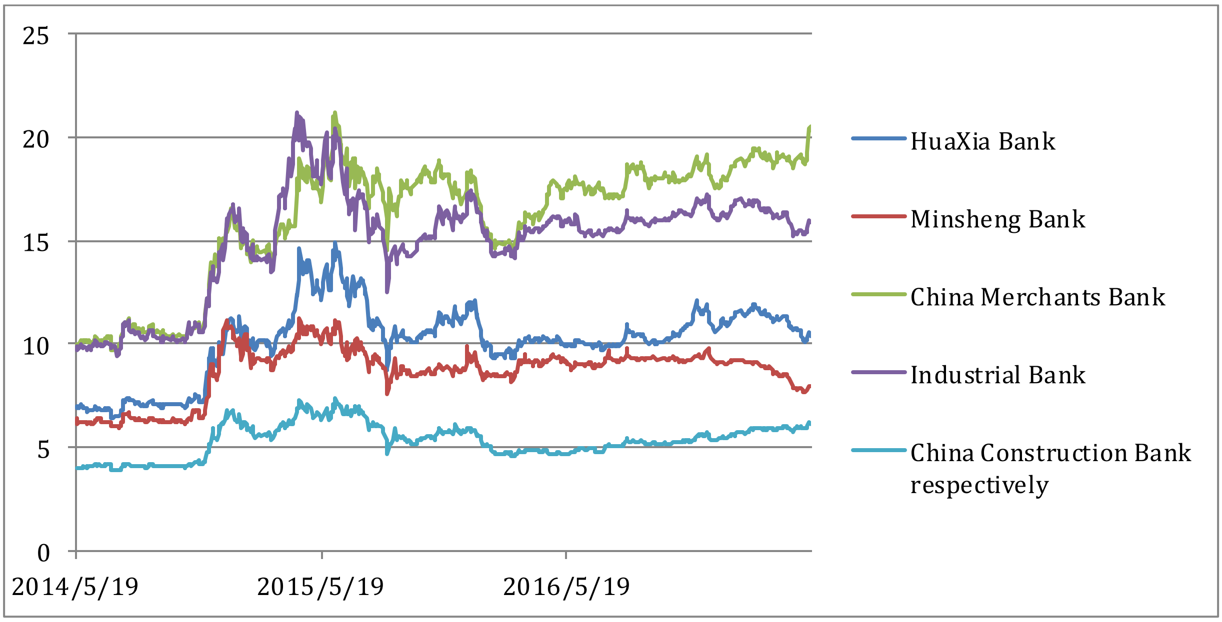
Perlin, M.S. (2009). Evaluation of pairs-trading strategy at the Brazilian financial market, *Journal of Derivatives & Hedge Funds,* vol. 15, no. 2, pp. 122-136.

Thomaidis, N., & Kondakis, N. (2006). Detecting statistical arbitrage opportunities using a combined neural network-GARCH model, *Available Online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1984252, [Accessed 24 Jan 2012]*

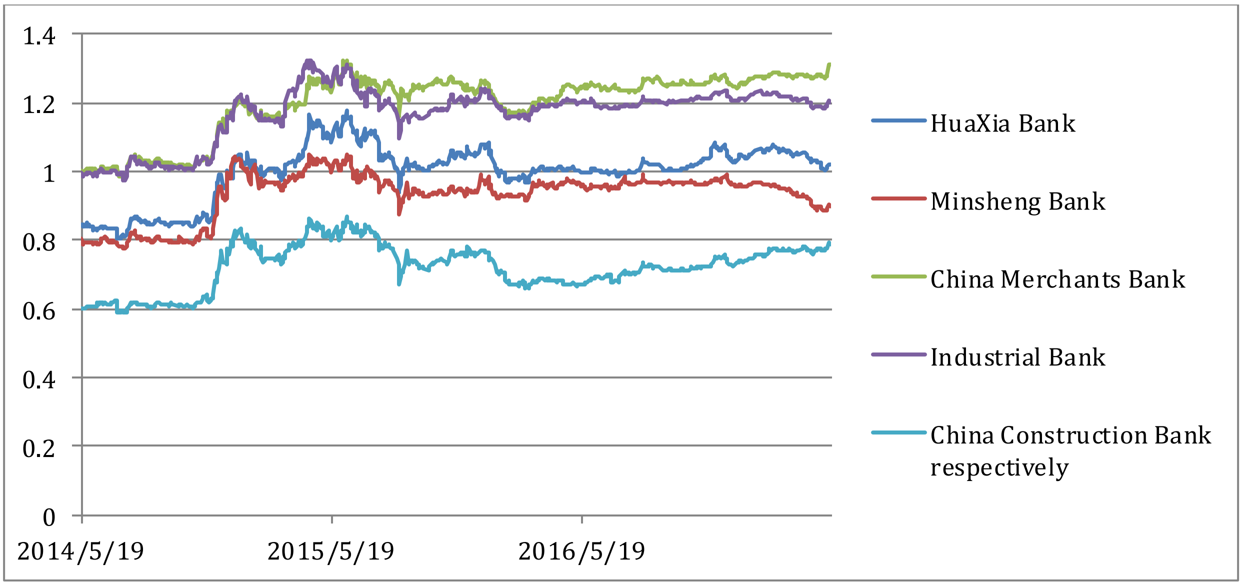
Thomaidis, N.S., & Dounias, G.D. (2006). An Intelligent Statistical Arbitrage Trading System, *Springer Berlin Heidelberg, Berlin, Heidelberg*, pp. 596-599.

Vidyamurthy, G. (2004). Pairs trading: quantitative methods and analysis, *John Wiley & Sons* Vol. *217*

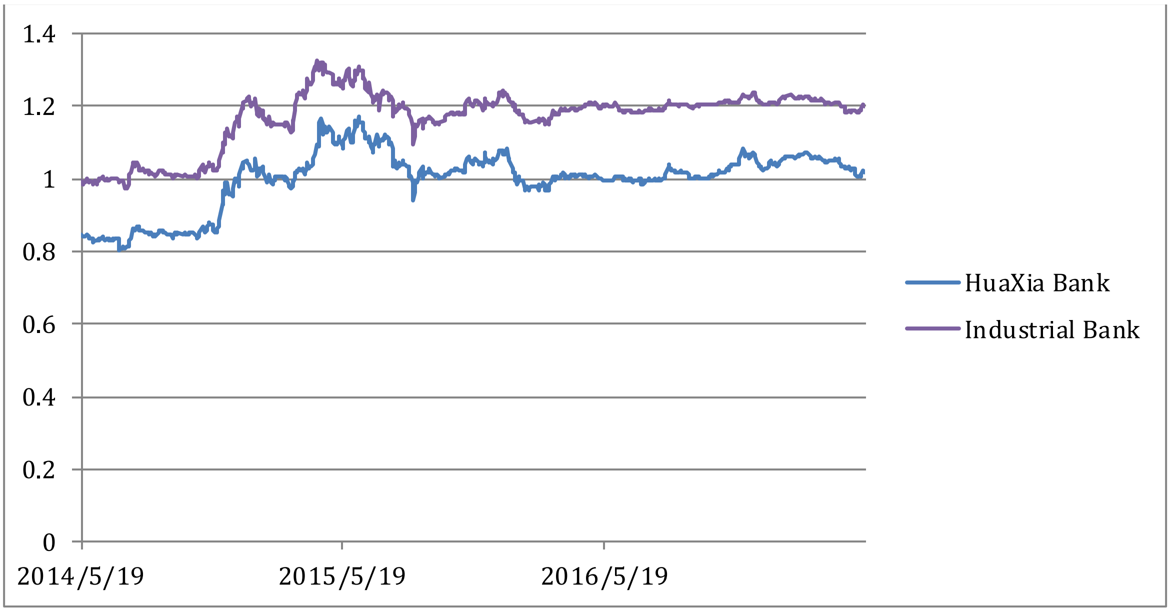
Appendix



Line chart 1: the price of the five banks



Line chart 2: the logarithm price for the five banks



Line chart 3: the logarithm price for HuangXia Bank and Industrial Bank

Null Hypothesis: X has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=17)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.668095	0.4469
Test critical values:	1% level	-3.443442	
	5% level	-2.867207	
	10% level	-2.569850	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: Y has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=17)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.703876	0.4287
Test critical values:	1% level	-3.443442	
	5% level	-2.867207	
	10% level	-2.569850	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(X) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=17)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-22.50676	0.0000
Test critical values:	1% level	-3.443469	
	5% level	-2.867219	
	10% level	-2.569857	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(Y) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=17)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-22.31833	0.0000
Test critical values:	1% level	-3.443469	
	5% level	-2.867219	
	10% level	-2.569857	

*MacKinnon (1996) one-sided p-values.

Result 1 ADF test for HuangXia Bank and Industrial Bank

Dependent Variable: Y
Method: Least Squares
Date: 05/18/17 Time: 19:52
Sample: 5/19/2014 5/19/2016
Included observations: 492

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.199162	0.009424	21.13272	0.0000
X	0.965504	0.009524	101.3738	0.0000

R-squared	0.954489	Mean dependent var	1.150123
Adjusted R-squared	0.954396	S.D. dependent var	0.094067
S.E. of regression	0.020088	Akaike info criterion	-4.973320
Sum squared resid	0.197731	Schwarz criterion	-4.956253
Log likelihood	1225.437	Hannan-Quinn criter.	-4.966619
F-statistic	10276.65	Durbin-Watson stat	0.083295
Prob(F-statistic)	0.000000		

Result 2: the regression of Y and X

Null Hypothesis: RESID01 has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=17)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.965132	0.0030
Test critical values:	1% level	-2.569680	
	5% level	-1.941469	
	10% level	-1.616266	

*MacKinnon (1996) one-sided p-values.

Result 3: the ADF test of residual resid01

Dependent Variable: Y
Method: Least Squares
Date: 05/18/17 Time: 19:56
Sample (adjusted): 5/20/2014 5/19/2016
Included observations: 491 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.008409	0.003608	2.330530	0.0202
X	0.820357	0.023569	34.80638	0.0000
X(-1)	-0.792095	0.025472	-31.09719	0.0000
Y(-1)	0.968634	0.012592	76.92600	0.0000

R-squared	0.996528	Mean dependent var	1.150437
Adjusted R-squared	0.996507	S.D. dependent var	0.093905
S.E. of regression	0.005550	Akaike info criterion	-7.541912
Sum squared resid	0.015001	Schwarz criterion	-7.507726
Log likelihood	1855.540	Hannan-Quinn criter.	-7.528487
F-statistic	46596.41	Durbin-Watson stat	1.891959
Prob(F-statistic)	0.000000		

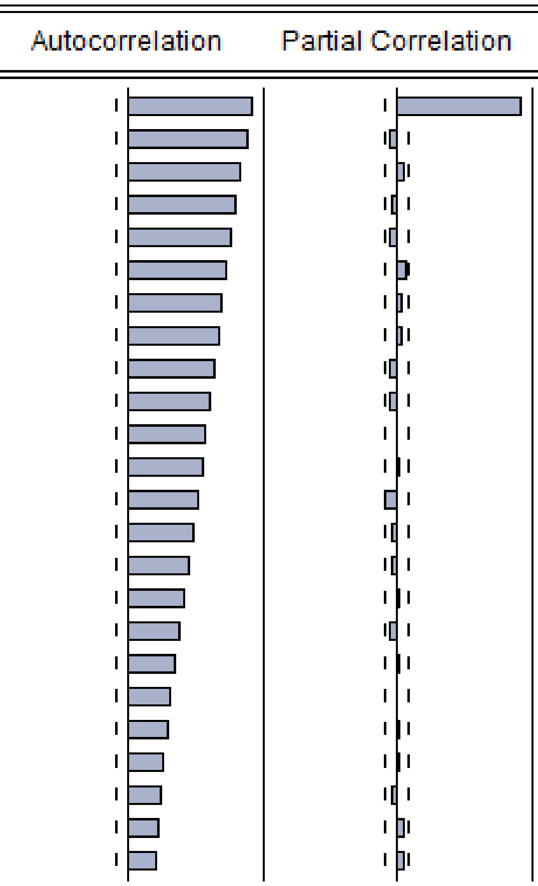
Result 4: the auto-regression of Y

Dependent Variable: E
Method: Least Squares
Date: 05/18/17 Time: 20:02
Sample (adjusted): 5/20/2014 5/19/2016
Included observations: 491 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
E(-1)	0.979048	0.008665	112.9896	0.0000

R-squared	0.928540	Mean dependent var	-0.020260
Adjusted R-squared	0.928540	S.D. dependent var	0.020993
S.E. of regression	0.005612	Akaike info criterion	-7.525865
Sum squared resid	0.015431	Schwarz criterion	-7.517318
Log likelihood	1848.600	Hannan-Quinn criter.	-7.522509
Durbin-Watson stat	1.905476		

Result 5: the auto-regression of residual of spread series



	AC	PAC	Q-Stat	Prob
1	0.958	0.958	453.96	0.000
2	0.913	-0.052	867.22	0.000
3	0.873	0.043	1246.3	0.000
4	0.833	-0.031	1592.1	0.000
5	0.791	-0.051	1904.1	0.000
6	0.756	0.073	2189.6	0.000
7	0.726	0.040	2454.1	0.000
8	0.701	0.034	2700.8	0.000
9	0.671	-0.062	2927.5	0.000
10	0.638	-0.063	3132.6	0.000
11	0.605	-0.003	3317.8	0.000
12	0.575	0.009	3485.4	0.000
13	0.538	-0.096	3632.1	0.000
14	0.499	-0.027	3758.7	0.000
15	0.462	-0.032	3867.3	0.000
16	0.429	0.020	3961.1	0.000
17	0.393	-0.052	4040.0	0.000
18	0.359	0.010	4106.2	0.000
19	0.329	-0.006	4161.8	0.000
20	0.303	0.021	4209.1	0.000
21	0.279	0.015	4249.1	0.000
22	0.252	-0.045	4281.9	0.000
23	0.230	0.051	4309.2	0.000
24	0.215	0.053	4333.2	0.000

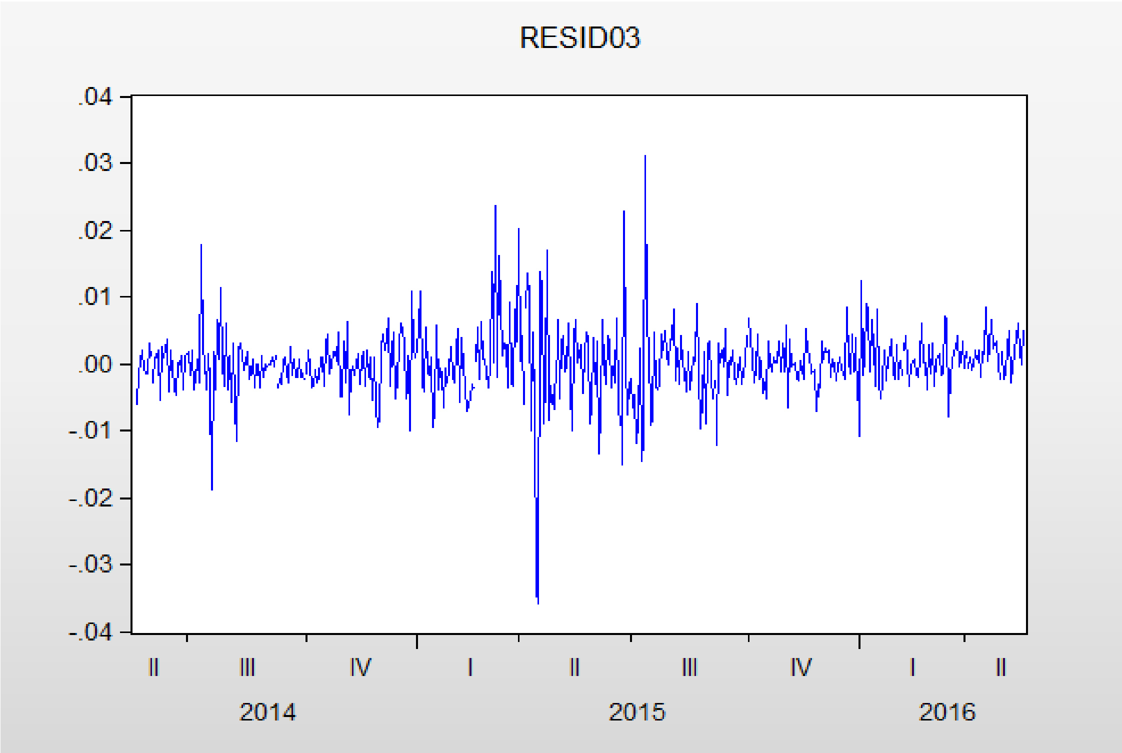
Result 6: the correlogram of spread series

Dependent Variable: SPREAD
Method: Least Squares
Date: 05/18/17 Time: 20:10
Sample (adjusted): 5/20/2014 5/19/2016
Included observations: 491 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.009101	0.003196	2.847439	0.0046
SPREAD(-1)	0.966037	0.012087	79.92052	0.0000

R-squared	0.928886	Mean dependent var	0.263740
Adjusted R-squared	0.928741	S.D. dependent var	0.020993
S.E. of regression	0.005604	Akaike info criterion	-7.526640
Sum squared resid	0.015356	Schwarz criterion	-7.509546
Log likelihood	1849.790	Hannan-Quinn criter.	-7.519927
F-statistic	6387.289	Durbin-Watson stat	1.890018
Prob(F-statistic)	0.000000		

Result 7: the auto-regression of spread series



Result 8: graph of resid03

Dependent Variable: SPREAD
Method: ML - ARCH (Marquardt) - Normal distribution
Date: 05/18/17 Time: 20:14
Sample (adjusted): 5/20/2014 5/19/2016
Included observations: 491 after adjustments
Convergence achieved after 19 iterations
Presample variance: backcast (parameter = 0.7)
GARCH = C(2) + C(3)*RESID(-1)^2 + C(4)*GARCH(-1)

Variable	Coefficient	Std. Error	z-Statistic	Prob.
SPREAD(-1)	1.001222	0.000724	1382.207	0.0000
Variance Equation				
C	1.66E-06	3.88E-07	4.276793	0.0000
RESID(-1)^2	0.269695	0.042621	6.327732	0.0000
GARCH(-1)	0.706942	0.038721	18.25711	0.0000

R-squared	0.927585	Mean dependent var	0.263740
Adjusted R-squared	0.927585	S.D. dependent var	0.020993
S.E. of regression	0.005649	Akaike info criterion	-7.880832
Sum squared resid	0.015637	Schwarz criterion	-7.846645
Log likelihood	1938.744	Hannan-Quinn criter.	-7.867407
Durbin-Watson stat	1.922493		

Result 9: GARCH(1, 1)