Implementing Pairs Trading Strategy on U.S. Stocks

Instructor: Dr. Vinci Chow
Tian Zongyue, 1155124541

Abstract

Pairs trading is a market neutral strategy which essentially revolves around three factors: how to select profitable pairs, when to enter and exit a trade, and how to allocate assets between pairs of securities to maximize profit. This empirical study focuses more on the previous two factors. It designed a pairs trading strategy and implemented the strategy on U.S. stock data to investigate its profitability.

1. Introduction

Pairs trading is a market neutral strategy that originated from Morgan Stanley in the mid-1980s (Vidyamurthy, 2004). It typically involves a portfolio of two securities, whose spread(difference) follows a mean-reverting process. When the spread diverges substantially from its mean, by going long on the relatively undervalued security and short on the relatively overvalued one in a predetermined ratio, a profit can be made by closing the position when the spread reverts back to the mean.

Like all arbitrage strategies, the more this strategy is applied, the less profitable pairs exist (Meissner, 2016). This empirical study tried to find profitable pairs among U.S. stocks using data from 2013-12-30 to 2018-12-30, and designed a trading model with the help of ARIMA model and LSTM model to test if the pairs are truly profitable or not using the data from 2018-12-30 to 2019-12-30.

2. Data

For the strategy's underlying security price series, we use the daily adjusted close price of 944 U.S. stocks. The training data set ranges from 2013-12-30 to 2017-12-30 (4 years).

The validation data set ranges from 2017-12-30 to 2018-12-30 (1 year). And the test data set range from 2018-12-30 to 2019-12-30 (1 year), as shown in Table 1.

Table 1.

Training set	Validation set	Test set
30/12/2013 - 30/12/2017 (1009)	30/12/2017 -30/12/2018 (251)	30/12/2018 – 30/12/2019 (250)

3.Pairs Selection

3.1 Basic Methods

There are 3 commonly-used basic methods to identify profitable pairs: the distance method, the correlation method and the cointegration method.

a) The distance method, proposed by Gatev, Goetzmann, and Rouwenhorst (2006), selects pairs which minimize the sum of squared distances between the two assets' normalized price series.

That is, select pair(i, j) by

$$argmin_{i,j} \sum_{t=0}^{T} (p_{i,t} - p_{j,t})^2 ,$$

where $p_{i,t}$ is the normalized price of asset i at time t. However, this method has a major drawback: The spread generated by this method tends to have the smallest fluctuation around its mean as well, making it less profitable.

b) The correlation method selects pairs who have the highest absolute Pearson correlation coefficient. When implementing this method, one needs to note that Pearson correlation is time-frame sensitive. Short time frame and long time frame may produce highly different results (Wilmott, 2009).

c) The cointegration method, suggested by Engle and Granger (1987), identifies pairs by checking whether there exists a linear combination of two assets' price series that is stationary. Since stationary process has a constant mean, variance, and autocorrelation, it is mean-reverting, making this method suitable for finding profitable pairs.

3.2 Pairs Selection Method in this Study

This study identifies pairs based on the cointegration method and four additional criteria to increase robustness. An exhaustive search is applied to the 944 stock data (446040 possible pairs), based on the four criteria below:

a) Firstly, a pair is selected if they are cointegrated. Cointegration is confirmed by applying the augmented Engle-Granger two-step cointegration test (Engle and Granger ,1987) with a p-value threshold of 0.01 on the training set. In order to approximate the linear combination of price series Y and X that is stationary, by normalizing the coefficient of Y to 1, we fit the spread

$$S_{X,Y\ t} = Y_t - \beta X_t$$

to an Ornstein–Uhlenbeck process by maximum likelihood estimation with 50 different values of β equally spaced in interval [0.05,1]. The β that gives the highest average log likelihood is selected to generate the spread $S_{X,Y}$ for pair (X, Y).

The Ornstein–Uhlenbeck process (OU process) is a commonly-used process when modeling stationary mean-reverting time series.

A time series Xt follows an OU process if it follows the probability density function

$$f^{OU}(x_i|x_{i-1};\theta,\mu,\sigma) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{(x_i - x_{i-1}e^{-\mu\Delta t} - \theta(1-e^{-\mu\Delta t}))^2}{2\tilde{\sigma}^2}\right)$$

, where $\tilde{\sigma}^2 = \sigma^2 \frac{1-e^{-2\mu\Delta t}}{2\mu}$, or its SDE follows

$$dX_t = \mu(\theta - X_t)\,dt + \sigma\,dB_t\,,$$

where θ is the long-term mean, μ is called the speed of reversion, sigma is called the size of noise, and Bt is a standard Brownian motion.

For each pair (Y,X), we denote its spread at time t with its own coefficient β by

$$S_t^\beta \equiv Y_t - \beta X_t$$

(Spread's subscript X,Y is omitted for simplification). To express the parameter values produced by maximum likelihood estimation, we make the following definition:

$$X_x = \sum_{t=1}^{n} S_{t-1}^\beta , X_y = \sum_{t=1}^{n} S_t^\beta ,$$

$$X_{xx} = \sum_{t=1}^{n} (S_{t-1}^\beta)^2, X_{yy} = \sum_{t=1}^{n} (S_t^\beta)^2, X_{xy} = \sum_{t=1}^{n} S_{t-1}^\beta S_t^\beta$$

The optimal parameter estimates are given by

$$\begin{aligned} \theta^* &= \frac{X_y X_{xx} - X_x X_{xy}}{n(X_{xx} - X_{xy}) - (X_x^2 - X_x X_y)}, \\ \mu^* &= -\frac{1}{\Delta t} \ln \frac{X_{xy} - \theta^* X_x - \theta^* X_y + n(\theta^*)^2}{X_{xx} - 2\theta^* X_x + n(\theta^*)^2}, \\ (\sigma^*)^2 &= \frac{2\mu^*}{n(1 - e^{-2\mu^*\Delta t})} (X_{yy} - 2e^{-\mu^*\Delta t} X_{xy} + e^{-2\mu^*\Delta t} X_{xx} \\ &\quad - 2\theta^*(1 - e^{-\mu^*\Delta t})(X_y - e^{-\mu^*\Delta t} X_x) + n(\theta^*)^2(1 - e^{-\mu^*\Delta t})^2) \end{aligned}$$

Then the optimal β for pair (Y, X) is given by

$$argmax_\beta \frac{1}{n} \sum_{t=1}^{n} ln(f^{OU}(S_t^\beta | S_{t-1}^\beta ; \theta^*, \mu^*, \sigma^*))$$

b) Secondly, since we always take a long position in one security and a short position in the other, negative βs are ignored in the previous process. Thus, the linear combination found in the first step may not be the correct one. We simply apply the Dickey–Fuller test with a p-value threshold of 0.01 on each spread to confirm its stationarity.

c) Thirdly, to increase the robustness of spread's stationarity(reduce the probability of spurious relationship),we separately calculate the mean of the spread in the training set and in the validation set. As the mean of a stationary process is constant, intuitively the smaller the difference of the means between the training set and the validation set, the more likely this spread is a stationary process. Therefore, we only keep the pairs whose difference of its means is above its 25th percentile and below its 75th percentile.

d) Finally, to increase profitability, we want the spread to cross its mean frequently. Because the more frequently a spread crosses its mean, the more opportunities there are to trade. For instance, if a spread only crosses its mean 2 times a year, then potentially there are only 2 opportunities to close a position. We define mean crossing at time i of a time series S from t=a to t=b (a <i ≤b) as an indicator function I:

$$I(S_a^b,i)=\begin{cases}1 \; if \; (S_{i-1}-\frac{\sum_{t=a}^{b}S_t}{b-a})*(S_i-\frac{\sum_{t=a}^{b}S_t}{b-a})\leq 0\\ 0 \; otherwise\end{cases}$$

Mean crossing time of series S from t=a to t=b (a<b) is defined by

$$MT(S_a^b)\equiv\sum_{t=a+1}^{b}I(S_a^b,t)$$

Let

$$SMT=\{MT(S_{training\,set})|S\in set\ of\ pairs\}.$$

We discard those pairs whose mean crossing time in its training set are below the 75th percentile of SMT. Lastly, for a more consistent stationary spread S,

$$MT(S_{training\,set})<=4*MT(S_{validation\,set})$$

should hold since the length of the training set is 4 times as long as the length of the validation set and the 4-year training set is using a long term mean. We discard the pairs that do not satisfy this condition.

After pair selection, we assume that the remaining pair’s spread is stationary, has a constant long-term mean, and cross its mean value frequently.

4. Trading Strategy

4.1 Predetermined framework

Every trading model for pairs trading essentially must focus on two things: deciding when to open and when to close a position. There are only two possible positions in pairs trading for any spread $Y_t - \beta X_t$:

a) Long: When spread is small and we expect it to become larger, we long the spread (long Y & short X)

b) Short: When spread is large and we expect it to become smaller, we short the spread (short Y & long X)
And ideally we close a position when the spread reaches expectation.

For the trading model in this study, no commission fee is imposed for both long and short position. However, there is an annually paid interest for shorting any stock. When shorting a stock at time t and close the position at time T, the interest is calculated by

$$(Price_t * Volume * (T-t) * 0.03) / 365$$

, where 0.03 is the fixed interest rate.

For each pair (Y, X) and its own β, it follows the same trading framework:

1. Each pair is traded under a finite time interval, e.g. for t = 0,1,2,…,T
2. At t=0, $10,000 is given exclusively to each pair.
3. When opening a position at time t, we first determine the integer volume of the stock to be longed, by the money we have divided by its present price

   $$\underline{Money_t / Price_t}$$

   ,where the underline gives the floor value(integer) of the value above it. Then if the stock to be shorted is X, its volume is determined by

   $$\underline{\beta Volume_Y} .$$

   If it is Y that is about to be shorted, its volume is $\underline{Volume_X \div \beta}$.
4. A pair's position can be only opened after all of its previous opened position are closed.
5. At the end of a year, all the interests incurred in this year must be paid.
6. If after closing a position, the account's money is less than 0, the account will be frozen permanently (No position will be allowed to open).
7. At t=T, any position that is open will be closed and the trading stops. The rate of return is calculated by $\frac{Asset_T}{10000} - 1$.

4.2 Determine the Open and Close Signal

We define the percentage change of spread S at time t by $\Delta_t = \frac{S_t - S_{t-1}}{S_{t-1}}$.

Define the expected percentage change of spread S at time t+1 by $\Delta'_{t+1} = \frac{S'_{t+1} - S_t}{s_t}$,

where S'_{t+1} denotes the expected value of S at time t+1.
Define strategy as a function from real number to action space [‘Go Short’, ‘Exit Short’, ‘Go Long’, ‘Exit Long’] by

$$strategy(\Delta'_{t+1} \mid Sn, Sp, Lp, Ln) = \begin{cases} Go\ Short & if\ \Delta'_{t+1} < Sn < 0 \\ Exit\ Short & if\ \Delta'_{t+1} > Sp \geq 0 \\ Go\ Long & if\ \Delta'_{t+1} > Lp > 0 \\ Exit\ Long & if\ \Delta'_{t+1} < Ln \leq 0 \end{cases}$$

Under this strategy, the goal is to find parameters (Sn, Sp, Lp, Ln) that will generate the largest rate of return for each pair i under the previous trading framework.

Index the selected pairs (and their spreads) using $i \in I \equiv \{1,2,3...\}$.
Denote the set of all percentage changes of spread i in the training set by $f_i(training)$, the set of all positive changes for spread i by $f_i^+(training)$, the set of all negative changes by $f_i^-(training)$.
Define function $Qf_i^{\pm}(training)(q)\ ,\ q \in [0,1]$ to be the (q*100)th percentile of $f_i^{\pm}(training)$.

Propose that by tuning parameters (sn, sp, lp, ln) according to the rate of return generated using

$$strategy(\Delta_{t+1}|(Qf_i^-(training)(sn), Qf_i^+(training)(sp), Qf_i^+(training)(lp), Qf_i^-(training)(ln))$$

, where $\Delta_{t+1} \in f_i(validation)$ and $t \in Validation\ Set_i$, one can determine the optimal (Sn, Sp, Lp, Ln) for the testing set of pair i as

$$(Qf_i^-(training)(sn), Qf_i^+(training)(sp), Qf_i^+(training)(lp), Qf_i^-(training)(ln))$$

The possible values for (sn, sp, lp, ln) used in this study is given by:

sn: [0.1,0.15,0.2,0.25,0.3,0.35], sp: [0,0.05,0.1,0.15,0.2],
lp: [0.9,0.85,0.8,0.75,0.7,0.65], ln: [1,0.95,0.9,0.85,0.8]

4.3 Expected Percentage Change Forecasting

After (Sn, Sp, Lp, Ln) for each pair i is determined, we only need to forecast the next day's expected percentage change in the testing set to see if our strategy works or not. This study proposes two models: LSTM model and ARIMA model.

Each parameter is tuned by training the model on the training set and testing the model on the validation set. For each pair i, the parameter that has the least mean squared error on the validation set is selected.

4.3.1 LSTM Neural Network

This model is composed of two LSTM layers and one dense layer.

The model written in python using Keras is shownin Figure 1.

Figure 1.

```
Inputs = Input (shape = (T, 1))
x = LSTM(lstm1, activation='relu',
return_sequences=True)(Inputs)
x = LSTM(lstm2, activation='relu')(x)
Output = Dense(1,activation='relu')(x)
```

The parameters to be determined for pair i are:

i. The size of the timestep: T
ii. Number of neurons for the first LSTM layer: lstm1
iii. Number of neurons for the second LSTMlayer: lstm2

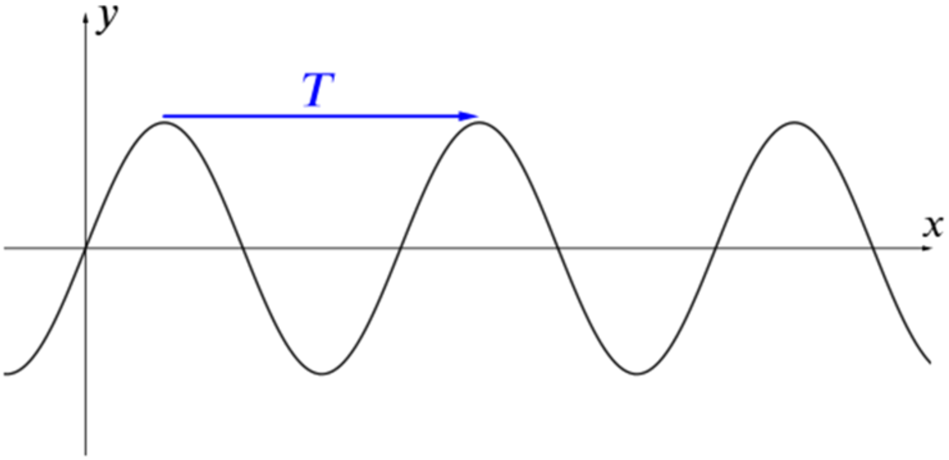
iv. Epoch: [50, 100,150,200 300]

For each pair i, define T_i by

$$\frac{|spread_i^{training}|}{2*MT(spread_i^{training})}$$

, namely, the cardinality of spread i in the training set divided by 2*Mean Crossing Time of spread i in the training set. Ideally, given a periodic mean-reverting time series, the mean is usually cross 2 times in one period, as in Figure 2. Hence, the length of one period is exact $T_{i.}$ Any timestep bigger than T_i might give redundant information.

Figure 2.



After input shape was determined for each pair i as Ti, possible lstm1i takes two values from set{16,32,64,128} that are the closest to Ti. For each lstm1i, lstm2i takes its possible value from set {lstm1i, lstm1i / 2}

4.3.2 ARIMA Model

A rolling ARIMA model (p,0,q) was used, where p is the order of autoregression, q is the order of moving average. From 3.Pair Selection Methods, all spreads are tested as stationary, hence the order of differencing is 0. Since the the order of differencing is 0, p is highly unlikely to be 0.

Hence all possible values of p and q are given by:

p: [1, 2], q: [0, 1, 2]

The rolling ARIMA model is designed as follows:

1. For each pair i, use all the data from Si0 to Sit to fit in ARIMA(p,0,q), then forecast the value Si t+1’
2. Repeat the first step, for every t from the last index of the training set to the second last index of the validation set to obtain [Si t+1’,Si t+2’,…,Si t+n’] (n being the size of the validation set)
3. Calculate the mean squared error using the series in ii. and spread i in the validation set to tune (p,q)

5.Results

5.1 Pairs Selection

After implementing *section 3.2’s* pair selection method on the training set of 449 stocks, 463 pairs were selected.

Due to computational limitation, only 20 pairs which are randomly drawn from the 463 pairs were investigated in *section 4.*

5.2 Open and Close Signal Selection

After implementing *section 4.2*‘s strategy to find the open and close signal, 16 pairs successfully found its optimal signal which generates positive rate of return on the validation set. The other 4 pairs can not found any signal that generates positive rate of return. It could be caused by drastic change between the distribution of return(next day percentage change) in the training set and validation set, resulting in the percentile values selected from the training set not suitable for those in the validation set. After separately computing each pair’s variance in the training set and the validation set, and taking the difference between the two variances, the average ranking(from the largest difference to the smallest one) of the 4 pairs is 5 (1, 2, 8, and 9 respectively). It favors the previous hypothesis, but the number of sample(20) is not sufficient enough to reach a conclusion.

5.3 The LSTM Model

After training the LSTM model proposed in *section 4.3.1*, the mean of mean squared error

on the validation set is 0.035337, with the highest one being 0.062894. It suggests the LSTM is likely to preform well when predicting the testing set values. After implementing *section 4,2*'s strategy using LSTM forecasted values on the 16 pairs with their own optimal open and close signals, the mean rate of return is 33% with the highest return being 186% and the lowest being -17%. There are in total 33 number of trades occurred in the testing set (Open then close on one pair is counted as 1 trade) and 21 of them are profitable trades (rate of return between the open position and close position is positive), whereas there are 1458 trades occurred in the validation set. There are in total 7 profitable pairs (total rate of return > 0), 4 unprofitable pairs and 9 pairs which never made any trade in the testing set.
After calculating the difference in variance between the real percentage change on the validation set and the forecasted percentage change on validation set, result shows that all pairs' percentage changes' variance became smaller when using forecasted values. The mean of the difference is 25, which may be the reason why there is a huge drop in the number of trades made between the validation set and the testing set (lower variance suggesting less expected percentage change value can reach the predetermined percentile as they are likely all concentrated around the 50th percentile).

5.4 The ARIMA Model

After training the ARIMA model described in *section 4.3.2*, *its* mean of mean squared error on the validation set is 0.035083, with the highest one being 0.064513, suggesting ARIMA model performed similarly to the previous LSTM model on these stationary time series. However, the training speed of ARIMA model is way faster than the training speed of LSTM model. After implementing *section 4.2*'s strategy using ARIMA forecasted values, the mean rate of return is -0.06%, with 18 pairs which never made any trade in the testing set and the lowest rate of return being -10%, the highest being 9%. There are in total 6 number of trades, 4 of which are unprofitable.
The drop in the total number of trades and the number of traded pairs suggests a larger decline in the variance of ARIMA forecasted percentage changes. After repeating the variance difference calculation process at the end of *section 5.3*, the mean difference is 29.5, in favor of the hypothesis.

6.Conclusion and Future Work

By following the pair selection process and the trading model design using an LSTM model on the 20 randomly selected pairs, it generated 33% average return in one year, with some pairs gained over 100% rate of returns. Further study could try the same trading model design on the remaining 443 pairs or extend the test set time frame further to see if the return is consistent. One can also investigate the factors causing negative returns to further refine the selected pairs and to decrease risk.

While predicting stationary time series appears to be less time consuming and more accurate than non-stationary ones, when it is being implemented into trading strategies, even small errors can incur losses. In this study, ARIMA model's forecast appears to be closer to its present value than LSTM model's. The huge drop in the number of trades between the validation period and the testing period suggests this trading strategy is likely to be sub-optimal. One could use different forecasting models to gain more accurate results or change the strategy described in *section 4.2* to increase profitability.

Reference

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies,* 19(3), 797–827.

Meissner, G. (2016). Correlation Trading Strategies:Opportunities and Limitations. *The Journal of Trading*, 11(4), 14–32. doi:10.3905/jot.2016.11.4

Vidyamurthy, G. (2004), *Pairs Trading: Quantitative Methods and Analysis*. Hoboken, NJ: John Wiley and Sons

Wilmott, P. (2009). Frequently Asked Questions in Quantitative Finance. Chichester, U.K.: John Wiley