Cointegration and Risk-factor Correlation-based Enhanced Stock-pair Trading Strategy

Abstract

In this project, we utilized the stock's cointegration to create a stock pair trading strategy. For enhancing the performance, we also incorporate correlation into the model. In the section 1, we walk through the models for this project, then in the section 2 we talk about the strategy in detail. Data we used for back test is given in section3, followed by our back-test results and analysis on them. Finally, we give out our conclusions as summary and our references.

Team members:

Name: Yifan Li ID: N14714308

Name: Youyuan Zhang ID: N10182594

Name: Hanyu Zhang ID: N17047193

1. Model

In this section we talk about the models we used in this project, including cointegration, covariance matrix estimation and Ornstein–Uhlenbeck process. In the project, we use cointegration to find trading opportunities, we use Ornstein–Uhlenbeck process to model process and we use correlation to enhance our strategy

1.1 Cointegration

1.1.1 Definition

Cointegration is a statistical property of a collection $(X_1, X_2, \dots, X_k)$ of time series variables. If all the series are integrated of order d and a linear combination of this collection is integrated of order less than d, then the collection is said to be co-integrated.

In statistics, the order of integration, denoted $I(d)$, of a time series is a summary statistic, which reports the minimum number of differences required to obtain a stationary series.

Formally, if (X, Y, Z) are each integrated of order d, and there exists coefficients (a, b, c) such that $aX + bY + cZ$ is integrated of order less than d, then (X, Y, Z) are co-integrated.

If two or more series are individually integrated (in the time series sense) but some linear combination of them has a lower order of integration, then the series are said to be co-integrated.

A common example is where the individual series are first-order integrated but some vector of coefficients exists to form a stationary linear combination of them.

1.1.2 Intuitions of Pair Trading Strategy

The most widely used model of stock price is Geometric Brownian Motion:

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

S_t is the stock price, μ is the average return, σ is the volatility and W_t is Brownian motion.

From Ito's Lemma, we have

$$dlog(S_t) = \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dW_t$$

$log(S_t)$ is a Brownian motion with drift, it is a I(1) process.Then we could apply the cointegration idea to $log(S_t)$ where S_t is the stock price.

Suppose we have two stocks 'well co-integrated', denote stock prices as X_t and Y_t.

$$Spread_t = log(Y_t) - (\alpha + \beta\, log(X_t))$$

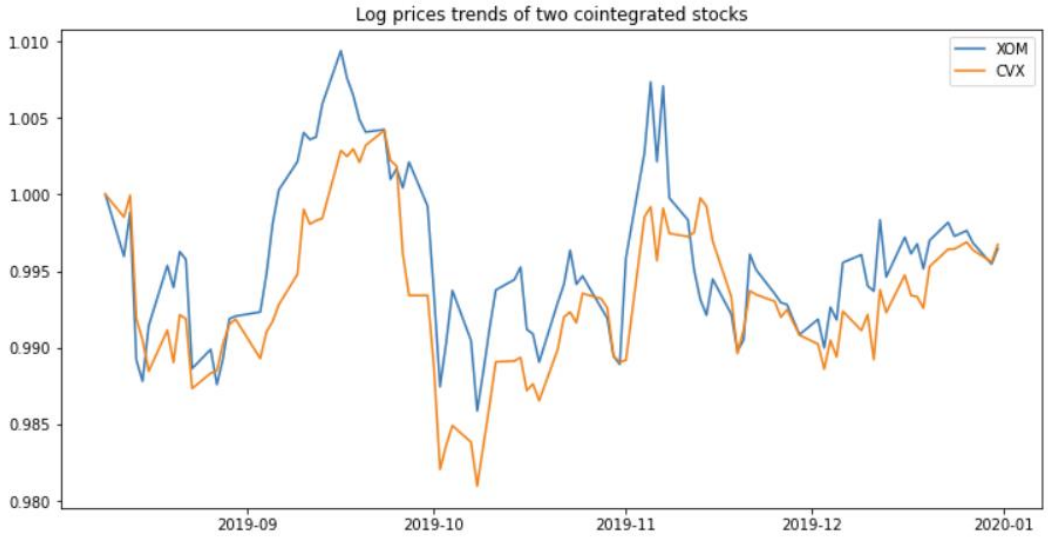
here we define 'well co-integrated' by existing coefficients α and β such that $log(Y_t) - (\alpha + \beta\, log(X_t))$ is a stationary time series.

Then we will have a simple pair trading strategy:

$$\begin{cases} if\ Spread_t > m\text{: } buy\ X_t\ and\ sell\ Y_t \\ \\ if\ Spread_t < -m\text{: } buy\ Y_t\ and\ sell\ X_t \end{cases}$$

$m > 0$ and it is the threshold in the pair trading strategy, generally decided by transaction fee and the expected profit. The profit of pair trading strategy is generated from the mean-reversion property of $Spread_t$ defined as above.

The following figure shows how two co-integrated stocks behave:



1.1.3 Cointegration Test

1.1.3.1 Engle–Granger Two-Step Method

The Engle-Granger two-step method contains two steps, the first step is OLS regression and the second step is to test whether the residuals from the regression are stationary.

Suppose we have two time series X_t and Y_t, and we need to test whether the two time series are co-integrated or not.

The first step is to do OLS regression

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

then we could calculate the residuals of the OLS regression

$$u_t = Y_t - \widehat{\beta_0} - \widehat{\beta_1} X_t$$

The second step is to test whether time series u_t is stationary or not. Here we could apply the Augmented Dickey-Fuller (ADF) test or Phillips-Perron test.

If u_t is stationary, then we could conclude that the two-time series X_t and Y_t are co-integrated; if u_t is not stationary, then we could conclude that the two-time series X_t and Y_t are not co-integrated.

However, the Engle-Granger test may suffer from bias when the number of time series is greater than two. It needs to explicitly define which series is to be used as the dependent variable in the regression. Finally, the OLS regression in the first step will lead to spurious estimators if the variables are not co-integrated. This makes the stationarity analysis on the residuals unreliable.

1.1.3.2 Johansen Test

The Johansen test is a procedure for testing co-integration of several I(1) time series, suppose we have k time series in total. This test permits more than one co-integrating relationship, so it is more generally applicable than the Engle-Granger two-step method.

There are two types of Johansen test, either with trace or with eigenvalue, and the inferences might be a little bit different.

For the trace test, the null hypothesis is that the number of cointegration vectors is $r = r^* < k$ and the alternative is $r = k$. Testing proceeds sequentially for $r^* = 1$ and then $r^* = 2$, etc. The first non-rejection of the null hypothesis is taken as an estimate of r.

The null hypothesis for the 'maximum eigenvalue' test is as for the trace test but the alternative is $r = r^* + 1$ and testing proceeds sequentially for $r^* = 1$ and then $r^* = 2$, etc. The first non-rejection of the null hypothesis is taken as an estimate of r.

The Johansen method is the maximum likelihood estimator of the so-called reduced rank model. Consider a Vector Autoregressive (VAR) process for a p-dimensional vector

$$X_t = \sum_{i=1}^{k} A_i X_{t-i} + u_t$$

The VAR process could be transformed into a Vector Error Correction Model (VECM):

$$\Delta X_t = \pi X_{t-1} + \sum_{i=1}^{k-1} \Delta X_{t-i} + u_t$$

where the multiplier matrix π could be decomposed into two matrices such that

$$\pi = \alpha\beta^T$$

The vector or matrix β represents the co-integration vectors, and α is the matrix of error-correction coefficients which measure the rate each variable adjusts to the long-run equilibrium. The number of co-integrating vectors is identical to the number of stationary relationships in the π matrix. Mathematically, the rank of π determines the number of independent rows in π, and therefore also the number of co-integrating vectors. The rank of π is given by the number of significant eigenvalues found in π. Each significant eigenvalue represents a stationary relationship. This makes the two types of Johansen test, either with trace or with eigenvalue, equivalent.

The Johansen test for co-integration is commonly regarded as superior to the Engle-Granger two-step method. This is particularly true when the number of time series is greater than two. The Johansen test could be used for both determining how many co-integration vectors there are and also estimating all the distinct relationships. And we use this method to test cointegration of stock pairs.

1.2 Covariance Matrix Estimation

Covariance Matrix Estimation is a technique about cleaning the covariance matrix for the purpose of reducing the noise in it. We incorporate this technique to our strategy because, according to report presented by Shuo Qu (in the reference section), stocks that are similar in risk factor spaces are more likely to move in tandem and less likely to diverge significantly, which can substantially reduce failure convergence in the out-sample investment periods. So taking the correlations into our consideration can help us better understand the relationship between assets and provide a consistent and interpretable framework for performance and risk attribution.

We know risk is complex and multidimensional, traditional risk decomposition methods like factor models just can provide limited information of risk. Due to this potential problem and the low accessibility of fundamental data of our target companies. We decide to use Covariance Matrix Estimation technique to support our trading strategy since it can decompose the risk in the most efficient way and give us the most important k risk factors (although unknown) to clean the covariance matrix.

More specifically, after deciding all cointegrated stock pairs, we calculate their estimated covariance matrix of 1 year historical data, then we do PCA on the matrix to extract the eigenvalues and the corresponding eigenvectors and we are considering the following method to reduce the noise in the matrix:

$$\Sigma_{Estimate} = \sum_{i=1}^{K} \lambda_i v_i {v_i}^T + \frac{\sum_{i=K+1}^{N} \lambda_i}{N-K} \sum_{i=K+1}^{N} v_i {v_i}^T \quad where$$

$$\{\lambda_i\} \, and \, \{v_i\} \, are \, eigenvalues \, and \, eigenvectors$$

Which is also called “Eigenvalue Clipping”

With the new covariance matrix, we also can get easily get the reduced correlation matrix which clarifies the risk factor correlation between different pairs. Based on the above, higher correlation indicates a higher probability of two stocks are exposed to same risk factors and therefore fundamentally similar, consequently their prices are more likely to move together in the near future in the absence of idiosyncratic events. We use this idea to enhance our pair-trading strategy, if the calculated correlation is low than a threshold, we will not consider this pair as a tradable pair in the incoming investment period. And once we open a trade, if in the process the correlation goes down by more than 20%, we will close the trade.

1.3 Ornstein–Uhlenbeck Process

1.3.1 Definition and Mathematical Properties

The Ornstein–Uhlenbeck process is defined by the following stochastic differential equation:

$$dx_t = \theta(\mu - x_t)dt + \sigma dW_t$$

where $\theta > 0$ and $\sigma > 0$ and μ is a constant, W_t denotes the Wiener process.

Assuming x_0 is constant, the mean is

$$E(x_t) = x_0 e^{-\theta t} + \mu(1 - e^{-\theta t})$$

and the covariance is

$$cov(x_s, x_t) = \frac{\sigma^2}{2\theta}(e^{-\theta|t-s|} - e^{-\theta(t+s)})$$

Then we will have

$$\lim_{t \to +\infty} E(x_t) = \mu$$

$$var(x_t) = \frac{\sigma^2}{2\theta}(1 - e^{-2t\theta})$$

$$\lim_{t \to +\infty} var(x_t) = \frac{\sigma^2}{2\theta}$$

These two results are the mean and standard deviation values for deciding whether we open a trade or not.

1.3.2 Modeling O-U Process

In order to model the O-U process on a computer, it is usual to discrete time and calculate samples at discrete time steps of width Δt.

$$dS_t = \theta(\mu - S_t)dt + \sigma dW_t$$

$$S_t - S_{t-1} = \theta(\mu - S_{t-1})\Delta t + \sigma dW_t$$

$$S_t = S_{t-1} + \theta(\mu - S_{t-1})\Delta t + \sigma dW_t$$

Gillespie (1996) points out that this simulation is only valid when the discrete Δt is sufficiently small.

An exact formula that holds for any size of Δt is:

$$S_t = e^{-\theta\Delta t}S_{t-1} + \left(1 - e^{-\theta\Delta t}\right)\mu + \sigma\sqrt{\frac{(1 - e^{-2\theta\Delta t})}{2\theta}}\, dW_t$$

And we will use this formula instead.

1.3.3 Parameter calibration of observed O-U Process

Here we will use Least Square regression for parameter estimation.

Firstly, we can take the updating formula above and simply turn it into a regression:

$$S_t - S_{t-1} = \theta(\mu - S_{t-1})\Delta t + \sigma dW_t$$

$$S_t - S_{t-1} = \theta\mu\Delta t - \theta S_{t-1}\Delta t + \sigma dW_t$$

$$y = a + bx + \varepsilon_t$$

If we regress a 'y' value of $S_t - S_{t-1}$ against an 'x' of S_{t-1} then we will recover $\hat{\theta}$ as $\frac{b}{\Delta t}$, and from there we can recover $\hat{\mu}$ as $\frac{a}{\hat{\theta}\Delta t}$.

Finally, we can recover $\hat{\sigma}$ as $\frac{sd(\varepsilon_t)}{\sqrt{\Delta t}}$.

If we use the exact updating formula,

$$S_t = e^{-\theta\Delta t}S_{t-1} + \left(1 - e^{-\theta\Delta t}\right)\mu + \sigma\sqrt{\frac{(1 - e^{-2\theta\Delta t})}{2\theta}}\, dW_t$$

$$y = a + bx + \varepsilon_t$$

we notice that we can now regress S_t against S_{t-1} , and derive

$$b = e^{-\hat{\theta}\Delta t}$$

$$\hat{\theta} = -\frac{ln\,(b)}{\Delta t}$$

$$(1 - b)\hat{\mu} = a$$

$$\hat{\mu} = \frac{a}{1 - b}$$

$$\hat{\sigma} = sd(\varepsilon_t)\sqrt{\frac{2\hat{\theta}}{1 - e^{-2\hat{\theta}\Delta t}}}$$

And the half-life of the process is given by $\frac{ln\,(2)}{\theta}$

1.3.4 Application of O-U Process

In our model, given two time series of asset prices, firstly we will calculate the return rates for each asset. Then we will do OLS regression, which means we regress the return rates of one asset against the return rates of the other asset. Then we calculate the time series of residuals of our regression.

According to the Engle-Granger two-step method for cointegration test, if the residuals are stationary, then the initial two time series of asset prices has a good cointegration between them. So we just need to explore the properties of the residuals.

Then we will regard the residuals as an O-U process and we use the above formulas to estimate the parameters of this O-U process. We could also use the parameters we get and the mathematical properties of O-U process to calculate the expectation and variance of the residual, as time t goes to infinity. These properties of the residuals could help us to decide the cointegration between the initial two time series of asset prices.

2. Strategy

2.1 How we find the trading signal

Firstly, at the beginning of a investment period, we evaluate all stock pairs in our dataset. Once a pair passes Johansen test and its correlation is higher than our threshold, we consider this pair of stocks as a tradable pair in the following investment period. And we do regression on $log(S_{t,1})$ and $log(S_{t,2})$:

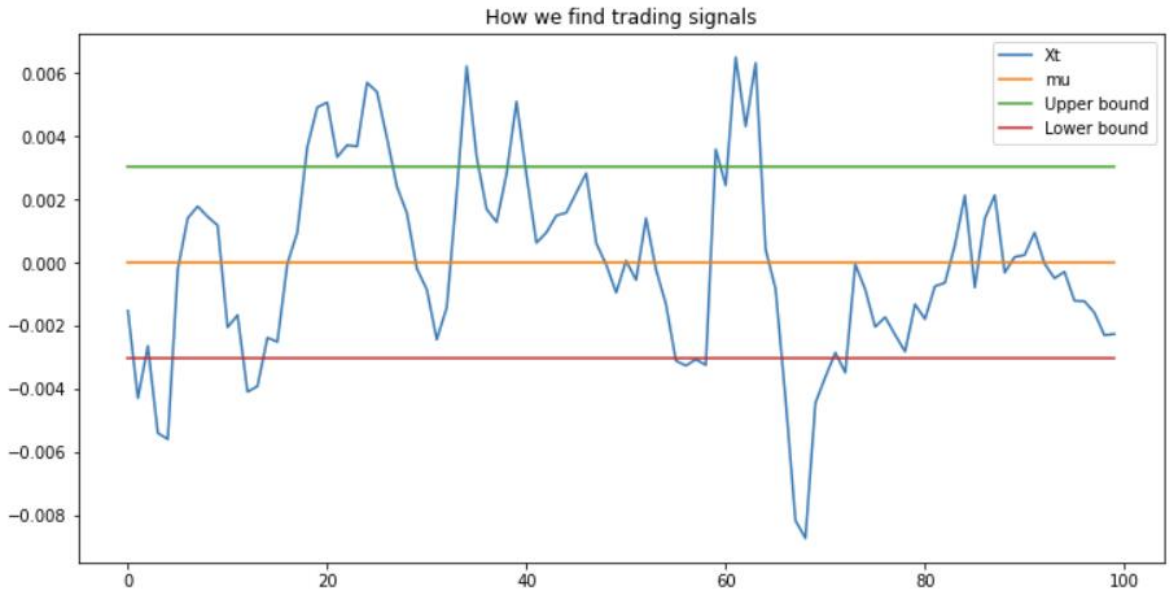
$$log(S_{t,1}) = a + b\, log(X_t) + \varepsilon_t$$

After using OLS to find a and b, we model the residual by OU process and calibrate the parameters θ, μ, and σ and calculate the mean $\mu_x = \mu$ and standard deviation $\sigma_x = \frac{\sigma}{\sqrt{2\theta}}$ of residual with the formulas proven in the Section 1.

Then during the period, we continuously track the value of

$$X_t = log(S_{t,1}) - a - b\, log(S_{t,2})$$

Once the value reaches the upper bound $\mu_x + 2\sigma_x$, if remaining trading time is greater than 2*half-time, we open a trade for this pair by shorting 1 unit of asset 1 and longing b units of asset 2; and we execute the opposite trade if the value reaches the lower bound $\mu_x - 2\sigma_x$.



2.2 How we quit a trade

We stop the trade of a pair once this pair satisfies one or more of the following requirements:

1. Xt has reverted within two boundaries;
2. After ten half-lives has passed since the position was opened;
3. At the end of 3-month trading period if the pair hasn't converged;
4. Correlation dropped by more than 20% post trade open;
5. Xt break through the stop-loss boundaries ($\mu_x \pm 3\sigma_x$).

2.3 Other details

When we consider trading cost, since all the companies in our asset pool are big cap companies and full of liquidity, we use 5 bps are trading cost for adjusting the open price and close price of a trade.

The threshold for correlation will be 0 at first, but we try different threshold to see if it works.

A pair can be reopened after it is closed in any given investment period.

3. Data and parameter settings

3.1 Data

Due to the limited accessibility of stock data, we select 20 stocks in different industries from Nasdaq, data range from 2010.01.01 to 2019.12.31. The details of our asset pool is given by the following table:

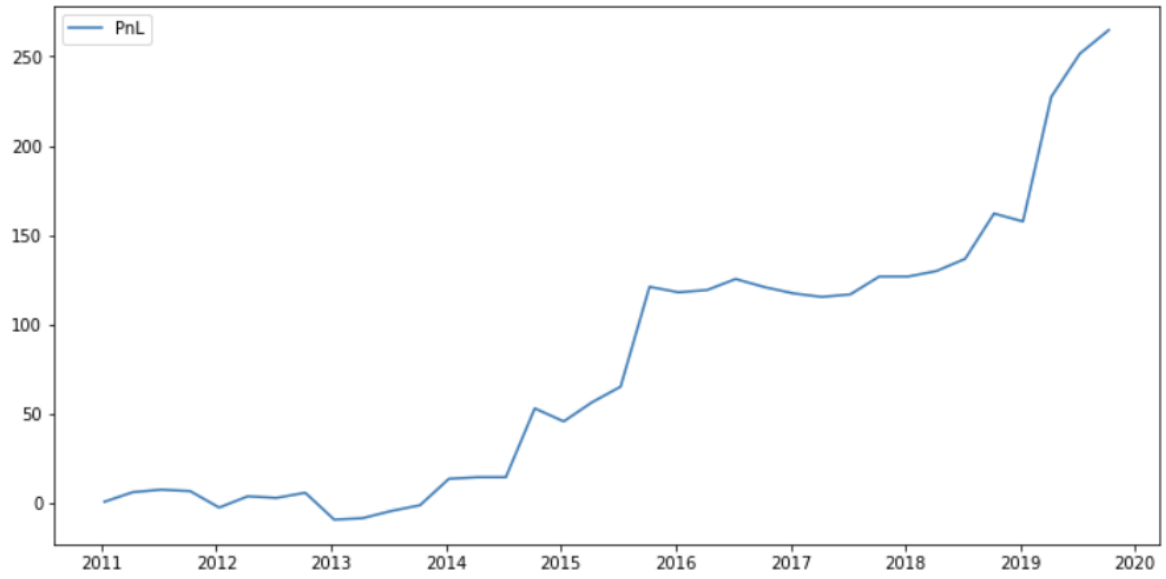
No.	Ticker	Stock	Sector
1	MSFT	Microsoft Corporation	Technology
2	JPM	JPMorgan Chase & Co.	Financial Services
3	JNJ	Johnson & Johnson	Healthcare
4	XOM	Exxon Mobil Corporation	Energy
5	WFC	Wells Fargo & Company	Financial Services
6	WMT	Walmart Inc.	Consumer Defensive
7	UNH	UnitedHealth Group Incorporated	Healthcare
8	CVX	Chevron Corporation	Energy
9	ADBE	Adobe Systems Incorporated	Technology
10	MDT	Medtronic plc	Healthcare
11	MMM	3M Company	Industrials
12	HON	Honeywell International Inc.	Industrials
13	GE	General Electric Company	Industrials
14	ABT	Abbott Laboratories	Healthcare
15	MO	Altria Group, Inc.	Consumer Defensive
16	TXN	Texas Instruments Incorporated	Technology
17	LLY	Eli Lilly and Company	Healthcare
18	KO	Coca-Cola Co	Consumer Staples
19	PEP	PepsiCo, Inc.	Consumer Staples
20	BA	Boeing Corporation	Industrial

3.2 Parameter settings

1. Length of data for calculating correlation and calibrating OU process:1 year
2. Length of every investment period: 3 month
3. Correlation threshold: -0.2, -0.1, 0 (default), 0.1, 0.2
4. Trading cost: 0 bps (default), 5bps
5. K=5 (in covariance matrix estimation)

4. Back-test results and analysis

Since at the beginning of each trading period, we need to use 1-year historical data to test cointegration and calibrate model parameters, our back-test start at 2011-01-10, we record the quarterly gains to calculate total PnL. The following figure shows the total PnL during the whole trading period under 0 trading cost condition:



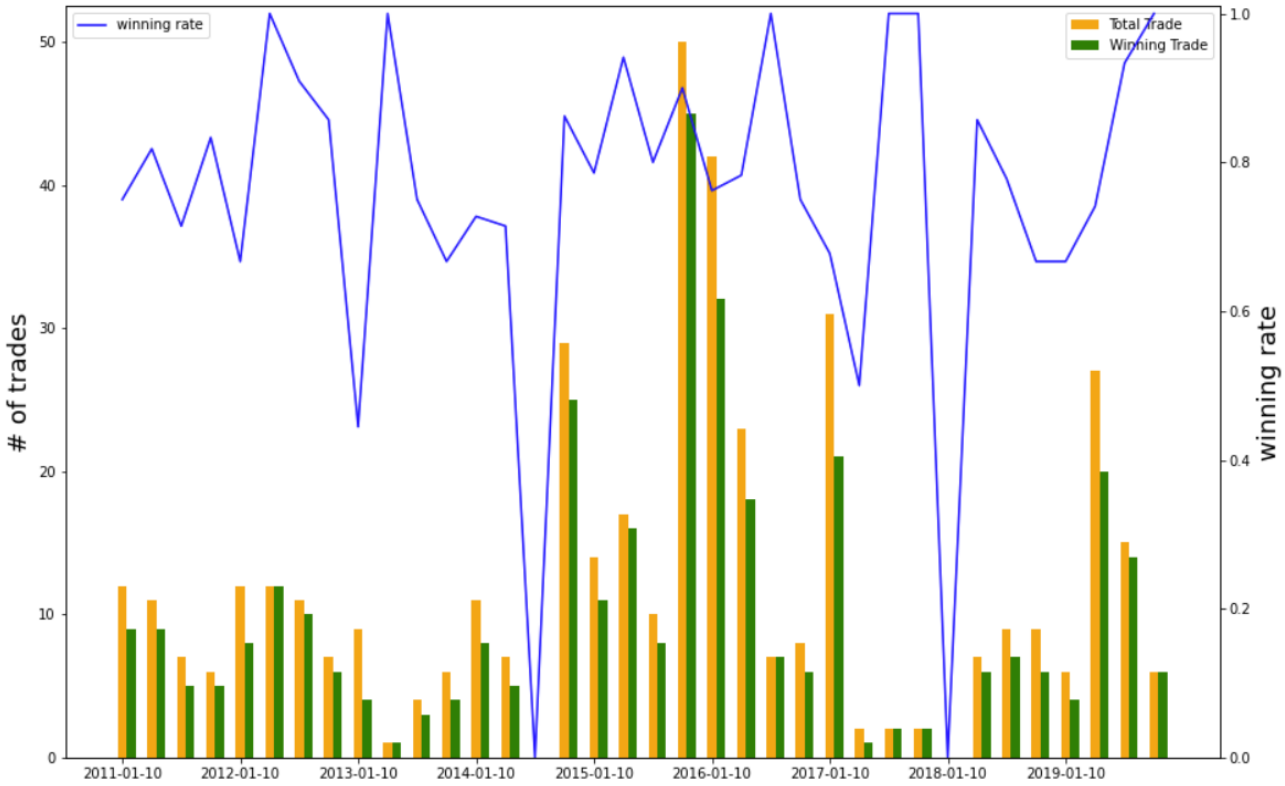
From this figure, we can see a desirable PnL curve, indicating the performance of our strategy is satisfying.

4.1 Winning rate

We define the winning rate of the strategy as below:

$$winning\ rate = \frac{\#\ trades\ with\ a\ postive\ net\ profit}{\#\ trades}$$

The following figure show how our strategy behave on winning rate:



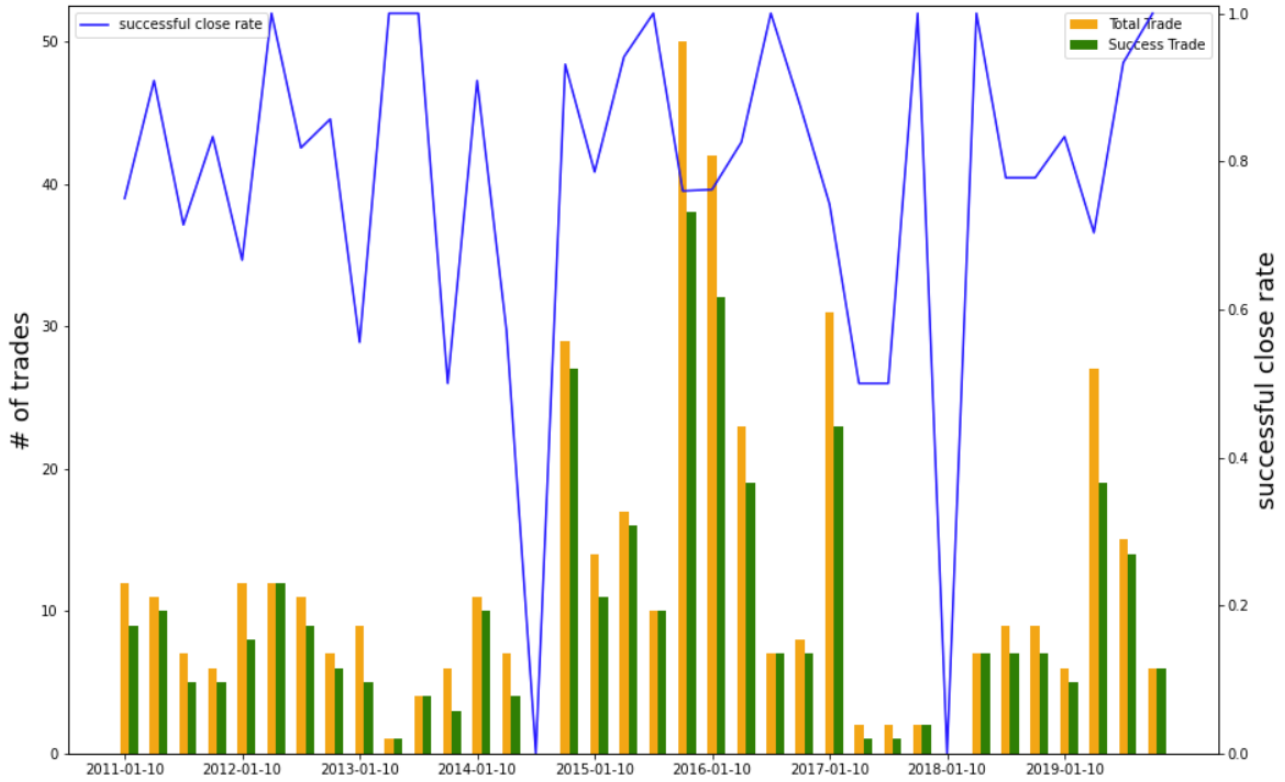
In the graph, the bar plot are total number of trades and the number of winning trades on different quarter, the line plot is trends of the winning rate. We notice that, even the rate is fluctuating, in most trading periods we get winning rate larger than 80%, and it is rarely lower than 50%. Two 0 winning rates in the graph are due to the number of total trades is 0. So from the perspective of winning rate, our strategy works well and does a great job.

4.2 Successfully close rate

We define the successful close rate as below:

$$successful\ close\ rate = \frac{\#\ trades\ closed\ due\ to\ mean\ reversion}{\#\ trades}$$

One big different between this rate and winning rate is, the successful close rate directly exams the correctness of our trading idea which mean-reversion by cointegration, sometimes the closed trade generate positive profit even it is not reversed to the mean (details are in Section 2.2), while the winning rate will consider this, the successful close rate will not consider this as a contribution from our strategy, so we think this rate can better measure the performance of the trading logic. The results is given below:



We notice that, when considering the successful winning rate, we get a similar result, indicating that our trading logic works well

More details about the closing condition of opened trades is given by the following table:

	Total	Winning	Success	Time due	Half-life	Lower Corr	Stop loss
1/10/2011	12	9	9	3	0	0	0
4/10/2011	11	9	10	0	0	1	0
7/10/2011	7	5	5	2	0	0	0
10/10/2011	6	5	5	0	1	0	0
1/10/2012	12	8	8	2	0	2	0
4/10/2012	12	12	12	0	0	0	0
7/10/2012	11	10	9	1	1	0	0
10/10/2012	7	6	6	1	0	0	0
1/10/2013	9	4	5	3	0	1	0
4/10/2013	1	1	1	0	0	0	0
7/10/2013	4	3	4	0	0	0	0
10/10/2013	6	4	3	0	0	3	0
1/10/2014	11	8	10	1	0	0	0
4/10/2014	7	5	4	3	0	0	0
7/10/2014	0	0	0	0	0	0	0
10/10/2014	29	25	27	2	0	0	0
1/10/2015	14	11	11	1	0	2	0
4/10/2015	17	16	16	0	0	1	0

7/10/2015	10	8	10	0	0	0	0
10/10/2015	50	45	38	11	0	1	0
1/10/2016	42	32	32	7	3	0	0
4/10/2016	23	18	19	4	0	0	0
7/10/2016	7	7	7	0	0	0	0
10/10/2016	8	6	7	1	0	0	0
1/10/2017	31	21	23	3	0	5	0
4/10/2017	2	1	1	1	0	0	0
7/10/2017	2	2	1	0	0	1	0
10/10/2017	2	2	2	0	0	0	0
1/10/2018	0	0	0	0	0	0	0
4/10/2018	7	6	7	0	0	0	0
7/10/2018	9	7	7	1	0	1	0
10/10/2018	9	6	7	1	0	1	0
1/10/2019	6	4	5	1	0	0	0
4/10/2019	27	20	19	2	1	5	0
7/10/2019	15	14	14	1	0	0	0
10/10/2019	6	6	6	0	0	0	0

More details of how we define each trade by closing conditions can be found in Section 2.2

4.3 Returns, volatility and Sharpe ratio

Since we always create long-short portfolio, the ways of calculating is not very straightforward and a little bit different from the usual way:

Assuming the value reaches the upper bound $\mu_x + 2\sigma_x$, and we are going to short 1 unit of asset 1 and long b units of asset 2. For calculating the net asset value (NAV) of your portfolio at the beginning, for the long side the NAV is the value of your stock holdings, and for the short side the initial NAV is zero since the cash proceeds from the sale balances the liabilities of the short holdings. Then we have:

$$NAV(t_1) = b * S_{2,t_1}$$

When we close the trade, we do opposite trade, and the NAV is equal to the initial cash proceed from the sale minus the current liability of the short position, which is the negative value of the stocks that are shorted. Then we have:

$$NAV(t_2) = b * S_{2,t_2} + S_{1,t_1} - S_{1,t_2}$$

Then we can calculate the return of this trade by:

$$R = \frac{NAV(t_2)}{NAV(t_1)} - 1$$

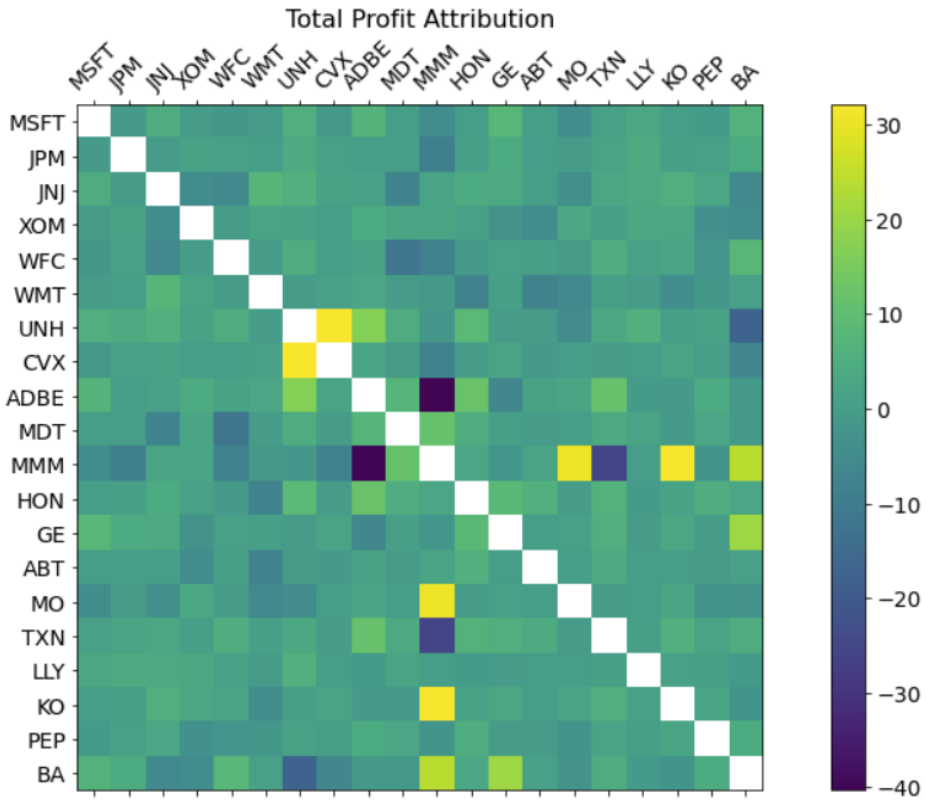
The maximum trading period is 3 month, so we set risk-free rate as 5 bps. We notice that, due to the different ways of calculation, we encounter some extreme return values (some even larger than 20), we are not sure whether these extreme values are normally caused by long-short investment or calculation error. We decide to give out the annualized means, volatilities and Sharpe ratios of returns in different quantile ranges:

	mean	std	Sharpe ratio
0<= quantile <= 1	0.009731449	3.620039901	0.002550096
0.01<= quantile <= 0.99	0.046359128	0.856618222	0.053535083
0.05<= quantile <= 0.95	0.042601929	0.065257013	0.645170951
0.1<= quantile <= 0.9	0.031314775	0.026053657	1.182742827
0.2<= quantile <= 0.8	0.047444059	0.013563772	3.460988516

From the above table, we notice all mean values of returns is larger than 0. When we consider all returns, due to the influence of extreme values, the volatility is very high, leading to a low Sharpe ratio 0.0025. After we remove the extreme values, the mean value becomes stable at around 3-4%, and the volatility is still going down when selecting less values in the middle. If the extreme values are not caused by error, I think these results indicate our strategy is not stable enough, it will give us extreme positive return and extreme negative returns as well, maybe a better constraint on stop loss can solve this problem, but we do not focus on this in the project. And relatively low Sharpe ratio is normal since we always construct a long-short arbitrage portfolio.

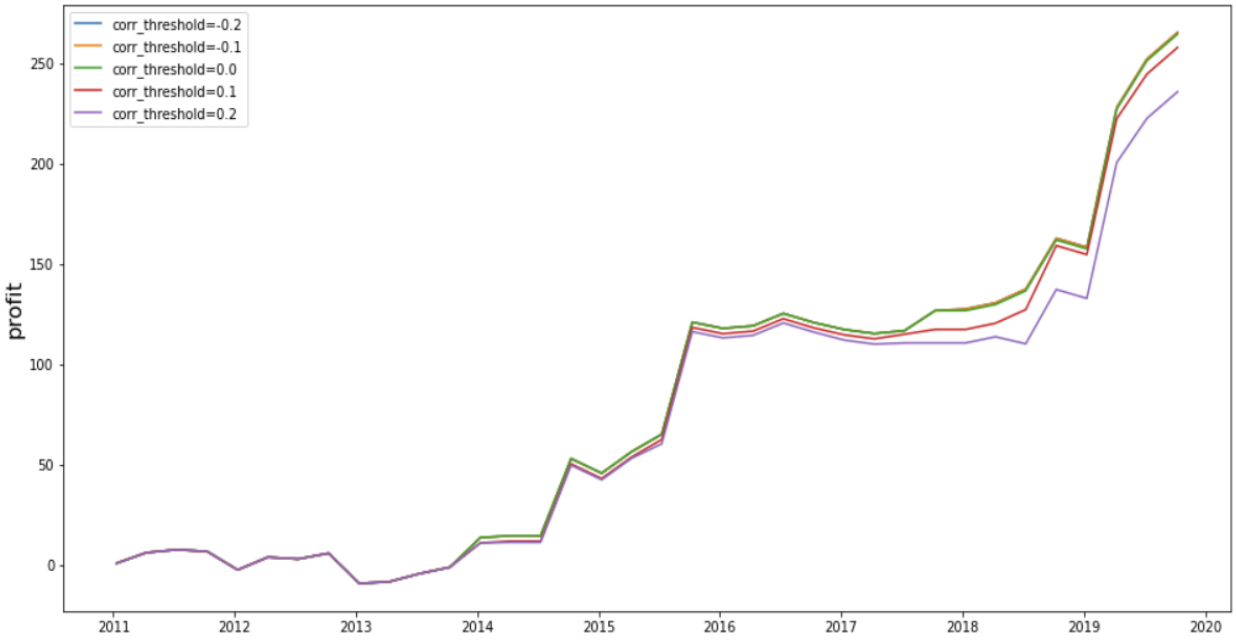
4.4 Pair contribution

We also attribute the total profit to each pair, the following figure gives out the details:



4.5 Correlation impact

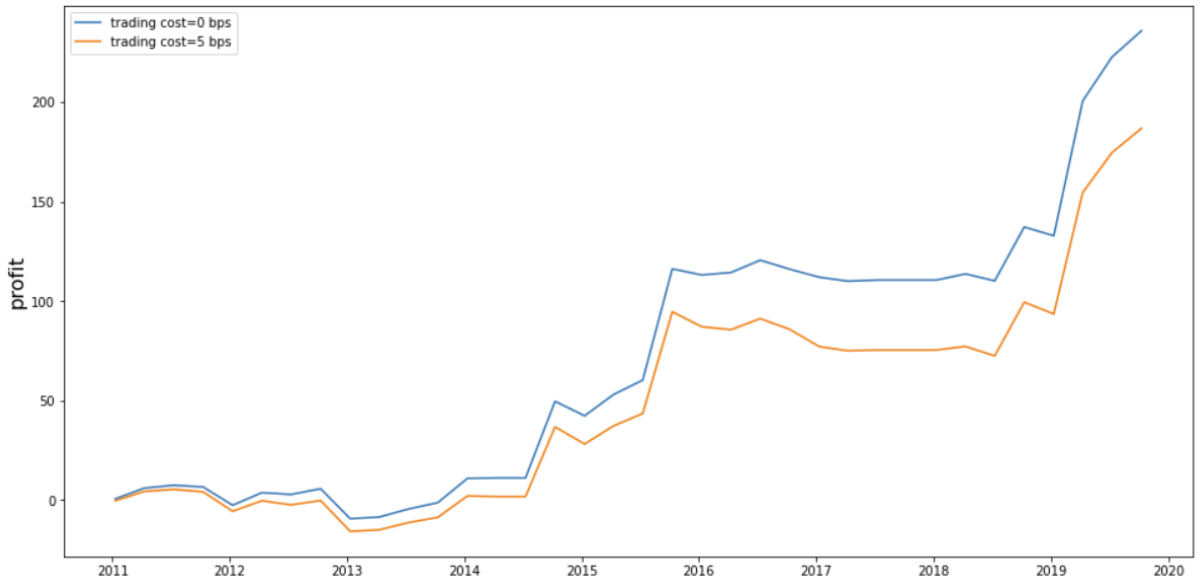
In this section we use different correlation threshold to see if it works in our strategy, the following figure clearly present the facts for the impact of correlation:



Surprisingly, the result is not as we expected. When threshold=0.0, we get the best result, when we increase the correlation threshold to 0.1 or 0.2, the total profit drops down a little bit. Although the total profits do not change too much with the thresholds, we concede that replacing risk factor model by PCA to evaluate pair correlation is not a valid idea. According to our reference, maybe a good interpretable factor model is the best way to enhance the cointegration pair trading strategy.

4.6 Trading cost impact

The trading impact given by the following figure:



We notice that, the trading cost does have great impact on our strategy profit. When trading cost= 5 bps, the profit of our trading strategy drops by over 20%, not to mention the facts that the real-trading trading cost is always higher than 5 bps and we will suffer more from this when trading frequency is higher or stock liquidity is lower.

5. Summary

In this project we develop a enhanced pair-trading strategy using Cointegration, Correlation Estimation and OU Process Calibration. Then we test our strategy on our target asset pool with 20 stock. In the 10 year back-test period, we achieve significantly positive profit from the strategy and trading logic is robust due to the high winning rate and successful close rate. And we realize the great impact of trading cost on strategy. Moreover, we find that he correlation-enhanced method does not provide us a significant improvement on the strategy performance, we attribute this problem to its low interpretability compared with a good risk factor model.

In all, we can conclude that our enhanced pair-trading strategy shows its effectiveness on our dataset by the satisfying back-test results.

6. References

1. https://commoditymodels.files.wordpress.com/2010/02/estimating-the-parameters-of-a-mean-reverting-ornstein-uhlenbeck-process1.pdf
2. Taewook Kim, Ha Young Kim. “Optimizing the Pairs-Trading Strategy Using Deep Reinforcement Learning with Trading and Stop-Loss Boundaries”
3. Haipeng Xing. “A singular stochastic control approach for optimal pairs trading with proportional transaction costs”
4. Shuo Qu, Ph.D. “Deutsche Bank Enhancing Pairs Trading News Analytics”