FRE-GY 7121 Statistical Arbitrage Final Project Report

Pair Trading Based on Kalman Filter

Author: Deyang Fan, Yu-Hsing Chang , Levi Lan

Supervisor: Daniel Totouom Tangho

Date: Oct 23, 2022

Abstract

The problem in Pair Trading is that the hedge ratio is constant in the holding period; however, the relation between two assets may not remain constant over time. In this project, we try to resolve this issue. Therefore, we implement the Kalman filter model to continuously monitor its value to adjust it accordingly. Besides, we also use rolling OLS and rolling TLS to make comparisons with Kalman Filter. Our Purpose of this project is to implement dynamic hedging ratio through Kalman filter to improve pairs trading strategy. We will compute the PnL, expected loss, expected gain, hit ratio and trading frequency to see our strategy performance.

1. Introduction

Pairs trading was first introduced in the mid-1980s and uses statistical and technical analysis to seek out potential market-neutral profits. A pairs trade strategy is based on the historical "stationary" correlation between two assets and trade when the prices move away from the mean correlation. The securities in a pairs trade must have a high positive correlation, which is the primary driver behind the strategy's profits. A pairs trade strategy is best deployed when a trader identifies a correlation discrepancy. Relying on the historical notion that the two securities will maintain a specified correlation, the pairs trade can be deployed when this correlation falters.

However, the possible issue within the area of Pair Trading is that the relation between the two assets (called hedge ratio) may not remain constant over the time, and we should be continuously monitoring its value to adjust it accordingly. Moreover, because of the noise that exists in daily prices, we should take that into account to prevent making abrupt changes in the hedge ratio that are not meaningful. Therefore, we try to resolve the issues mentioned above through this project.

To extract useful information about the dynamic hedge ratio from an observation (daily asset prices), we will leverage the Kalman filter to estimate the hedge ratio (hidden states in Kalman filter procedures).

In this paper, we will first explore a pair that can be used in pairs trading. We first fetch all the tickers in S&P500, and then do the conitegration test to explore paris having high cointegration with each other. Then, we conduct Kalman Filter to estimate the dynamic hedging ratios and construct a portfolio. Besides, we also compare Kalman Filter with rolling OLS and rolling TLS methods to see the difference of hedge ratios.

Then we are going to do pair trading on this 'stationary portfolio', i.e., long the portfolio when its value is below a certain threshold and short it when its value is above.

2. Explore Stock Pairs

2.1 Data Sample

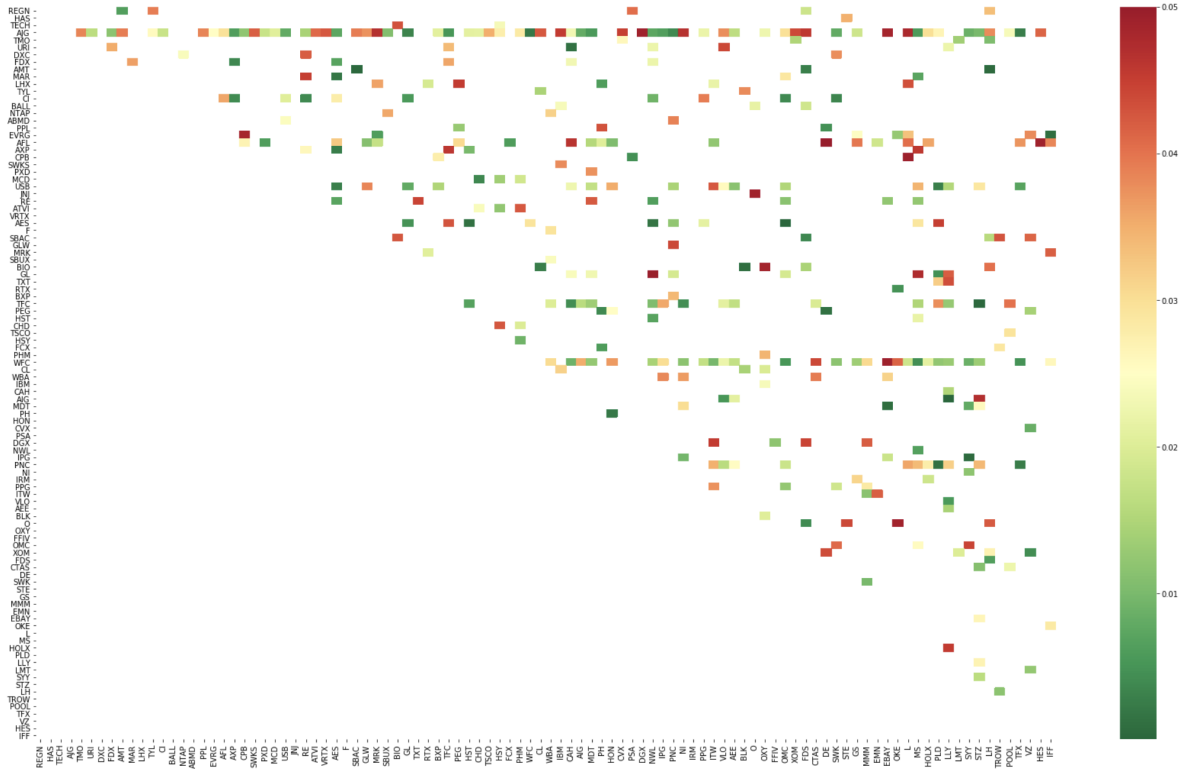
We first use stocks in S&P 500 index and fetch their historical price data from 2000 to 2014. Then, we transform the price data into log price for testing cointegration because what we care about is the stock return, which the log price represents. After doing the data cleaning and transformation, we have a total of 365 stocks with a period from 2000-2014. In addition, in order to test the different market environments, we divide this period into 3 intervals: 2000-2004, 2005-2009, and 2010-2014.

2.2 Cointegration

For our strategy, the securities in a pairs trade must have a high cointegration, which is the primary driver behind the strategy's profits. In order to explore pairs for pair trading, we need to

test cointegration between each stock pair. We set the significant level at 5% in the cointegration test, which means that a pair has high cointegration if its p-value is smaller than 5%.

We randomly choose 100 stocks for the test. There are a total of 4950(n(n-1)/2) pairs to be tested in 3 periods. Under these 3 time intervals, we select the pairs having p-values smaller than 5%. The result shows that there were 716 pairs in the period 2000-2004, 359 pairs in the period 2000-2004, 430 pairs in the period 2000-2004.The below heat map shows the p-values of the cointegration test. The darker the green point is, the higher cointegration the pair has.



Then, we select the intersection pairs in these 3 periods, meaning that all the intersection pairs can show high cointegration in different market conditions. 6 pairs are selected to construct our example pairs. For the next part, the intersection pairs will be used to estimate the dynamic hedge ratio and show our strategy performance.

	Pairs Name	P-value(2000-2004)	P-value(2005-2009)	P-value(2010-2014)
0	[AJG, DGX]	0.011129	0.049063	0.037598
1	[WFC, ITW]	0.031275	0.009654	0.045848
2	[PPG, ITW]	0.004449	0.037465	0.011682
3	[DGX, ITW]	0.023911	0.045391	0.021341
4	[AJG, TYL]	0.041160	0.025752	0.002181
5	[MDT, SYY]	0.000404	0.007979	0.011229

3. Model details

Our next step is to build a hedging portfolio using three methods—OLS, TLS, and Kalman filter—after filtering and obtaining the pairs. The spread is then traded. We will use the log price relationship between two stocks to make it compatible with the outcome of the cointegration test and to make trading the pair simpler.

Let y1t and y2t denote the prices(log-prices) of two stocks. If they are cointegrated, then the centered spread zt=y1t−γy2t - μt (i.e., with mean removed) is stationary and can be expressed as the following.

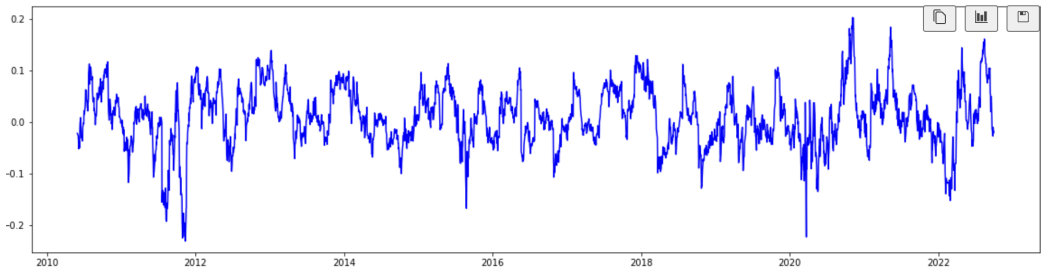
$$y_{1t} = \mu + \gamma y_{2t} + \epsilon_t$$

$$z_t = y_{1t} - \gamma_t y_{2t} - \mu_t$$

The portfolio is then built using a linear combination. The spread, whether constant or time-varying, should theoretically represent the expectation of the combined portfolio. The expectation of a centered spread should be zero, as we can assume.

$$E[z_t] = E[\epsilon] = 0$$

Therefore, as it departs from the historical center, we will enter the market and trade the spread. The portfolio we constructed is as the following:



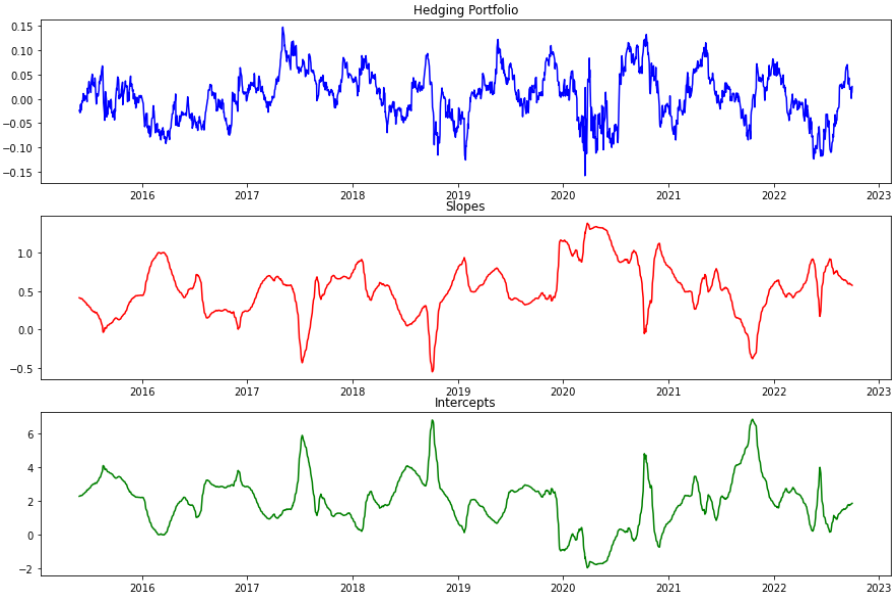
To estimate the mu and gamma parameters, we will first apply the regression and thereafter the kalman filter. Additionally, because they were time-varying, we performed regression on a rolling basis.

3.1 Rolling OLS

Rolling OLS is the first estimation technique. In order to do regression analysis on the data included within the window, we defined the look-back window T.

$$\underset{\mu_{t_0},\gamma_{t_0}}{\text{minimize}} \sum_{l=t_0-T_{\text{lookback}}+1}^{t_0} (y_{1t} - (\mu + \gamma y_{2t}))^2.$$

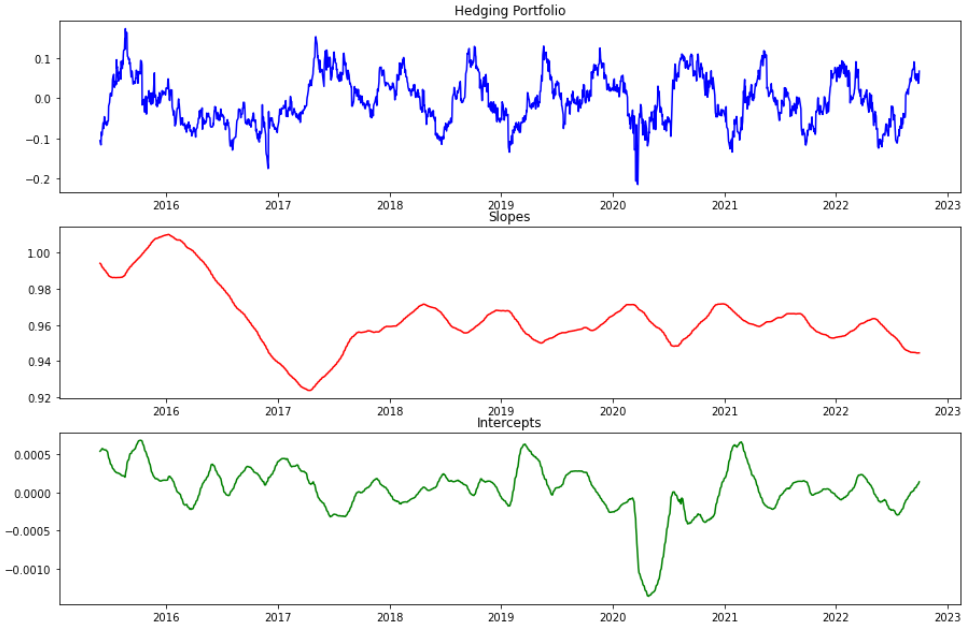
The rolling OLS is simply an OLS but performed on a rolling window basis defined by the window length T_lookback. We did the rolling analysis on the pair [‘ADI’, ‘AME’], slopes and intercepts are depicted using rolling estimates. Our log price portfolio is shown as the blue line.



It only fluctuates in relation to the mean, as is expected.

3.2 Rolling TLS

TLS (total-least-squares) will be used as the next comparison. Similar to the rolling OLS, but we use TLS as the estimation method. We once more construct our portfolio and plot the slopes and intercepts.



However, if we merely use linear regression to estimate those two parameters, the fundamental problem is that we must choose a lookback window, conduct the estimation, and presumptively believe that they will continue to have this connection in the near future. Moreover, we expect that the spread will eventually reach its long-term equilibrium. But in reality, even in the near future, they are not constants. They aren't market observables either. On the other hand, a long-term partnership may end abruptly.

In the research that follows, we model the spread using a Kalman filter. This adaptive filter estimates the slopes and intercepts dynamically while updating itself iteratively. We apply the method for calibrating the covariance matrices over the training period from the Python library pykalman.

3.3 Kalman Filter

Kalman filter is a two-step (prediction and correction) estimator algorithm. The Kalman filter is most used in tracking and control systems to provide accurate estimates in the presence of uncertainties, but it can be adapted for use in several different applications, from finance to computer vision.

The Kalman filter solves a problem when we need to estimate some unknown variable (called state), based on a set of measurements observed over time, but with noise and other inaccuracies. When trying to apply Kalman Filter to Pairs Trading, we are estimating the dynamic hedging ratios between two correlated stocks.

3.4 Kalman Equations

In the pair trading case, the hidden states are the slopes and the intercepts. And the new observations are daily prices. Basically, we make the assumption that the next day's slope is the same as the previous day. And from the kalman equation, we calculate the model predicted prices and actual price, and then based on the uncertainty measurements, we update the estimations iteratively and daily.

The state and measurements are given by the following state transition equation and observation equation.

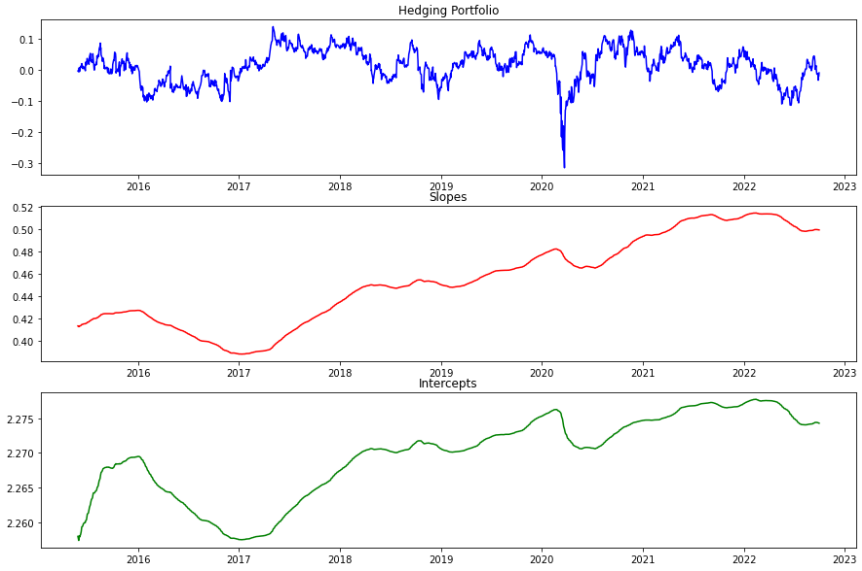
$$\boldsymbol{\alpha}_{t+1} = \mathbf{T}_t\boldsymbol{\alpha}_t + \mathbf{R}_t\boldsymbol{\eta}_t$$
$$y_{1t} = \mathbf{Z}_t\boldsymbol{\alpha}_t + \epsilon_t$$

Where

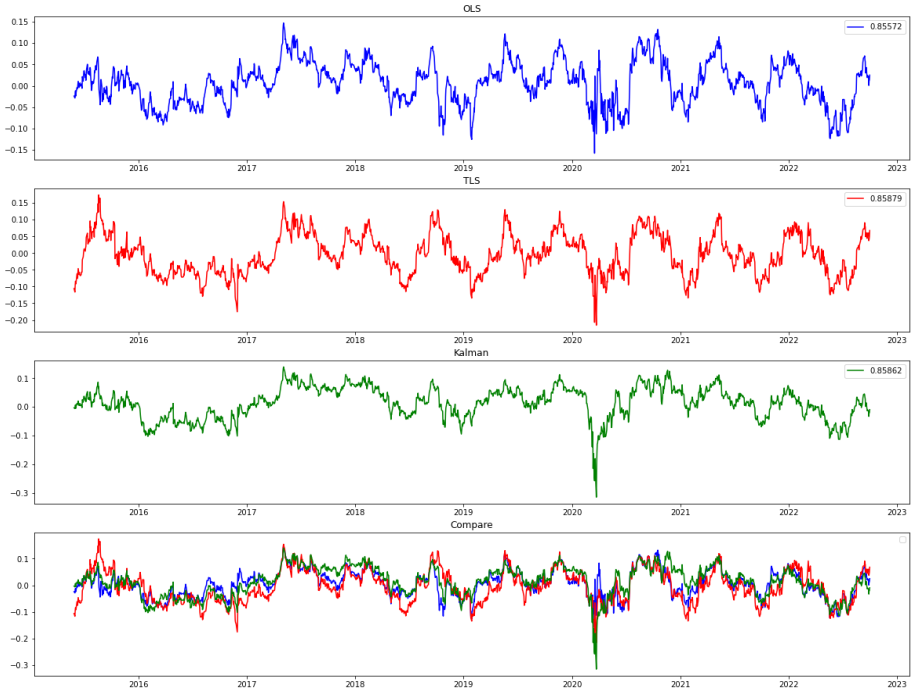
- $\boldsymbol{\alpha}_t \triangleq \begin{bmatrix} \mu_t \\ \gamma_t \end{bmatrix}$ is the hidden state (with $\boldsymbol{\alpha}_1 \sim \mathcal{N}(\mathbf{a}_1, \mathbf{P}_1)$)
- $\mathbf{T}_t \triangleq \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is the state transition matrix
- $\mathbf{R}_t \triangleq \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$ is the i.i.d. state transition noise with $\mathbf{Q}_t = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$
- $\mathbf{Z}_t \triangleq \begin{bmatrix} 1 & y_{2t} \end{bmatrix}$ is the observation coefficient matrix
- $\epsilon_t \sim \mathcal{N}(0, h_t)$ is the i.i.d. observation noise (with $h_t = h$).

Note that this equation relates the current state with the state at a previous time step plus an external (optional) control. The second equation (measurement equation) relates the state with the measurement.

We also use the previous pair as a demonstration. After applying this technique, we get our hedging ratio and intercepts and constructs the hedging portfolio. As we can see that the hedging ratio is smoother compared to the previous method. (This could be due to the fact that we assume the transition matrix to be the identity matrix.)



Here we compare the three portfolios that are constructed using different methods. Their std are generally the same, with the TLS having the slightly largest volatility.



4. Strategy Implementation

4.1 Strategy

Based on a series of central spread derived from Kalman filter, original least square and total least square, we transform into a series of standard deviations with selected window size. If standard deviation is greater than a threshold, the trading signal is -1. In this case we will short the pair, which is to short the first stock and long the second stock. On the other hand, If standard deviation is lower than a minus threshold, the trading signal is 1. In this case we will long the pair, which is to long the first stock and short the second stock.

To simplify, if the trading signal is -1, we will keep the total position as short 1 pair. For example, if the current position is long 1 pair, we will short 2 pairs. If the current position is short 1 pair, we will do nothing. If the current position is 0 pair, we will just short 1 pair. Same logic applies to the situation when the trading signal is 1. We will keep the total position as long 1 pair. For example, if the current position is long 1 pair, we will do nothing. If the current position is short 1 pair, we will long 2 pairs. If the current position is 0 pair, we will just long 1 pair.

4.2 Taking Pair of PPG and ITW and Hedge Ratio Derived from Kalman Filter as Example

4.2.1 Split Dataset

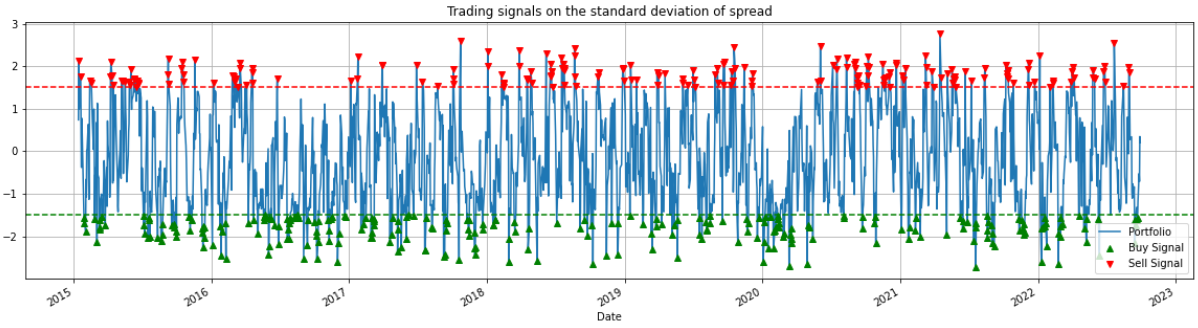
- Training Period: 2010/1/1 - 2014/12/31
- Testing Period: 2015/1/1 - 2022/9/30

4.2.2 Tuning Parameters on Training Data

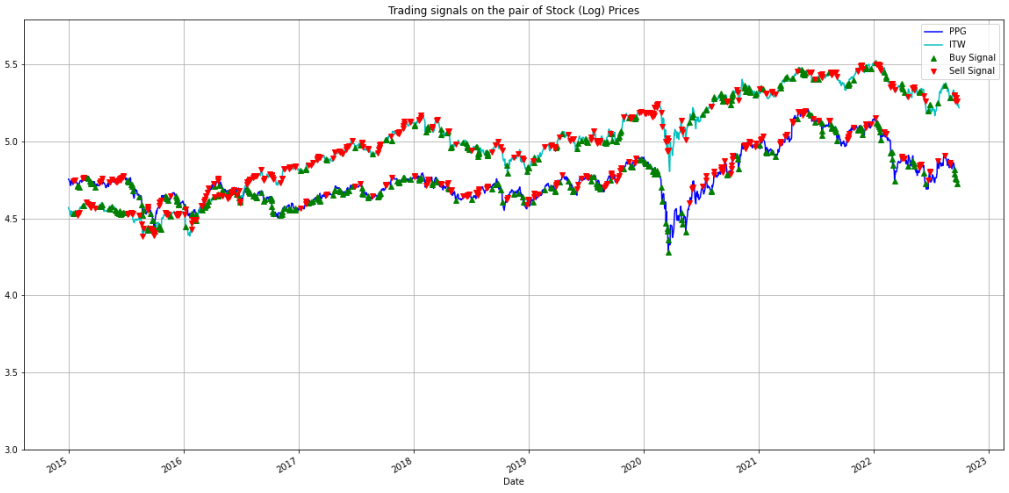
We will tune two parameters, window size to calculate standard deviation and standard deviation threshold to determine trading signal. To prevent overfitting, we choose Expected Returns as evaluation target instead of cumulative PnL. The result we get is "Best Expected Return is 0.0998% with threshold 1.5 and window size 10".

4.2.3 Strategy Performance

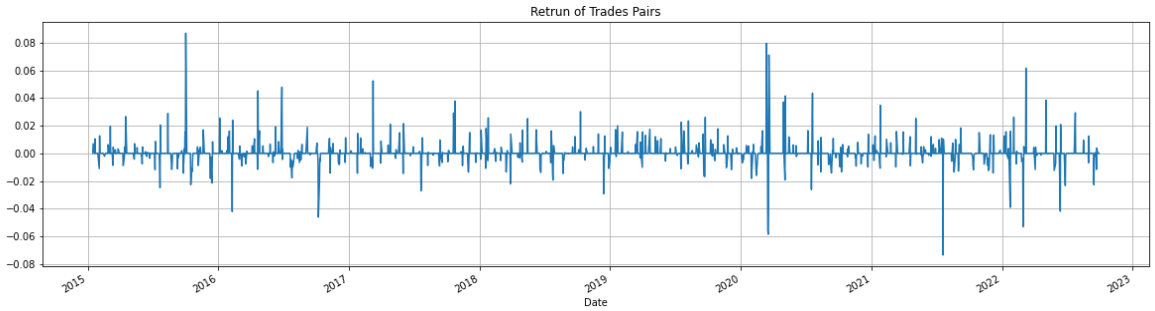
The first chart shows the trading signals on the standard deviation of spread.



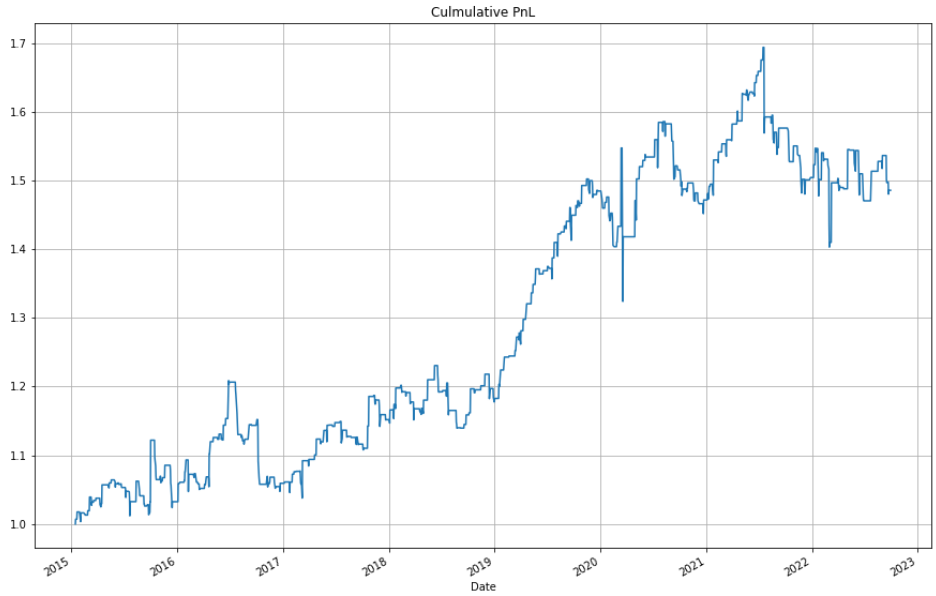
The second chart shows the trading signals on the pair of stocks. Note that the price here is log price.



The third chart shows daily returns of traded pairs.



The fourth chart shows cumulative PnL.



4.2.4 Comparison among three hedge ratios derived from Kalman Filter, original least square and total least square.

As you could see in the table, TLS provides the best performance. It reaches 100% cumulative performance, highest sharpe ratio, highest hit ratio and highest expected return. Kalman Filter is the second best, and OLS is the worst.

Hedge Ratio	Kalman Filter	OLS	TLS
Cumulative PnL	48.56%	-52.32%	100.92%
(Annualized) Sharpe Ratio	0.49	-0.73	0.81

Maximum Drawdown	-17.16%	-63.11%	-18.85%
Hit Ratio	48%	35%	50%
Conditional Expected Return	1.15%	1.55%	1.25%
Conditional Expected Loss	-0.89%	-1.20%	-0.95%
Expected Return	0.09%	-0.21%	0.16%
Trading Frequency	456	323	447

4.3 Comparison all selected pairs on three three hedge ratios derived from Kalman Filter, original least square and total least square.

In the table below, it shows the expected return for all 6 pairs under three methods. It is obvious that OLS always performs the worst. Kalman Filter and TLS perform better.

Expected Return	Kalman Filter	OLS	TLS
PPG & ITW	0.09%	-0.21%	**0.16%**
AJG & DGX	-0.12%	-0.22%	**-0.08%**
WFC & ITW	-0.14%	-0.15%	**-0.03%**
DGX & ITW	**0.003%**	-0.11%	-0.04%
AJG & TYL	**0.01%**	-0.12%	-0.19%
MDT & SYY	**0.10%**	-1.07%	0.05%

5. Further Improvement

Through the cointegration test, we can explore pairs having high cointegration from the market, and then apply them to the pair trading strategy. However, there are 95703 pairs that should be tested for stocks in S&P 500, which takes a lot of time for computing the cointegration. This problem should be considered if we have more time. Besides, we can try other methods such as distance approach to explore pairs and compare with the cointegration method.

For the strategy, there are more considerations like transaction cost, slippage to mimic the real trading process. Also, more advanced strategies could be applied. For example, do not limit ourselves to having only one maximum long or short pair. When we see two consecutive short signals, we could short the second pair to increase the exposure and maximize the profits. Lastly, we have to acknowledge not all pairs from the first part have good trading results. So far, we know that the performance is related to the cointegration. The better cointegration testing result, the better performance it will be. But there may be more factors that could influence the trading results.

6. References

- *Welch, Greg, and Gary Bishop. "An introduction to the Kalman filter." (1995): 127-132.*
- *Chan, E. (2013, May 28). Algorithmic Trading: Winning Strategies and Their Rationale (1st ed.). Wiley.*
- *Dynamic Hedge Ratio Between ETF Pairs Using the Kalman Filter | QuantStart. (n.d.). Retrieved October 18, 2022, from https://www.quantstart.com/articles/Dynamic-Hedge-Ratio-Between-ETF-Pairs-Using-the-Kalman-Filter/*
- *Build software better, together. (n.d.). GitHub. Retrieved October 18, 2022, from https://github.com/QuantConnect/Research/blob/master/Analysis/02+Kalman+Filter+Based+Pairs+Trading.ipynb*
- *GitHub - KidQuant/Pairs-Trading-With-Python. (n.d.). GitHub. Retrieved October 18, 2022, from https://github.com/KidQuant/Pairs-Trading-With-Python*
- *Cordeiro, M. (2022, January 6). Implementing Kalman Filter in Python for Pairs Trading | Analytics Vidhya. Medium. Retrieved October 19, 2022, from https://medium.com/analytics-vidhya/understanding-and-implementing-kalman-filter-in-python-for-pairs-trading-9b8986d79b2d*