Pairs Trading Strategies the Optimization in Decision-making Processes

Hungwei Chang
Jiaxin Li
Tianpei Zhu
Xiaohan Cheng
Yanlin Chen
Yuanhang Zhao

April 30, 2021

Contents

1 Introduction

An investment strategy or portfolio is considered market neutral if it seeks to avoid some form of market risk entirely, typically by hedging and derives its returns from the relationship between the performance of its long position and the performance of its short positions, regardless of whether this relationship is done on the security or portfolio level[1].

Market Value Neutrality	Share Neutrality	Sector Neutrality	Beta Neutrality	Factor Exposure Neutrality
Buying equal amounts of long and short investments so that the market risk is equal on each side of the portfolio.	Balancing a trade with an equal number of long shares and short shares.	Neutral strategy within sectors to prevent the exposure of certain sectors from being too concentrated.	Market benchmark Beta is zero. Regardless of market ups and downs, we can still obtain excess profits from a combination of long and short	Exposure on certain factors is 0.

Figure 1: Types of Market Neutrality

According to Ganapathy Vidyamurthy (2004)[2], pairs trading is a typical market neutral strategy in its most primitive form. The market neutral portfolios are constructed using just two securities, consisting of a long position in one security and a short position in the other, in a predetermined ratio. At any given time, the portfolio is associated with a quantity called the spread. This quantity is computed using the quoted prices of the two securities and forms a time series. The spread is in some ways related to the residual return component of the return already discussed. Pairs trading involves putting on positions when the spread is substantially away from its mean value, with the expectation that the spread will revert back. The positions are then reversed upon convergence.

We choose pairs trading as our main strategy. We generally divide pairs trading strategy into 4 stages, that is, pair selection, parameters calculation, optimization and risk management. In each stage mentioned above, several methods are used to increase the performance of the whole strategy. In addition, to hedge the risks from the macroeconomic factors, we set up a macroeconomic pairs trading strategy on gold to augment our portfolio return.

As a pairs trading strategy, the most important part is to find a "perfect" pair. At the pair selection stage, PCA, DBSCAN and OPTICS are used to find the clusters in a specific equity set. Within the clusters, proper stock pairs can be filtered by different criteria, such as Hurst exponent, correlation, cointegration and fundamental analysis. The output of the first stage are equity pairs.

With the pairs selected from the first stage, we then use OLS and Kalman Filter to implement the pairs trading strategy and Zscore to generate the buy or sell signal at the second stage.

[1] Douglas S. Ehrman, The handbook of pairs trading strategies using equities, options, and futures, 2006

[2] Ganapathy Vidyamurthy, Pairs Trading : Quantitative Methods and Analysis, 2004

At the third stage, to optimize the Sharpe Ratio, we use grid research to backtest different combinations of parameters. We also add necessary risk management methods like Value at Risk to control loss.

The following section is to describe our data set, which will introduce the pairs we choose. And then, we will present our strategy in details, which will include both theoretical basis and results.

2 Data Description

2.1 ETFs

An exchange traded fund (ETF) is a type of security that tracks an index, sector, commodity, or other asset, but which can be purchased or sold on a stock exchange the same as a regular stock. An ETF can be structured to track anything from the price of an individual commodity to a large and diverse collection of securities. ETFs can even be structured to track specific investment strategies.

There are a wide range of kinds of ETFs, but they all have one thing in common: they are designed to track a pre-existing index of some sort. Here are probably the most mainstream ones:

- Stock Index ETFs

These ETFs track an existing stock index and attempt to replicate its performance. For example, SPY tracks the SP 500, DIA tracks the Dow-Jones Industrial Average, the QQQ tracks the Nasdaq and IWM tracks the Russell 2000. There can be several ETFs that track the same index, since ETFs are issued by individual companies, some companies may want to track the same index as another.

- Commodity ETFs

There are also ETFs designed to follow a basket of commodities. These ETFs are very popular with investors that would like to buy oil, for example, but do not want to start trading commodity spot contracts or futures. Some ETFs in this category are OIL for Oil, GLD for Gold, and SLV for Silver.

- Volatility ETFs

Volatility ETFs are much more complicated; they are based on the "fear" of the market at any given time. Volatility ETFs are generally based off the VIX volatility index, which measures how much investors expect the market to move over the next 30 days. These are more complex financial instruments, and while anyone with a brokerage account can buy them, they are more difficult to manage and use.

In our project, we test the following stocks included in ETFs. Here is a part of the stock list, and the complete list is showed in Appendix:

[MMM, AOS, ABT, ABBV, ABMD, ACN, ATVI, ADBE, AAP, AMD, AES, AFL, A, APD, AKAM, ALK, ALB, ARE, ALXN, ALGN, ALLE, LNT, ALL, GOOGL, GOOG, MO, AMZN, AMCR, AEE, AAL, AEP, AXP, AIG, AMT, AWK, AMP, ABC, AME, AMGN, APH, ADI, ANSS, ANTM, AON, APA, AAPL, AMAT, APTV, ADM, ANET, AJG, AIZ, ...]

2.2 S&P500 Components

The S&P 500 Index is a market-capitalization-weighted index of the 500 largest publicly-traded companies in the U.S. It is not an exact list of the top 500 U.S. companies by market capitalization because there are other criteria to be included in the index. The index is widely regarded as the best gauge of large-cap U.S. equities.[3]

A company must meet the following criteria to be selected by the Index Committee and be included in the S&P 500 index:

- The company should be from the U.S.
- Its market cap must be at least $8.2 billion.
- Its shares must be highly liquid.
- At least 50% of its outstanding shares must be available for public trading.
- It must report positive earnings in the most recent quarter.
- The sum of its earnings in the previous four quarters must be positive.

2.3 CCS (Chinese Concept Stock)

China Concepts Stock is a set of stock of companies whose assets or earnings have significant activities in Mainland China. The People's Republic of China is undergoing major financial transformation, many leading mainland-based companies choose to list themselves overseas to gain access to investor capital. Currently, there are China Concepts Stock listed in several major stock exchange around the globe, which includes: SEHK, SGX, NYSE, NASDAQ, AMEX, LSE, Euronext, TSE.

Logically, the general performance of Chinese Concept Stocks will be largely affected by Sino-American relationship. For example, trade war can cause the

[3]Details about S&P 500 Components: https://www.spglobal.com/spdji/en/indices/equity/sp-500/data

stock price to fall greatly. This kind of policy risk is hard to predict. Therefore, pairs trading as a market neutral strategy provides a good way to avoid downside risk caused by foreign policies.

Here is a part of Chinese Concept Stocks, and the complete list is showed in Appendix.

[YI, VNET, QFIN, JOBS, BABA, AMBO, JG, ATHM, BIDU, BZUN, GLG, BGNE, BILI, BLCT, BRQS, BEDU, CSIQ, CBAT, CMCM, CAAS, CCRC, JRJC, CGA, HGSH, CJJD, COE, CPHI, CREG, SXTC, CXDC, CNET, CD, CLPS, CCM, DADA, DQ, ...]

3 Pairs Trading Strategy

3.1 Pair Selection

3.1.1 Dimension Reduction

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables, the principal components. Each component can be seen as representing a risk factor.

Considering that an Unsupervised Learning algorithm will be applied to these data, the number of features should not be large. In this paper, we applied PCA in the normalized return series, defined as below:

$$R_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}} \tag{1}$$

We reduce over 700 daily stock returns (2018/01/02-2021/04/15) to 50 variables while trying to keep as much variance as possible. Each of the resulting principal component can be seen as representing a risk factor, and the stocks will be clustered based on these components.

3.1.2 Clustering

a. t-SNE t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each data point a location in a two or three-dimensional map. In this paper, we will use this approach later to visualize our clusters.

Given a set of N high-dimensional objects $x_1, ..., x_N$, t-SNE first computes probabilities p_{ij} that are proportional to the similarity of objects x_i and x_j. For $i \neq j$, define:

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\Sigma_{k \neq i} exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{2}$$

Then, we define the joint probabilities $p_{i,j} = \frac{p_{j|i} + p_{i|j}}{2N}$.

t-SNE aims to learn a d-dimensional map $y_1, ..., y_N$ (with $y_i \in R^d$ that reflects the similarities p_{ij} as well as possible. To this end, it measures similarities q_{ij} between two points in the map y_i and y_j, using a very similar approach. Specifically. For i $\neq$ j, define:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\Sigma_k \Sigma_{l \neq k} (1 + \|y_k - y_l\|^2)^{-1}} \tag{3}$$

The locations of the points y_i in the map are determined by minimizing the (non-symmetric) Kullback–Leibler divergence of the distribution P from the distribution Q, that is:

$$KL(P|Q) = \Sigma_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{4}$$

The minimization of the Kullback–Leibler divergence with respect to the points y_i is performed using gradient descent. The result of this optimization is a map that reflects the similarities between the high-dimensional inputs.

b. DBSCAN The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm detects clusters of points based on their density. The reason why we can easily detect clusters of points is that there is a typical density of points within each cluster which is considerably higher than outside of the cluster. To accomplish that, two parameters need to be defined: ϵ, which specifies how close points should be to each other to be considered "neighbors", and minPts, the minimum number of points to form a cluster.

The DBSCAN execution can be described in a simplified manner as follows:

1. Find the points in the -neighborhood of every point and identify the core points with more than minPts neighbours, where minPts is a parameter to be tuned.

2. Find the connected components of core points on the neighbour graph, ignoring all non-core points.

3. Assign each non-core point to a nearby cluster if the cluster is an -neighbor, otherwise assign it to noise.

Figure 2 shows that after applying DBSCAN to clustering our SP500 stocks universe based on the 50 principal, we were able to find 13 clusters and 485 pairs to be evaluated.

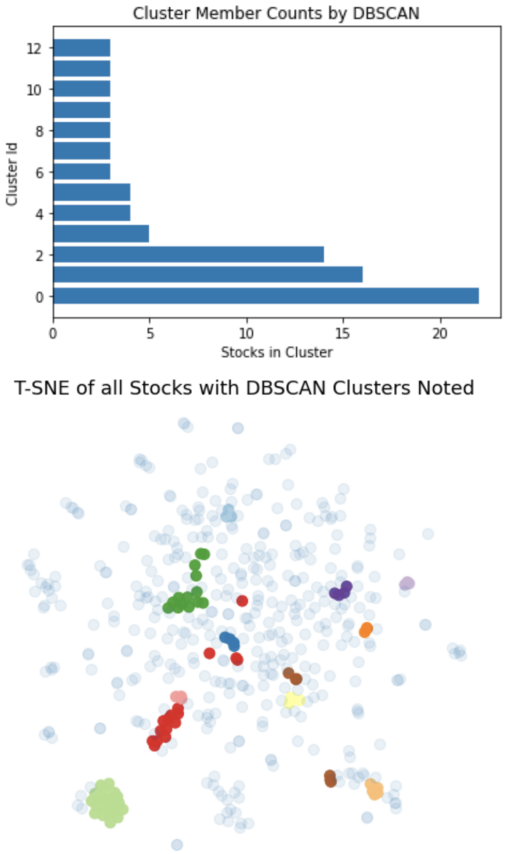

Figure 2: Clustering Result by DBSCN

c. OPTICS The OPTICS algorithm proposed by Mihael Ankerst in 1999 is based on DBSCAN, with the introduction of some important concepts that enable a varying implementation. It addresses the problem DBSCAN has when regions in space have different densities. The OPTICS algorithm is capable of detecting the most appropriate for each cluster, which specifies how close points should be to each other to be considered neighbors. In this enhanced setting, the investor is only required to specify the parameter minPts. Therefore, in this paper, we compare these two clustering algorithms with our datasets and then propose using OPTICS not just to account for varying cluster densities but also to facilitate the investor's task.

Figure 3 shows that applying OPTICS to clustering our SP500 stocks universe based on the 50 principal, we were able to find 24 clusters and 111 pairs to be evaluated.

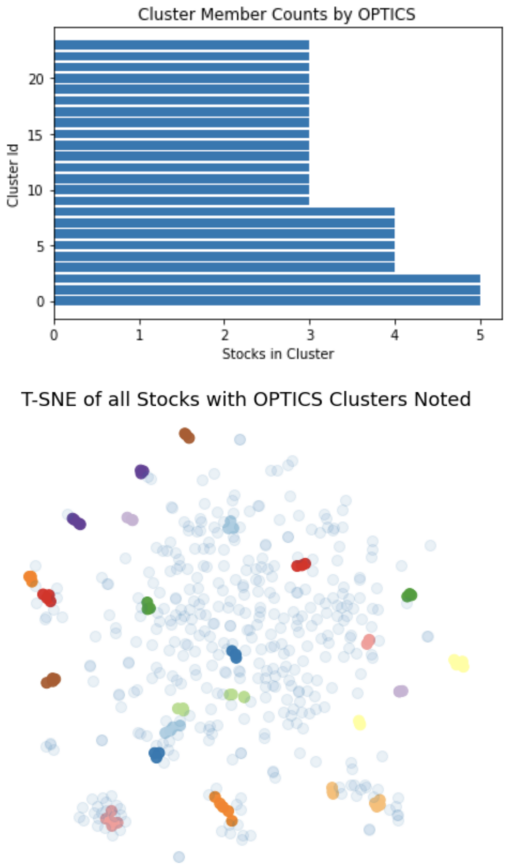

Figure 3: Clustering Result by DBSCN

In order to evaluate the integrity of the clusters, we also show the composing

price series.In this paper, we select eight clusters and represent the logarithm of the price series of each stock.Figure 4 shows the generated clusters display a tendency to group subsets of stocks from the same category while not impeding clusters containing stocks from different ones.

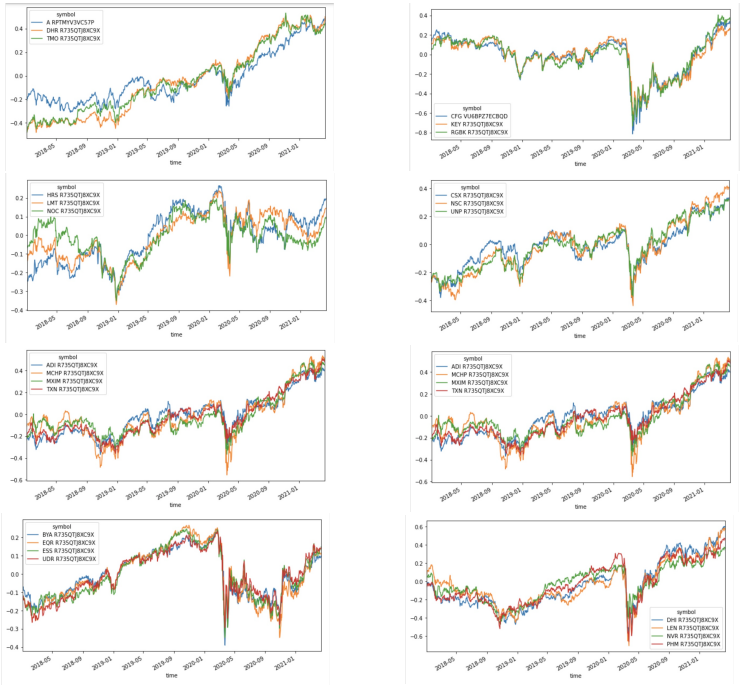

Figure 4: Price series composition of some clusters

3.1.3 Hurst Exponent

The first criterion is Hurst exponent. We require that the Hurst exponent of spread is less than 0.5, which means the spread of a selected pair is mean reversing.

The Hurst exponent is used as a measure of long-term memory of time series. It relates to the autocorrelations of the time series, and the rate at which these decrease as the lag between pairs of values increases.

The Hurst exponent is referred to as the "index of dependence" or "index of long-range dependence". It quantifies the relative tendency of a time series either to regress strongly to the mean or to cluster in a direction.A value H in the range $(0.5, 1)$ indicates a time series with long-term positive autocorrelation, meaning both that a high value in the series will probably be followed by another high value and that the values a long time into the future will also tend to be high. A value in the range $(0, 0.5)$ indicates a time series with long-term switching between high and low values in adjacent pairs, meaning that a single

high value will probably be followed by a low value and that the value after that will tend to be high, with this tendency to switch between high and low values lasting a long time into the future. A value of $H = 0.5$ can indicate a completely uncorrelated series, but in fact it is the value applicable to series for which the autocorrelations at small time lags can be positive or negative but where the absolute values of the autocorrelations decay exponentially quickly to zero.

Hurst Exponent	Time Series
$H = 0.5$	random walk
$H < 0.5$	mean reversion
$H > 0.5$	persistent trend

Table 1: Hurst Exponent Criterion

The Hurst exponent, H, is defined in terms of the asymptotic behaviour of the rescaled range as a function of the time span of a time series as follows

$$E\left[\frac{R(n)}{S(n)}\right] = Cn^H \; as \; n \to \infty \tag{5}$$

where

- $R(n)$ is the range of the first n cumulative deviations from the mean
- $S(n)$ is the series (sum) of the first n standard deviations
- $E[x]$ is the expected value
- n is the time span of the observation (number of data points in a time series)
- C is a constant.

To estimate the Hurst exponent[4], one must first estimate the dependence of the rescaled range on the time span n of observation.A time series of full length N is divided into a number of shorter time series of length $n = N, N/2, N/4, ...$ The average rescaled range is then calculated for each value of n.

For a (partial) time series of length n,

$$X = X_1, X_2, \ldots, X_n$$

the rescaled range is calculated as follows:

1. Calculate the mean;

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6}$$

[4]Implementation of Hurst exponent in Python: https://github.com/Mottl/hurst

2. Create a mean-adjusted series;

$$Y_t = X_t - m \; for \; t = 1, 2, \ldots, n \tag{7}$$

3. Calculate the cumulative deviate series Z;

$$Z_t = \sum_{i=1}^{t} Y_i \; for \; t = 1, 2, \ldots, n \tag{8}$$

4. Compute the range R;

$$R(n) = max\left(Z_1, Z_2, \ldots, Z_n\right) - min\left(Z_1, Z_2, \ldots, Z_n\right) \tag{9}$$

5. Compute the standard deviation S;

$$S(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - m\right)^2} \tag{10}$$

6. Calculate the rescaled range

$$\frac{R(n)}{S(n)} \tag{11}$$

and average over all the partial time series of length n. The Hurst exponent is estimated by fitting the power law

$$E\left[\frac{R(n)}{S(n)}\right] = Cn^H \tag{12}$$

to the data. This can be done by plotting

$$\log \frac{R(n)}{S(n)} \tag{13}$$

as a function of $\log n$, and fitting a straight line; the slope of the line gives H (a more principled approach fits the power law in a maximum-likelihood fashion)

3.1.4 Filter by criterias

a. Correlation and Cointegration With the selected pairs in clustering, by calculating the correlation and cointegration parameters between 2 stock in every pairs, we can rank them with the results of the two parameters above.

As aspect of correlation, we rank them by the value of correlation coefficient. Higher correlation means that the specific pair is more correlated. Table 2 is the results:

As aspect of cointegration, we do the cointegration test for every single pair in the selected pairs, and rank them with the p value. lower p value means that the specific pair is higher cointegrated. Table 3 is the results:

Pairs	Correlaition
DHR, TMO	0.994439
DNB, MHP	0.992253
CMS, WPH	0.984816
AEE, CMS	0.976660
ESS, UDR	0.975056

Table 2: Ranked by Correlation

Pairs	P-Value
MCHP, MXIM	0.000904
CHV, XON	0.002567
ALK, LUV	0.004910
DHR, TMO	0.007147
P, XON	0.007320

Table 3: Ranked by P-Value in Cointegration Test

b. Fundamental Data After filtering out the stocks using PCA, different methods of clustering, and hurst, we try to analyze the fundamental side of the data. By fundamental data, we refer to the trading microstructure data such as prices, volumes, dollar volumes. Other aspects of the fundamental data we consider are the basic industry-level or company-level characteristics. Also, we incorporate the data from the financial statements into our analysis.

First, we use the metric $DollarVolume$ to filter out stocks with similar trading patterns. Dollar Volume is a widely-used metric to evaluate the liquidity of the shares.

$$DollarVolume = Price \times TradingVolume \tag{14}$$

Usually, stocks with higher $DollarVolume$ tend to have a tighter bid-ask spread[5]. Since we implement the pairs trading strategy and we incur two-sided trades on every entry and exit position, we emphasize on the minimization of the bid-ask spread. By controlling the $DollarVolume$, we can keep track of the transaction costs.

Second, we account for the fundamental sectors among different stocks for our stock paris target. After clustering and more technical-based analyzing methods to select stocks, we expect to see some stock pairs consisting of stocks from very different industries. However, different sectors such as consumer cyclical and technology are fundamentally different and might not be an ideal pair when we assume the stocks are inherently similar in pairs trading.

Hence, we analyze the fundamental sectors for our selected pairs. The sector-wide classification is defined[6] as follows (Table 4). On QuantConnect, sector-level is not the only classification method to characterize the fundamentals of stocks. In fact, sector-level is the broadest category among all. We employ

[5]https://www.investopedia.com/terms/d/dollar-volume-liquidity.asp
[6]https://www.quantconnect.com/docs/data-library/fundamentals

Sector Name	Sector Code
MorningstarSectorCode.BasicMaterials	101
MorningstarSectorCode.ConsumerCyclical	102
MorningstarSectorCode.FinancialServices	103
MorningstarSectorCode.RealEstate	104
MorningstarSectorCode.ConsumerDefensive	205
MorningstarSectorCode.Healthcare	206
MorningstarSectorCode.Utilities	207
MorningstarSectorCode.CommunicationServices	308
MorningstarSectorCode.Energy	309
MorningstarSectorCode.Industrials	310
MorningstarSectorCode.Technology	311

Table 4: Sector Classification

the sector-level classification instead of a finer-industry level classification because we try to preserve the most pairs that satisfy the minimum fundamental similarity. If we define a relatively narrow universe selection criteria, we may miss out on many potential trading opportunities.

With the above framework, we implement the filtering methods on QuantConnect. From the period of April 1, 2021 to April 15, 2021, we filter out the stocks with the top 500 highest dollar volume each day (there are 10 trading days in this period) by the coarse selection method. At the same time, we use the fine selection to filter the sector. Hence we can log the stock symbol with the sector code into a text file. After collecting the top 500 highest dollar volume stocks for 10 days, we grouped the stocks into sectors so we can briefly understand what types of stocks are more likely to be traded during this short period.

Symbol	Number of Days on the top 500
KLAC R735QTJ8XC9X	10
CSCO R735QTJ8XC9X	10
EBAY REC37LGZ79T1	10
PLTR XIAKBH8EIMHX	10
PINS X3RPXTZRW09X	10

Table 5: Company Classification

Figure 5 shows that among all the top 500 dollar volume stocks in this period, stocks from certain sectors are more likely to show up in the top 500 dollar volume trading list. For example, stocks from industry code 311 and 102 (technology and consumer cyclical, respectively) are most likely to be on the top 500 list.

Other than the sector-wise analysis, we also analyze the same 10 days trading period by the company-level (Table 5). For simplicity, only the first 5 rows are shown here in Table 5. Then, We calculate the number that each company

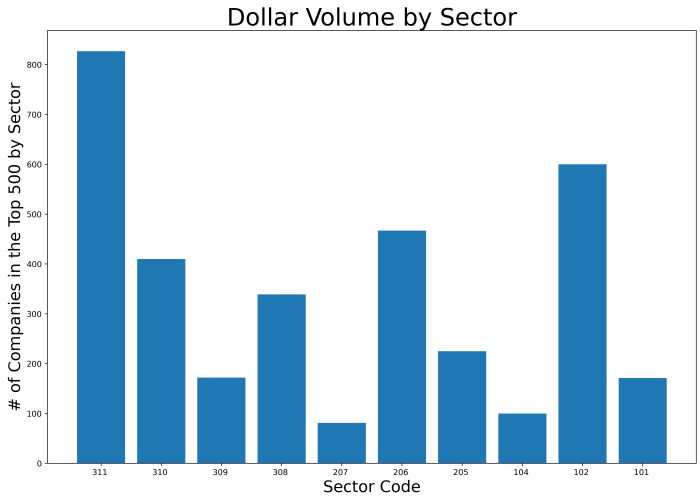

Figure 5: Dollar Volume by Sector During the Period

appeared on the top 500 dollar volume list and find the following. By definition, every company should at least appear once and at most ten times in this 10 days period, because we collect the top 500 dollar volume stocks every day during this period. There are 540 companies who have appeared on the top 500 list at least once in this short period, with 102 companies that just appeared on the list once and 220 companies that have appeared 10 times (every day in this period). Hence, we may conclude that for most of the time, stocks with high dollar volume usually preserve liquidity during a short period of time.

Another important aspect of the fundamental data is the information on financial statements. However, because of the limitations of financial data, we cannot obtain the company-level enterprise value, sales number, forecast growth, or margin rate on QuantConnect platform. However, we still collect the data from other sources and consider the EV-based multiple to evaluate the fundamental strength of a company. Enterprise value (EV) is calculated by market value data as equation 15.

$$EV = MV_{equity} + MV_{debt} - Cash \tag{15}$$

For future work, we expect to include the financial ratios measured by market values such as EV/ Sales (enterprise value divided by sales revenue) or P/E ratio to help filter the stocks. For example, one possible extension is to combine financial ratios with the dividend discount model.

For research purposes, we test this idea by collecting the company-level financial earnings data from Yahoo Finance. Using the sample from the components of the $S\&P500$ index, we collect the EV/ Sales, growth rate in the last

quarter, β, and the margin rate for every stock in the index. With the DCF, the EV/ Sales ratio can be interpreted as follows (Damodaran, 2012).

The discounted cash flow model (DCF) can be written as equation 16 with the on-going assumption of an enterprise.

$$V_0 = \frac{FCF_1}{1+r} + \frac{FCF_2}{(1+r)^2} + \frac{FCF_3}{(1+r)^3} + ... \tag{16}$$

In particular, the free cash flow (FCF) in each period is determined by company's earnings and dividend policy. We consider the notation of RIR = Reinvestment Rate, g = growth, r = cost of capital, τ = tax rate,
M = Margin Rate, g_n = growth rate after n years forever.

$$FCF = EBIT \times (1-\tau) \times (1-RIR) \tag{17}$$

If we substitute equation 17 into equation 16 (Damodaran, 2012), we can obtain the fundamental components (RHS in equation 18) of the earnings multiple EV/ Sales.

$$EV/Sales = \tau M \times \frac{(1-RIR)(1+g)(1-\frac{(1+g)^n}{(1+r)^n})}{r-g} + \frac{(1-RIR)(1+g)^n(1+g_n)}{(r-g_n)(1+r)^n} \tag{18}$$

Hence,

$$EV/Sales = f(\tau, M, RIR, r, g) \tag{19}$$

To test the idea of using the earnings multiple with DCF, we regress the EV/ Sales multiple on Margin, β (a proxy of the cost of capital r), and growth rate.

	coef	**std err**	**t**	**P-value**	**[0.025**	**0.975]**
Intercept	5.6890***	0.719	7.917	0.000	4.276	7.102
Margin	14.5466***	1.643	8.852	0.000	11.316	17.777
beta	-1.5736***	0.576	-2.733	0.007	-2.706	-0.442
g	-0.0501	0.048	-1.053	0.293	-0.144	0.043

Table 6: Coefficients of the Margin, Beta, Growth

$$EV/Sales = 5.69 + 14.55M + -1.57\beta - 0.05g \tag{20}$$

We find the coefficients on both the margin and beta are significant as the first column in Table 6 shows, but the growth rate is not significant. One possible reason is the growth rate we use in the regression is the historical growth rate, while usually practitioners use the forecasted growth rate in the DCF model. In this regression, we show that it is possible to use the component of earnings multiple to predict EV/ Sales. In the future, if we have access to a more comprehensive fundamental data set, we can predict the $EV/Sales$ of a company from $\tau, Margin, RIR, r, g$ and use it as a filtering method.

For our project, we implement the fundamental analysis described above after clustering. In other words, fundamental analysis serves as a final status check criteria. The advantage of performing the fundamental analysis in the last step of the filtering process is we allow the possibility to generate creative pairs from different sectors, and then we can decide whether the pairs should be implemented. The steps of the filtering process are described as follows.

For example, on April 26, 2021, we first obtained 106 pairs from clustering or other methods. Using the 106 pairs as the basis, we analyze whether each stock in the 106 pairs has been traded as a top 500 dollar volume stock in the testing period (April 1, 2021 to April 15, 2021). The stock has to be ranked as the top 500 dollar volume stocks at least once in this period to meet our criteria of liquidity. Then, we examine the sector of the stock. Interestingly, for those stocks that meet our criteria of liquidity, we seldom encountered the situation when stocks in a given pair from clustering are from different sectors. We infer that for the highly traded stocks, clustering itself almost ensures that the stocks in a given pair are from the same sector.

Lastly, we employ our experimental idea to apply financial ratios with the DCF. We compare the fitted value of EV/ Sales from equation 20 and the observed value of EV/ Sales in our data. If the observed value is less than the fitted value, we treat the stock as undervalued[7] because our regression model predicts that it may be greater on the scale of others. We assume the stocks with undervalued EV/ Sales (shown in the last column of Table 7) have a higher probability to be traded by the market when the market is rather efficient (stocks considered here has demonstrated sufficient liquidity), and we might detect pairs with potential higher converging speed.

In summary, we check the fundamental status in the last step of our filtering process. We identify the liquidity (those that have been on the top 500 dollar volume at least once in the period) and sector status. In our research, we find that if both stocks in the pair are highly liquid, clustering is likely to select stocks from the same sector. In the future, we expect to employ our experimental idea to analyze financial ratios with DCF models.

Old Index[8]	Symbol	Num[9]	Sector	Sector String	Value Status
38	AEP	4	207	Utilities	undervalued
39	CMS	2	207	Utilities	undervalued
44	AEP	4	207	Utilities	undervalued
45	ETR	1	207	Utilities	undervalued
80	CMS	2	207	Utilities	undervalued
81	ETR	1	207	Utilities	undervalued
208	DHI	10	102	ConsumerCyclical	undervalued
209	LEN	6	102	ConsumerCyclical	undervalued

Table 7: Selected Pairs from Fundamental Persepective

[7]Undervalued status is shown in the last column of Table 7.

3.1.5 Pairs Selected

Equity 1	Equity 2
DHR	TMO
DNB	MHP
CMS	WPH
NSC	UNP
CFG	RGBK

Table 8: S&P 500 Pairs Selected

Equity 1	Equity 2
VOO	VV
VEA	VXUS

Table 9: ETF Pairs Selected

Equity 1	Equity 2
PTR	SHI
BABA	JD
HNP	HOLI

Table 10: Chinese Concept Stocks Pairs Selected

[8]Two consecutive numbers indicate a pair. For example, no. 38 and no. 39 are a pair from our clustering methods.

[9]The number in this column indicates how many times the stock has appeared on the top 500 dollar volume list in this period. By definition, the maximum number is 10, and the minimum number is 1.

3.2 Trading Strategy: Parameters

3.2.1 OLS

For a equity pair P and Q, we assume a regression slope β (also called hedge ratio), regression residual ϵ, the standard deviation of ϵ σ and initial capital for a certain pair C. The linear regression equation relates to the log prices is as follows:

$$\log P_t = \beta \log Q_t + \alpha + \epsilon \tag{21}$$

$$zscore = \frac{\epsilon}{\sigma} \tag{22}$$

Here we first use OLS to estimate all the parameters. The portfolio of these two stocks will have absolute market value weights of

$$W_{tP} = \frac{1}{1+\beta} \; for \; P_t \tag{23}$$

$$W_{tQ} = \frac{\beta}{1+\beta} \; for \; Q_t \tag{24}$$

Weights of P and Q will always be the opposite because we generally long one and short the other. The trading strategy is as follows:

- Every trading day t, if no open position and $zscore > entry\ threshold$, short the spread, that is, short $\frac{W_{tP} \times C}{P_t}$ shares of P, and long $\frac{W_{tQ} \times C}{Q_t}$ shares of Q.
- Every trading day t, if no open position and $zscore < entry\ threshold$, long the spread, that is, long $\frac{W_{tP} \times C}{P_t}$ shares of P, and short $\frac{W_{tQ} \times C}{Q_t}$ shares of Q.
- If $zscore > exit\ threshold$ and we already longed the spread, liquidate all the positions.
- If $zscore < exit\ threshold$ and we already shorted the spread, liquidate all the positions.

3.2.2 Kalman Filter

Another method to estimate the hedge ratio dynamically is the Kalman Filter.[10]

The Kalman Filter is a unsupervised algorithm for tracking a single object in a continuous state space. Given a sequence of noisy measurements, the Kalman Filter is able to recover the "true state" of the underling object being tracked.

The advantages of Kalman Filter are:

- No need to provide labeled training data
- Ability to handle noisy observations

[10] Documentation of Kalman Filter package pykalman: https://pykalman.github.io

The disadvantages are:

- Computational complexity is cubic in the size of the state space
- Parameter optimization is non-convex and can thus only find local optima
- Inability to cope with non-Gaussian noise

As figure 6 shows, Kalman Filter can be viewed as a three-step process of prediction, observation, and correction[11].

$$corrected\ state = predicted\ state + k(observation\ -\ prediction) \tag{25}$$

(observation - prediction) is called the observation innovation. A fraction of the observation innovation is added as a correction to the predicted state. The value of this fraction k is known as the Kalman Gain. k is decided such that the corrected state has the least amount of error variance associated with it. k is indeed optimal in the case where the mathematical models of state and observation are both linear and the errors are drawn from independent Gaussian distributions.

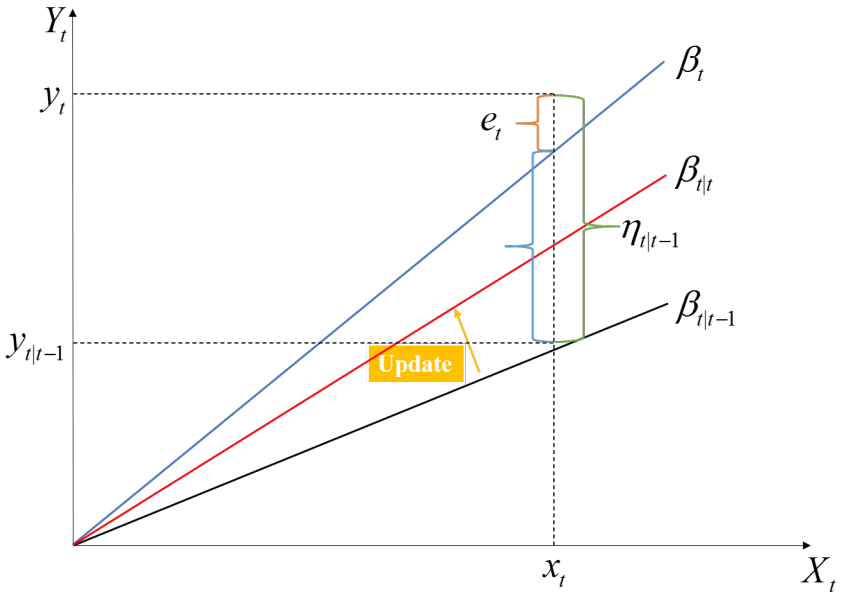

Figure 6: Kalman Filter as a Three-step Process of Prediction

Unlike most other algorithms, the Kalman Filter is traditionally used with parameters already given. The KalmanFilter class can thus be initialized with

[11] Implementation of Kalman Filter in Python: https://github.com/pykalman/pykalman

any subset of the usual model parameters and used without fitting. Sensible defaults values are given for all unspecified parameters (zeros for all 1-dimensional arrays and identity matrices for all 2-dimensional arrays).

A Kalman Filter is fully specified by its initial conditions (initial state mean and initial state covariance), its transition parameters (transition matrices, transition offsets, transition covariance), and its observation parameters (observation matrices, observation offsets, observation covariance). These parameters define a probabilistic model from which the unobserved states and observed measurements are assumed to be sampled from.

To make notation concise, we refer to the hidden states as x_t, the measurements as z_t, and the parameters of the KalmanFilter class as follows,

Parameter Name	Notation
initial state mean	μ_0
initial state covariance	Σ_0
transition matrices	A
transition offsets	b
observation matrices	H
observation offsets	d
observation covariance	R

Table 11: Parameters of the Kalman Filter

In words, the Linear-Gaussian model assumes that for all time steps $t = 0, \ldots, T-1$ (here, T is the number of time steps),

- x_0 is distributed according to a Gaussian distribution
- x_{t+1} is an affine transformation of x_t and additive Gaussian noise
- z_t is an affine transformation of x_t and additive Gaussian noise

These assumptions imply that that x_t is always a Gaussian distribution, even when z_t is observed. If this is the case, the distribution of $x_t|z_{1:t}$ and $x_t|z_{1:T-1}$ are completely specified by the parameters of the Gaussian distribution, namely its mean and covariance. The Kalman Filter calculates these values, respectively.

Formally, the Linear-Gaussian Model assumes that states and measurements are generated in the following way,

$$x_0 \sim \mathcal{N}(\mu_0, \Sigma_0) \tag{26}$$

$$x_{t+1} = A_t x_t + b_t + \epsilon_t \tag{27}$$

$$y_t = H_t x_t + d_t + \xi_t \tag{28}$$

$$\epsilon_t \sim \mathcal{N}(0, Q) \tag{29}$$

$$\xi_t \sim \mathcal{N}(0, R) \tag{30}$$

The Gaussian distribution is characterized by its single mode and exponentially decreasing tails, meaning that the Kalman Filter work best if one is able to guess fairly well the vicinity of the next state given the present, but cannot say exactly where it will be. On the other hand, these methods will fail if there are multiple, disconnected areas where the next state could be.

This process is stated mathematically as follows:

1. Evaluate $\hat{X}_{t|t-1}$ and $\hat{P}_{t|t-1}$ using the state equation.

$$\hat{X}_{t|t-1} = A\hat{X}_{t-1|t-1} \tag{31}$$

$$\hat{P}_{t|t-1} = A\hat{P}_{t-1|t-1}A^T \tag{32}$$

2. Find the observation Y_t and R by observing the system. Note we have the matrix H defined as follows:

$$Y_t = HX_t + v_t \tag{33}$$

3. Compute the Kalman gain K_t.

$$K_t = \hat{P}_t H^T (H\hat{P}_t H^T + R)^{-1} \tag{34}$$

4. Evaluate $\hat{X}_{t|t}$ given by

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + K_t(Y_t - H\hat{X}_{t|t-1}) \tag{35}$$

5. Evaluate $\hat{P}_{t|t}$

3.3 Risk Management

a. Margin Constraints: Since the cash generated from the short position can be used to purchase long stock position, pairs trading strategy requires very little cash or capital in theory. In this paper, we set the margin of our account to be 50%, and enter when $(n_P P + n_Q Q)m <= E$, where we long P and short Q, and the total account equity is E.

b. Stop-loss Limit: In this paper, we use two stop-loss approaches. One is that if an upper (lower) threshold is crossed($|Z| \geq 4$) by our spread of a pair before reverting back to the mean spread after opening a position, then we close the position of our trade and incur a loss on that pair. For another, if the total loss on one pair is great than the stop-loss limit (0.15) of the initial capital on that pair, we close the position of it and stop trading that pair any more. In both cases we avoid a rampant loss due to the spread never reverting back to the mean.

c. Value-at-Risk: VaR is a probabilistic measure and is defined as the worst loss over a target horizon given a level of confidence such that $P[L \geq VaR] = 1-\alpha$. The equation states that if the confidence is. e.g. 95% then the probability of exceeding that loss is 5% or less. VaR can be computed by either a parametric approach where the parameters are assumed to be known, or a non-parametric method based on historical data or Monte Carlo simulations. In this paper, we use historical data with a 30 days' look-back period to calculate the daily VaR under a 95% confidence level. We impose a limit of VaR with no more than 30,000 to reduce our position sizes.

3.4 Performance Measurement

a. Sharpe Ratio:

$$SR = \frac{R - R_f}{\sigma_i} \tag{36}$$

where R_f is riks-free rate and σ is volatility or standard deviation of returns. The Sharpe ratio helps us determine whether a portfolio's excess returns are due to investment decisions that do not incur too much risk. A Sharpe ratio greater than one is considered acceptable.

b. Total Return:

$$TR = \frac{V_t}{V_0} - 1 \tag{37}$$

where V_t is the value of our portfolio at time t and V_0 is the initial value of our portfolio

c. Max Drawdown:

$$MaxDD_t = max_{u \in [0,t]}(M_u - S_u) \tag{38}$$

where S_u is the value of an asset at time t and M_t is the running maximum given by $max_{u \in [0,t]} S_u$. The maximum drawdown is defined as the largest drop of the asset price from the running maximum up to time t.

3.5 Optimization on Parameters

Our optimization method is grid research. We backtest all the combinations of parameters to maximize the Sharpe Ratio with in-sample data from 2018/01/02 to 2021/04/15. We use QuantConnect to do the grid research. The detail is as follows:

Parameter	Min	Max	Step Size
enter	1	3	0.5
exit	0	0.5	0.1

Table 12: Grid Research Detail

Figure 7: Optimization on Parameters

The optimized enter/exit boundary for OLS regression is 2.0/0.1, and the highest Sharpe ratio is 1.427. The optimized enter/exit boundary for Kalman filter is 1.0/0.5, and the max Sharpe ratio is 1.334. In this case, OLS leads to a slightly higher back-testing Sharpe ratio via the optimization on parameters.

Method	Sharpe Ratio	Enter	Exit	Return	MaxDD
OLS	1.427	2.0	0.1	16.479%	2.9%
Kalman	1.334	1.0	0.5	25.314%	4.0%

Table 13: In Sample: Optimal Back-testing Statistics on Selected SP500 Pairs

Figure 8: In Sample: Optimal Back-testing Result with OLS

Then we applied optimized parameters obtained in our in-sample period to our out-of-sample period - from 2016/01/02 to 2017/12/29. As Table14 shows, during the out-of-sample period, the Sharpe ratio obtained from backtest with OLS is 0.549, and Sharpe ratio obtained by Kalman is 1.485. We can see that in the out-of-sample period, strategy with Kalman filter performs much better than that with OLS, and it can help us bring a more steady return.

which is lower than that of the in-sample period. The reason is that our clustering models are trained via a dataset beginning with 2018/01/02, selected pairs were not necessarily related well before that date.

Method	Sharpe Ratio	Enter	Exit	Return	MaxDD
OLS	0.549	2.0	0.1	3.345%	3.1%
Kalman	1.485	1.0	0.5	10.420%	2.9%

Table 14: Out of Sample: Back-testing Statistics with Optimized Parameters

Figure 9: In Sample: Optimal Back-testing Result with Kalman Filter

Figure 10: Out of Sample: Back-testing Result with OLS and Optimized Parameters

Figure 11: Out of Sample: Back-testing Result with Kalman and Optimized Parameters

We also plot the 6-month and 12-month portfolio rolling beta[12]. 12-month rolling portfolio beta is almost within $(-0.005, 0.005)$ and we believe it is beta neutral, which means our strategy will not be strongly affected by market ups and downs. The long-short exposure also shows share neutrality.

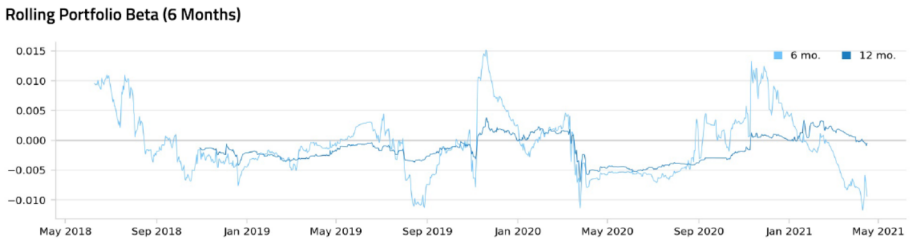

Figure 12: Rolling Portfolio Beta

This strategy also performs well during Covid-19 Pandemic 2020. Our profit are not affected by large market benchmark drawdown.

[12]The completed report can be found here: https://github.com/cxh1996108/MATH590-02/tree/main/Project

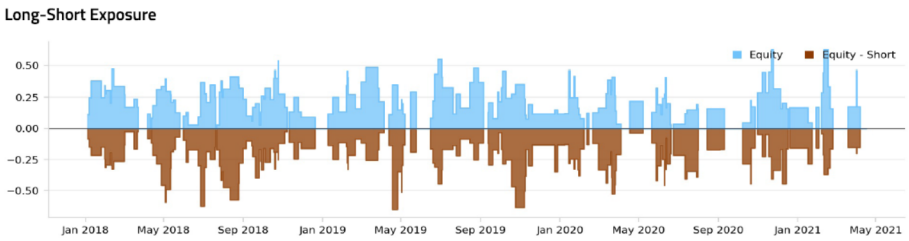

Figure 13: Long-Short Exposure

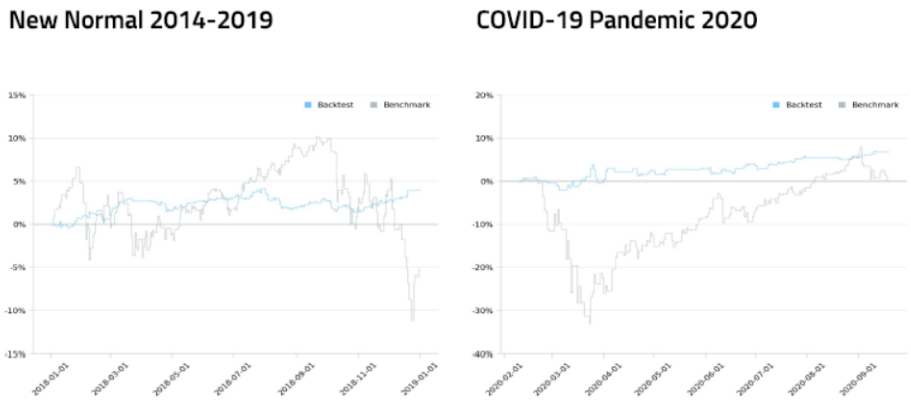

Figure 14: Covid-19 Pandemic 2020

3.6 Sub Strategy: Hedging Macro risks

3.6.1 Gold Trading Strategy

In order to increase returns and hedge macro risks, we have added a gold trading sub-strategy: long gold when the $\frac{gold}{S\&P\ 500\ Index}$ ratio is greater than the 5-day exponential moving average of the ratio and long S&P 500 index when the $\frac{gold}{S\&P\ 500\ Index}$ ratio is lower than the 5-day exponential moving average of the ratio.

For this sub-strategy, we selected the S&P 500 index ETF"SPY" and the best performing gold ETF "SGOL" in the second quarter of 2021[13] as the investment targets.

3.6.2 Basic Mechanism

S&P500 index is a very ideal investment target for the following reasons. First, S&P 500 index funds track the S&P 500, which is one of the best representations of the stock market as a whole. That means if the stock market as a whole is doing well, the S&P 500 index fund will likely be performing well, too. Of course, this also means that if the stock market takes a turn for the worse, the index fund will take a hit as well. But because the market has always recovered from every downturn it's ever experienced, there's a very good chance the investments will also bounce back. Second, Index funds are one of the most affordable investments out there, making them an excellent option for those who don't have much spare cash to invest. Especially right now when money is tight for millions of households, you may be primarily focused on paying the bills or building an emergency fund but still want to invest for retirement. Even if you only have a few dollars to spare, investors can invest that money in an S&P 500 index fund and then sit back and let it grow. Third, When investing in an S&P 500 index fund, you're actually investing in 500 different stocks at once. That level of diversification substantially lowers risk, because if a few of those stocks don't perform well it won't tank the entire portfolio. Of course, if the S&P 500 itself experiences a downturn, the your index fund will as well. But, again, the stock market historically has always recovered from its crashes, so the index fund will, too.

Although S&P 500 seems very profitable, we need other asset to hedge the risk, and here we choose gold.

Gold is a safe haven for investors when the market performs bad. Gold linked to the US dollar has a certain hedging effect, so there is a certain negative correlation between the trend of gold and the SP 500, and the negative correlation will accelerate the inflow of hedging funds to a certain extent.

In particular, the lower the ratio of $\frac{gold}{S\&P\ 500\ Index}$ goes, the more expensive the S&P 500 Index becomes relative to the price of gold, which would normally be indicative of a risk on environment or an outperformance in stock environment where investors will pile money into risk assets and prefer risk assets over

[13]https://www.investopedia.com/articles/etfs/top-gold-etfs/

and above safe havens assets. And vice versa, when there is a strong move higher of the ratio, the gold is getting more expensive relative to the price of S&P 500 Index, which is indicative of a risk-off environment where investors prefer safer investments to equities to protect their portfolio.

Also, the gold can be seen as an inflation hedge for several reasons. First, gold a is real, durable, tangible, relatively transportable and universally acceptable asset. Thus, an expected increase in the CPI may motivate investors to convert their current assets into gold to be protected from inflation. Second, when the expected inflation increases, there would be a rise in nominal interest rates (Feldstein, 1980). This leads to a rise of the required rate of return of holding gold (or its opportunity cost) and so its prices. Third, Levin and Wright(2006) assume that changes in gold extraction costs are led by inflation. Thus, in the long term, gold prices would rise to compensate this cost increase. Some studies such as Artigas (2010), Shahbaz et al. (2014), and Bampinas and Panagiotidis (2015),document that gold is an efficient hedge against inflation. We can also find the basis directly from real world. For example, in 2009 when central bank intervened with quantitative easing, the fears of inflation started to creep in , as a result, the gold gained a lot attraction from investors. Therefore, in 2009-2011, the gold withstood inflation and its prices rose sharply.

Figure 15: In Sample: Optimal Back-testing Result with Gold Trading Strategy using OLS

Method	Sharpe Ratio	Enter	Exit	Return	MaxDD
OLS	1.629	2.0	0.0	25.79%	5.6%
Kalman Filter	1.491	2.0	0.5	23.38%	4.1%

Table 15: In Sample: Optimal Back-testing Statistics on Selected SP500 Pairs with Gold Trading Strategy

Figure 16: In Sample: Optimal Back-testing Result with Gold Trading Strategy using Kalman Filter

Method	Sharpe Ratio	Enter	Exit	Return	MaxDD
OLS	1.431	2.0	0.0	10.62%	2.9%
Kalman Filter	2.025	2.0	0.5	13.74%	2.1%

Table 16: Out of Sample: Back-testing Statistics with Optimized Parameters and Gold Trading Strategy

Figure 17: Out of Sample: Optimal Back-testing Result with Gold Trading Strategy using Kalman Filter

3.7 Live Trading

We started our live trading on April 24, 2021. Figure 18 shows the performance of our ongoing strategy.

Figure 18: Live Trading Result

4 Future Work

We will implement this strategy in Chinese A-share stock market and Hong Kong stock market.

The backtesting results of both ETFs and Chinese Concept Stocks are not ideal. We will try to figure out whether our selection method is flawed or the ETFs and CCSs are not suitable for pairs trading.

Period	Equity	Method	Sharpe Ratio	Enter	Exit	Return	MaxDD
In	ETF	OLS	-0.64	2.0	0.0	-1.39%	1.4%
In	ETF	Kalman	-0.256	2.0	0.5	-2.57%	5.8%
Out	ETF	OLS	-0.71	2.0	0.0	-0.60%	0.8%
Out	ETF	Kalman	0.365	2.0	0.5	0.94%	1.3%
In	CCS	OLS	0.6	2.0	0.0	15.90%	7.5%
In	CCS	Kalman	-0.11	2.0	0.5	-5.58%	22.2%
Out	CCS	OLS	0.429	2.0	0.0	10.40%	17.9%
Out	CCS	Kalman	1.095	2.0	0.5	17.63%	4.1%

Table 17: Back-testing Statistics with Optimized Parameters using ETF and CCS

References

Damodaran, A. (2012). Investment valuation: Tools and techniques for determining the value of any asset. John Wiley & Sons.

Simao Moraes Sarmento , Nuno Horta, Enhancing a Pairs Trading strategy with the application of Machine Learning, Expert Systems with Applications, vol 158, 15 November 2020, 113490

Gatev, E. G., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs Trading: Performance of a Relative-Value Arbitrage Rule. Review of Financial Studies, 19(3), 797C827.

Mario, C. B. , De la Orden De la Cruz Carmen, & Camilo, P. R. . (2018). Pairs trading techniques: an empirical contrast. European Research on Management and Business Economics, 24(3), 160-167.

Sarmento, S. M. , & Horta, N. . (2020). Enhancing a pairs trading strategy with the application of machine learning. Expert Systems with Applications, 158, 113490.

Taewook Kim and Ha Young Kim, 2019.

Ganapathy Vidyamurthy, Pairs Trading : Quantitative Methods and Analysis, 2004

Kissell, R. (2013). The science of algorithmic trading and portfolio management (pp. 87-128). Academic Press.

Feldstein, M. . (1983). Inflation, tax rules, and the prices of land and gold. NBER Chapters, 14(3), 309-317.

Levin, E. J. , & Wright, R. E. . (2006). Short-run and Long-run Determinants of the Price of Gold. world gold council.

Shahbaz, M. , Tahir, M. I. , Ali, I. , & Rehman, I. U. . (2014). Is gold investment a hedge against inflation in pakistan? a co-integration and causality analysis in the presence of structural breaks. The North American Journal of Economics and Finance, 28(APR.), 190-205.

Georgios, Bampinas, Theodore, & Panagiotidis. (2015). Are gold and silver a hedge against inflation? a two century perspective. International Review of Financial Analysis.

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

https://pykalman.github.io/

Appendix

The selected ETFs list is: [MMM, AOS, ABT, ABBV, ABMD, ACN, ATVI, ADBE, AAP, AMD, AES, AFL, A, APD, AKAM, ALK, ALB, ARE, ALXN, ALGN, ALLE, LNT, ALL, GOOGL, GOOG, MO, AMZN, AMCR, AEE, AAL, AEP, AXP, AIG, AMT, AWK, AMP, ABC, AME, AMGN, APH, ADI, ANSS, ANTM, AON, APA, AAPL, AMAT, APTV, ADM, ANET, AJG, AIZ, T, ATO, ADSK, ADP, AZO, AVB, AVY, BKR, BLL, BAC, BAX, BDX, BRK.B, BBY, BIO, BIIB, BLK, BA, BKNG, BWA, BXP, BSX, BMY, AVGO, BR, BF.B, CHRW, COG, CDNS, CZR, CPB, COF, CAH, KMX, CCL, CARR, CTLT, CAT, CBOE, CBRE, CDW, CE, CNC, CNP, CERN, CF, SCHW, CHTR, CVX, CMG, CB, CHD, CI, CINF, CTAS, CSCO, C, CFG, CTXS, CME, CMS, KO, CTSH, CL, CMCSA, CMA, CAG, COP, ED, STZ, CPRT, GLW, CTVA, COST, CCI, CSX, CMI, CVS, DHI, DHR, DRI, DVA, DE, DAL, XRAY, DVN, DXCM, FANG, DLR, DFS, DISCA, DISCK, DISH, DG, DLTR, D, DPZ, DOV, DOW, DTE, DUK, DRE, DD, DXC, EMN, ETN, EBAY, ECL, EIX, EW, EA, EMR, ENPH, ETR, EOG, EFX, EQIX, EQR, ESS, EL, ETSY, RE, EVRG, ES, EXC, EXPE, EXPD, EXR, XOM, FFIV, FB, FAST, FRT, FDX, FIS, FITB, FRC, FE, FISV, FLT, FLIR, FMC, F, FTNT, FTV, FBHS, FOXA, FOX, BEN, FCX, GPS, GRMN, IT, GNRC, GD, GE, GIS, GM, GPC, GILD, GPN, GL, GS, GWW, HAL, HBI, HIG, HAS, HCA, PEAK, HSIC, HES, HPE, HLT, HFC, HOLX, HD, HON, HRL, HST, HWM, HPQ, HUM, HBAN, HII, IEX, IDXX, INFO, ITW, ILMN, INCY, IR, INTC, ICE, IBM, IFF, IP, IPG, INTU, ISRG, IVZ, IPGP, IQV, IRM, JBHT, JKHY, J, SJM, JNJ, JCI, JPM, JNPR, KSU, K, KEY, KEYS, KMB, KIM, KMI, KLAC, KHC, KR, LB, LHX, LH, LRCX, LW, LVS, LEG, LDOS, LEN, LLY, LNC, LIN, LYV, LKQ, LMT, L, LOW, LUMN, LYB, MTB, MRO, MPC, MKTX, MAR, MMC, MLM, MAS, MA, MXIM, MKC, MCD, MCK, MDT, MRK, MET, MTD, MGM, MCHP, MU, MSFT, MAA, MHK, TAP, MDLZ, MPWR, MNST, MCO, MS, MSI, MSCI, NDAQ, NTAP, NFLX, NWL, NEM, NWSA, NWS, NEE, NLSN, NKE, NI, NSC, NTRS, NOC, NLOK, NCLH, NOV, NRG, NUE, NVDA, NVR, NXPI, ORLY, OXY, ODFL, OMC, OKE, ORCL, OTIS, PCAR, PKG, PH, PAYX, PAYC, PYPL, PENN, PNR, PBCT, PEP, PKI, PRGO, PFE, PM, PSX, PNW, PXD, PNC, POOL, PPG, PPL, PFG, PG, PGR, PLD, PRU, PEG, PSA, PHM, PVH, QRVO, QCOM, PWR, DGX, RL, RJF, RTX, O, REG, REGN, RF, RSG, RMD, RHI, ROK, ROL, ROP, ROST, RCL, SPGI, CRM, SBAC, SLB, STX, SEE, SRE, NOW, SHW, SPG, SWKS, SNA, SO, LUV, SWK, SBUX, STT, STE, SYK, SIVB, SYF, SNPS, SYY, TMUS, TROW, TTWO, TPR, TGT, TEL, TDY, TFX, TER, TSLA, TXN, TXT, BK, CLX, COO, HSY, MOS, TRV, DIS, TMO, TJX, TSCO, TT, TDG, TRMB, TFC, TWTR, TYL, TSN, USB, UDR, ULTA, UAA, UA, UNP, UAL, UPS, URI, UNH, UHS, UNM, VLO, VAR, VTR, VRSN, VRSK, VZ, VRTX, VFC, VIAC, VTRS, V, VNO, VMC, WRB, WBA, WMT, WM, WAT, WEC, WFC, WELL, WST, WDC, WU, WAB, WRK, WY, WHR, WMB, WLTW, WYNN, XEL, XLNX, XYL, YUM, ZBRA, ZBH, ZION, ZTS]

The CCS list: [YI, VNET, QFIN, JOBS, BABA, AMBO, JG, ATHM, BIDU, BZUN, GLG, BGNE, BILI, BLCT, BRQS, BEDU, CSIQ, CBAT, CMCM, CAAS, CCRC, JRJC, CGA, HGSH, CJJD, COE, CPHI, CREG, SXTC, CXDC, CNET, CD, CLPS, CCM, DADA, DQ, DTSS, DOGZ, LYL, DXF, EVK, SFUN, FANH, FAMI, FUTU, FTFT, FFHL, GDS, BTBT, GRNQ, GSX, GURE, HLG, HX, HOLI, HNP, HTHT, HUYA, IQ, ITP, JD, JKS, KNDI, KBSF, BEKE, KC, LX, LI, LLIT, LITB, LU, LKCO, MDJH, MNSO, MTC, MOMO, MOXC, NTP, NTES, EDU, NEWA, BIMI, NIO, NIU, NOAH, SEED, OSN, FENG, PDD, PME, PT, PLAG, PHCF, QD, QTT, RCON, SOL, RENN, RETO, REDU, SECO, AIHS, SGOC, TYHT, SVA, SOGO, SOHU, TAL, TANH, TAOP, TEDU, PETZ, TME, NCTY, TCOM, TC, TOUR, MYT, UTSI, UXIN, VIOT, VIPS, WB, WEI, XIN, XPEV, XNET, YSG, YRD, YUMC, YY, ZLAB, ZH, ZKIN]