Pairs trading using K-means clustering and cointegration

28-06-2021

Deepanshu Jain
Computer Science
School of engineering
Instructor - Anirban chakraborti

Abstract

The goal of this paper is to develop a model that will give us better returns than the S&P 500 stock market returns by K-means clustering and pairs trading strategies . First of all , we will cluster the stocks on the basis of important features (such as P/E ratio , Market Cap) by using K- means Clustering algorithm . We will find optimal value for K by creating the elbow graph using our fundamental data features . After finding K, we will create our K-Means algorithm and select stocks from all Clusters to test whether or not we could identify tradeable relationships in different combinations of pairs of particular clusters . We'll create a method that will allow us to iterate over our pairs, compute the slope and then perform the CADF test. The pairs that are cointegrated will be stored in a list . If two pairs are cointegrated then they will drift towards and apart from each other around the mean . It means their spreadness is stationary or converged . When the series diverges from one another, we say that the *spread* is high. When they drift back towards each other, we say that the *spread* is low. We need to buy one security and short the other . We long the security that is underperforming and short the security that is over performing .As the pairs are cointegrated the stock which is underperforming now, will overperform and the stock which is overperforming now, will underperform at some point of time to maintain the spreadness . At that point of time , We will do trading in reverse order then the previous trading . By doing so in cointegrated pairs we will get good returns as compared to the overall market .

Introduction

K-means clustering - Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Predefined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group . It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more similar the data points are within the same cluster.

Pairs trading - Pairs trading is one of the most popular types of trading strategy. In this strategy, usually a pair of stocks are traded in a market-neutral strategy, i.e. it doesn't matter whether the market is trending upwards or downwards, the two open positions for each stock hedge against each other. The key challenges in pairs trading are to:

1.Choose a pair which will give you good statistical arbitrage opportunities over time

2.Choose the entry/exit points

For pair selection we will use cointegration.

For entry/exit points we will use mean reversion .

Cointegration - The most common test for Pairs Trading is the cointegration test. Cointegration is a statistical property of two or more time-series variables which indicates if a linear combination of the variables is stationary.

Let us understand this statement above. The two-time series variables, in this case, are the log of prices of stocks A and B. Linear combination of these variables can be a linear equation defining the spread:

As you know, Spread = log(a) – nlog(b), where 'a' and 'b' are prices of stocks A and B respectively.

For each stock of A bought, you have sold n stocks of B.

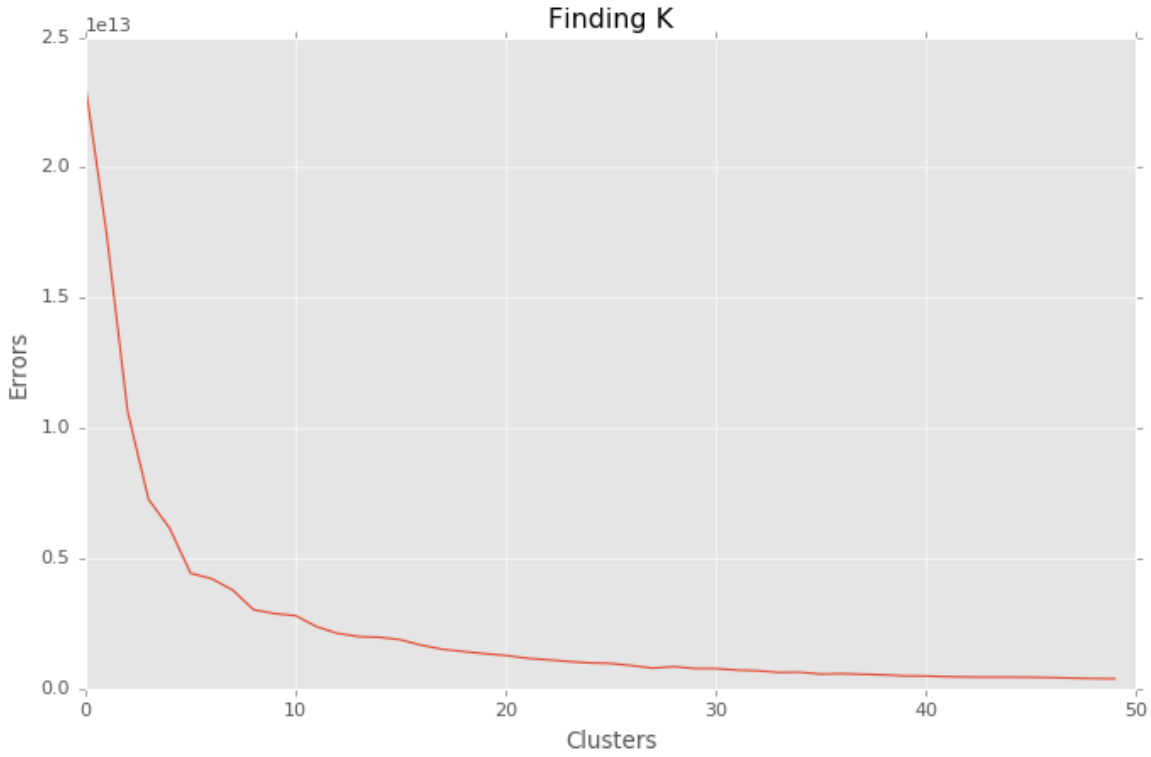
If A and B are cointegrated then it implies that this equation above is stationary. A stationary process has very valuable features which are required to model Pairs Trading strategies. For instance, in this case, if the equation above is stationary, that suggests that the mean and variance of this equation remains constant over time. So if we start with 'n', which is called the hedge ratio, so that spread = 0, the property of stationary implies that the expected value of spread will remain as 0. Any deviation from this expected value is a case for statistical abnormality, hence a case for pairs trading!

Mean reversion - Mean reversion, or reversion to the mean, is a statistical principle that can be used in finance and investing. It states that volatile prices and historical fluctuations will eventually return, or revert, to the long-term average, or mean, of a dataset.

Methodology

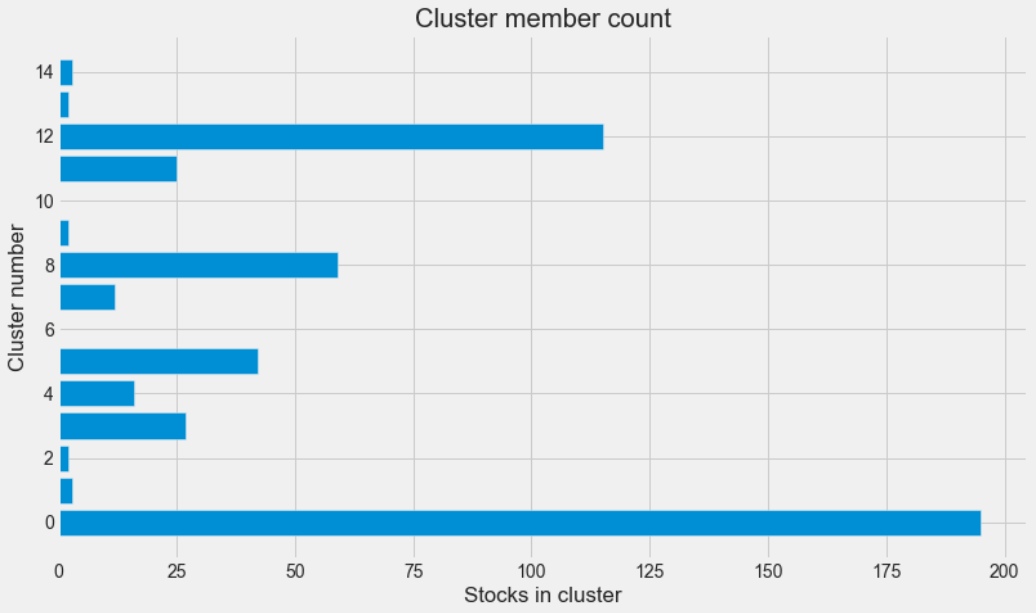
Stock Data - S&P 500

1. ***Clustering -*** By using K-means clustering , we will form all possible clusters . K value is estimated by minimising error (by elbow methode) .Clustering is done on two columns - P/E ratio and Market capital . First of all we will find out the value of K by the elbow method of error minimization .



From this graph we will select k value as 15

Using this k value , we will use k means clustering and make different clusters of S&P 500 stocks



2. ***Pairs formation***

After forming the clusters we will use 6 stocks from the 0th cluster .

```
symbol_list=['AOS','AYI','AAP','AMD','AES','AKAM']
```

Then from these pairs we will form all possible pairs

```
#creating method to identify each possible pair
def create_pairs(symbolList):
 #creating a list to hold each possible pair
    pairs=[]
 #initializing placeholders for the symbols in each pair
    x=0
    y=0
    for count,symbol in enumerate(symbolList):
        for nextCount,nextSymbol in enumerate(symbolList):
            x=symbol
            y=nextSymbol
            if x!=y:
                pairs.append([x,y])

    return pairs
```

--> [['AOS', 'AYI'], ['AOS', 'AAP'], ['AOS', 'AMD'], ['AOS', 'AES'], ['AOS', 'AKAM'], ['AYI', 'AOS'], ['AYI', 'AAP'], ['AYI', 'AMD'], ['AYI', 'AES'], ['AYI', 'AKAM'], ['AAP', 'AOS'], ['AAP', 'AYI'], ['AAP', 'AMD'], ['AAP', 'AES'], ['AAP', 'AKAM'], ['AMD', 'AOS'], ['AMD', 'AYI'], ['AMD', 'AAP'], ['AMD', 'AES'], ['AMD', 'AKAM'], ['AES', 'AOS'],['AES', 'AYI'],['AES', 'AAP'], ['AES', 'AMD'], ['AES', 'AKAM'], ['AKAM', 'AOS'], ['AKAM', 'AYI'], ['AKAM', 'AAP'], ['AKAM', 'AMD'], ['AKAM', 'AES']***1***

3. ***Selection of cointegrated pairs***

First of all we will use intraday trading (Closing price) data of the last 5 years for these 6 stocks to find the cointegration .

training period 28-06-2016-->28-6-2020

testing period 19-03-2021-->17-06-2021

```
#creating our training data dataframe using our training period start and end dates
training_df=get_training_data(original_data,symbol_list,'2016-06-28','2020-06-28')
training_df
```

	AOS	AYI	AAP	AMD	AES	AKAM
0	41.840000	244.660004	156.979996	5.120000	11.79	53.049999
1	42.799999	246.610001	161.009995	5.130000	12.03	54.560001
2	44.055000	247.960007	161.630005	5.140000	12.48	55.930000
3	43.605000	250.250000	164.330002	5.070000	12.33	55.500000
4	43.264999	246.990005	161.190002	4.960000	12.20	54.250000
...	...	...	...	...	...	...
1002	47.189999	89.320000	151.600006	54.759998	13.98	101.589996
1003	47.470001	90.529999	149.970001	53.990002	13.84	100.879997
1004	45.110001	85.809998	143.000000	52.389999	13.84	100.970001
1005	44.740002	86.820000	142.910004	51.930000	13.88	102.889999
1006	44.450001	84.449997	139.990005	50.099998	13.75	105.370003

1007 rows × 6 columns

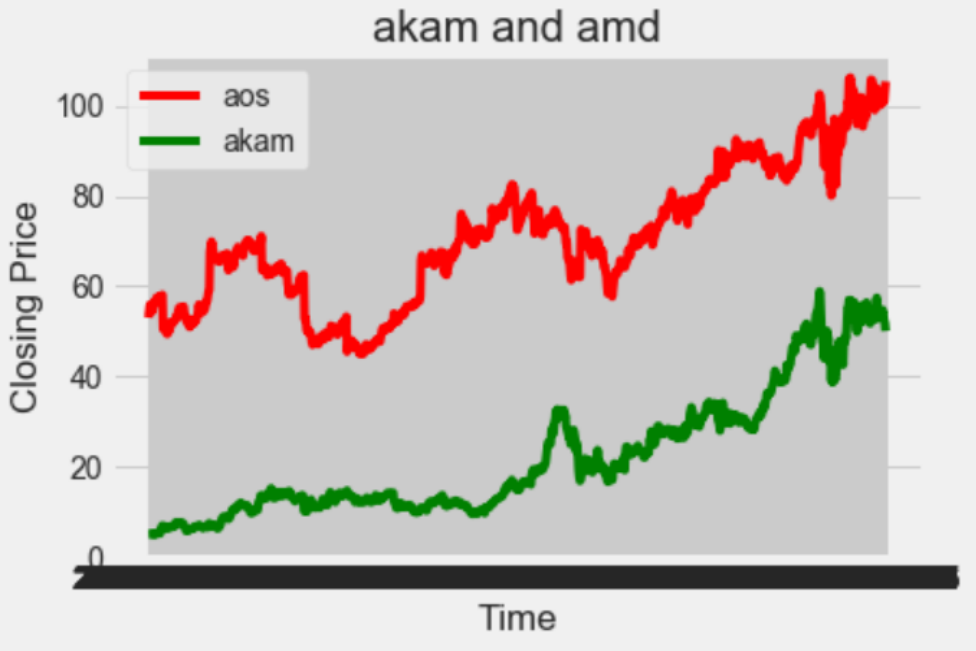
From all possible pairs we will select the cointegrated pairs by doing a cointegration test. So, .We'll create a method that will allow us to iterate over our pairs, compute the slope and then perform the CADF test. The pairs that are cointegrated will be stored in a list.

After doing the test , we will get only one pair that is cointegrated with a confidence interval of 90%.

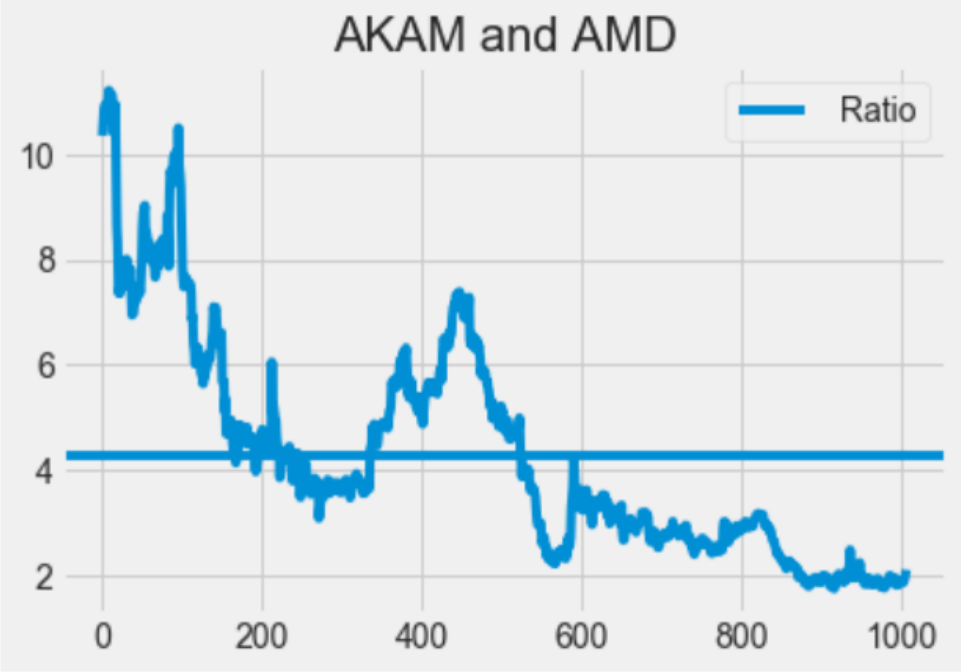
['AKAM', 'AMD']
Pair Cointegrated at 90% Confidence Interval

4. ***Mean reversion on a cointegrated pair***

Now, For this cointegrated pair we will first plot the prices vs day curve .



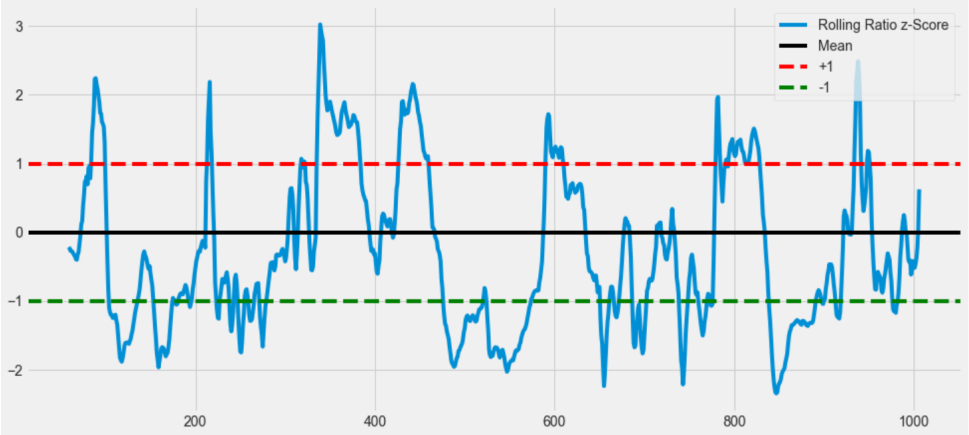
Let's plot their ratios on a graph to see what's going on.



What we need to do next is to try to standardize the ratios because the absolute ratio might not be the most ideal. We need to use z-scores.

The z score is calculated by:

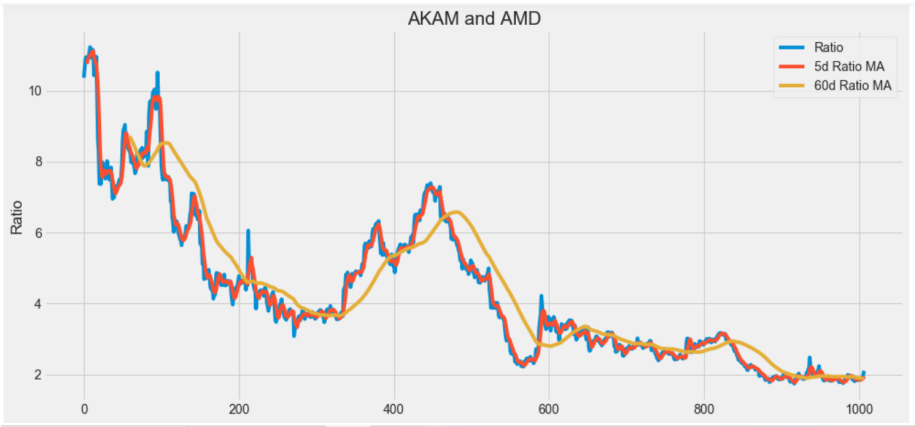
###Z Score (Value) = (Value - Mean) / Standard Deviation



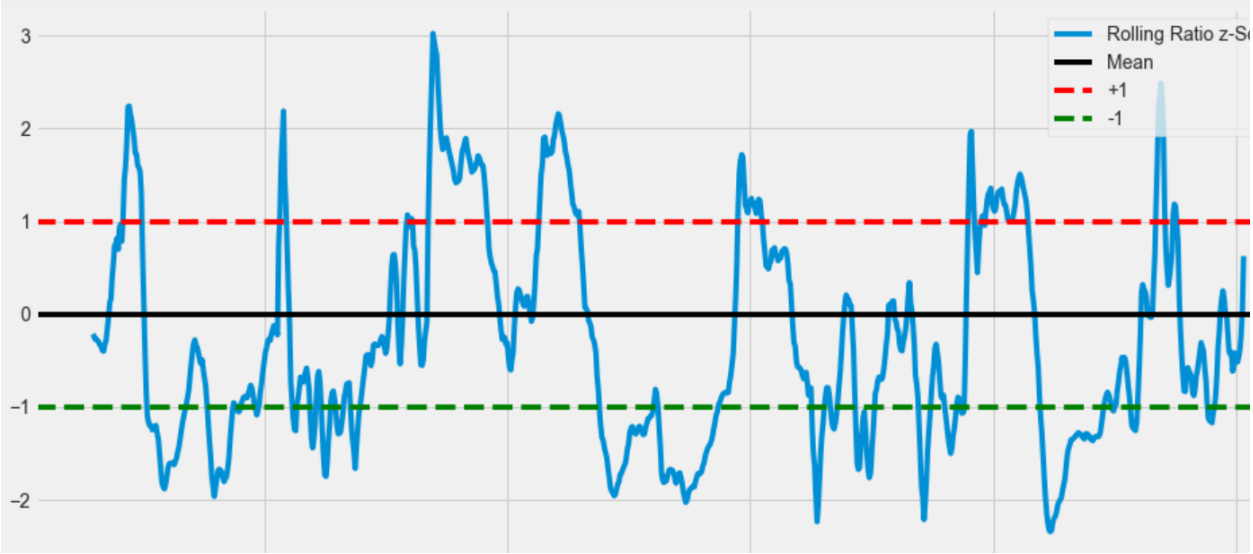
We need to find out what features are actually important in determining the direction of the ratio moves. Knowing that the ratios always eventually revert back to the mean, maybe the moving averages and metrics related to the mean will be important.

Let's try using these features:

- 60 day Moving Average of Ratio
- 5 day Moving Average of Ratio
- 60 day Standard Deviation
- z score



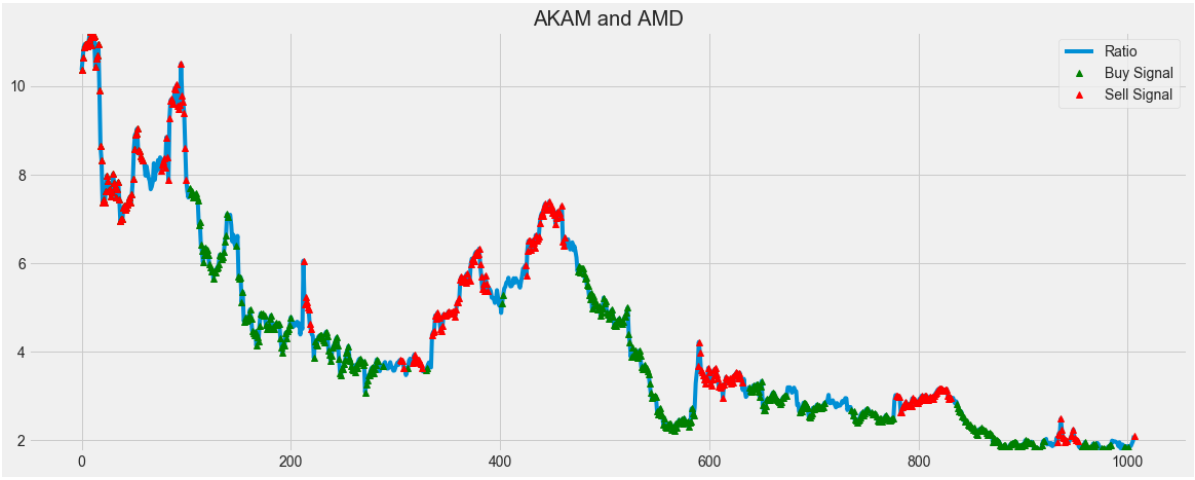
Let's also take a look at the moving average z-scores.



5. ***Creating a model***

Taking a look at our z-score chart, it's pretty clear that if the absolute value of the z-score gets too high, it tends to revert back. We can keep using our +1/-1 ratios as thresholds, and we can create a model to generate a trading signal:

- Buy (1) whenever the z-score is below -1.0 because we expect the ratio to increase
- Sell (-1) whenever the z-score is above 1.0 because we expect the ratio to decrease



These are the trading signals for the ratios

Now , we will plot it for stocks

Now we can clearly see when we should buy or sell on the respective stocks.

Results

```
trade(training_df['AKAM'], training_df['AMD'], 60, 5)
```

```
Percentage returns is  43.26076372920325
```

By doing so we get a returns of 43.26%

If we compare this returns with the overall S&P 500 returns which is

(Price of S&P 500 in june 2020 - price of S&P 500 in june 2016)/ price of S&P 500 in june 2016 X 100%

=(3236-2344)/2344 X 100 %

= 38.05 %

So , By doing pairs trading , we get returns higher than S&P 500 stocks returns .

References

1."Machine Learning in Computational FInance " by Victor Boyarshinov.

2..Article on "Applying Machine Learning Ensembles to Market Microstructure to Achieve Portfolio Optimization"

3. Article on "Employing Machine Learning for Pairs Selection"
by Aaron Debrincat

4."sktime: A Unified Interface for Machine Learning with Time Series" by Markus Löning, Anthony Bagnall, Sanjay Surya Ganesh, Viktor Kazakov, Jason Lines, Franz J. Király

5.Article on "pairs trading basics :Correlation,cointegration and pairs trading strategy" by anupriya Gupta

6.Github : "Pairs trading" by jerry tiger