Comparing Different Methods for Identifying Good Pairs Trades

Miao-Chin Yen
Department of Management Science & Engineering
Stanford University
miaochin@stanford.edu

1 Introduction

Pairs-trading is considered as the precursor of statistical arbitrage. The notion is to identify a pair of stocks whose price performance behave similarly by examining the historical data. If investors believe the price history would repeat, they could utilize the naive "Buy Low and Sell High" strategy to make profit when the spread of the time series between the pair widens. To the best of our knowledge, pairs trading is a popular short-term speculation strategy on Wall Street and worth studying. We further use figure 1 to explain the idea. There are three pairs (Stock A, Stock B), (Stock A, Stock C) and (Stock B, Stock C) in the figure. (Stock A, Stock B) is the pair with similar performance. Our goal in this project is to identify the good pairs trades. A good approach should generate pair (Stock A, Stock B) among the three candidates in figure 1 for us to do pairs-trading.

According to [1], the methods proposed to identify good pairs can approximately be categorized to five approaches - Distance, Cointegration, Time Series, Stochastic Control and others. Distance approach leverages various distance metrics to identify pairs. Cointegration approach utilizes cointegration test to choose pairs. Time series approach models the spread as a mean-reverting process. Stochastic control approach aims at identifying the optimal portfolio constructed by a pair. Other approach includes copula approach, Principal Components Analysis (PCA) approach and some machine learning related approaches.

In this project, we put emphasis on existing Distance, Correlation, Cointegration and Granger-Causality methods to identify the pairs. Furthermore, we propose a time series approach based on Autoregressive Process. We apply these approaches to generate pairs and test them on the data not used to find pairs to do comparison. We show that the mean-reverting property would be the most important factor for identifying good pairs trades if our target is stable performance. The remainder of this project is organized as follows: In section 2, we cover the five different methods for selecting pairs. In section 3, we run experiments for methods stated in section 2. Finally, section 4 concludes and summarizes directions for further research.

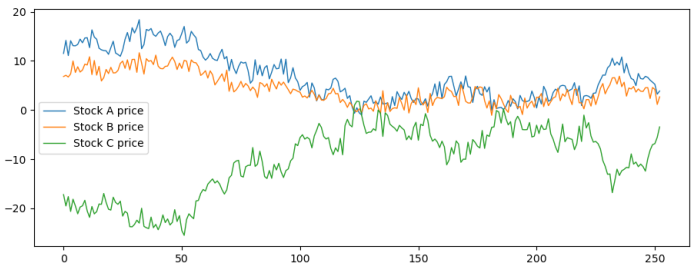


Figure 1: Price Time Series of Stock A, B and C

2 Pair Selection

Pair selection identifies the pair of stocks we believe would have similar performance in the market. Suppose we have n stocks, and their price time series are $\{P_{i,t}\}_{t=0}^{T}$, $i = 1, 2, ..., n$. We then have $\frac{n(n-1)}{2}$ different pairs. In this section, we state five different methods we would apply to select the pairs.

2.1 Method I : Selected by Distance

Distance approach selects the pair with the shortest euclidean distance between two time series. Here, we compare the cumulative return time series to avoid ignoring the pair having same trend for cumulative return, but one of the stock price is extremely higher than the other. If one stock has higher cumulative return and the other performs the same way, we should view them as a possible pair. Let $\{R_{i,t} = \frac{P_{i,t}-P_{i,t-1}}{P_{i,t}}\}_{t=1}^{T}$ for $i = 1, 2, ..., n$. We choose pair with the shortest euclidean distance:

$$(i^*, j^*) = \underset{i,j,i\neq j}{\arg\min} \sqrt{\sum_{t=1}^{T}(R_{i,t} - R_{j,t})^2}$$

For the identified pair (i^*, j^*), we construct the portfolio consisting of two positions of equal capital allocation. We long position in stock i and short position in stock j. Therefore, we consider time series $\{z_t^{(1)}\}_{t=1}^{T} = \{R_{i^*,t} - R_{j^*,t}\}_{t=1}^{T}$ in our experiment.

2.2 Method II: Selected by Correlation

Correlation approach selects the pair with the highest linear correlation. We again use the cumulative return time series to do comparison. Since the core idea of pairs trading is to expect the pair to have similar performance, we then expect the relationship to be positive not negative. Therefore,

$$(i^*, j^*) = \underset{i,j,i\neq j}{\arg\max}\, \rho_{ij} = \frac{\sum_{t=1}^{T}\left(R_{i,t} - \bar{R}_i\right)\left(R_{j,t} - \bar{R}_j\right)}{\sqrt{\sum_{t=1}^{T}\left(R_{i,t} - \bar{R}_i\right)^2}\sqrt{\sum_{t=1}^{T}\left(R_{j,t} - \bar{R}_j\right)^2}}$$

$$s.t. \qquad 0 \leq \rho_{i,j} \leq 1$$

where $\bar{R}_i$ is the mean of time series $\{R_{i,t}\}_{t=1}^{T}$.

For the identified pair (i^*, j^*), we construct same portfolio as in the distance approach. We long position in stock i and short position in stock j with same capital allocation. Therefore, we consider time series $\{z_t^{(2)}\}_{t=1}^{T} = \{R_{i^*,t} - R_{j^*,t}\}_{t=1}^{T}$ in our experiment.

2.3 Method III: Selected by Cointegration

Cointegration enables us to describe the relationship between stock price time series instead of cumulative return time series. We first give some terminology. An $I(1)$ or integrated of 1 series is a random walk. An $I(0)$ or integrated of 0 series is a weakly stationary time series.

Given the above terminology, we say that two price time series $\{P_{i,t}\}_{t=0}^{T}$ and $\{P_{j,t}\}_{t=0}^{T}$ are cointegrated if $\{P_{i,t}\}_{t=0}^{T}$ and $\{P_{j,t}\}_{t=0}^{T}$ are $I(1)$ series and

$$\exists\ \beta \text{ s.t. } \{z_t = P_{i,t} - \beta P_{j,t}\}_{t=0}^{T} \text{ is an } I(0) \text{ series}$$

Here, we believe that a good pair of stocks is cointegrated and cointegration thus become a representation of good pairs. For the cointegrated time series $\{P_{i^*,t}\}_{t=0}^{T}$ and $\{P_{j^*,t}\}_{t=0}^{T}$, we long one share of stock i^* and short β shares of stock j^*. Therefore, we consider time series $\{z_t^{(3)}\} = \{P_{i^*,t} - \beta P_{j^*,t}\}_{t=0}^{T}$ in our experiment.

2.4 Method IV : Selected by Autoregressive (AR) Process

In autoregressive process approach, we fit the cumulative return spread $\{X_t^{(i,j)} = R_{i,t} - R_{j,t}\}_{t=1}^T$, $i, j \in \{1, 2, ..., n\}, i \neq j$ as an AR process. The notion is that the current cumulative return spread would depend on the previous cumulative return spread. We first give some terminology. An order-p autoregressive process is defined as

$$X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + W_t \qquad \text{where } W_t \sim WN\left(0, \sigma_w^2\right)$$

We say an AR(p) process is stationary if all roots of the characteristic polynomial are not on unit circle. i.e., $|z| \neq 1$ for

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0$$

Since the selected pair should have the same performance in the long run, the fitted AR process should be mean-reverting (stationary). If the pair is a perfect one, the mean of the cumulative return spread should be 0 or it should not deviate from 0 too much for a long time. Therefore, only the pairs with stationary fitted AR process would be considered. Furthermore, we choose the pair with the highest mean-reverting speed. We could use half-life to characterize the speed of mean-reverting. Shorter half-life means that the stationary AR process is expected to halve its distance to the stationary mean faster. For a stationary AR(p) process $\{X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + W_t\}_t$, half-life is defined as

$$h = -\frac{\ln 2}{\ln \sum_{k=1}^{p} \phi_k}$$

We give an AR(1) example to illustrate the notion of half-life. For higher order p, please reference [4] for formal mathematical formulation.

Example. Half-life of stationary AR(1)

Let $X_t = \phi_1 X_{t-1} + W_t$ where $|\phi_1| < 1$ denote a weak-stationary AR(1) process. We further let $E[X_t] = \mu$ for all t because of stationarity. By computation, we derive $\mu = 0$. We then plug μ into the AR(1) process and get

$$X_t - \mu = \phi_1 \left(X_{t-1} - \mu\right) + W_t$$

We define $Y_t = X_t - \mathbb{E}[X_t]$ to be the the distance between X_t and the stationary mean. By definition of half-life, we want to find h such that

$$\mathbb{E}_t \left[Y_{t+h}\right] = \frac{1}{2} Y_t$$

Since $\mathbb{E}_t \left[Y_{t+h}\right] = \phi_1^h Y_t = \frac{1}{2} Y_t$,

$$h = -\frac{\ln 2}{\ln \phi_1}$$

The formal optimization problem for AR approach is as follows. We choose the pair with shortest half-life where the fitted AR(p) process is stationary:

$$(i^*, j^*) = \arg\min_{i,j,i\neq j} -\frac{\ln 2}{\ln \sum_{k=1}^{p} \phi_k^{(i,j)}}$$

$$\text{s.t.} \qquad X_t^{(i,j)} \text{ is an stationary AR}(p) \text{ process}$$

For the identified pair (i^*, j^*), we construct same portfolio as in the distance and correlation approaches. Therefore, we consider time series $\{z_t^{(4)}\}_{t=1}^T = \{R_{i^*,t} - R_{j^*,t}\}_{t=1}^T$ in our experiment.

2.5 Method V: Selected by Granger Causality

Granger Causality is a test used to verify the usefulness of one variable to forecast another. The formal definition of Granger Causality is as follows. We let Ω_t denote all relevant and available information up to time t and denote $X^t_{t+h}(\Omega_t)$ as the optimal h-step-ahead predictor of process $\{X_t\}$ given all information in Ω_t. The corresponding mean squared error (MSE) is noted as $P^t_{X,t+h}(\Omega_t)$. The process $\{Y_t\}$ is said to Granger-cause $\{X_t\}$ if MSE is reduced when Y_t is included, i.e.,

$$P^t_{X,t+h}(\Omega_t) < P^t_{X,t+h}(\Omega_t \backslash \{Y_s \mid s \leq t\})$$

In this method, we select pair (i^*, j^*) where $\{R_{i^*,t}\}_{t=1}^T$ is Granger-cause $\{R_{j^*,t}\}_{t=1}^T$ and we consider time series $\{z_t^{(5)}\}_{t=1}^T = \{R_{i^*,t} - R_{j^*,t}\}_{t=1}^T$ in our experiment.

3 Experiments

We use the Vanguard Small Cap Value ETF (VBR) as stock universe and select 165 stocks from this universe, resulting in $\frac{165 \cdot 164}{2} = 13530$ different pairs. We use the data from Jaunary 1, 2019 to December 31, 2020 (formation period/historical data) to select pairs and the data from January 1, 2021 to December 31, 2021 (trading period) for testing the performance of the pairs. We use log price to do the experiments.

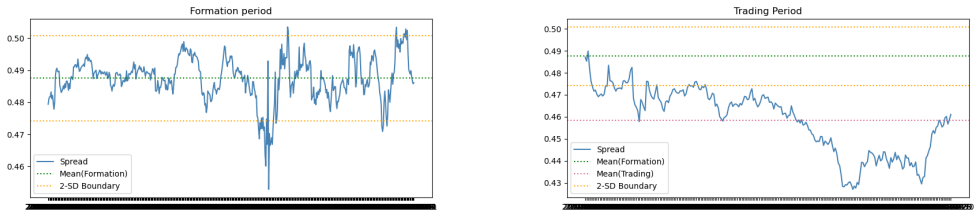
For each of the five above methods, we choose top five pairs from the historical data. We expect the selected pair should have similar performance in the trading period. We use the portfolios specified in Sec. 2 to test the performance for each method.

We first define some criteria to decide whether the selected pair is good or not. Let $t = 1, 2, ..., T_1$ be the formation period and $t = T_1 + 1, T_1 + 2, ..., T_2$ be the trading period. Further, we define $\hat{\mu}_F$ and $\hat{\sigma}_F$ to be the sample mean and sample standard deviation of $\{z_t^{(j)}\}_{t=1}^{T_1}$, $j = 1, 2, 3, 4, 5$ for each method. Similarly, we define $\hat{\mu}_T$ and $\hat{\sigma}_T$ to be the sample mean and sample standard deviation of $\{z_t^{(j)}\}_{t=T_1+1}^{T_2}$, $j = 1, 2, 3, 4, 5$ for each method.

3.1 Criteria

(1) For the constructed portfolios, if the pair we select is a good one, $\{z_t^{(j)}\}_{t=T_1+1}^{T_2}$ should lie within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$ with high probability.

Example. In figure 2(a), we notice that $\{z_t^{(3)}\}_{t=1}^{T_1}$ seems to be a mean reverting process. Hence, we believe the selected pair is good. However, in figure 2(b), only part of $\{z_t^{(3)}\}_{t=T_1+1}^{T_2}$ is within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$. Therefore, it is not a good pair according to criteria (1).



(a) Formation (b) Trading

Figure 2: Cointegration: OGS-POR

(2) Sample mean within trading period should not be too different from sample mean within formation period. To compare these two sample mean, we calculate MD $:= |\frac{\hat{\mu}_{\mathrm{T}} - \hat{\mu}_{\mathrm{F}}}{\hat{\sigma}_{\mathrm{T}}}|$.

Note that the criteria are reasonable because we do experiments on portfolios consisting of a pair of stocks. If it is a good pair, the portfolios should be an offset.

3.2 Comparison between different methods

In table 1, we list the statistics of top five pairs from distance method and correlation method. We notice that there are some low values for criteria 1. OGS-IDA pair from method I only has 9.9% within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$ and UA-UAA from both methods only provides 24.7% for criteria 1.

Method I	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD	Method II	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD
EGP-FR	87.2	0.58	UA-UAA	24.7	2.17
NWE-BKH	80.0	0.26	PEB-PK	100.0	0.67
CENT-CENTA	99.6	1.10	CENT-CENTA	99.6	1.10
UA-UAA	24.7	2.17	GPMT-ARI	100.0	0.94
OGS-IDA	9.9	1.91	ROIC-SITC	58.9	2.84

Table 1: Distance vs. Correlation

We next study the cointegration approach. Since there are many pairs cointegrated in our experiments, we further choose by distance and correlation respectively ($P_{i,t}$ vs. $\beta P_{j,t}$). From table 2 and 6, we notice that if we first do cointegration and sort the pairs by correlation, the overall performance is worse than sorted by distance. We may deduce that distance would be more effective than correlation for selecting pairs if we have considered cointegration effect.

Method III-Distance Sorting	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD	Method III-Correlation Sorting	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD
OGS-POR	9.56	1.97	PTEN-CLR	53.3	0.66
NWE-BKH	78.0	0.25	FULT-BRKL	99.6	0.30
IDA-WTM	99.6	0.35	SFNC-FCF	7.2	3.58
OGS-SR	45.8	2.39	PEB-ASB	5.9	2.93
FR-EGP	99.6	0.63	ASB-PK	7.5	2.64

Table 2: Cointegration- Distance Sorting vs. Correlation Sorting

We further study the autoregressive process approach. We test four orders of AR process. Overall, we could expect there are over 35% days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$ for this method for 4 different orders according to table 3 and 4. It is a huge improvement compared to distance, correlation and cointegration approaches where there will be some extreme low values for criteria 1. AR process approach gives us more stable performance.

Method IV-AR(1)	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD	Method IV-AR(2)	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD
GPMT-SABR	36.25	1.35	GPMT-SABR	36.25	1.35
IDA-WTM	45.81	0.35	OFC-MGEE	99.60	0.69
FOR-WH	43.82	0.99	H-BXMT	90.83	0.35
NWE-BKH	78.08	0.25	FOR-WH	43.82	0.99
H-UMPQ	39.84	1.63	H-UCBI	59.76	2.78

Table 3: Autoregressive Process- AR(1) vs. AR(2)

Method IV-AR(3)	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD	Method IV-AR(4)	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD
OFC-MGEE	99.60	0.69	OFC-MGEE	99.60	0.69
FOR-WH	43.82	0.99	FOR-WH	43.82	0.99
SBRA-FHI	58.56	1.25	IDA-WTM	45.81	0.35
GPMT-SABR	36.25	1.35	WH-HOMB	75.29	0.36
IDA-WTM	45.81	0.35	SBRA-FHI	58.56	1.25

Table 4: Autoregressive Process- AR(3) vs. AR(4)

In table 5, we show the experimental results for Granger Causality approach. Since this method can only identify the pairs with Granger Causality, we further select the pairs by sorting based on distance and half-life as the same idea in cointegration approach. Note that we do not know whehter $\{R_{i,t}\}_{t=1}^{T}$ is Granger-cause $\{R_{j,t}\}_{t=1}^{T}$ or $\{R_{j,t}\}_{t=1}^{T}$ is Granger-cause $\{R_{i,t}\}_{t=1}^{T}$. We need to test both versions. The pairs chosen by half-life overall performs better than by distance based on table 5 and 6. And we further notice Granger Causality approach with half-life sorting gives us the same result as AR approach with order 4. We deduce if the pair is Granger Causality, the fitted AR spread process would be stationary with high mean-reverting speed.

Method V-Half-life Sorting	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD	Method V-Distance Sorting	% Days within $[\hat{\mu}_{\mathrm{F}} - 2\hat{\sigma}_{\mathrm{F}}, \hat{\mu}_{\mathrm{F}} + 2\hat{\sigma}_{\mathrm{F}}]$	MD
OFC-MGEE	99.60	0.69	NWE-BKH	78.08	0.25
FOR-WH	43.82	0.99	IDA-WTM	45.81	0.35
IDA-WTM	45.81	0.35	BRKL-FULT	97.60	0.51
WH-HOMB	75.29	0.36	ISBC-CATY	15.13	2.79
SBRA-FHI	58.56	1.25	IDA-TMP	78.80	0.15

Table 5: Granger Causality- Half-life Sorting vs. Distance Sorting

(Stock A - Stock B implies A is granger cause of B)

In table 6, we list the average statistics for all 5 methods. We observe that Method II gives us the best result for criteria 1. Method IV-AR(4) and Method V-half-life sorting give us the best result for criteria 2. Hence, correlation, Granger Causality and mean-reverting all play important roles when selecting pairs.

	Method I	Method II	Method III-Distance Sorting	Method III-Correlation Sorting	Method IV-AR(1)
Average Criteria 1 (%)	60.28	76.64	66.51	34.70	48.76
Average Criteria 2	1.20	1.54	1.11	2.00	0.91
	Method IV-AR(2)	Method IV-AR(3)	Method IV-AR(4)	Method V-Half-life Sorting	Method V-Distance Sorting
Average Criteria 1(%)	66.05	56.80	64.61	64.61	63.06
Average Criteria 2	1.23	0.92	0.72	0.72	0.81

Table 6: Average Performance of 5 methods

3.3 Factor Analysis

The five methods all emphasize different factors. According to the experimental result, it seems that mean-reverting property would be the most important when selecting pairs if we focus on the stability of the performance of the pairs. However, if we only consider the average performance of criteria 1, correlation method would be suffice to generate useful pairs.

4 Conclusion

In this project, we study five approaches for pairs selection. Distance approach selects the pair with the shortest euclidean distance between their cumulative return time series. Correlation approach selects the pair with the highest linear correlation between their cumulative return time series. Cointegration approach chooses the pair where their price time series are cointegrated. In AR approach, we fit the cumulative return spread as AR process and use mean-reverting property to choose pairs. In Granger-Causality approach, we select the pairs based on if there is granger cause effect between two cumulative return time series. Except for the AR process method, all the other methods are existing and have been tested for a long time. From the experiments, we conclude that mean-reverting would be principal factor if we consider the stable performance of the pairs. For other not studied methods like PCA, it also focuses on the factor of stationary. However, we still notice that correlation method selects a pair with 100% for criteria 1. Therefore, each method has the potential to provide strong and useful pair. Thus, the best identifying way may be to incorporate all these factors and give them different weight. Nevertheless, we still have trouble assigning the weight and it would be a possible future direction.

References

[1] Krauss Christopher, 2015. "Statistical arbitrage pairs trading strategies: Review and outlook," FAU Discussion Papers in Economics 09/2015, Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics.

[2] Alexander, C., Giblin, I. and Weddington, W., 2002. Cointegration and asset allocation: A new active hedge fund strategy. Research in International Business and Finance, 16(5), pp.65-90.

[3] Marco Avellaneda and Jeong-Hyen Lee, 2009. Statistical arbitrage in the US equities market. Quantitative Finance, Vol. 10, No. 7, 761-782.

[4] Morshed, A.K.M. Mahbub; Seong, Byeongchan; and Ahn, Sung K., "More Sources of Bias in Half-life Estimation" (2005). Discussion Papers. Paper 36.